June, 1998

A Bird's Eye View of Authority Control in Cataloging

Karen S Calhoun, *University of Pittsburgh - Main Campus*

A Bird's Eye View of Authority Control in Cataloging

Karen Calhoun
Cornell University Library

Introduction

I have to admit I had misgivings when the planning committee invited me to speak about authority control in cataloging to a conference involving the taxonomic community. What in the world could I have to say that would be useful to you? I thought maybe the key to figuring that out was examining what it is about authority control that has kept *me* fascinated for over 15 years.

So, after reflecting on that, annoying my friends by forcing them to tell me about taxonomy (nobody knew much), reading the proposal to the National Science Foundation for this workshop about six times,[1] exploring some of your Web sites,[2] and searching BIOSIS for abstracts of some of your papers, I came to the conclusion that I really *do* have something to say.

Specifically, I'm going to tell you the story of the rise of cooperative authority control in cataloging, drawing parallels as I go to the problem domain of the systematics community and biological information managers. I'm going to describe what I think made community-wide authority control possible in library catalogs, and I'm going to give you a high-level view of how it works, both from a systems perspective and from the perspective of a cataloger.

Along the way I'll share with you what I see as the limitations of the current system in libraries and where we need to go next. I'll conclude with what I see as our communities' shared challenges for the future.

The Rise of Community-Wide Authority Control in Cataloging

Twenty years ago, the Council on Library Resources, a very important agency in the United States library community, was taking a keen interest in establishing the pieces of a national library and information service network. Many found the Council's position persuasive—that a critical component of an information network is an integrated, consistent authority file.[3] Many also agreed that the existence of such a file would not only reduce operational costs in libraries, but raise the quality of library catalogs.

The Council on Library Resources has provided leadership, funding opportunities, and has championed key research projects over the years, so when the Council talks, libraries and the organizations that serve them generally listen. Perhaps the role that the Council played in the late 70s and early 80s with respect to authority control in libraries can be likened to the current initiative Systematics Agenda 2000, in promoting systematics science and its role in biodiversity science and conservation. I see from browsing the Web that the IUBS and many other organizations are taking a keen interest in Systematics Agenda 2000.[4]

As the 1970s came to a close, the Council on Library Resources was able to leverage its influence to bring together organizations with very different agendas—the Library of Congress, the National Library of Medicine, the National Agricultural Library, and three shared cataloging systems (the Research Libraries Group—RLG, the Online Computer Library Center—OCLC, and the Western Library Network—WLN)—and get them to agree on three points:

1. it is feasible and important to establish a shared authority file
2. it is feasible and important to establish procedures and requirements for a nationwide authority service, and
3. it is feasible to develop a set of design elements for an authority control system

After that initial meeting of minds, suffice it to say a lot happened, and it wasn't linear, fully logical, or apolitical. It has taken some time to work everything out, but a great deal of progress has been made toward a fully integrated, consistent authority system.

Some of the barriers that had to be overcome at the time included:

- the historical lack of a strategy to coordinate the creation of a shared file
- the need to automate the largest and most important authority control file in the nation, which resided at the Library of Congress
- the lack of a common format for sharing authority records in an online mode
- a lack of consensus among the key players as to the importance and role of authority files

I imagine some of these sound familiar to the taxonomic community, since its organizations appear to have compiled nomenclatural look-up files independently of other organizations, and various individuals and groups are beginning to explore how they might leverage their organizations' investments in taxonomic data.

There were four other key elements in the plot of this story. These were (1) the existence of the *Anglo-American Cataloging Rules*, second edition, adopted by the Library of Congress (and in turn by the nation's libraries) at the beginning of the 1980s, (2) the widespread use of the Library of Congress Subject Headings system in U.S. libraries, (3) the MARC (**ma**chine-**r**eadable **c**ataloging) format for bibliographic data, which is a communications standard for encoding catalog records, which can then be exchanged among computer systems, and (4) the trend in libraries away from card catalogs toward online catalogs.

Together, these four elements created a state of readiness—an incipient infrastructure—that made it feasible and reasonable to invest in a shared, community-wide authority control system.

The second chapter of this story opened with the gradual implementation, through the 1980s, of the Linked Systems Project, or LSP. LSP established a single authoritative source—one coherent file—for the authority control of names and some titles, from which all file copies and updates flowed, and in close to real time. While the original technological infrastructure of the Linked Systems Project, which was based on OSI (open systems interconnection) protocols for linking disparate computer systems, has been superseded, the design of the current system for exchanging authority data among the host organizations is essentially the same.

