Clustering Fiction Works to Improve Online Catalog Displays

Allyson Carlyle
Associate Professor
Information School
University of Washington
MGH 370, Suite 370
Seattle, WA 98195-2840
(206) 543-1887

Final Report
1999 OCLC/ALISE Library and Information Science Research Grant Project

6565 Frantz Road, Dublin, Ohio 43017-3395 USA
http://www.oclc.org/

Suggested citation:
Carlyle, Allyson. 2006. "Clustering Fiction Works to Improve Online Catalog Displays." Final Report of a 1999 OCLC/ALISE Library and Information Science Research Grant Project. Published electronically by OCLC Research. Available online at: http://www.oclc.org/research/grants/reports/1999-carlyle.pdf

Executive Summary

Online catalogs frequently fail in their attempts to retrieve and display bibliographic records representing works of fiction that appear in many manifestations. It is likely that these works are sought frequently by users of online catalogs, so research improving searching and display of these works is desirable. Making improvements that can be done automatically is also desirable. The project reported on here has two parts: first, it proposes automatic methods of expanding the number of work records identified in a search based on multiple attributes present in authority and bibliographic records; second, it proposes improving displays by using automatic clustering methods for shortening long displays.

Four works of similar size (from 386-784 records in the work sets) were studied: *Bleak House*, by Charles Dickens; *Kidnapped*, by Robert Louis Stevenson; *Little Women*, by Louisa May Alcott; and *Three Musketeers*, by Alexandre Dumas. Records representing English-language editions of these works were provided by OCLC using their automatic identification program. Additional records representing other editions not provided by OCLC were identified using various methods, for example, keyword searching on words from author names and titles, and manual searching of the *National Union Catalog*.

The record identification study proposed multiple attributes for use in identifying editions of works. Author names, titles (including uniform titles and titles proper, and name-title added entries), and Library of Congress numbers were used to retrieve multiple types of work records in multiple languages. The results of the record identification methods were highly successful for the English language works studied – 95% of *Bleak House*; 94% of *Kidnapped* records; 77% of *Little Women* records and 94% of *Three Musketeers* records were retrievable using the automatic identification methods proposed.

Recommendations based on the results of the identification study include:

- Augment the current work identification algorithms being used by OCLC, including the FRBR Work Set algorithm, with methods suggested here, especially:
    - use of information contained in authority records, i.e., alternate forms of author name and work title,
    - use of LCC number (especially for works published originally in a language other than English), and
    - use of titles proper harvested from bibliographic records containing uniform titles.
- Do further research of this type on works of non-fiction.
- Do further research to determine the types of situations in which minimal human intervention in a mostly automated process would be fruitful.
- Do further research on including collected works (that is, editions entitled “Collected Works,” etc.) in automatic identification algorithms.
- Do further research on identification of records representing related works.
- Do further research to investigate the nature of works – are difficulties related to identification similar from work to work? if not, how are they different?
- Refine exclusion criteria by testing on a wide variety of works.

The cluster research formulated cluster types suggested by previous research, including user-based research. Relationship-based clusters used in this study include: illustrated editions, non-English language editions, editions in collections of 2 or more works, nonbook format editions, editions with amplifications such as introductions, abridged editions, large print or other orthographically variant editions, editions comprised of parts or selections, and English language editions without other cluster features.

The results for the clustering methods were not as encouraging as results for work identification methods. Editions clustered very unevenly. Each work had an extremely large cluster representing illustrated editions (47-76% of records would fit into an illustrated editions cluster). Many of the other clusters were represented by very few editions. For example, each of the following clusters were near or under 10% for each work: large print and other orthographically variant editions, parts, abridgements, nonbook editions. Most records were, however, clusterable using automatic methods. The largest number of unclusterable records represented collections (46 records), parts (41 records), and amplifications (26 records).

Recommendations based on the results of the cluster study:

- Investigate different clustering or sub-clustering algorithms that would create more evenly sized clusters, for example, using record-set-types based on Svenonius (1988) or FRBR-based displays of expression and manifestation sets.
- Do further research on clustering of records representing related works.
- Do further research to investigate the nature of works – are types of clusters similar from work to work? If not, how are they different?
- Include records representing related works in future research.

