Moving to a Single Bibliographic Record: Issues and Opportunities

SUS Technical Services Planning Committee

August 18, 2008

This report is intended to be an exploratory document that outlines the complex issues associated with moving to a shared record structure and describes the experience of other large shared systems. Included are cost considerations and an assessment of the sustained commitments required from CSUL to support such an effort. Ongoing planning work by the TSPC and other committees as well as local and consortial labor-intensive efforts by TS staff across the SULs in database preparation and cleanup before merging would necessitate a multi-year timeframe for implementation of a single bib Aleph architecture.

Aleph architecture options for sharing bibliographic records

When we discuss moving to "a single bibliographic record" what is really meant is a change to the architecture of our Aleph system. Currently there is a separate instance of Aleph for each campus, with no shared bibliographic records of any type. Based on a survey of large shared systems in North America the following are the possible Aleph architectures currently used for sharing. The options for sharing or not sharing of the patron file are not included. Shared patron file options will require research and testing.

The Ex Libris Shared Systems recommendation states that the HOL library has to be shared if the BIB is shared. After more research it looks like that is no longer the case. There is a table in Aleph called tab_library_relations that now says that there can be individual HOL libraries if there are individual ADMinistrative libraries.

Architecture options currently in use are:

- One bibliographic (BIB) record file, one merged bibliographic record per title, multiple ADMinistrative files (ADMs). This is the model that CCLA uses. Ex Libris recommends this architecture. This allows for the option of using either PDQ or ILL for unmediated borrowing, although CCLA found issues with PDQ (an Ex Libris product for unmediated borrowing) and doesn't use it. Separate ADMs means that acquisitions, items maintenance and circulation financials are kept separate.

- One BIB file, one merged bibliographic record per title, one ADM. This is Maryland's architecture. It was designed for them by ExLibris because Maryland wanted to share patrons and circulation. Maryland struggled for years to try and figure out how to keep things like Acquisitions and items processing separate. Ex Libris does not recommend this architecture. This model allows for use of PDQ.

- One BIB file, multiple bibliographic records, multiple ADMs. PALS in Minnesota uses this model. Ex Libris recommends this architecture. This model retains the local copy of bibliographic record so it doesn't save space by de-duping bibliographic records, nor reduce time in batch loading.

- One bibliographic file, with some merged/shared bibliographic records, some separate, multiple ADMs. This is used by VCCS in Virginia. They only share a small percent of bibliographic records, i.e. large batch jobs, purchased centrally. This is a variation of the PALS model discussed above. FCLA is asking for more information about this model.

Points in favor of a single bibliographic record:

- Unmediated borrowing can be instituted. This was the Maryland Consortium's major reason for moving to a single bibliographic record.
- Simplification of shared storage maintenance. Instead of operating a separate instance of Aleph for the storage unit, it would be a separate ADM.
- Possibility of shared authority work.
- Shared database maintenance.
- A smaller bibliographic database.
- Full and partial updates of discovery tool will be much faster and more efficient.
- Record displays in the discovery tool will be cleaner since minor variations in headings will be eliminated that now lead to apparent duplication of name and subject fields.

Points against having a single bibliographic record:

- Loss of some local data. This could have particular impact on special collections.
- Loss of local autonomy in catalog construction. A shared bibliographic record will require close and specific adherence to national cataloging standards without wiggle room for local needs/wants.
- Greater need for coordinated database maintenance.
- Proprietary records will require special handling and may require considerable vendor negotiation and in some cases extra money expended in order to include them in a statewide database.
- There has not been a statewide standard for cataloging. This has led to divergent cataloging practices, the consequences of which are now manifested in the current Endeca union view record. Permanently merging the records would mean that there would be no way to fix some of these problems.

Important considerations

- Based on discussion with personnel from the Maryland consortium (USMAI) and previous experience in the SUL with migration, it will probably take two years of planning for the database merge, and will require considerable ongoing maintenance. Five years into their merge process, Maryland is still making important, high level decisions on serial displays, for example. These are not minor cataloging issues but decisions which have substantial impact on users.
- In order to truly create a union catalog, there must be an ongoing commitment from the directors to support intensive committee work, including more face-to-face meetings by technical service personnel.
- There needs to be a strong commitment to maintaining database quality.
- The upgrade to V.19 and changing the Aleph architecture are two processes that need to take place independently of each other.

