May, 1996

Characteristics of Member-Established Headings in the OCLC Database

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

Authority Control in the 21st Century: An Invitational Conference

Characteristics of Member-Established Headings in the OCLC Database

Karen Calhoun

OCLC

Table of Contents

Before taking the job of manager of the Product Planning Department about a year ago, I was the manager of OCLC's Online Data Quality Control Section, which was one of staff areas that worked on OCLC's massive project to develop automated authority control software for the OCLC database and Harvard University.

In the part of the project in which I was involved, the software made about 5.6 million corrections to personal and corporate names and LC subject headings in the OCLC database, and over 774,000 corrections for Harvard. That project was a tremendous learning process for all of us who were fortunate enough -- or unfortunate enough, some days -- to have a role in it. This paper an extension of my learnings from that project. It offers some early insight into a different but related area of authority control.

Background

- **OCLC research on heading errors, 1992**
- **OSU/OCLC authority control seminars**
- **"Gap" headings**

In late 1992 we were busy with analyses of all types related to the automated authority control project. One of the analyses we ran in the Online Data Quality Control Section was an estimation of the rates and types of heading errors in Harvard's records and in personal and corporate name headings in the OCLC database. We ran the analysis to get baseline measures of the quality of the headings and to help us prioritize our work on the heading correction algorithms.

Also at that time, OCLC staff were engaged in a series of seminars on authority control topics with Jennifer Younger and several of her staff at the Ohio State University. In a couple of those meetings, we had several long discussions about what we called "gap" headings. "Gap" headings are ones that are uncontrolled--or at least unauthorized--by LC-level authority work. As we saw it, there are two kinds of "gap" headings -- those that are simply wrong, and those that appear to be correct but are not authorized in the LC AF. In other words, the heading is okay but there is no LC-level authority record for it.

Personal Name Headings in the OCLC Database, 1992

	Percent of Total

This is a chart from the late 1992 research on heading errors. I've only included the results for personal names, because that's what I'll be discussing today.

About 17% of the headings in our sample of about a thousand personal name headings from the OCLC database contained errors. Many of those have since been cleaned up by the authority control software.

About 40% were perfectly fine headings, but not covered by the LC Authority File.

About 43% of the headings exactly matched authorized headings from the LC Authority File.

It was at this point that we started getting curious about those correct headings. These are headings established mostly by member libraries applying AACR2 practices. We began wondering about the differences between these headings and LC-established headings (with "LC established" meaning headings represented by records in the LC Authority File, regardless whether a NACO or LC cataloger had created the authority record).

We were familiar, for example, with Lotka's law of bibliometrics and follow-on studies, which suggest that 60% or more of author entries are unique (in other words, occur only once in a catalog or database), so we thought that many of the correct headings would be single postings.

Personal Name Coverage

	Percent of Total
Covered by AF	51
Not Covered by AF	49

Using the same sample data, we evaluated how many of the incorrect headings were represented in the LC Authority File. We found that a little less than half of the incorrect headings were variants of LC-established headings, and the rest were variants of correct member-established headings. Therefore, as you can see from the immediately preceding chart, our sample suggests that 51% of personal name headings in the OCLC database are covered by the LC Authority File, and 49% are not.

The actual percentage of headings not covered by the LC Authority File is probably higher, because our sampling methodology was biased toward the selection of prolific authors. Suffice it to say there are many, many member-established headings that are not covered by LC-level authority records.

Do these headings merit full-level LC style authority control? What kind of authors do they represent? How are these authors different from the ones that end up with headings in the LC Authority File? How much should we worry that they aren't covered in the LC Authority File? We didn't know, and we still don't, but we have made some progress toward getting the answers.

Significance of Research

- **Headings for which authority work can be skipped**
- **Member-established vs. LC-established**
- **Member headings for machine-derived authority records**
- **NACO expansion**

If we knew the answers to these questions, we might be able to predict, for instance, the attributes of first authors whose headings would be unique and remain unique in a large database until the end of time. If we could predict that somewhat reliably, we could skip authority work for those authors, except for checking for conflicts. Conversely, we might be able to predict up front the attributes of authors for whom LC-level authority control is desirable from the beginning.

As part of our automated authority control project, we built an OCLC internal corrections database of LC-established and

not represented in the LC Authority File, we could make better decisions about what kind of member headings to select for LC-level authority control. Perhaps we could identify a useful subset of the member headings in OCLC's correction files for the creation of machine-derived authority records, or perhaps we could select new NACO libraries to concentrate on important headings now poorly covered in the LC Authority File, or perhaps we could do both.

