

FACILITATING RETRIEVAL OF FICTION WORKS IN ONLINE CATALOGS

Allyson Carlyle

Assistant Professor
Information School
University of Washington
Box 352840
Seattle, WA 98195-2840
Phone: 206 / 543-1887
Fax: 206 / 616-3152
Email: acarlyle@u.washington.edu

Sara Ranger

Information School
University of Washington
Box 352840
Seattle, WA 98195-2840
Phone: 206 / 221-6444
Email: saranger@u.washington.edu

ABSTRACT

Systematic retrieval and meaningful display of bibliographic records that represent works of fiction appearing in many editions such as Lewis Carroll's *Alice's Adventures in Wonderland* have not been achieved in online catalogs or other information retrieval systems. One means of achieving systematic retrieval and meaningful display of these works is to discover attributes that could be used to classify records as belonging to a particular work set. Author name, title, and Library of Congress classification number may be used to successfully classify records as belonging to a particular work set. These attributes may be discovered automatically using software that would employ existing information in both authority and bibliographic records. A simulation of automatic attribute identification and record classification records shows that automatic methods of work record classification may be largely successful in achieving their purpose.

1. INTRODUCTION

Manually created document representations such as library cataloging records or indexing records are constructed specifically and painstakingly to facilitate the retrieval and display of documents of interest to users of information systems. However, in current systems even manually created document representations may not facilitate retrieval as much as might be desirable or helpful to a user. A search in a library catalog for a work of fiction that has appeared in many editions illustrates this well. For example, a title search on "alice in wonderland" may not retrieve editions of Lewis Carroll's work published under the titles *Alice's Adventures in Wonderland*, *The*

Annotated Alice: Alice's Adventures in Wonderland and Through the Looking Glass, The Complete Illustrated Works of Lewis Carroll or Alicia en el País de las Maravillas. A search such as this is also likely to result in a strictly alphabetical display of titles, which may or may not consist solely of Lewis Carroll's work.

Automatic methods of enhancing retrieval and display of manually created records in information retrieval systems have been investigated. For example, Larson (1991) experimented with clustering records based on classification numbers in existing cataloging records to enhance retrieval and display of results. McGarry and Svenonius (1991) also explored automatic methods of improving catalog results display using information existing in records. They proposed compression of LC subject headings with multiple subdivisions existing in cataloging records in an attempt to shorten extremely long subject heading displays.

As is the case in prior investigations, the study proposed here is also conducted using information already available in bibliographic records. Consequently, no further investment need be made in record construction. Rather, emphasis is placed here on developing more sophisticated automatic means for effecting the classification of sets of work records. The automatic assembly of work record sets has the potential to form the basis of intelligible and intelligent displays of retrieval results.

Research to discover the effect that incomplete or inconsistent search results has on users, such as the result of a search on *Alice in Wonderland* described above, has not been done despite the fact that the such results demonstrate the failure of current online catalogs to fulfill the second objective. The second objective requires the catalog to perform two functions. First, an effective catalog facilitates a catalog user's ability to ascertain "which editions of a particular work are in the library (International Federation of Library Associations, 1971, xiii). Second, it "relate[s] and display[s] together the editions which a library has of a given work" (Lubetzky, 1960, ix). If implemented as articulated, the second objective would help ensure that users coming to libraries knowing only that they are interested in a particular work such as Carroll's *Alice in Wonderland* would be made aware of all of the different editions and related works available so that they may select the one or ones that best meet their needs.

The research conducted here attempts to discover the effectiveness of using classification to retrieve as well as "relate and display" records for items representing works published in many editions. In information retrieval systems designed to achieve classification, attributes that identify a record as an edition of a particular work could be used to assemble records representing a particular work into a work class or set. A work class could then be retrieved and displayed in response to a query.

