The Use of Automatic Indexing for Authority Control

Martin DILLON: University of North Carolina at Chapel Hill; Rebecca C. KNIGHT: Wichita State University, Wichita, Kansas; Margaret F. LOSPINUSO: University of North Carolina at Chapel Hill; and John ULMSCHNEIDER: National Library of Medicine.

Thesaurus-based automatic indexing and automatic authority control share common ground as word-matching processes. To demonstrate the resemblance, an experimental system utilizing automatic indexing as its core process was implemented to perform authority control on a collection of bibliographic records. Details of the system are given and results discussed. The benefits of exploiting the resemblance between the two systems are examined.

INTRODUCTION

It is not often realized how close the relationship is between automatic indexing using a thesaurus, on the one hand, and automatic authority control, on the other. Making the connection is worthwhile for many reasons. The first has to do with terminology. Though one would be naive to hope for a reduction in specialized vocabulary, it is helpful to appreciate that what is called a thesaurus in one application is referred to as an authority file in the other; that the two have virtually the same structure, similar working parts, and play the same role in controlling the content of fields in a bibliographic file in their creation and, at least potentially, during retrievals by users.

A second reason emerges in system development. Below we discuss the various ways that a library can implement authority control. They range from a fully manual system, where the authority file exists only in card form, to online, automatic authority management. There are intermediate points as well. For each of the automated implementations, the system investment in software can be great. Recognition of the close parallel in function of these two library needs allows for parallel development of software for any of these stages.

A third reason looks to the future. Successful system-patron interaction

Manuscript received April 1981; accepted September 1981.

ought not to depend upon a patron's knowledge of the authorized entry forms currently in use for a library. First, the concept of a controlled vocabulary is far too narrow: authority control should encompass *all* fields available for searching. But the patron need not be aware of complicating details: substitutions of recognized variants for authorized forms ought to be carried out automatically during patron retrievals (with due regard, of course, for the intent of the patron).

This article describes a project in authority control in a specialized system environment, one that is increasingly typical in many of its features. The file of records is relatively small, currently below 10,000, and has a potential for growth not exceeding 100,000. The collection, derived from the Annabel Morris Buchanan Collection of American religious tune books at the University of North Carolina (Chapel Hill) Music Library, has many similarities with standard book collections, but its details vary greatly and cataloging conventions have been developed locally. Its use for scholarly research is similar to that for any standard collection of bibliographic records.

A great many such nonstandard collections exist—the morgue file in a newspaper, machine-readable data files, even properties marketed by cooperatives of real estate agencies. Developing automated retrieval systems for such collections are similar enterprises, sharing similar goals and problems. In particular, all require extensive authority control similar to that required by a tune-book collection.

The important feature of the method of authority control described here, one that makes it likely to be of interest to others, is its use of the same structures and software that are used for general vocabulary control. The three major software components we will refer to below are: thesaurus maintenance, automatic indexing, and automatic updating. These components antedated our effort to implement a similar system for authority control. When the problems that dealt with authority control per se were investigated, it was discovered that the system already available for subject control could be used exactly as it stood for authority control as well. Initial experiments confirmed this relationship.[1]

Authority Control and Automatic Indexing

Automatic authority control has been approached largely as a unique problem requiring special software development for its implementation. But authority control shares common ground with automatic subject indexing. Both are term-matching activities based on a list of preferred terms plus a much larger list of match terms. Each preferred term is tied to a number of match terms, but each match term is tied to only one preferred term. In the indexing environment, document text is examined for certain terms; these "free text" (uncontrolled vocabulary) terms are tied to equivalent (controlled vocabulary) terms in a thesaurus. When an uncontrolled vocabulary term is encountered in a document, its associated controlled

vocabulary term is posted to the document as a descriptor. In authority control, document text is also examined for certain terms, e.g., author names. These "free-text" author names (i.e., names just as they appear on a title page) are tied to their authoritative name form (controlled vocabulary) in an authority file. When a "free-text" author name is encountered, the authoritative name is posted to the document or book (i.e., assigned as a heading or entry point).