Costs (Will be difficult to estimate until we get to the point of negotiating for costs)

- Aleph is licensed based on the number of staff users, not on the number of instances. It would not cost more money to add an instance for the shared storage facility. There would however, be a cost in terms of server resources including disk space.
- To implement a shared bibliographic database there would be considerable costs in terms of staff time both at the individual institutions and at FCLA.
- There will be no savings in Endeca costs unless there is a substantial decrease in the number of bibliographic records. Our Endeca license is for a set number of records (eight million). We currently use 7.5 million records. We will be adding more records for digital libraries soon, as well as the CRL records.

See the appended spreadsheet and the FCLA responses by Michele Newberry, nos. 3 and 4, for more analysis on costs.

Decisions that must be made before proceeding

- Which Aleph architecture?
- What preliminary cleanup must happen before merge?
- What timeline?
    - After v.19? (FCLA estimates v. 19 upgrade in Summer 2009 and probable migration to the new architecture in Summer 2011)
    - When we change systems? (either to URM, open source or another commercial product)
- Would all the SULs merge at once or would some set of SULs go first, with the others transitioning later?
- Policy for record merging:
    - Shared bib (merge together with some algorithm for what fields to retain over others to create the shared record and then what local info. to retain.)

    - One bib (choose one campus's record, still could retain some local info.)
- Decisions on setup of merged Aleph parameter tables (in the BIB library, such as indexes)
- What would happen with proprietary records purchased by one institution, would we have to renegotiate with record vendors and pay for access for all?
- How would post-merge authority maintenance be performed?
    - use new aleph functionality to do locally
    - send db out to vendor for initial and ongoing maintenance
    - send db out to vendor initially, ongoing using aleph software
- Centralized loading: how much of batch loading should be done at FCLA? What about approval plans?
- Ongoing Cataloging/authority policies
- How should unmediated borrowing be realized?
    - Aleph ILL versus PDQ
    - How would the patron file be set up in new environment?
- Would there be retention of local Aleph files (Keep frozen files? For how long?)

Further issues requiring discussion and exploration

- It is recommended that there be thorough research into different methods of implementing unmediated borrowing. So far, there has not been any discussion known to the TSPC of the various technical possibilities for implementation. There may be a way which would be faster and cheaper to implement unmediated borrowing than through a shared bibliographic record file.
- Acquisitions workflows
- Circulation and ILL
- OPAC Group
- Special Collections
- Shared Remote Storage Facility

See also the appended FCLA responses by Michele Newberry.

Submitted by: Amy Weiss, Cecilia Botero, Jean Phillips, Jeffrey Bowen and Susan Heron

Cost estimates for single bib-related tasks

Activities	Staff time internally (hrs)	Staff time statewide (hrs)	FCLA staff time (hrs)	Other costs	Time Estimates
Setting guidelines for cataloging for merger and in the future	High	High	Low	Travel/Conference calls	2 years
Determining merge specifications	High	High	High		6 months
OCLC reclamation program	Medium	Low	High		6 months
Identifying and resolving problem records from OCLC reclamation	High	Medium	Low		1 year
Setting guidelines for acquisition records in the future	High	High	Low	Travel/Conference calls	1 year
Setting guidelines on what local fields will be retained	Medium	Medium	Low		6 months
Setting guidelines for serial records	High	High	Low	Travel/Conference calls	2 years / ongoing
Determining what will be done with proprietary records. Renegotiating and implementing contracts with record vendors	Low	High	Low	Negotiating what will be done by schools that would have to pick up purchasing of records that they never had to pick up before. Picking up additional subscriptions to proprietary records across all SULs if required	3 months / ongoing
Generating Aleph reports for clean up	Medium	Low	Medium		6 months
Setting guidelines for authorities	Low	High	Medium		2 years
Identifying and moving copy-specific or local information to Holdings records	High	Low	Medium		1 year
Testing	High	High	High		1 year
Setting up new architecture	High	High	High		6 months