Classification of work records may be accomplished in two ways. First, it may be done manually by creating links between all of the records representing editions of the work. Second, it may be done automatically by creating a program that would identify work attributes and use those attributes to classify work records. Manual creation of links is time-consuming and, thus, costly. In the current era of retrenchment in library cataloging, it would be likely to be low on an agenda that is largely driven by cost reduction. It is thus highly desirable to develop an innovative automatic method for improving retrieval and display, as is proposed in this study.

This paper presents the results of a research project investigating the extent to which records representing fiction works published in many editions may be identified and classified automatically, thus creating a means of retrieving and displaying them together as a group. The specific research questions pursued are: What attributes of bibliographic records may be used to automatically identify and classify bibliographic records representing editions of fiction works published in many editions? How might the attributes used for identification be discovered automatically? To what extent is automatic classification using these attributes successful in actually identifying and classing bibliographic records representing editions of fiction works published in many editions?

2. RATIONALE FOR THE PROJECT / RELATED RESEARCH

Previous research indicates that access to works that have been published in many editions may be problematic for users. Carlyle (1996) reports the lack of organization in records retrieved in author searches in online catalogs. Records representing editions of the works of an author are frequently not listed together in any content or relationship-based order, but are intermingled haphazardly as a result of strict title alphabetization. Such arrangements do little to help users identify the works retrieved in their searches. Furthermore, as previously noted, title searches frequently retrieve *particular* titles, but because works of fiction manifested in many editions are often published under multiple titles, the retrieval of these works is incomplete.

Works published in many editions may also be problematic because catalog users report difficulty with large retrieval sets (Matthews, Lawrence, & Ferguson, 1983). A possible outcome of this difficulty is that they may not bother to look through all of the records retrieved in a search, particularly a search that retrieves many screens of results (Wiberley, Daugherty, & Danowski, 1995). A long-term research project led by Kilgour (1995) seeks to increase the effectiveness of known-item searching by limiting searches to a combination of author surnames and title keywords. Although these limitations do tend to result in small retrieve sets, the question as to whether or not such sets contain all and only records associated with the work sought is not addressed in this research. In addition, although his findings indicate that 92.8 percent of surname/title keyword combination searches lead to small retrieval sets, the searches that retrieve large sets remain a problem, particularly searches for well-known and frequently-sought works. For example, one of the works investigated in Kilgour's research (1995) is Charles Darwin's *Origin of Species*, which, when searched in the OCLC Online Union Catalog (OLUC) at the time, retrieved 354 records.

One of the solutions to the problem of large retrieval sets is to summarize results by classifying retrieved items based on attributes that characterize the results (e.g., Carlyle & Summerlin, 2000). O'Neill develops a record identification program for an OCLC project designed to classify bibliographic records representing English-language editions of fiction works published in many editions (1994). The software developed consists of algorithms that matched the following attributes: author name (MARC 100 field¹, subfield a), title proper (245 field, subfield a) and uniform title (240 field, subfield a). O'Neill's project is a first step toward automatic classification of work records. However, although reporting success, the project does not actually report the

¹ All fields discussed in this paper are MARC fields. For reference, consult the MARC 21 format, available at: <http://lcweb.loc.gov/marc/>.

extent to which it was successful, nor does it report any research on records that may have been missed. Consequently, more work needs to be done to investigate automatic classification as a method of assembling work records in library catalogs.

The project reported in this paper selects four fiction works identified in O'Neill's original research for a detailed analysis. It extends O'Neill's research by identifying and analyzing records that either failed to be identified in the 1994 project or were not included, for example, non-English-language versions and sound recordings. One goal of the project is to expand the attributes used in O'Neill's project to make record identification more successful. Another is to include attributes that can be discovered with minimal or no human effort.

3. METHODOLOGY

In this study, records representing editions of four voluminous fiction works were analyzed manually to discover the attributes present in them that could be used for automatic work identification and classification, and the extent to which those attributes would be successful in identifying records comprising a work set. The four fiction works analyzed were:

- 1) *Bleak House*, Charles Dickens
- 2) *Kidnapped*, Robert Louis Stevenson
- 3) *The Three Musketeers*, Alexandre Dumas
- 4) *Little Women*, Louisa May Alcott.