An automatic authority control system, then, is realizable by applying standard automatic subject-indexing software, which exploits the resemblance between the two processes. The input would consist of a thesaurus (in this case, an authority file) and bibliographic records; the indexing discovers matches between the list of possible terms in the thesaurus (variants of author names) with the "free-text" terms (title-page author names), and posts the appropriate controlled thesaurus terms (authoritative author name form) whenever a match occurs. (See figure 1.)

THE TUNE-BOOK PROJECT

An experimental version of an authority control system using automatic indexing was implemented to test the feasibility of automatic indexing as

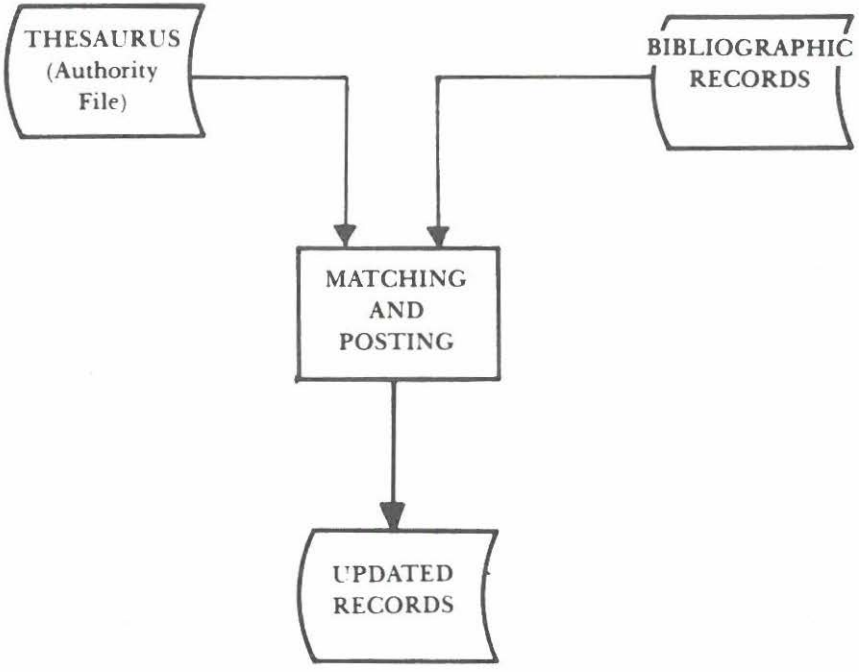

Fig. 1. Authority Control by Indexing.

the core process for authority control. The goal was automatic authority control for the Buchanan Collection index, the first step in work on a more comprehensive project, an index of American religious tune books, in particular, the shape-note tune books.

For the study of American cultural and musical history it is important to be able to trace the dissemination of these hymn tunes and texts, but the absence of a comprehensive index of American hymn tune books severely constrains such studies. Many factors have discouraged scholars from constructing an index, among them the magnitude of the repertory. Using computers to sort, file, and print reduces many of the problems associated with the size of the repertory, but does not address those created by the diverse forms of names and texts used by the tune-book compilers. Correct hymn titles and especially accurate composer attributions were not important to the compilers of the tune books. Consequently, although many tune-book compilers *did* attempt to indicate who had composed the work, the names of the composers appeared in various forms. For example, the name "Israel Holdroyd" might appear as simply "Holdrad" or "Holdrayd" with no first name given, or a first initial might be added, or an abbreviated first name, such as "Is." might be used with one of several forms of the family name. Automatic authority control over these names is necessary to the study of this collection, since only automatic means can address the problems of magnitude encountered in approaching the index as a whole.

The database now contains about 6,000 records for these tune books. They are stored in MARC format with variable-length fields giving a variety of information about each tune.