Authority clean up by outside vendor	Medium	Low	High		8 weeks

FCLA Responses to Questions Raised

by the Shared Bib Task Force via Email July 15, 2008

Michele Newberry

July 24, 2008

1. The model that Jean Philips has proposed is eleven ADMs, one for each of the SULs. Is this what you have envisioned? If so could you please give us a comparison of the costs of this model with the present one of seventeen instances of Aleph. Jean has told us that it depends on number of staff users; this contradicts what we had heard from other sources who indicated that it was based on instances. We need some clarification and pricing.

   [FCLA] Yes, Jean's description of the model is what we at FCLA have envisioned as the best Aleph architecture needed to support a shared/single bibliographic file. We based this decision on our own understanding of Aleph and on information we've gleaned from talking to other Aleph consortia that share single bib libraries. Sharing a bib file will let us collapse, over time, the 11 separate Oracle instances into one. However, a shared/single bib model should continue to have separate ADMs for each university to allow for optimal independence of acquisitions and item data. This would require keeping the 11 ADMs that we have now. If we go to a shared/single bib model for the 11 SUL catalogs, we will continue to need separate Aleph instances for special functions as we currently have them. Currently they are: DLU01 (the old QF records), UXU01 (the Endeca merged file), CRL01 (CRL testing), LCA10 (LC resource authorities) and MSH12 (MeSH resource authorities) each serving an identifiable purpose that benefits the SUL. The parallel indexing project requires the 17th instance, xxU01, where we copy all the bib/hol/adm data for each university in order to rebuild your indexes. In all likelihood, it is probable that future requirements to support external data sources to feed into MANGO may require additional Aleph instances to facilitate data workflow and management. Nothing in this analysis presumes that reducing the SUL catalog Aleph instances will allow/necessitate elimination of the others we have created for a variety of reasons. It may be possible to do so but that will require more analysis. Thus let us assume for the moment that we would be reducing from 17 instances to 7.

   What Jean has told you regarding our contract with Ex Libris is correct. Our ongoing Ex Libris costs are based on the number of concurrent users (public and staff) and not at all on the number of Aleph instances. The contract does not speak to this variable at all. We did not originally pay based on number of instances therefore we cannot reduce our costs based on that. So, regarding software costs, there would be no difference whether we stay at 17 instances, grow to more than that, or reduce to 7 instances. We have asked our Ex Libris sales rep for confirmation of our interpretation of the contract and we are waiting for his answer.

   Independent of this initiative, we are working with Ex Libris to develop a better usage tracking capability to see if, in fact, we can reduce the number of concurrent user licenses so that we could reduce costs that way. This would be totally independent of the decision to reduce Aleph instances by sharing a bibliographic library.

2. We assume that programmer time will be needed to achieve the consolidation of records onto a single record. Could you give us an estimate of how many FTE programmers, for how long, and at what rate?

[FCLA] All we have for as long as it takes. ;-) Seriously, the fact of the matter is that we've never done this before so projecting time and cost is quite difficult. I would view this process to be equivalent to what we went through in the original NOTIS to Aleph migrations albeit mainly with the bib data this time. What will your specifications be for the consolidation of the records? You've raised the issue of distinguishing between single and shared bib records. In either case, we have to program an algorithm for choosing the base/master record. If single means we keep one and throw the others away, the level of effort is quite different from whether we have to program to identify pieces of bibliographic data and merge them into the chosen bib record and/or store them in the 'superholdings' like Maryland. Some work in this regard has been done for Mango but, as was mentioned on the TSPC call, that merging of data isn't permanent and we can change it with each new forge of Endeca. Creating a single bib library in Aleph will be permanent and, once you begin sharing the work to maintain this single library, we won't be able to go back and do it over again.

Consolidating the bib data is just one of the tasks that we'd have to do to make this Aleph architectural change. Programming will also be needed to reconcile all of the record keys for linking the HOL, ADM, ITEM, etc… records to the bib record that is retained. Since all of these ancillary records now link to your separate bib records, this change will be mandatory and we have to get it right. Considerable analysis and testing will be needed to be certain and these iterative activities are part of the development process.