1. Introduction

Fiction works associated with many bibliographic records may cause significant problems for users who perform catalog searches for them. In the first major study of online catalog use, scanning long displays was ranked as the fifth most problematic aspect of using online catalogs (Matthews, Lawrence and Ferguson, 1983). A more recent study showed that online catalog users report overload problems when more than 100 records are retrieved (Wiberley, Daugherty and Danowski, 1995).

Catalog displays are often composed of long lists of records that do little to shed light on the composition of the retrieved record set or aid the user in record selection. Record clusters that organize retrieved record sets into intelligible categories may communicate search results more quickly and effectively than current catalog displays. Record clustering is, thus, a critical area of research because it can be used to alleviate the information overload problem inherent in information retrieval systems.

This research analyzed OCLC MARC bibliographic records for fiction works associated with large numbers of manifestations (the terms "manifestation" and "edition" are used interchangeably in this report) to discover how existing records may best be identified and clustered into subgroups automatically.

The research questions addressed in this project are:

- To what extent can existing automatic record identification programs for fiction works be improved to enhance recall?

- To what degree can automatically clustered records shorten long displays and make retrieval sets more intelligible to users?

In a previous study, Edward T. O'Neill identified fiction works held by large numbers of libraries (1994). As a by-product, O'Neill identified fiction works associated with large numbers of manifestation records. These records were identified automatically and were limited to English-language manifestations.

In this project, four fiction works found in O'Neill's research were analyzed. Those works were:

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

These works were selected from the list of highly-manifested fiction works found in the O’Neill project and represent the most highly manifested works associated with approximately 300 manifestations. The selection of works involves a tradeoff; namely, that the larger the work, the fewer works may be selected for study. In the present study, we opted for large works because they were likely to be associated with a number of different types of manifestations. However, it is also possible that different works are associated with different types of manifestations and thus it is desirable to analyze more works rather than fewer. For example, a work such as *Little Women* has as its primary audience children, whereas a work such as *Bleak House* does not. This study does **not** analyze works to determine these types of differences. In addition, we intended to exclude works with more than 300 manifestations so that more works could be integrated into the study, without sacrificing the wide variety of manifestation types. Even so, as it turned out, the works studied were larger than expected (between 386 and 784 manifestations).

In the project proposal, a fifth work was included, Geoffrey Chaucer's *Canterbury Tales.* It was not included because when records were actually received from OCLC, the work set sizes were much larger than had been reported in O'Neill's original research. Over 1,000 records were received for Canterbury Tales, and the number of records received for the first four works was, for each, well over 300. To keep the project within a reasonable timeframe, only the first four works were analyzed for the final project.

This project extends O'Neill's research by identifying and analyzing records that were either unidentified in the 1994 project or were not included intentionally — specifically, non-English-language versions. A major purpose of the analysis was to determine how records missed in O'Neill's automatic record identification program might be included in an expanded program. In addition, the project studied the extent to which manifestation records may be clustered into groups based on item attributes and relationships to improve online catalog displays for works.

The project, outlined in the methodology section below, includes the following steps:

- identify records for manifestations missed or not retrieved in the O'Neill study;
- analyze all records to determine means for identifying them automatically;
- examine all records to create operational definitions for clustering based on item attributes;
- determine the success of operational definitions in automatically identifying records to be included in the clusters;

2. Related Research

O'Neill’s study of fiction works is closely related to this project and, in part, provided the inspiration for it (1994). An object of O'Neill's research was to develop software that would identify all manifestations of a work of English language fiction. In this research, records representing 10,000 widely held fiction works were initially identified using an automatic identification program. Once the 10,000 works were identified, the automatic program was made less restrictive and additional manifestations were discovered. The project reported in this paper expanded the methods employed in the O'Neill study by using a variety of techniques, including using Library of Congress Classification (LCC) numbers and variant titles appearing in authority records associated with works.

In other closely related research, the experimental BOPAC and BOPAC2 systems developed at the University of Bradford identified and clustered work records to simplify catalog displays (Ayres, Nielsen, and Ridley, 1995; Ayres, Nielsen, and Ridley, 1998). In BOPAC2, the more recent project, clustering and subclustering of records related to a work was performed automatically using information available in MARC records. Initial clustering was work-based; that is, sets of records sharing the same author and title information were clustered together. Once a user selected a work set, subsets based on title (e.g., title proper), author (e.g., translator, editor), language, physical format, or source target could be selected. Within a subcluster, records could be further clustered by edition, publisher, date, physical format, series title, or language. The creators of BOPAC2 acknowledged that clustering at the work level was not perfect, so they allowed users to combine clusters at this level if they recognized that two or more clusters represented the same work.