Creation of the Authority File

A thesaurus of authority records for the Buchanan Collection was manually created and placed in an online file. The initial authority file comprises a selection of composers whose names are present in conflicting forms in the present database. These were obtained by analyzing the file sorted by tune names, noting those tunes for which it appeared that the name of the same composer was given in more than one form. All forms of the name found were entered on cards along with the name of the tune (or tunes) through which the relationship was established. We used an explicit algorithm as a guide in determining which names were actually forms of the same name (see appendix for details). This process resulted in a list of 266 distinct composers, each with one to four different name forms. All were compared with the list sorted by composers, noting additional forms. These names were then checked in several reference works, and authoritative forms (with dates) were established when possible.

IMPLEMENTATION

Software Systems

File processing for the tune records and the authority thesaurus was

accomplished using a local software product, Bibliographic/MARC Processing System (BPS). BPS is a general-purpose software package for the manipulation of MARC-format records. This experiment used BPS subsystems for creation of MARC-format records, sorting and formatting, and file updating (i.e., updating a master file with the contents of a transaction file).

The automatic indexing program used here was intended as part of a thesaurus-based document query system.[2] It is compatible with BPS, but utilizes generalized automatic indexing principles—its compatibility depends only on properly formatted thesaurus and bibliographic records. It includes file-processing programs for the thesaurus (authority file) and the bibliographic records (tune records) and a matching program that performs the indexing. Posting of the authoritative name forms to the proper MARC record is done with standard BPS updating procedures using output from the matching program.

Automatic Authority Control Process

As input the system uses a thesaurus and the text of fields selected from MARC-format document records. The thesaurus consists of pairs of terms: the first of each pair is the term searched for in a document, the second is the authority term assigned to the document, whenever the first term is found. Figure 2 gives examples.

The text may be abstracts, titles, or the contents of any field selected from the documents for authority control. In this case, the text is derived from the composer field; for authority work in general, any field requiring authority control would be input.

The first step in authority control is as follows. The text sample and a stop-word list are input to the initial text-processing program. The incom-

VARIANT		AUTHORITY FORM
Cole, J.	/	Cole, John 1774-1855
Clarke, Thos.	/	Clark, Thomas
Coles, Geo.	/	Coles, George
Cuzens, B.	/	Cuzens, Benjamin
Ball, S. B.	/	Ball, R. F.
Holrad	/	Holdroyd, Israel
Holroyd	/	Holdroyd, Israel
.		.
.		.
.		.

Fig. 2. Thesaurus/Authority File Format.

ing text (in this case, composer names) is separated into individual words. The stop-word list is used to remove designated words from the input, which in authority control might be titles of address and so on—terms such as "Miss," "Elder," or "Reverend." (Automatic indexing uses the stop-word list to eliminate similarly noncontributory terms, such as conjunctions and prepositions.) The processing program can also convert plurals to singulars if desired. The purpose of this option in automatic indexing is to pare down variants in order to increase matches by standardizing term forms. However, plurals are not converted in authority control, since names are usually distinguished from one another by their full forms. The processing produces a list of individual terms. Each term is given once along with the number of words in the term, then broken up with the document number attached to each piece.

The thesaurus authority records are edited by the thesaurus processing program into specially formatted matched pairs of variant and authoritative forms. Input is the match-term/variant-term file (figure 2) and the same stop-word list used for document processing. The stop-word list eliminates all unwanted words in the list of variant name forms. Output is a file containing all possible name forms (variants), the number of terms in each name and their positions in the name, and the authoritative name form, as in figure 3.

Next the two files are used as input to a matching program that creates an inverted file of the processed document text, then compares each match term from the prepared thesaurus with the inverted file. A match is discovered according to one of the following criteria:

1. *Exact match:* Match term and document term are the same words, in the same order, and adjacent.
2. *Stop word exact match:* Words are the same in match term and in document term, and in order, but deleted stop words may intervene between words in the document term.
3. *Any order match:* Term must be the same words and adjacent (i.e., without intervening words) and may be in any order.

VARIANT	#WORDS	RELATIVE POSITION	AUTHORITY FORM
Hastings, Thos.	2	1 2	Hastings, Thomas 1784-1872
Hastings, Thos Dr	3	1 2 3	Hastings, Thomas 1784-1872
Holdrad	1	1	Holdroyd, Israel
Holdroyd	1	1	Holdroyd, Israel
Houser, W	2	1 2	Hauser, William 1812-1880
.	.	.	.
.	.	.	.
.	.	.	.

