Technical Requirements and Prospects for Authority Control in the SIBIS-Archives Database

Richard V. Szary

My presentation focuses on what the SIBIS system, as implemented on Geac hardware and software, will be able to support in terms of authority control for the system's archives users. Based on those capabilities, staff can begin to think about how they fit in with and can enhance their particular databases and work flow, and plan for their use.

The contract under which SIBIS was purchased and subsequent exchanges of requirements and specifications between the Smithsonian and Geac, outlines the authority control capabilities that the system is committed to supporting. The basis for authority control requirements in the contract is the traditional implementation as practiced in libraries, but enhancements were also included as requirements that define a relatively flexible and sophisticated system.

While our discussion of how the SIBIS-Archives repositories should implement authority control must not be dictated by the capabilities of the Geac system, that system does exist and is being used, and implementation decisions will be constrained by system capabilities. The purpose of my presentation is to outline what capabilities we can expect from Geac as part of our contract, and our best estimate as to their current status and prospect for eventual implementation.

Other speakers have addressed some of the issues involved with implementing authorities, both as independent and shared undertakings. In doing so, they have discussed many of the conceptual and work flow requirements of authority control. I will try to confine my remarks to the technical aspects of authority control implementation. While I am convinced that the major obstacles to improved description and retrieval of archival and other historical materials are cultural and psychological within our professions rather than technical, I will try to restrict my remarks to the technical issues.

As a prelude to my remarks, let me briefly review the current system environments in which we work. The current implementation of SIBIS uses the Geac Library Information System which runs on a Geac 8000 minicomputer. That software and that machine do not meet the requirements of the contract signed between Geac and the Smithsonian. In 1984, at Geac's request, we agreed to satisfy the contract requirements with the development of new software, now known as BPS (Bibliographic Processing System), on a different machine, the Geac 9000. I will discuss exclusively the authority control capabilities that we are expecting on the BPS/9000 system, as we have no plans to implement authority control on the current 8000 system. I will describe briefly eight authority control functions and capabilities that are addressed by the contract and where we stand regarding their development and implementation.

Capability 1: *Number of authority files*

Our contract requires the ability to support multiple authority files in the same system, as there is no way to collapse all of the specialized terms that the various repositories will need into one authority file. While we will need to return to the question of how the system will be able to use these multiple authority files for retrieval purposes, the contract calls for the system to support 10,000 logically independent authority files.

In implementing multiple authority files, each authority file will contain records for one type of heading (e.g., personal name, subject) from a common source (e.g., *Library of Congress Subject Headings*, the *Art and Architecture Thesaurus).* In practice, this means that, for example, even though the Library of Congress Name Authorities file contains both personal and corporate names, when that file is loaded into SIBIS, it will split into two local authority files, one for personal names and one for corporate names. The capability also exists to have multiple authority files for the same type of heading (e.g., two personal name authorities). I might mention in passing that the system will not be restricted to applying authority control only to the traditional library access points (names, subjects, uniform titles, etc.). The system will allow creation of an authority file for any type of heading, including form and genre terms, geographic place names, and specialized subject fields.

Geac has completed development and implementation of the capability for multiple authority files. Both Geac and the Smithsonian Institution (SI) expect problems to develop as the number of authority files grows, due to limits on the size of the tables that define these files, but are confident that the current implementation will satisfy immediate and mid-range needs.

Capability 2: *Configuration*

The capability to support multiple authority files, and the likelihood that there will be different combinations of authority headings that repositories will want to use for their records, required that the system provide a flexible way of defining which headings in which records would be controlled by which authority files. The system uses a set of tables to define these parameters for a given set of records. The linking of a heading in a bibliographic record to one in an authority file requires three pieces of information: (1) the repository that owns the record, (2) the type of record (book monograph, serial, archival collection, etc.), and (3) the field in which the heading is located. Based on these three pieces of information, the system will know which authority file to look in to check the heading's validity.

Currently, only the type of record and the field number are used. The capability to use the owner is outstanding.

Capability 3: *Linking of headings to authority records*

A long-range benefit of authority control is the ability to not only verify headings as they are entered into a record, but also to update those headings automatically as they change over time. To do this, the system must establish an explicit link between a heading in a bibliographic record and an authority record. Establishing this link allows the system to find all headings verified against an authority record to be found and changed when the authority record is changed. The system also uses the presence of the link to prevent unauthorized changes to the heading after it has been verified. Once a heading is linked to an authority record, it cannot be changed unless: (1) the authority record is changed, in which case the change will migrate to the headings in bibliographic records in an overnight process; or (2) the link is broken.

This capability is complete and operational.

