# Development of An Authority Control System for the Smithsonian Institution Libraries

*Thomas Garnett*

In this paper, I present the development and the non-development of an automated authority control system at the Smithsonian Institution Libraries (SIL), defining the prerequisites for the system, the processes of evaluation, and our findings. Since I address the idea of authority control, I will briefly mention our current manual authority system which will require a simple description of our cataloging processes. However, for an audience of archivists, it will be useful to sketch our library's bibliographic "universe" which in several key areas differs from that commonly found in archival repositories.

SIL catalogs primarily books and serials. Thousands, sometimes millions of copies of these items exist. Very few items in our collection are totally unique. Though SIL may have only one copy of a book, thousands of copies usually exist in other libraries. This situation is reversed for archives which primarily catalog unique items.

Little ambiguity exists among members of the library community as to what constitutes a book or a serial, although we may get picky about such issues, for example, as the difference between a monographic series and a serial. Librarians also have many valid and substantial disagreements about the description and handling of items, for example, uniform titles, bound-with items, form of conference names. But in comparison with the archival universe, by and large, SIL deals with items that are standardized to a remarkable degree. If an archival repository receives the papers of Dr. Simonize J. Pseudopodia, it could be receiving almost anything — manuscripts, letters, calendars, objects, notes, notebooks, or invoices.

Librarians catalog items primarily to enable a user to uniquely identify and find the item. With the exception, perhaps, of rare books and manuscripts, we do not catalog items to provide information not germane to retrieval and identification. This is partly a legacy of library practices that have existed for most of the twentieth century and partly due to lack of staff. SIL, and I suspect most North American libraries, does not have the time to delve into the contents of each book in its collection. We have an average of less than 2 subject terms per bibliographic record. Examining a library bibliographic record will not tell you much about the content of the item. Archival printed finding aids, on the other hand, contain a wealth of information not immediately related to finding or collocating items such as biographical information, corporate history, and explanation of a discipline.

Since SIL does large amounts of copy cataloging which I shall explain below, and because we are members of a national bibliographic utility called the Online Computer Library Center (OCLC), we adhere to national standards. OCLC is a shared cataloging resource. Thousands of North American libraries contribute their cataloging to a shared database of over ten million bibliographic records. To contribute records to this resource requires agreement by participating institutions that they will follow certain standards in cataloging. If there is a disagreement between the way we would rather handle a situation and the way indicated by standard cataloging practices, we choose to follow the national standards. The differences could lie in the form of a personal name, the level of specificity a cataloger might like to use in describing a work, or what should be entered as the title of a work.

We use the *Library of Congress Name Authorities* (LCNA) and *Library of Congress Subject Headings* (LCSH) in determining the forms of names and subjects in our bibliographic records. We use them *not* because we consider them the best but because almost all other North American libraries use them and because they are easily available.

Because of budgetary and personnel constraints, our cataloging units face certain trade-offs. We acquire items faster than we can catalog them. If

we do poor cataloging, the items will later be difficult to find. If we spend too much time cataloging, many items will take too long to reach the shelves. Yet, we are in sight of the day when our entire collection will be fully cataloged. This hope conditions or drives the manner in which we handle current cataloging.

Our users' access to the information in libraries beyond ours depends on items being described in a similar manner. In common with many other libraries and in contrast to most archival repositories, a significant portion of what we provide to our users is borrowed from other institutions.

In contrast to most Smithsonian Institution (SI) archival units and, I suspect most archival institutions in general, a portion of our staff devotes itself exclusively to cataloging. These people do not assist users in locating items except indirectly with consistent cataloging, they do not physically retrieve and store items, nor do they make decisions about what to acquire or what to weed.

When we receive an item, we search on the large OCLC database to see if some other institution has already cataloged it. We may find:

1. The Library of Congress (LC) has cataloged the item.
2. Another institution has cataloged the item.
3. No institution has cataloged the item.
4. LC has cataloged a different version or edition of the item.
5. Another institution has cataloged a different version or edition.