Varieties of Tomatoes

So what do tomatoes have to do with all of this? I have a friend here at OCLC who is an avid gardener. The other day in the lunch room we were looking out the windows, longing for spring to come, and we spent almost the whole lunch period talking about the seedlings that now fill his dining room area. He named off the tomato plant varieties -- like Beefsteak, Big Boy, Yellow Pear, Viva Italia, Saint Pierre, and Arkansas Traveler. Mind you, I'm not even going to mention the usual grocery-store variety of tomatoes--we'll leave them out of this discussion entirely.

Each type of tomato that my friend is growing has its special characteristics. Some are extremely popular, like Beefsteak and Big Boy, and some are directed to special uses such as container gardening, growing in hot and dry conditions, using to make tomato paste, and so on.

On the other hand, most of us have had the experience of looking down one late spring day and seeing a "volunteer" tomato plant growing in some likely or unlikely spot. We have little idea what kind of fruit that volunteer plant is going to bear -- might be good, might not.

Varieties of Names

- **Most used names**
- **Other useful names**
- **"Volunteers"**

I think the problem of sorting out the headings for which the labor of a fuller form of authority control is justified is comparable to dividing the garden varieties of tomatoes from the volunteers. The Beefsteak and Big Boy tomatoes can be compared to the headings that LC and NACO catalogers work on. They are authors associated with our main-stream publishers. On the other hand, Yellow Pear and St. Pierre tomatoes, for which there are specialized uses, can be compared to some subset of member-established headings (we don't know which subset, but we suspect there is one) that is deserving of a fuller form of authority control than is now available for them. Volunteers can be compared to the estimated 60% or more of authors whose headings will remain single postings -- that is, unique -- forever.

Methodology

- **Samples from OCLC internal correction file for personal names**

LC-established and member-established names from our internal database of personal name headings. He selected 103 LC headings and 124 member headings at random. Once we had the headings, Susan Walker of ODQCS searched and downloaded all the associated bibliographic records from the OCLC database for me. Next, Robert Bremer created a dBase file from the bibliographic records, and John Ogden, an OCLC research assistant, generated descriptive statistics for me from that file. Based on the descriptive statistics, I prepared a number of hypotheses about the characteristics of member headings vis a vis LC-established headings. Finally, I asked John to run tests of statistical significance on the hypotheses. As you may recall, significance tests help the researcher to sort out results that are likely to be attributable to chance from results that are unlikely to have occurred by chance.

Preliminary Findings

- **Holding libraries**
- **Postings**
- **Language**
- **Subjects**
- **Dissertations**
- **Date of publication**
- **Physical format**

I cannot stress too much that our results are preliminary. In April a year ago, I changed jobs and became the manager of the Product Planning Department. The person in that position doesn't do much individual research and I'm no exception. But here is what our results suggest so far. Further research on this topic may prove us gloriously right or miserably wrong.

I'll start with the result that is the least surprising: the average LC author in our sample has 282 holding symbols associated with his or her heading, compared to an average of 22 for each member author. Our tests indicate this result and the next ones I'll mention are statistically significant.

Our LC authors have an average of 13 postings in the OCLC database; member authors average 4. Member authors are more likely to havejust one posting (56% as opposed to 24% of the LC authors).

LC authors are more likely to have publications in multiple languages. 27% of LC authors have publications in multiple languages, vs. 4% of member authors.

Our Member authors are more likely to have written exclusively in English.

Member authors are more likely to have written exclusively on science topics or topics in the field of medicine.

Not surprisingly, our member authors are more likely to be associated with a publication that is a dissertation. 33% of the member authors had a dissertation posting, compared to 7% of the LC authors.

Our member authors tend to be associated with older works. The mean date of publication of member authors' works is more likely to be before 1960. 30% of member authors' works have mean publication dates before 1960, vs. 15% of LC authors' works.

Our member authors are more likely than LC authors to be associated exclusively with non-book formats. 13% of member authors are associated exclusively with non-book formats, vs. 4% of the LC authors in our sample.

Next Phase

- **Further evaluation of sample data**
- **What groupings of attributes make sense?**
- **Which first authors require "just in case" authority control?**

The technique we were going to try is called multidimensional scaling. I first worked with the technique while I was a research assistant at Drexel University's College of Information Studies, years ago. Researchers at the school had used multidimensional scaling to do analyses of co-cited authors, thus revealing "invisible colleges." The technique has also been used in marketing research to identify sets of consumers with a specific set of tastes and preferences.

Mapping the attributes of the member-established headings might help us identify the "volunteers," for which we might want to consider doing authority control "just in time." Conversely, we could try applying multidimensional scaling to see if we can discover the attributes associated with headings that deserve a fuller level of authority control (or access control, as the case may be) than we are currently giving them. Thank you for your attention.

Return to Proceedings Home Page