These works were selected from a list of fiction works published in many editions discovered in the OCLC OLUC through automatic procedures in the O'Neill project. The study was limited to records representing *editions* of these works only; it did not investigate records for works related to these works, such as children's adaptations, videorecording versions, criticism, etc. This limit was imposed primarily because of time constraints on the project; the number of records to be analyzed would have increased by two, three, or even more times if related works had been included. The study was limited to works of fiction because the number of works that could be included in the study was small, and it was assumed that the variability among different types of works such as fiction and non-fiction could be large. Furthermore, the selection of a particular type of work allows comparisons to be made when further research is done investigating other types of works.

The records analyzed in this study came from two sources. The first source was the OCLC Office of Research. Records from this source represent all of the English language editions identified through the use of the automatic work-record identification program developed in the O'Neill project. The second source of records was the OLUC. Additional records for editions not discovered in the O'Neill project, including records for non-English language and non-book editions, were retrieved manually for this study, using a variety of strategies. First, authority records associated with the works studied were retrieved. Variant titles identified in the authority records were searched in the OLUC. Second, each work was searched in the National Union Catalog (Library of Congress, 1953- ; 1968-1981) to discover additional variant titles to be used in searching the OLUC. The following strategies were also used in searching the OLUC: last name author and title word keyword searches, and general author searches.

For purposes of this study, classification is defined as a process comprised of two interrelated actions: 1) the identification of a bibliographic record as an edition of a particular work; and 2) the assembly of those records into a single set. Defined in this way, classification involves the selection of an attribute or a combination of attributes that identify the record as a member of a particular work class, and the subsequent assembly of the records exemplifying those attributes to comprise the class. O'Neill selected two attribute combinations to assemble records representing editions of fiction works: (1) author name (100 field, subfield a) and title proper (245 field, subfield a) and (2) author name (100 field, subfield a) and uniform title (240 field, subfield a). Some variations in title proper were manually identified and then incorporated into the work identification program.

Because the goals of this study were somewhat different from that of the O'Neill project, this study began with an analysis of records that were known to comprise the work set. Each record identified for the study was analyzed manually to discover attributes existing in the record that could successfully identify it as a member of a work set, and classify it with other records for editions of the same work.

4. RESULTS

4.1 Attribute Identification

The first research question investigates the attributes present in bibliographic records that may be used to automatically identify bibliographic records representing editions of fiction works published in many editions. Attributes identifying a work that were discovered in the record analysis consist of specific MARC bibliographic fields that have specific field content. Attributes leading to the identification of works discovered in the analysis include:

- Author name – an author name as it appears in a subfield *a* of a 100 field or a 400 field in a MARC authority record (*Name*);
- Standard title – a title as it appears in a subfield *t* of a 100 field or a 400 field in a MARC authority record (*Title*);
- Library of Congress classification number signifying a particular work of fiction (*LCC#*).

Single attributes and attribute combinations that could be used to automatically identify and classify bibliographic records as members of a work set are limited to their appearance in specific fields. Those used in this research include:

- The combination of two attributes in two separate fields: (1) *Name* in a 100 field, subfield *a* and (2) *Title* in a 240, 245, 246, or 740 field, subfield *a* (*Name + Title*);

- The combination of two attributes in a single field: (1) *Name* and (2) *Title* in a 700 (name-title added entry) field, subfields *a* and *t* respectively, that also has a second indicator of 2² (*Name-Title Added Entry*);
- A single attribute in a single field: (1) *LCC#* in an 050 or 090 field (*LCC*).

This research expands the number of attributes used to identify work records in the O'Neill project by making use of information about authors and works present in authority records. Cross reference forms of name and title (400 field forms) identified in name and work authority records are used to expand the search for relevant authors and titles. This research also adds LC classification number as an attribute that identifies an edition of a work. The LC classification number may be particularly useful for identifying records such as translations, which may contain non-standard author or title information. Attribute combinations are also expanded by searching the 246 and 740 title fields in bibliographic records in addition to 240 and 245 fields; and by searching for the presence of *name* and *title* in 700 fields.