At the same time, libraries had started getting their own mini- and mainframe computers back in the 70s and early 80s, when they began mounting online catalogs of their collections. Over the next few years, personal computers, easy-to-use graphical interfaces such as Windows, and more and more affordable telecommunications technology became widespread. Add to this picture the fact that libraries were early implementers of the Internet. As a result of all this, the opportunity costs for cooperatively sharing data and files fell like a rock—in other words, it became possible for libraries to significantly reduce their costs, eliminate duplicative efforts, and provide higher quality service to their clienteles by mounting shared online information services.

We are all aware of the outward shift in the demand curve for online information. The demand for online authority records grew at the same time and for similar reasons. And, as those who took part in the early efforts to build a cooperative authority file managed to lower the barriers to participation, the supply of online authority records grew also.

Growth of the Authority File

In 1980, the Library of Congress loaded its file of 180,000 machine-readable authority records into its name authority file. In 1982, they finished the conversion of their paper-based authority file and loaded another 1.1 million records into the file. By this time, the Library of Congress had also established cooperative arrangements with the Government Printing Office, the University of Chicago, and several others. However, the means for contributing authority records was paper-based, and involving the manual typing, mailing, and re-keying of worksheets.

In 1984, the MARC format for authority data was finished and the existing 1.4 million records were converted to comply with it. This was an important milestone, because it provided all potential contributors with a standard for the intellectual content and structure of shared records.

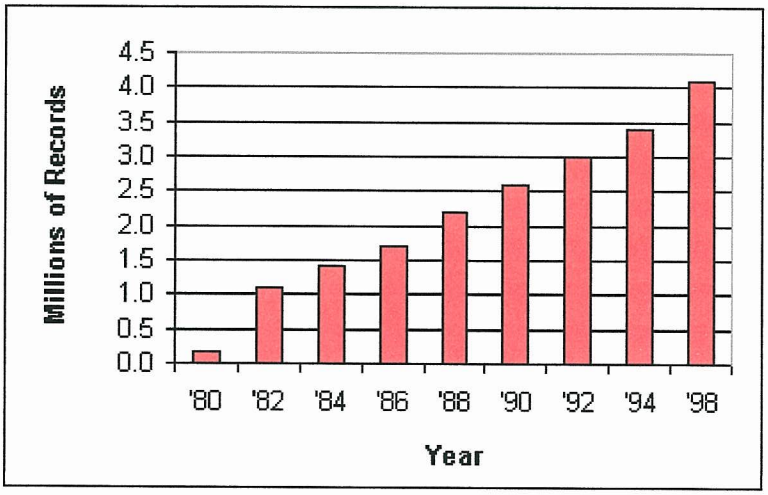

Figure 1. Growth of the authority file.

In 1987 and 1988, the Linked Systems Project made it possible to ditch the laborious and costly

paper-based contribution method in favor of online contribution and updating of the file. Yale and Indiana University led the way (although if you talked to them at the time, they might have said they were more on the "bleeding edge" than the "leading edge"). Once the kinks of early implementation were worked out, the foundation was laid for rapid growth, so that from 1988 on, each year has seen significant growth in the number of contributors to the shared file. By 1995, about one-third of the newly contributed records were coming from libraries other than the Library of Congress. That trend has continued, and today the name authority file contains over 4.1 million records.

What Made it Possible

To summarize, I think the initiative to build a community-wide authority control service succeeded because the key implementers had the *will to cooperate*—that is,

- CLR, by championing the concept of an integrated file, served as the catalyst the community needed
- The nation's largest libraries could clearly see the potential of a shared file for reducing their costs and enhancing the quality of their public catalogs
- The concept of a shared authority file was compatible with the missions of the major shared cataloging networks—OCLC, RLG, and WLN
- Over time, the key players were able to allocate the necessary human and machine resources for getting the job done

In addition, the key implementers had the *means to cooperate*—that is,

- the Library of Congress was in a position to mount and administer a cooperative program
- the shared cataloging networks could recover the costs of developing their parts of the shared authority service from their cataloging operations
- the Library of Congress was in a position to coordinate the establishment of procedures and standards, as well as training for new participants
- the whole enterprise took maximum advantage of current or emerging technology

How It Works

There are two ways to look at how community-wide authority control has been deployed in libraries. One is a systems perspective, in which we look at processes, inputs and outputs. Another operational view is that of a cataloger—that is, how community-wide authority control supports the intellectual process of creating catalog and authority records.