Endeca programming may have to be done as well especially if you decide you want to store bib data in holdings records. Even if, as Jean mentioned, Ex Libris is reducing support for this function and the Aleph OPAC might not continue to understand the existence of such data, Mango may be able to handle it. But that will take programming – and that will take a functional specification and time to implement it.

So, with these caveats in mind, I think we would expect to use 3 FTE programmers and 3 FTE librarians for 18-24 months to get us from 11 Aleph instances to one. We would not be able to devote any significant time to this project until after the completion of the V.19 upgrade which, in all likelihood, would be implemented in Summer 2009. Given the realities of the academic calendar, the most probable timeframe in which we'd want to undertake such a significant data migration with be Summer 2011.

3. Would there be any savings in hardware and/or storage costs?

[FCLA] We are starting an analysis process to tighten up our initial projections but our quick take on this right now is that:

Servers: No significant savings in server hardware at all. The volume of activity in the single instance should be very comparable to that in the shared instances so I don't see any likelihood that we would reduce costs because of this activity. The continuing reduction in costs of servers we would realize as we normally upgrade hardware overtime will have a bigger impact.

Storage: For the bib data alone, we think there might be a 25% savings to reduce from 11 instances to one instance assuming that you will want to keep some local data in a merged bib record and/or superholdings. We are basing this on extrapolations from the merged file we use to feed Endeca. The

bib records are just one piece of the whole. There are many related files of data tied to the bib records including all the indexes. One set of files we don't currently have are institutional logical bases. These files will offset some of the savings realized in reducing to one bib file. And, none of the ancillary data will be reduced. That means that we will continue to have the same volume of data for the HOL (xxU60) and ADM (xxU50) libraries and their unique indexes as well as all of the Oracle re-do logs that track the online and batch transactions for changed records and the batch temp files for the hundreds of batch jobs that run everyday and/or week. Many of those jobs are Sublibrary dependent and that will not change with the shared/single bib library. So, bottom line, we think that the overall savings across all of the Aleph data will only be about 18%. With the cost of disk these days, that savings is relatively insignificant.

Both: For the 18-24 months of the transition, we will require more server capacity because of the amount of testing we will need to do and the iterative process we will undergo to get the merged bib library created. Creating the shared Aleph instance with the new, merged bib record plus all of your ancillary data and indexes will require almost doubling our disk storage to make the transition work. Therefore our costs will actually increase during this period. And, if you will want continued (non-update) access to your separate Aleph instances for some period after the transition is complete, that doubling of disk storage will continue until we can delete the individual instances. For the original Aleph migration, that window was 1-3 years depending on which phase you were in. How long would you expect to have access to your separate Aleph instances after this migration?

4. What savings would we expect in FCLA support?

[FCLA] None for the 18-24 months that it will take to make the transition. In fact, the level of support will increase during this time period because we will continue to do everything we already do for the 11 Aleph instances (and all the other Aleph/Endeca-related activities) as well as work on the transition. All Aleph configuration tables will have to be merged and tested. All existing Aleph functionality will have to be tested in a multi-ADM environment to determine how it will change system behavior. Given that FCLA will probably not have any additional staff, this means finding some compromises in what will be expected of the current staff during this transition. Once completed, we will realize some savings in a few areas. These include the level of effort and time needed to:

- ***install service packs and implement version upgrades***
- ***replicate Aleph licenses across 11 instances***
- ***maintain 11 sets of index tables***
- ***managing the common elements of bib and auth library parameter tables and MARC tables***

Those savings will not be mathematical since SPs and version upgrades have many components that are ADM and Sublibrary driven to insure that table configurations and other parameters are still functioning correctly. Since we currently have approximately 2 FTE devoted to the former and our experience with the latter doesn't give us a lot to extrapolate on, it is hard to predict what the savings might be. At the most, 1 FTE. I don't see any significant savings in all other areas of Aleph support. Day-to-day operations will continue, your acquisitions and circulation volume isn't tied to a particular architecture, the daily and weekly batch jobs will continue, configuration tables will still exist and will, in fact, be even more complicated to maintain because they will be combined. These projections are based on duplicating the existing functionality in the one bib environment. New functionality would have to be reviewed to see if it can be included in the project or not. We already know from CCLA's

experience that there are issues when some tables get very large and we have to expect that some of ours will be larger than theirs due to the complexities of the SULs.