4.2 Automatic Attribute Identification

The second research question looks at how attributes used for identification could be discovered automatically. Automatic identification of work attributes may be accomplished two ways: first, using attributes harvested from name authority records in the Library of Congress Name Authority File (NAF), a file of authority records containing information about controlled author and work names; and second, using attributes harvested directly from bibliographic records.

Name and *Title* attributes could be harvested from NAF authority records. Works that appear in many editions are usually represented in work authority records in the NAF. Once an initial work-clustering program such as the one used in the O'Neill project gathered an initial selection of sets of bibliographic records representing works, author names appearing in 100 fields of those records could be automatically searched in the NAF. Names appearing in the a subfields of 100 and 400 fields in those authority records could be automatically collected, searched, and matched against 100 and 700 fields in bibliographic records. Titles matching 245 field titles appearing in t subfields of 100 and 400 fields in those records could be automatically collected, searched, and matched against 240, 245, 246, and 740 fields in bibliographic records.

LCC# could be harvested directly from 050 and 090 fields in bibliographic records. Because not all works are represented by single *LCC#*s, an algorithm would have to be developed that would: (1) identify the most commonly occurring *LCC#* in 050 and 090 fields in the initial work record set; (2) search that *LCC#* in the catalog; (3) accept an *LCC#* as a work attribute only if the majority of records retrieved contained the *name* and *title* attributes of the work sought.

4.3 Automatic Clustering Results

Automatic clustering was simulated in this project using the bibliographic records representing editions of the works studied. This was done by manually analyzing each record and recording the

² The second indicator of 2 is necessary because it unambiguously identifies the presence of a work that is contained within the item being cataloged; the name-title added entry field is also used to indicate that the work cataloged is related to another work.

attribute or attributes present in each record. Many records that represented works related to the works studied were incidentally retrieved in the course of assembling records for the study.³ Some of these records were retrieved by the O'Neill program; some were retrieved in the search for relevant records in the OLUC. These records pose a problem for automatic clustering because they frequently contain the same attribute combinations that work records contain, and as a result could be misidentified as belonging to the work set. The success of a work-clustering program is thus a product of two things: its ability to identify records that belong in the work set *and* its ability to exclude records that do not belong in it.

In the simulation, records representing works related to the work studied were identified and thus excluded using the following criteria:

- Presence of the stems “dramat*”, “adapt*”, or the word “paraphrase” in a 245, 500, or 520 field (excludes dramatizations, adaptations, and paraphrases);
- Presence of the stem “simpl*” or the words “retold” or “retelling” in a 245 field (excludes simplifications and retellings);
- Presence of a “g”(projected medium, e.g., videorecording or motion picture) or “j” (musical sound recording) in the Type fixed field;
- Presence of the stem “comic*” in a 650 field (excludes comic strip versions);
- Presence of the term “kit” in a 245 subfield h (excludes kits, which frequently contain adapted versions of texts);
- Absence of a 100 field or presence of a 100 field with a name other than the name associated with the work in question, when a *Name-Title Added Entry* attribute combination was not present.

While these criteria worked well for the four works studied here, they will need refinement in future research, as it is possible that they would exclude too many records. For instance, the stem “comic” in a 650 field could exclude a record for an edition of a work assigned the subject heading “Comic literature”, which should otherwise be included as part of a work set.

Results of the research simulating the identification of work records (Tables 1-4) show that the majority of records for all four works can be identified using the attributes and attribute combinations described in the *Automatic Attribute Identification* section above. Eighty-six to 98% of all records for a work are correctly identified using the methods described here, with only two to 14% unidentifiable.⁴ Between zero and 15 records are incorrectly identified as members of a particular work set.