Systems Perspective: Record Contribution and Distribution

The master copy of the authority file is housed on the Library of Congress computing system. Synchronized copies of the file are held and maintained at OCLC and RLG, two shared cataloging systems with a combined membership of thousands of libraries, some of whom participate in the cooperative program to build and maintain the authority file. Catalogers at participating libraries sit at their workstations in their libraries, online to their library's catalog or to one of the shared cataloging systems, or perhaps they have sessions

running on both.

In the course of cataloging, catalogers discover the need to create new authority records, change existing ones, or capture copies of records for use in their libraries' catalogs. The OCLC and RLG cataloging services support all of these actions. To contribute a record, a cataloger connects to either the OCLC or RLG system, adds the new record or modifies the existing one, then issues commands that result in the contribution of the new or changed record to the master file at LC.

The OCLC and RLG systems collect the contributions and send them to the LC system once a day. In turn, the LC system processes the contributed records from both OCLC and RLG, updates the master copy of the authority file, and then distributes the day's new and changed records (which include records from LC, OCLC, and RLG catalogers) to OCLC and RLG.

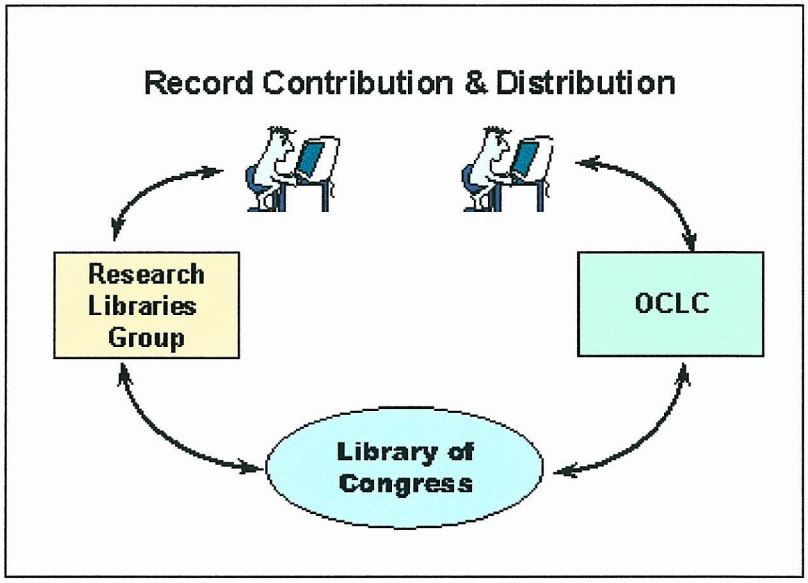

Figure 2. Record contribution and distribution.

The loop is complete when the OCLC and RLG systems collect the day's distributions from LC and update their copies of the authority file. Thanks to the reliability of the host systems, the copies of the file held at OCLC and RLG are rarely more than 48 hours out of synchronization with the master file at LC.

Catalogers' Perspective: How the Authority Files Are Used

The following section provides an overview of the four functions of the name and subject authority files in libraries:

1. **Authority function** - support consistency of headings
2. **Finding function** - provide links from variants and other authorized forms of headings
3. **Information function** - show usage and scope of headings
4. **Maintenance function** - support manual & automatic error detection and correction

The authority function: Like systematics organizations, libraries want to establish consistency in the nomenclature used in their catalogs. Achieving this consistency is extremely time consuming and requires highly trained staff.

The finding function: Taxonomists refer to different spellings, synonyms, homonyms, hierarchies, and so on. The idea is to record these somehow, and link them to authorized nomenclature. Catalogers do the same thing when they provide references from variant or related forms of a name or topic. When authority control is then deployed in a library

catalog, the references lead searchers from variant to authorized headings in the catalog. I'll provide an example shortly.

The information function: Authority records usually contain documentation about the sources used to establish the name or subject heading (that is, literature references). They may also contain information the cataloger discovered in the course of researching the name or subject. Or, the record might contain information about the scope or usage of the heading.

The maintenance function: Authority data is also used to support the detection and correction of errors in library catalogs. In some cases, the quality control of a library catalog can be a manual affair. However, many library databases are too large for that to be an effective approach, so increasingly, larger libraries contract with organizations called *authority control vendors.* The vendors use the authority file, software programs they have written for the purpose, and support staff to do both one-time and ongoing automated clean-up of their customers' catalogs.