If LC has cataloged the item we hold, we customize the existing record, and transfer the record into our catalog database. This is called "copy cataloging". We do as little checking of the bibliographic accuracy of this record as possible by simply accepting what a reputable cataloging source has provided. Approximately 85 percent of our cataloging is performed in this manner. While occasionally mistakes and inconsistencies are found in LC copy, we have determined that the time spent reviewing each LC copy record in full detail is not worth the few corrections it might bring. We take this approach (as do many other North American libraries) for both budgetary and quality control reasons, since LC provides high-quality cataloging with usually authorized forms of names and subjects. Thus, LC does the work for us. Minor authority control problems occasionally result such as conflicts between Anglo-American Cataloging Rules, 1st Edition (AACR1) and AACR2 forms of names, but in large part this type of cataloging creates a database that consists of headings established uniformly by the Library of Congress insofar as the Library of Congress is uniform. (More on that subject later). To archival units, I wish to emphasize the obvious — that it is possible for the SIL and other North American libraries to do copy or "derived" cataloging because we predominantly catalog published and thus not unique material.

For cases where LC copy is unavailable there exist a variety of possible procedures ranging from modifying member copy, to original cataloging from scratch. I suspect the latter most closely resembles cataloging work in an archive. When we can find no copy to modify, we must describe the item in hand. To determine the correct (i.e. "authorized") form of names, our staff determines whether an LC established form of the name exists by searching the LCNA, which is mounted on OCLC as a search-only file. We also search our manual file of locally established names. If we cannot find an established form of the name, following AACR2 we "establish" a form of the name and create another entry in our local file of authorized names. Note that the last step, the one we avoid at all costs, is to invest the time establishing a name ourselves. This step requires time and intelligence — both of which mean money in that higher-paid staff are required to perform this work.

Very rarely, we cannot find an appropriate subject term in the LCSH, more I suspect due to our adherence to standards than to any intrinsic merit of LCSH. When this occurs we establish a local subject heading but place it in the local 690 tag rather than the 650 tag in the bibliographic record. The new heading will then be added to a manual file of locally established subject headings. The only exceptions to this are in adding a period or form subdivision onto an LC established topical subject heading which we do not count as a locally established heading.

The above is a sketch of the bibliographic "context" in which we operate. When we started planning for automated authority control we had to consider what did we want and why did we want it, the purpose and functional requirements for our automated authority system.

1. *Controlled vocabulary for access points* — From the above description of our cataloging, one might think that the large majority of entries for any given author or subject are uniform. For a variety of historical reasons, this is not the case. Users of our online catalog can face a wide array of forms of names for the same author. This is due to varying quality of records loaded over many years, changes in cataloging practice from AACR1 to AACR2, data input errors, and so forth. Remember that we have over 400,000 machine-readable

records. When searching under the author "Jung" for works by Carl Gustav Jung, a user will discover an index containing authors, "Jung, C.J." and "Jung, Carl Gustav". Fortunately, these sort closely together so a user might be led to look at both. For corporate headings this is not always as likely. Ultimately, our goal is to have one form of name for all entries containing the name and one form of subject heading for all entries pertaining to that subject.

2. *Cleanup of past mistakes and prevention of future mistakes* — Since many different forms of the same access point appear in our data base, we need a method to convert existing unauthorized forms of headings to authorized forms. Incoming headings that are not authorized must be flagged, blocked, highlighted, and reported in some way. Of course, this will only work if the unauthorized forms in our data base or in a load, match the unauthorized forms in an authority record. However, in cases where we have multiple instances of the same unauthorized form, the authority record could be edited to contain the instance of our data base in a 4xx field. This would require some form of batch authorization.
3. *Batch authorization* — Since the bulk of our bibliographic records are loaded from the large bibliographic utility OCLC, and because we already have a large database, we wanted the capability for the system to check each record in a given batch against our own authority file, report errors, and change, or "flip", unauthorized headings to authorized. A batch might be a portion of our data base, our entire data base, or a group of bibliographic records loaded from an external source such as OCLC.
4. *Automated support for syndetic structure* — One of the fundamental differences between authority control and controlled vocabulary is that authority control provides for cross referencing. If a user enters a search string that matches an unauthorized form, and *if the unauthorized form exists in the 4xx of an authority record,* the user should in some way be referred to the correct form of the name or to bibliographic records containing the correct form of the name. The user should also be able to see related headings and, *if so coded,* broader and narrower terms.
5. *Ability to make global changes* — When we wish to change all forms of a given name or subject term, we want to be able to make the change in one place (the authority record) and have the change migrated, not to the entire database but to all occurrences of the name or subject term that are linked to the authority record. To change an entry record by record would require a massive effort if the heading occurs in several thousand records. An authority control system should be able to change all the occurrences of the non-preferred form in the appropriate records to the preferred form.