³ Records representing works related to the works studied were not systematically searched for in the OLUC and included in the pool of records assembled for the study; thus, the results of the study indicating the number of records mis-identified as work records may be higher than is reported here.

⁴ In an actual catalog, the percentage of unidentifiable records is likely to be somewhat higher, given the possibility of totally unrelated records containing the attributes being searched.

Table 1. *Bleak House*

Attribute Combinations	No. of Records	% Identified
<i>Name + Title</i>	363	96
<i>Name-Title Added Entry</i>	3	1
<i>LCC</i>	2	1
Correctly ID'd as work	368	98
Not identifiable:	11	3
• <i>Title not present</i>	11	3
TOTAL Work Records	379	101*
Misidentified as work	8	NA
TOTAL Records Analyzed	387	NA

*Not 100% due to rounding error.

Table 2. *Kidnapped*

Attribute Combinations	No. of Records	% Identified
<i>Name + Title</i>	419	95
<i>Name-Title Added Entry</i>	2	0*
<i>LCC</i>	6	2
Correctly ID'd as work	427	97
Not identifiable:	12	3
• <i>Name not present</i>	1	0
• <i>Title not present</i>	11	3
TOTAL Work Records	439	100
Misidentified as work	0	NA
TOTAL Records Analyzed	439	NA

* 0.5%

Table 3. *Little Women*

Attribute Combinations	No. of Records	% Identified
<i>Name + Title</i>	635	93
<i>Name-Title Added Entry</i>	3	0*
<i>LCC</i>	8	1
Correctly ID'd as work	646	94
Not identifiable:	39	6
• <i>Title not present</i>	33	5
• <i>Name not present</i>	4	1
• <i>Title spelled incorrectly</i>	2	0**
TOTAL Work Records	685	100
Misidentified as work	10	NA
TOTAL Records Analyzed	695	NA

* 0.4%; ** 0.3%

Table 4. *Three Musketeers*

Attribute Combinations	No. of Records	% Identified
<i>Name + Title</i>	440	65
<i>Name-Title Added Entry</i>	1	0*
<i>LCC</i>	145	21
Correctly ID'd as work	586	86
Not identifiable:	93	14
• <i>Title not present</i>	84	13
• <i>Name not present</i>	1	0**
• <i>Title spelled incorrectly</i>	8	1
TOTAL Work Records	679	100
Misidentified as work	15	NA
TOTAL Records Analyzed	694	NA

* 0.1%; ** 0.1%

In most cases, *Name+Title* is enough to correctly identify and classify a record as belonging to a particular work set. *LCC* is an effective attribute combination for identifying work records that do not contain a correct *Name+Title*. The ability of the *LCC* to identify a work record is particularly important for *The Three Musketeers*, which is comprised of records that contain many varying or incorrect *names* and *titles*, thus, limiting the ability of *Name+Title* to identify the record.

In the results reported in Tables 1-4, only a single attribute combination was tallied, even though many records contained more than one attribute combination that would allow them to be classified as members of a particular work set. For example, some records contained both *LCC* and *Name+Title*. Preference was given to authorized forms (authority record 100 forms) of name that appeared in 100 and 700 fields, respectively; and to authorized forms of title that appeared in 240, 245, 246, and 740 fields, respectively. Cross reference forms of name and title (appearing in 400 fields in authority records) were tallied as identifiers only if authorized forms did not appear in the records.

5. DISCUSSION

The results of this exploratory study of the ability of bibliographic records to be automatically identified and classified as members of a particular work set indicate that further research in this area is desirable. The success of a small number of automatically identifiable attributes used together in attribute combinations to automatically classify work records for the four works studied varies from 86 to 98 percent, which is relatively high. If one disregards *The Three Musketeers*, success for the three remaining works is 94-98%.

One reason why *Three Musketeers* records may not have been as easily classifiable as records for the other works is that it is a work published originally in French and has been translated more frequently than the other works. Of 679 records classifiable as editions of *The Three Musketeers*, 287 records are in English.