Record Examples—How It Fits Together

Suppose you were cataloging the book *One flew over the cuckoo's nest.* Here's an abbreviated view of data drawn from the MARC cataloging record for that title from the Library of Congress' catalog on the Web. If you saw this record in MARC format, it would be much harder to read, since it is encoded to facilitate machine processing and exchange among computer systems.

Sample Catalog Record
Author: Kesey, Ken.
Title: One flew over the cuckoo's nest, a novel.
Published: New York, Viking Press [1962]
LC Call No.: PZ4.K42On
Subjects: Psychiatric hospital patients—United States —Fiction.
Control No.: 62008602

You see various items of information about the title, the authorized form of the author's name (Kesey, Ken), and a subject heading (Psychiatric hospital patients—United States —Fiction).

Authority records support catalog records. Here's what the name authority record looks like for the author, Ken Kesey. If you saw this record in MARC authority record format, you would see a great deal of encoded data to facilitate machine processing and record exchange.

Sample Name Authority Record
Heading: Kesey, Ken.

Notes: His One flew over the cuckoo's nest, 1962. b. 1935
Control No.: n 50044585

The heading field shows the authorized form of this author's name—Kesey, Ken. This is the heading form that is valid for retrieving, sorting, and displaying catalog records associated with Ken Kesey's works. In this case, no references were made from variant forms of Kesey's name. The field that is labeled "Notes" is called by catalogers the "Source Citation." This citation is quite a brief one, containing only the title and publication date of the work in hand when the authority record was first made, plus an indication of Kesey's year of birth, probably discovered when the cataloger was researching Kesey's name for the purpose of establishing the authorized heading.

Subject authority records are also encoded in the MARC format for authority data. Here's the subject authority record for the topic "Psychiatric hospital patients." This is the preferred heading for this concept in libraries that use the Library of Congress Subject Headings system. In the Medical Subject Headings system, the preferred heading for this concept could be different, and probably is. The "Make Reference From" field indicates a variant form that will appear in the library's catalog as a cross reference. The "Make See Also Reference" field indicates a hierarchical relationship to another authorized heading that is broader in scope, "Mentally ill." This will appear as a "see also" note in the catalog.

Sample Subject Authority Record
Heading: Psychiatric hospital patients
Make Reference From: Patients in psychiatric hospitals
Make See Also Reference From Broader Term: Mentally ill
Control No.: sh 85108352

You may recall that I said authority control has an information function. Here is an abbreviated version of the subject authority record for the heading "Fiction," which in the catalog record I showed you, is used as a subdivision under the main heading "Psychiatric hospital patients." The "Complex Reference" field indicates it is okay to use "Fiction" as a subdivision under a topical main heading—this is a special instruction to help catalogers successfully construct Library of Congress subject headings in catalog records.

Another Sample Subject Authority Record
Heading: Fiction
Complex Reference: Use the subdivision Fiction under names of countries, cities, etc., names of individual persons, families, and corporate bodies, and under classes of persons, ethnic groups, and topical headings ...
Control No.: sh 85048050

Where Do We Go From Here?

I may have given you the impression that the library community has succeeded in creating an ideal and complete authority control system, but I assure you that's not so. Let me share with you where I think the library community is going and will need to go from here.

Right now, the efforts of the community are focused on continuing to develop and refine the work we started at the end of the 1970s. The program has grown strong in the United States. Efforts to recruit non-U.S. libraries have been partially successful, but there are many important cultural, linguistic, and practical distinctions between national authority files, and I personally doubt that the present U.S. model can be scaled up to support a truly international system for exchanging authority data. The global library community may be too large and diverse for one all-encompassing, monolithic authority file to be workable in the long run. A new strategy and system model is perhaps needed, perhaps something along the lines of the European Union's AUTHOR experiment, which has tested search and retrieval across a federation of interoperable authority files from five national libraries.[5]

Another refinement of the current system features the development of various desktop applications to accelerate and ease the process of creating authority records. Gary Strawn, who is speaking tomorrow, has been the master of such tools, which do things like automatically generate a preliminary authority record from a heading in a catalog record. Strawn has also developed a tool to allow a cataloger to key an authority record in one system, then easily pass it to OCLC for contribution to the Library of Congress.

The present shared authority control system is really a service for catalogers; there has been relatively little community-wide progress on deeply integrating authority data into end-user information retrieval systems (e.g., providing support for user queries in catalogs, like mapping a user's searching vocabulary into the vocabulary of the database being searched). More work could be done in that area, but I am not aware of any initiatives at present.