Capability 4: *On-line authorization*

On-line authorization allows a cataloger/processor to sit at a terminal and automatically check the headings in a record against those in an authority file. The user will call up a record containing a heading to be checked, and ask the system to check it. Based on the configuration discussed above (record owner, record type, field), the system will look in the appropriate authority file and try to match the heading in the record. The system will respond in one of three ways:

(1) if an identical match exists, the system will allow the user to create a link between that heading and the authority record, so that whenever the authority record is changed, the system will change the linked heading as well;

(2) if the heading exists as a non-preferred form of another heading, the system will display the valid heading and allow flipping of the heading to that form; or

(3) if the heading does not exist, the system will let the user browse the headings in the authority file and link to the one that the user chooses. This will be particularly useful in the case of misspellings, incomplete names, and the like.

If the user does not find an appropriate heading in the authority file, the system will allow the user to create a new authority record from the heading and link to it immediately. That new heading then becomes available to others using the authority file as well.

This capability is complete and operational.

Capability 5: *Batch authorization*

On-line authorization works well for checking headings in records as they are entered or for working on specific problem headings. Efficiency, however, requires that the system perform as much authority work as possible automatically. Batch authorization is the capability of the system to work within specified parameters to assist in the verification, linking, and creation processes of authority control.

There are two areas in which batch authorization can be effective. In the first, existing records in the database may contain headings that are not authorized, but need to be, against the current authority file. This situation may exist either when authority control is first introduced, or if major updates to the authority file are implemented. On-line authorization of all these headings is impractical as a staff member would be forced to call up each record and each heading in the record explicitly and search the authority file for a match. Batch authorization allows the system to scan the entire database and (1) link headings which are identical matches for authority records; (2) report instances where headings partially match existing authority records; and (3) report instances where headings have no match in the authority file and depending on the repository's desires, leave them unlinked or create new authority records from the unmatched headings.

Batch authorization also permits the creation of a new authority file from headings in an existing database. In this case, the system scans the database, creates an authority record for every unique instance of a heading, and then links the two. This operation has obvious constraints based on how clean and consistent the headings in the existing database are.

Both functions — initial or re-authorization of a database and creation of an authority file from existing headings — are supported by the same system capability. In the second case, the authority file against which headings are matched is simply empty to start with so there will be no matches and new authority records will be created for each heading.

This batch authorization capability does not yet exist within BPS. Geac and the Smithsonian are currently discussing a schedule for specifications development and eventual implementation.

Capability 6: *Sources of authority records*

Authority records should be able to enter the system through a variety of means. The above discussion of on-line and batch authorization described methods of creating authority records from existing or new headings in bibliographic records. In cases where an authority file is created or maintained separately from description of collections, manual keying of individual records is possible, similar to the manner used to enter new bibliographic records. If the authority file already exists in machine-readable form (such as the *Library of Congress Name Authorities)*, it can be loaded en masse from tape. Tape loads require authority records that are in the MARC Authorities format, but programs exist or can be written to convert other types of machine-readable files into forms suitable for loading. Similarly, existing authority records can also be downloaded individually from an authority source, such as Online Computer Library Center (OCLC) or Research Libraries Information Network (RLIN), in much the same way currently possible with bibliographic records.

All of these features appear to be developed and operational (with exception of headings derived from existing bibliographic records, as discussed above), but further testing of non-MARC loading needs to take place.

Capability 7: *Subfield-level validation*

In the case of compound headings, such as a subject heading with a geographic subheading or a name heading with an occupational qualifier, parts of the authority record itself may be controlled by other authority records. For example, an authority file for occupations may exist from which occupational qualifiers in personal name authorities are taken. In these cases, authority records for occupation headings would not only control headings in bibliographic records (MARC bibliographic tag 656, subfield a), but also headings in authority records (MARC Authorities tag 100, subfield c). In these instances, when a change is made in the authority record, the system must update not only bibliographic records containing that heading, but also any other authority records where the heading is used, as well as bibliographic records where the compound authority record is used.

This capability has been implemented partially in the BPS system, but while the implementation works, it is awkward to use and inefficient. Each combination of headings requires an individual authority record. Neither Geac nor the Smithsonian are satisfied, and work will continue on better implementation.

Capability 8: *On-line catalog capabilities*

Traditional authority control, in manual systems, is more directed towards control of the cataloging process than as a direct aid to the retrieval process. In these systems, the consistency which authority control imposes on the choice and form of headings makes the catalog a more standardized retrieval tool, thus indirectly aiding the user. Although selected cross-references from the authority file may also be included in the catalog to guide users to the vocabulary used in the catalog, users rarely have direct access to the authority file to help in constructing their search strategy.

The same approach can also be implemented in an automated on-line catalog, but the technology offers the opportunity to give users more insight into how the catalog is constructed, giving them a better chance to search efficiently. Some of the features could include: (1) automatic switching of search terms from non-preferred to preferred forms; (2) allowing users to view and search all terms related to the original search term (e.g., Paleontology, see also Paleobotany, Extinct animals); (3) an extension of the previous capability to allow users to view thesaurus relations (broader terms, narrower terms, and related terms) in those files which are so constructed, and modify searches to use those terms; and (4) incorporation of reference headings, not used in bibliographic records, but containing information, rules, and examples for how to construct searches that deal with the topics represented by those headings — for example:

Social registers.
Registers of a particular place are entered under the name of the place with subdivision, Social registers.