The above points served as our starting vision of a much improved cataloging environment. We thought we would not only have increased capabilities for our staff and users, but the amount of work and maintenance on the catalog would decrease. Our rosy initial vision has changed considerably.

The Smithsonian Institution signed a contract with Geac Inc., of Canada in September, 1983, for an automated integrated bibliographic and library system called the Smithsonian Institution Bibliographic Information System (SIBIS). Written into the contract from the start, and elaborated at many meetings, were requirements for automated authority control. SIL was to take a lead role in the SI to implement authority control on the new system. By April 1984, we had created a MARC database from archival tapes and by June 1984 had a full online catalog and cataloging module. However, the contractor was unable to provide automated authority control at that time.

At this time, the SIL organized a temporary planning group to analyze existing procedures with authorities, extract what was functionally necessary, determine what could be either automated or aided by automation, identify the unknown or problematic, develop requirements, and create a systematic set of procedures for operating in an automated environment. To clarify what was needed, both for the contractor and ourselves, we prepared a detailed study of authority control, both automated and manual. The effort took the form of:

1. A review of available literature.
2. Staff solicitation for input on what they did, how they did it, and why they did it. The latter was often a matter of detective work.
3. Determination of the magnitude of the project. How "messy" was our database? How many headings falling under authority control are loaded in an average day?
4. Identification of specifics. Exactly what tags and subfields in the MARC record for a given format do we want controlled? How do we keep "in synch" with LC if we do local editing of LC-derived authority records? How do we handle form and period subdivisions for LCSH since these are often not included in the LCSH record? How do we handle geographic names

coded in a 651 tag and place subdivisions in a LCSH authority record? Or names used as subjects?

In our study, we described a planned user/staff scenario for using the system. For the contractor as well as ourselves, it was necessary to continually remember the context in which all this would be occurring. Trade-offs between simplicity and increased capabilities are often necessary. We did not want to design an ivory tower system, beautiful to describe but impossible to use. In concert with other SI staff and other Geac users, we met with Geac in the spring of 1985 to iron out any misunderstandings. Many meetings and discussions followed.

I retell this bit of history because I want to stress that defining requirements, specifications, and procedures took time and was absolutely necessary. Although we had very bright catalogers who had been using manual authority control methods for years, until we started this internal study, we underestimated the complexity of automating the process. Even if the contractor had software ready to go, until we had done this thorough internal study, we would not have been able to start.

Geac returned with a prototype or test authority system that we loaded on our Geac 8000 minicomputer system, with a sample of several thousand bibliographic records from our live database, as well as several thousand LC-derived authority records. Geac also provided a very clear piece of documentation that described how the system worked and what it was supposed to do and not do. Such documentation, by the way, can save days of wasted effort during system implementation.

The same group of SIL staff who had prepared the study tested the system. As with the study the testing was partly a process of self-education. Testing took several months as we discovered many significant system bugs. The steps we followed in testing took an iterative pattern:

1. Comprehend and master what was delivered.
   a. Does the system do what it is supposed to do? If not, report a bug.
   b. If the system does what it is supposed to do, why does it do it?
   c. Does the system have functions that we did not ask for? Are they desirable? Do they fit our situation? Do they require changes to our specifications?
2. Compare our specifications with the product. If differences appear:
   a. Review our specifications. Are they unrealistic? For the perceived advantage to be obtained, are they worth it? Are they incomplete? Do they demand simplicity, when the intricacies of authority cataloging must allow for complexity? (We had a lot of these). Do our specifications ask for mutually exclusive things?
   b. If necessary, revise our specifications.
   c. If the product is incomplete or incorrect, report the need for upgrade or improvement.
   d. Clarify and negotiate with the vendor for change.
3. Install changes or upgrades and repeat the process.

This process, while not exactly profound, was what was nevertheless needed. While the SI archival units will not have to go through this entire process, they *will* need to master the details of the system and thoroughly review and revise their internal procedures.