Fig. 3. Processed Authority File.

4. *Stop word any order match:* Terms must have the same words and in any order, but intervening stop words are ignored.
5. *Any match:* Any word of the match term may be in any part of the document text in any order.

These match criteria are similar in intent to the criteria for deciding composer-variant forms/composer-authority form match mentioned above and presented in the appendix. An interesting possibility is to use such match criteria to discover variant author name forms in creating the authority file, since many variant forms result only from misspellings, title attributions, and so on. Pseudonyms would not be detected, but such a procedure would be useful in collating forms morphologically similar.

The experiment used criterion two, one of the most restrictive; the "free-text" composer name must match exactly and with its parts in the same order (except that stop words, such as "Miss" or "Elder," may intervene) as the variant author name before an authoritative form is posted. This seems the most reasonable choice for this project; presumably more flexibility could be achieved by adding criteria to the match process or by allowing Boolean combinations of criteria analogous to those outlined in the appendix.

The final output from the match module is three files: a print file of all match terms, a file of all unmatched authority names, and a file constructed for the update of the bibliographic records, giving the document and field to be updated and the update term.

The print file is a record printed for each term matched. The record gives the variant form matched, its field type, the proper authoritative name form as given in the thesaurus, and the identifier numbers of the documents in which the term is found. Field type is an identifying code assigned to each term in the prepared thesaurus, not necessarily the same as those identifiers in the MARC-format authority file; here, the field type is Preferred Composer Name (PCN). An example of the printed output file is in figure 4.

The update file is for use in an update program that posts the authoritative name form assigned by the indexing. It contains the document identifier number in which a match was found, the field type of term found (PCN), and the authoritative name form. The update program uses this file to add the authoritative composer name form as a new bibliographic data field to the appropriate bibliographic record, assigning as a field identifier the field type identifier accompanying it. Figure 5 gives the new records with added fields.

During the update process, a file containing all records *not* receiving a new authority-name field is generated. These records may contain a new variant of an authoritative name already in the file or a name altogether new to the file; in either case the unmatched author name would have to be added to the authority file and tied to an authoritative name form. The output also assists in tracing erroneous name-form assignments.

```
MATCH TERM: Walker              TYPE: PCN
AUTHORITY TERM: Walker, William 1809-1875
LOCATIONS: AA-1059, AA-1144, AA-1273,...

MATCH TERM: Davidson            TYPE: PCN
AUTHORITY TERM: Davisson, Anaias 1780-1857
LOCATIONS: AA-1035

MATCH TERM: Handel              TYPE: PCN
AUTHORITY TERM: Handel, George Frideric 1685-1759
LOCATIONS: AK-1045, AA-2093

MATCH TERM: Everett             TYPE: PCN
AUTHORITY TERM: Everett, E. G.
LOCATIONS: AA-1015, AA-1090, AA-1105, A1-1023, AK-1060,
           AK-1111, AI-1397...

MATCH TERM: Pond                TYPE: PCN
AUTHORITY TERM: Pond, Sylvanus Billings 1792-1871
LOCATIONS: AB-1054, AB-1660, AD-1248, AQ-1336, ...
                        .
                        .
                        .
```

Fig. 4. Update File.

Results

Table 1 gives some statistics on the experimental runs. In the 5,788 bibliographic records, 760 distinct composer names were present, the remainder (one composer per record) being duplicate forms; many of these are simply "anon," where the composer was not known. Earlier test runs on a subset of the file had fewer duplicates, and additions to the full database show few new composer name forms. Thus the database is nearing a stable state with an exhaustive list of composers; this stability contrib-