The present authority control system covers only about half of the name headings that appear in the databases of shared cataloging networks like OCLC. This represents millions of names, but it would be prohibitively expensive to make authority records for all of these headings using current work methods. Some progress has been made that could make it feasible to expand the authority file more economically, perhaps through the automatic generation of base authority records.

For many reasons, the MARC format for authority records needs to be continually amended to better support new ways of deploying authority control in library catalogs and to take maximum advantage of technological advances and progress in computer science. This work goes forward through various American Library Association groups.

The successful authority control system of the future may require something beyond a refinement of what we have now. This may be blasphemy to my library colleagues, but we are so heavily reliant on the United States' implementation of the MARC formats—and I ask myself whether these are the data structures of the future. Should we be examining the context of relational databases and data management, so important to the systematics community? As far as I know, little work is being done in this area.

Our present system is poorly integrated with the authority control conventions of the abstracting and indexing community—e.g., organizations like BIOSIS. I have always

wondered whether we should we try to do more in this area, but taking any significant action would surely require a rethinking of the library community's current model of authority control.

Common Ground

The exercise of putting together this talk has led me to discover that the taxonomic and library communities have much in common, at least as far as authority control is concerned. Where I once saw clear separation of our problem domains, I now see many shared issues and challenges.

Both of our communities are seeking strategies to eliminate duplicative work and improve the quality of what we produce for use in information retrieval systems. The challenges are significant and many:

- we need federated, interoperable data from many sources;
- the task requires a global approach, but not one that over-simplifies the data, so that valid distinctions are lost;
- we need to continue working on our data structures and record exchange methods, so we can take maximum advantage of the technologies available to us; and
- to help those who use the results of our work, we need to seek the deep integration of authority control in our information systems

One of the documents I scanned when I was preparing this talk included remarks about the intrinsic difficulty of sharing data across systems.[6] The authors advised that the best way to overcome this difficulty is to pick a place to start and start doing it, as soon as possible.

Although I can't state categorically that this sort of "can do" attitude provided the collective backbone the library community needed to create an integrated, shared authority file, it surely helped us to persevere. As I already mentioned, the process was anything but linear, fully logical, or apolitical, but the library community managed to make it work anyway.

I am honored to have had the opportunity to address you, and I hope this overview of authority control in cataloging will help you evaluate and decide upon your next steps.

Transcript of Discussion

Biographical Information

Karen Calhoun has been the head of the cataloging department in Central Technical Services at the Cornell University Library for the last year and a half. Before that, she worked for OCLC (Online Computer Library Center), a nonprofit membership organization that provides computer services and research to support libraries worldwide. At OCLC, Karen was involved in many initiatives to advance community-wide authority control and to develop new applications of automated authority control. Karen began her professional career as a catalog librarian at the University of Oregon. Quite active professionally, she is currently involved in supporting and promoting the Program for Cooperative Cataloging, about which you will hear more in the program. Karen holds an M.S. in library and information science from Drexel University and an M.B.A. from Franklin University.

[1] Blum, Stanley D., principal investigator, "A Workshop on the Development, Management, and Dissemination of Taxonomic Authority Files." A proposal to the National Science Foundation (Award DEB-9726045)

[2] For example, BIOSIS' "TRITON: Taxonomy Resource and Index to Organism Names," http://www.york.biosis.org/triton/backgr.htm and the International Organization for Plant Information's "Provisional Global Plant Checklist," http://bgbm3.bgbm.fu-berlin.de/iopi/gpc

[3] Council on Library Resources, Bibliographic Service Development Program. "An Integrated Consistent Authority File Service for Nationwide Use." *Library of Congress Information Bulletin* 39 (July 11, 1980): 244-8.

[4] See, for example, T. Younes' report of IUBS' 1996 activities, "International Union of Biological Sciences," at http://www.lmcp.jussieu.fr/icsu/Report96/AR_SUM/iubs.html

[5] See, for example, Bourdon, Francoise and Sonia Zillhardt, "AUTHOR: towards an European network for name authority data," paper presented to the 62nd General Conference of the International Federation of Library Associations, August 1996, Beijing; and Ede, Stuart, "Libraries and technology in the European Union: soldering the connections," *Information Technology and Libraries* (June 1996): 117-22.

[6] BIOSIS abstract of Tuttle, M.S. and S.J. Nelson's "The role of the UMLS in 'storing' and 'sharing' across systems," *International Journal of Bio-Medical Computing* 34 (1994): 207-37.