During this period, it became clear both to SI and to Geac that no amount of modification could make the existing 8000 system fully comply with all requirements for the contract. Geac proposed new software and hardware as a solution to our needs and several other large customers, most notably the Biblioteque Nationale in France. Initially dubbed BNSI (for Biblioteque Nationale/Smithsonian Institution), later changed to the Bibliographic Processing System (BPS), this was to be a complete rewrite of their catalog, cataloging, and authority control software and would run on a much larger computer, the Geac 9000.

We ended testing of the 8000 authority control system and eagerly awaited delivery and installation of BPS. BPS prototype systems were installed in late 1986 and have been undergoing testing and evaluation since then. Planning/requirements discussion and documents continued to pass between SI and Geac during this time. In 1987 we sent a copy of our bibliographic data base to The Computer Company, Inc. to compare headings with LCNA and LCSH and to receive a tape of matched MARC authority records. Though problems with the matching algorithms, we received tapes of 127,000 LC name and subject authority records to load into our new BPS system. Soon after this, we decided to purchase the entire LCSH tapes of approximately 137,000 subject authority records to load on our system.

Software development is a slow process and the development of BPS was and is no exception. Large portions of the BPS software are still not available. However, enough has been delivered for SIL to begin working with authority control on the 9000. Because we were testing and learning about not just an authority control add-on for an already familiar system but an entirely new system, testing has progressed slower than we had hoped.

At this point, we still do not have a functioning authority control system because, at least for SIL, significant portions of the BPS software necessary for authority control such as batch authorization are still not ready. However, we are far enough along to have learned something about how automated authority control will work in our situation.

One of the most disturbing findings is that our staff will need to spend *more* time on authority-control-related work than they currently spend. Since we do not ordinarily review every single name and subject heading entered in our database, implementing automated authority control will certainly increase our work load by "requiring" a more uniform database. The BPS software will lessen the necessary extra work by:

1. Flagging unauthorized headings.
2. Indicating possible partial matches.
3. Flipping bibliographic headings that exactly match a non-preferred form coded in a 4xx field in an authority record. Unless we perform significant additions to LC-derived authority records (which is a work-load increase), this flip will be a rare occurrence. For names, LC authority records have few 4xx cross-references. Most of the 4xx cross-references LC does have for names are forms of the name established under AACR1 and since superceded by an AACR2 form of the name that appears in the 1xx field.
4. Allowing creation of provisional authority records based on headings from bibliographic records that do not match.

However, use of the system will still be more work for our staff. Remember that the bulk of our records will still be loaded from OCLC. Records loaded from OCLC (or any other external source) must be matched by a batch process against the authority records and exceptions reported out. Staff will have to clean up the exceptions. "Clean up" of batch authorized records in this context proves a rather complicated process involving at least the following actions:

1. Is the non-matched heading from the bibliographic record miscoded (e.g. in the wrong tag)? If so, correct it.
2. Is the non-matched heading an obviously misspelled version of an authorized heading? ("Obviously" is a tricky word here.) If so, correct it either by editing the heading in the bibliographic record or by linking it to the authority record.
3. Does the non-matched heading match a valid heading from an LC authority record not in our authority database? If so, download the authority record.
4. Does the non-matched heading match an unauthorized heading in a bibliographic record in our database? If so, create a non-LC derived authority record and link to the heading in both records. This procedure is the most complex and subject to error. Often, it will require that the piece be in hand. However, since the bibliographic record may have been downloaded from OCLC one or two days before, the item may well be on the shelves already or borrowed by a user. Creating a name authority record means, among other things, establishing the name following AACR2. This is not always simple.