Table 1. Implementation Statistics

File Statistics:	
Total number of bibliographic records	5,788
Number of composer names in biblio records	760
Average number of compositions per composer	13.2
Total number of authority name forms (in authority file)	266
Total number of variant and authority names (in authority file)	599
Run Statistics:	
Total number of variant thesaurus names matched	372
Total number of variant thesaurus names unmatched	213
Average number of documents per matched term	5.87
Average number of documents per term	3.61
Total number of records updated by authority form	2,110

```
DOC ID:     AF-1147
ANTHOLOGY: The Union Harmony
IMPRINT:    selected by George Hendrickson
TUNE NAME: jerusalem
FIRST LINE:Jesus, my all to heav'n is gone,
PCN:        Walker, William 1809-1875
COMPOSER:   Walker, Wm


DOC ID:     AA-1358
ANTHOLOGY: The Sacred harp
IMPRINT:    by B. F. White, E. J. King [and D.P. White]---
                    4th ed.---Atalnta : D. P. Byrd, 1870
TUNE NAME: the hill of zion
FIRST LINE:The Hill of Zion yields,
PCN:        White, Benjamin Franklin 1800-1879
COMPOSER:   White, B. F.


DOC ID:     AB-1100
ANTHOLOGY: The Dulcimer
IMPRINT:    or, The New York collection of sacred music / by
                    I. B. Woodbury. --- New York :
                    F. J. Huntington
TUNE NAME: Carson
FIRST LINE:Jesus and shall it ever be,
PCN:        Bradbury, William Batchelder 1816-1868
COMPOSER:   Br, W. B.
```

Fig. 5. Updated Records.

utes to decreasing errors and fewer unmatched composer names in the automated authority control process.

The total number of thesaurus records matched applies to variant forms, authoritative forms (matching occurs for these also), and for those few forms that have no variants. The unmatched terms (213) are largely variants not in the database but gleaned from reference sources in anticipation of their occurrence, and authority forms, most of which do not occur in the database. The 2,110 matched represent the total number of composer names matched of the original 5,788 names. Most of the unmatched names are the "anon" entries (more than 2,000); the remainder are unanticipated forms not detected in the initial manual construction of the authority file. These unanticipated forms become new variants added to the authority file as described above.

CONCLUSIONS

Automated authority control as presented here has a number of advantages, either for libraries with their own processing facilities or for the management of information collections outside the standard library environment. Unifying the processes of subject control and authority control by using the same procedures and software for both simplifies the tasks of

systems personnel and information managers. Where catalog access is online, the patron benefits by applying subject access facilities to other searches. Ideally, substitutions for all variants would occur automatically, accompanied by an alert to the patron where it was felt necessary. At a minimum, the same command structure would be available for referencing names as would be normally available for consulting an online thesaurus. In either case, the difficulties of the patron are reduced, both in comprehending how the system works, and in acquiring a facility for using system commands.

REFERENCES

1. Gordon Ellyson Jessee, "Authority Control: A Study of the Concept and Its Implementation Using an Automated Indexing System" (Master's paper, School of Library Science, University of North Carolina at Chapel Hill, 1980).
2. Margaret S. Strode, "Automatic Indexing Using a Thesaurus" (Master's thesis, Department of Computer Science, University of North Carolina at Chapel Hill, 1977).

APPENDIX

Rules for Decisions on Similar Names

The following conditions may exist:

A = identical tune name
B = identical surname
C = identical first initial
D = same first letter of surname and close match of the rest of the surname. (55 percent match of latters in content, not in order. Such a similarity is presumed to represent a similarity in sound.)
E = similar tune name (same criteria as in D for percentage of match). EXCEPTION: words "new" and "old" cancel any presumed relation between similar tune names.
F = information in CMP subfield *x* field is identical in content

The following combinations of conditions indicate the same person, expressed in decreasing order of reliability:

1. A & B
2. B & C
3. A & D
4. C & D
5. B & E
6. C & D & E
7. D & E
8. F & (B or D)

Note: points seven and eight are regarded as tentative, and matches using these combinations are flagged for later checking.

Martin Dillon is associate professor of library science at the University of North Carolina at Chapel Hill. **Rebecca C. Knight** is administrative services librarian at Wichita State University, Wichita, Kansas. **Margaret F. Lospinuso** is music librarian at the University of North Carolina at Chapel Hill. **John Ulmschneider** is library associate at the National Library of Medicine.