5. One thing that we discussed was the advisability of having FCLA resume doing dataloads for larger numbers of records if we go to a one-bib model. What would that cost?

[FCLA] This seems logical for some data especially where that data is shared by multiple institutions and can be acquired through a consortial process but I wouldn't want to leave this open-ended. Obviously, daily OCLC downloads will continue to be local. Some files loaded through GenLoad may continue to be handled locally with the caveat that it might require some GenLoad modifications to be even more careful that duplicates are not introduced and that holdings are attached appropriately. This means programmer time. And, obviously, this would mean an expansion of FCLA support in this area.

Here are Mary Ann's comments on this matter:

> ***It would make sense to centrally load shared and recurring files, such as Marcive, netLibrary, and records for consortial eresource packages. Serials Solutions records would be an interesting case because the records may vary from institution to institution; however I think the complexity of the load would call for central loading. There are additional issues of quality control and standardized practices that may call for centralized loading for other types of files.***
>
> ***I see the smaller loads being run by the libraries. These may just be OCLC loads, as expertise and understanding of dataload principles and issues varies from school to school.***
>
> ***One interesting category is the vendor brief bibs for orders. UF has been working through the issues in their single bib project. The only matchpoint is probably ISBN or ISSN, but that works pretty well. When a library's OCLC record replaces another library's order record, the wrong holding may be updated and the item may link to the wrong holding (unknown how much of this will occur with separate ADMs and a shared xxu60).***
>
> ***I think GenLoad load options will need to be locked down quite a bit more than they are now. I see FCLA assuming more responsibility for writing GenLoad profiles even for OCLC record loads, then distributing tailored profiles to the SUL-wide. What will be the role of SUL library staff who are GenLoad experts?***
>
> ***I think we'll need to make more use of the Aleph loaders. GenLoad uses the pc_server. Under a single Aleph instance, the traffic from the GUI client, Mango, and GenLoad that's now distributed across 11 instances will all be hitting the pc_servers for one instance. It would be good if we could move a lot of the routine, shared dataloads to Aleph loaders to relieve some of that traffic. However, we'd exchange the pc_server traffic issues for the complications of scheduling batch loads when there's just one xxu01 batch queue for all 11 schools. I see a lot of loads scheduled for night and early morning to avoid the daytime contention. I'm sure there are more issues that'll come up over time.***

6. We are sure that we haven't covered everything, so please advise on what other costs or savings are involved.

[FCLA] Over time, the cost of building the Endeca file from the Aleph data will be reduced because many of the merge/dedup problems will be resolved permanently and won't have to be handled in the ongoing processing. From FCLA's perspective the significant savings would be in the shared bibliographic and authority headings maintenance that the libraries would realize. The other benefits would be an improved, more efficient unmediated borrowing capability that could take advantage of all your holdings being aggregated on a single bib record to determine availability and an equitable rota functionality.

Having said this, single bib architecture will need greater coordination and consensus on Aleph matters among the SULs thus FCLA will need to expend more effort identifying the things that need coordination/consensus, educating the SULs about the issues, and working with SUL reps to achieve a good end. It will take more effort for the libraries to reach consensus SUL-wide than institution-wide. There are a lot of little things for cataloging -- labels for indexes, local tags and how they're used, what's in the GUI search dropdown & in what order, call number displays, bib update error resolution... Not a huge deal, but it will be continuing overhead for FCLA and the libraries.

CONCLUDING REMARKS

The above may sound like FCLA isn't open to moving in the direction of a shared bib record architecture. That is not the case. We just want the rationale for doing so to be based on the true benefits that might accrue to the SULs and not on perceived savings in FCLA's costs.