Shared Cataloging Program Steering Committee

Minutes, April 11, 2001

Present: B. Culbertson, L.Gibbs, S. Layne, E. Fulsaas, P. French (recorder); Guests: K. Coyle, L. Barnhart

The purpose of this meeting was to explore new record merging options which could help create consolidated displays for records representing electronic resources in the redesigned Melvyl Union Catalog supported by the ExLibris Aleph system. Efforts on the part of the current SCP Steering Committee and past TFER1 and TFER2 groups to create CDL Cataloging Guidelines which provide a consistent presentation of bibliographic access to CDL titles in Melvyl have struggled with the fact that there are differing cataloging approaches in practice for electronic serials in particular. The question of what should merge and not merge is a longstanding issue with historical roots in the wider multiple versions cataloging question. The Steering Committee invited Karen Coyle to discuss with us the basis for record merging in the current Melvyl system and the prospects for merging multiple versions records in the new system.

CDL is beginning the planning process to redesign union catalog indexes, displays, etc., under the Aleph system. Melvyl transition task groups have been formed and have begun meeting. K. Coyle will lead the Database and Technology Group, which will work on indexes, merging algorithms and displays. The current plan calls for a test merged database to be available in Fall 2001 and a new union catalog named Melvyl Transition to be available by Summer 2002; Melvyl Transition will run in parallel with the existing Melvyl Catalog for approximately 6 months, after which the current Melvyl catalog will be completely replaced be the new system under the original name "Melvyl". At this time, some information is known about the new database design and but there are still many unknowns.

The Melvyl II database will store each campus record separately. A search will retrieve a group of separate records which will be merged dynamically at the time of display. Merging and dedupping will be based on defined "equivalency groups". At this time, ExLibris plans to emulate the current Melvyl catalog merging algorithms and will work with CDL to refine and adjust them through testing in the test database. Other ExLibris customers are planning to use the CDL merging algorithm when it is completed. Because the merging and dedupping will be dynamic, it will be possible to adjust the algorithm as we gain experience with its application in the combined database so long as the data elements used are included in indexes.

Under the current Melvyl Cat and PE merging algorithms, the goal is to keep different physical formats apart in separate record clusters. The current algorithm was defined approximately 20 years ago and at that time people felt strongly about keeping multiple versions apart. At that time, the most pressing multiple version situation was between microforms and paper copies. The long-standing debate over treatment of multiple versions has resurfaced with new intensity with the proliferation of electronic formats, including CD-ROMs and remote access to materials over the Web. Many titles now exist in three or more formats.

A second goal of the current merging algorithm is to merge as many records correctly as possible without causing incorrect, accidental merges. In a database of 20 million records, small idiosyncrasies caused by a single ambiguous element in the merge algorithm can cause many unintended record merges, some of which are never discovered due to the sheer size of the database. It's important to be careful and conservative in specifying data elements, which cause a merge.

The merge algorithm examines data elements in specified indexes. It does not examine the MARC record directly. The algorithm is tied to the UC Minimum Record Standard (198x), which

specifies required bibliographic data elements to support accurate union catalog merging. Typically, a merge algorithm looks for specified fields and compares their values looking for matching values. It is more problematic to create an algorithm based on "or" logic. In redesigning the merging algorithm, we will be able to choose which data elements the algorithm compares, including different values from the ones currently in use. For example, the current algorithm looks for a combination of values in the Leader and to 008 to keep records for different physical formats apart. The Leader values would have to be taken out of the formula if we wanted to bring formats together. There are many other data elements that would need to be studied carefully to predict the extent to which records for electronic and non-electronic versions would come together under any new formula.

In choosing new data elements to use in a merging algorithm, CDL will need to clearly define the goal of merging. It will also need to accommodate the needs and desires of all union catalog constituencies. A single algorithm must work for all instances and in all records. It is not possible to design special algorithms that apply to only a subset of records. The working goal is to achieve as close to 100% accuracy as possible. In a database of 20 million records, even a percentage point or two of error can represent a large number of "lost" records. An accuracy rate of 95% may be more realistic than 100%.

Successful merging must be based on bibliographic elements which are consistently applied and are found in all records. There is always a difficulty with inconsistent or non-existent coding. It is not known now whether data located in the leader will be usable for merging. Numerical or coded values are usually more reliable, however these can be problematic also. For example, not all records have OCLC numbers and they can be a problem because they change and are replaced by new numbers. The 776 (Other physical format) field could be examined as a basis for matching multiple versions but it does not always contain the numerical subfield values that could be used as a match point ($w LCCN or OCLC number, $x ISSN, $z ISBN). Whatever elements are chosen, they must be identified in advance so that they are indexed appropriately.

The 856 field is a poor merging element because there are too many variations in the form of the URL or persistent identifier. PIDs or PURLs are unique to each separate PURL resolver. Within the UC system there is interest in expanding the use of the CDL PID server located at UCSD but there are a number of technical issues that must be explored and resolved before that can happen. The CDL staff person who set up the PID server is no longer at CDL so we have lost some expertise. It may be necessary to move to a different URL resolution software package to support a more complex, cooperative PID server. UCSF has expressed interest in beginning to use the CDL PID server, and this may trigger renewed study of how to make this feasible.

As one of the constituencies of the union catalog with a special interest in merging, the Shared Cataloging Steering Committee will want to participate as much as possible in identifying potential ways to merge separate catalog records which represent access to the same material. As members of the new Database and Technology Group for the Melvyl transition, Becky and Pat will keep the steering committee members informed of questions, issues and progress. The committee may want to be more proactive in developing ideas for potential record merging. This will be an agenda topic for one of the next steering committee meetings.

Document owner: **HOTS webmaster**
Last reviewed: November 6, 2003