Although a detailed analysis has not yet been completed for the records that were not classifiable as work records, preliminary analysis indicates that records for some foreign language editions not identifiable as editions of the work in question because they do not contain the standardized or uniform title for the work, nor do they contain a cross reference form of the work title identifiable via a work authority record. Other records were not identifiable as members of a work set because of typographic errors and variant author names.

Another measure of success of the automatic classification methodology proposed here is the low number of records misidentified as work records. In fact, all of the records in this category represent editions of works that are related to the works studied; in other words, they are children's adaptations of the works studied, sequels to the works, etc. These records failed to be recognized in the part of the simulation that attempted to exclude all records representing works related to the work studied.

6. CONCLUSION

Anecdotal evidence and common sense suggest that it is likely that works published in many editions are among the most frequently sought works in online catalogs. As discussed above, evidence from previous research suggests that records representing editions of these works are poorly organized in such systems. In light of the findings of this study, a possible reason for the poor organization of records is that inadequate attempts have been made to identify or classify them as members of a particular work set. The results of the research performed in this project suggest that an avenue of exploration for enhancements to online catalogs employing automatic classification would be fruitful. Such enhancements would facilitate the retrieval and display of records representing works of fiction published in many editions. Although the automatic record identification and classification methods developed as a result of this study will not prove adequate for all works, or even all fiction works, nonetheless, the results do provide a foundation for future research aimed at improving retrieval and display in online catalogs and other information retrieval systems.

7. ACKNOWLEDGEMENTS

The OCLC/ALISE Library and Information Science Research Grant Award supported much of the initial data collection and early stages of the research. We would also like to acknowledge Joel Summerlin and Collette Davis for their assistance with data collection, and Lisa M. Fusco for editorial assistance. Comments from the anonymous reviewers are also gratefully acknowledged.

REFERENCES

- Carlyle, A. (1996). Ordering author and work records: An evaluation of collocation in online catalog displays. *Journal of the American Society for Information Science*. 47 (7), 538-554.
- Carlyle, A. & Summerlin, J. (2000). Transforming Catalog Displays: Record Clustering for Works of Fiction. *Dynamism and Stability in Knowledge Organization: Proceedings of the Sixth International ISKO Conference, 10-13 July 2000, Toronto, Canada*. Eds. C. Beghtol, L.C. Howarth, & N. J. Williamson. Wurzburg: ERGON Verlag, 320-326.

Kilgour, F.G. (1995). Effectiveness of surname-title-words searches by scholars. *Journal of the American Society for Information Science*. 46 (2), 146-151.

International Federation of Library Associations. 1971. *Statement of principles adopted at the International Conference on Cataloguing Principles, Paris, October, 1961*. Annotated edition with commentaries and examples by Eva Verona. London: IFLA Committee on Cataloguing.

Larson, R.R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog, *Library Quarterly*, 61 (2), 133-173.

Library of Congress. (1953-). *National Union Catalog*. Washington, D.C.: Library of Congress.

Library of Congress. (1968-1981). *The National Union Catalog, pre-1956 imprints*. London: Mansell.

Lubetzky, S. (1960). *Code of cataloging rules: Author and title entries. An unfinished draft*. American Library Association.

Matthews, J.R., Lawrence, G.S., & Ferguson, D.K., eds. (1983). *Using online catalogs: A nationwide survey: A report of a study sponsored by the Council on Library Resources*. New York: Neal-Schuman.

McGarry, D. & Svenonius, E. (1991). More on improved browsable displays for online subject access. *Information Technology and Libraries*. 10 (3), 185-191.

O'Neill, E.T. (1994). Manifestations of fiction works. *Annual review of OCLC research*. Dublin, OH: OCLC Office of Research, 11-15.
Also available at: <http://www.oclc.org/research/publications/arr/1994/part1/ficworks.htm>

Wiberley, S.E., Daugherty, R.A., & Danowski, J.A. (1995). Displaying online catalog postings: LUIS. *Library Resources & Technical Services*, 39, 247-264.