I am not sure to what degree the above process applies to Smithsonian archival units since you are predominantly doing original cataloging and, presumably, will be "authorizing" as you catalog directly on SIBIS. As a result of our testing, SIL plans to use the following process:

1. Enter the proposed bibliographic heading in the appropriate field and subfields of the bibliographic record.
2. Enter the "find link", command to search the appropriate authority records that are designated as controlling the particular tag in the particular format being worked on. BPS attempts to do a match on existing authority records. If it finds one, link it, if not, browse or search the authority file looking for candidate headings. Perhaps, the heading was entered incorrectly.
3. Having established that the heading from the bibliographic record is not in the authority file, search the LCNA or LCSH. If found, download the record for future linking (perhaps by batch) with this heading from the bibliographic record.
4. If the heading is not found in the LCNA or LCSH, determine whether or not to create a local authority record on SIBIS. However, we have decided that if the heading refers to a personal name and the personal name occurs only once in our database, we will not create an authority record for the heading. Any other heading will require an authority record.
5. If it is appropriate to create a local authority record, with the heading from the bibliographic record enter the "new authority" command. Following parameters defined by each site, the BPS system will create a local authority record using data from the bibliographic record.

6. Since the local authority record created from a heading that occurs in a bibliographic record contains only the heading and some fixed field data, it may be necessary to edit the recently created local authority record. "Edit" here could mean:
   a. Augment the system-supplied information that appears in the authority record.
   b. Add cross references so that users or catalogers entering one form of the heading can be referred to the preferred form.
   c. Add see-also references or broader/narrower terms so that users can be referred to related items.
   d. Add scope notes. For locally created subject terms, this will provide an extremely important aid the next time catalogers need to establish a term.
   e. Add other notes. For corporate names, one might wish to clarify the history of bodies that preceded the one referred to in the authority record. For personal names, information about the occupation of the individual may be important, especially in an archival setting.

The actual process may involve several more steps. I mention this detail to emphasize the necessary complexity involved. Observe that I have not even dealt with the companion intellectual processes involved which might include:

1. Determining the correct form for a name. What uniquely identifies this form of the name from other names, without adding excessive data?
2. Determining intervening levels of organizational bodies to include in the name of a corporate body that may be subordinate to a larger body.
3. Rationalizing conference name headings.
4. Determining the appropriate level of specificity to use in the assignment of topical subjects.
5. Clarifying the relationship of new subject terms to existing subject terms in the authority file.
6. Splitting or merging subject terms.

This list could continue for many more pages. Other than providing easier access to certain data, automation of authority control will not diminish the "intellectual" problems associated with cataloging. In fact, it increases the importance of those intellectual components and thereby requires *more* time in certain areas of cataloging. Since automated authority control assumes a uniform database, deviations from the uniformity stand out more. Implementing automated authority control raises substantial cataloging issues:

1. Linked authority systems provide cross references from non-preferred forms to preferred forms, if they are coded in the 4xx of authority records. Once users start seeing some cross references in an online catalog they may naturally come to expect cross references throughout the catalog. Browsing in some of our current, non-authority controlled indexes, alerts a user to the existence of multiple forms of what appear to be the same heading because we do not now have a uniform, "clean" database. This will not be so obvious in a catalog where many, but not all, headings are under authority control. One way to avoid user search headaches is to make it clear that the user must proceed gingerly through the catalog. Another way, is to insure uniform headings throughout the catalog so that collocation is possible. An in-between state puts great pressure to move towards a "pristine" catalog of uniform headings. This implies more work than is currently being done.

   A user who enters a search string and gets no matches may easily learn or be trained to try searching in some other manner. However, if a user *sometimes* gets referred from the entered non-preferred form of a heading to the preferred form and sometimes does not, we can naturally expect the user to ask for more cross references. This implies more local addition of 4xx fields to authority records.

   Geac's BPS capabilities will allow us to define a given tag in a given MARC format as under authority control yet still have unauthorized headings in the tag. We could, if we choose, for instance, have the 110 tags in monographs authority controlled by LCNA yet still have individual monographic records with unauthorized (albeit flagged) 110 headings. Whether or not this is desirable in an automated environment or not is a different matter. How ragged can a catalog be and still be worth the time and effort required for full authority control? If we are to increase both the complexity and amount of our authority work to use an automated authority control system, it appears that we will not be able to do it halfway as we do in our existing manual system of authority control which accommodates varying levels of consistency.
2. BPS provides a command called "REF" or "reference" to see related headings or see-also headings. BPS *may* in the future provide a broader/narrower terms display with the REF command. However, this command will not be very useful unless see-also references are coded in the authority records. Users seeing some

related terms by using the REF command will naturally ask for more. Staff will have to code these in.