In an automated system with multiple authority files, the capability for users to search these varying and potentially conflicting sets of terms requires a method of either integrating or differentiating among them and choosing the set most appropriate to their needs. Carol Mandel's report[1] points out the four options for handling this situation: segregated files (forcing users to choose which set of terms to use), mixed vocabularies (mixing the terms from multiple sets into one index which is searched as a unit), integrated vocabularies (mixing the terms from multiple sets, and providing some method of resolving conflicts in the way particular terms are used in different sets), and front-end navigation (providing a way for the user to convey to the system the types of terms to be searched).

The BPS system will handle many of the traditional presentation features, and most of these have already been implemented. Automatic switching, cross-reference display and selection have been implemented; reference authorities and thesaurus displays are still in development. Support of multiple authority files for the public user is still an outstanding question. Current plans call for following the segregated files option, but implementation hinges on the delivery of the capability to qualify searches by a predefined portion of the catalog, with such portions to be defined by their use of shared authority files.

In reviewing the status of the forgoing requirements, it becomes clear that most of the capabilities needed to support traditional authority control features, on both the maintenance and retrieval sides, have been implemented and are ready to use. The major outstanding capability is batch authorization and there is no firm schedule for its development. Other capabilities that are still under development, but which we feel will be delivered without major problems include reference authorities in the on-line catalog, and thesaurus displays.

Some of the features needed for more complete and flexible authority control are present, but many are still outstanding. The two areas in which the most development work remains are subfield level validation and multiple authority files. As mentioned, the ability to combine terms from different authority files into one heading is implemented, but leaves much to be desired in terms of work flow and efficiency. While Geac recognizes the problem, no resolution has yet been decided upon.

Multiple authority files present complex problems. The resolution of the technical issues depends on defining an approach that addresses the complex conceptual problems that they pose, particularly as they relate to the retrieval features of the public catalog. In many ways, these conceptual issues depend on resolution of a larger issue — what is the role, and purpose of a public catalog for archival and other historical materials, and how will users approach it. Much of the detailed study of user expectations, attitudes, and needs has yet to be undertaken, so that we can use that information to design useful retrieval tools.

The SIBIS-Archives system, as it will eventually be implemented in the Geac BPS system, will not be the ultimate answer to the question of how to provide access to historical materials. Its development and implementation over the past few years, however, along with the development of similar systems by other archives and networks, have forced us to articulate issues inherent in harnessing technology to information about the cultural resources under our care.

Implementation of authority control for the SIBIS-Archives catalog, as much as possible, will strive to improve our ability to manage and disseminate information about our holdings in the short run, while recognizing the current limited state of our understanding and not foreclosing options for future development.

NOTES

(1) Mandel, Carol A. *Multiple thesauri in online library bibliographic systems.* A report prepared for Library of Congress Processing Services, Cataloging Distribution Service. Washington, D.C. : Library of Congress, 1987.

AUDIENCE DISCUSSION

AUDIENCE SPEAKER

Is there the capability for tracking?

RICH SZARY:

There are some transaction-logging features on the system that we have not used as yet, so I'm not sure how extensive they are. My impression is that it will count number of hits within paricular indexes.

MARION MATTERS:

Will it tell you the specific words that searchers use?

TOM GARNETT:

It shows which indexes are used by terminal over a span of time. So, the library could learn that people search by title 20 percent more frequently than they search by subject or more often than they search by author.

FRED STIELOW:

Up to now, with the exception of a couple of comments in the last session, we have been talking about authority control for bibliographic records from the perspective of the researcher. I am wondering if you would like to comment or expand on the idea of using authority controls from an IRM (Information Resource Management) perspective or from an administrative perspective.

RICH SZARY:

That's what I was going to discuss when we broke last time, when we started talking about the utility of reference files and whether we can market them. Again, this is information that archivists and museum curators already have. They know the history of the organizations that comprise their collections, and they just haven't recorded it in most cases in a systematic fashion.

I can see using reference files in an IRM fashion to maintain an organization's changing bureaucratic structures, charts, staff assignments, whatever, on a continuing basis. I can see using them to control bibligraphic headings and retrieval of bibliographic materials. I can also see using them as an independent resource that allows the organization to manage itself better.

I'd like to offer a more concrete example or possibility. The Smithsonian Archives keeps a record of exhibits that the Smithsonian has produced over the years, in which the titles of exhibits are used as access points to records of various divisional files, museum files, and exhibits files. The possibility of augmenting exhibit titles with information on what the exhibt was about, where it was located, the types of artifacts in it, the educational level it was aimed at, and so forth, and maintaining that as an independent file, on a current basis, would give the Smithsonian an exhibits calendar that the archives would maintain.

Now, that's a new role for the archives to assume. I think it's a logical extension of its current work. And I think it's work that would raise the visibility of an archival unit, would contribute to the functioning of the parent institution, and also would help with bibliographic retrieval down the line.