3. BPS provides the ability to globally change authorized headings. While this is obviously desirable, any change to an authority record must be far more carefully monitored than in a manual environment. Where changing an authorized heading once meant typing a change to a manual card file and wishing for the ability to migrate the change throughout the database or card file, now it requires a quick change to a machine-readable record, followed by automatic overnight migration. A gross spelling error in a frequently used heading could result in a disaster.
4. Since authority work at SIL, and I assume in archival units, does not occur in isolation but as part of ongoing cataloging and accessioning, increasing the complexity of one of its aspects will make the overall cataloging work flow more complex. For batch-loaded records from OCLC, every name and subject heading will be "examined" by the system, and all non-matches will be unforgivingly reported. For locally originated records or, as we call it, "original cataloging", every heading will need to be checked at the time of creation because to do so later would prove too inefficient.
5. The overall complexity of working with and administering a local system increases with automated authority control. It is much more difficult for a cataloger to estimate the results of her/his actions when updating one authority record could change thousands of bibliographic records. This problem can be overcome to some degree by far more extensive training. Difficulty in administering the system, however, affects staff and users because modifying the system takes longer and requires more detailed testing and debugging after every change. In the long run, the increased complexity will require not "dumber" staff, as automation optimists often claim, but "smarter" staff.

Do not feel entirely dismayed, however. With increased functions, inevitably comes increased complexity. No free lunch. What are the courses in this non-free lunch? To answer this, I must leave discussing what we at SIL have tested and can definitely predict and give you our impressions. Remember that we do not as yet have our entire database under authority control. All we have is a small test database and a 5 foot stack of listings of headings from our database that do not match LCNA or LCSH headings. We have many months testing and learning ahead. Some pieces of software for the BPS authority control are still not even written much less tested and implemented. What follows are my impressions of what it will provide:

1. In the long run, automated authority control will result in a cleaner catalog, a catalog with one form of name for a given author, one form of subject for a given subject. I say "in the long run" because it will take more than a year with current staffing levels to clean up the list of no matches and partial matches compiled by the Computer Company. Staff must review each of the unmatched headings to determine the correct heading. In some cases, this might involve an easy correction of obvious misspelling or miscoding. In other cases, it may take significant time.
2. Future mistakes will be harder to make and easier to catch.
3. The most important benefit goes not to our staff but to our users: automated support for syndetic structure. How much time have users wasted looking for something in our collection that they cannot find because the headings they used for searching for the item did not match the headings used to catalog it? How often have our users walked away with "none" or "some" when they might have had "all"? We do not have hard data on this, but our impression is that this occurs all too often. Given all the gloomy warnings I have mentioned, given all the extra work it will take to fully exploit the system, automated authority control will allow for far easier and more precise access to our collections by our users. We view this as by far its major advantage.

As testing and implementation continue, we hope to share our findings with other units here. Below are major areas I have not touched on that remain pressing:

1. Implementation of logically multiple authority files.
2. Shared authority cataloging among SI units, both SIL and the archival units.
3. Use of authority records as information resources in themselves.
4. Links to and compatibility with the Smithsonian Institution Collections Information System (CIS).

I hope that as we and other units become operational with automated authority control systems we may together begin to explore these issues.

## AUDIENCE DISCUSSION

*FRED STIELOW:*

I hear, I think, a basic theoretical problem from the library side. You're looking for pristine catalog, and pristine methods. You're looking for a platonic ideal as though a real one exists. Archivists approach this issue from a very relative standpoint. We want to build up our catalogs. In fact, your sloppiness translates into new access points to us. And I think the library world needs to learn from the archival world that this one ideal does not reflect a reality. It reflects a false reality. We can use the sloppiness, these other terms, as a way towards greater access.

*TOM GARNETT:*

Again, I would go back to the point of the user. I agree, as a platonic ideal, that's ferstunk. And what is your preferred term may not be my preferred term. An automated system should provide some way to resolve these conflicts. I am concerned about the poor user who wants to find something and who doesn't know how we described it. If we can achieve some uniformity in how it is described in our system, that's a way for the user to get from his or her method of thinking about a term to what we have. I don't really care what is the best term. I just want the user to be able to get the material. I don't want the user to spend any more time than she or he has to in our catalog. It's purpose is to make materials accessible.