The clusters identified in this study reported in this paper were based in part on a taxonomy of bibliographic relationships developed by Tillett (1987). Several types of relationships included in the Tillett taxonomy were incorporated: equivalence, some derivative, whole-part, and some accompanying relationships. While Tillett was interested in all bibliographic relationships, here, we included only those relationships shared by manifestations of the same work.

In summary, this research project contributes to our understanding in the following ways:

- it proposes novel methods of identifying records representing manifestations of works automatically, and includes use of LCC numbers and work authority records;
- it proposes that relatively reliable clustering of work records may performed automatically, using the existing information in MARC records;
- it proposes that manifestations in multiple languages can be identified and clustered relatively reliably;
- it bases primary record clustering not on document-centered attributes such as publisher and date, but on attributes arising from studies of user behavior, bibliographic relationships among items, and traditional library filing practices;
- it results in highly detailed specifications for identifying and clustering records representing manifestations of works;

3. Part 1: Record Identification

3.1 Methodology

The first part of the research investigated the potential for automatic identification of bibliographic records representing voluminous works. The steps necessary to complete the research are summarized below:

1) assemble records representing editions of works (section 3.2)
2) determine work attributes (section 3.3);
3) create work identifiers using work attributes (3.4)
4) discover how work identifiers can be discovered automatically (3.5)
5) exclude non-work records from the initial assembly of records (3.6)
6) analyze records to see how effective automatically discovered work identifiers are in discovering work records (3.7)

3.2 Assembling sample work records for analysis

The first step in the research was to assemble bibliographic records for all manifestations of the works studied. A manifestation was operationally defined as any edition of a work that should receive the same main entry citation (author and title or uniform title) as the work itself, as prescribed by AACR2R and including application of Chapter 25, Uniform Titles. When the researchers were not sure that a record represented a manifestation of a work or was a related work, it was treated as a related work, and excluded from the study. Details of the exclusion process are described below in section 3.6. Researchers also excluded complete works of the authors in the study, although presumably the works studied would be included in them. However, records representing items that contained multiple works (most often, two works published together in a single volume) that contained the work studied were included.

Manifestations of the works discovered in O'Neill's study of the OCLC Online Union Catalog (OLUC) were limited to English language manifestations and were largely discovered through automatic procedures. OCLC provided these records for this study and they were included in the analysis. In addition, manifestations not previously discovered in the OLUC through automatic procedures were identified using other means, described below. Manifestations in non-English languages were also identified and included in this analysis.

Manifestations not provided by OCLC were identified manually using a variety of strategies:

1) Authority records associated with the works studied were retrieved. Variant titles identified in the authority records were collected and searched in the OLUC.
2) Each work was searched in the National Union Catalog to discover additional variant titles used. These variant titles were also collected and searched in the OLUC.
3) The OLUC was searched using the following strategies:
    a) last name author and title word keyword searches were performed to identify additional manifestations, including those contained in collections;

   b) author searches were performed and records were reviewed to discover additional manifestations.

4) Records for all translations with uniform titles were reviewed to discover variant titles proper that had not been identified in the previous steps.

3.3 Determining Work Attributes

Once all of the work records were assembled, the next task was to review those records to determine the attributes that could identify bibliographic records as representing editions of particular works. Each bibliographic record in the sample was analyzed to determine these attributes: (a) the type of field content and (b) the actual field content itself that would be necessary to identify that record as belonging to a work set (see Appendix 1 and 2 for complete coding information).

Three types of field content were discovered that were necessary to determine whether a record belonged to a work set or not:

1) author name;
2) title;
3) Library of Congress Classification (LCC) number.

Author name and title as identifiers are self-evident, as these comprise the work citation and are the standard identifiers of works. However, three of the works studied (*Bleak House*, *Kidnapped*, and *Three Musketeers*) had unique LCC numbers assigned to them, and, thus, this number could also be used to identify work records.

Field content discovered for author name included both authorized and non-authorized forms of name. Non-authorized forms of name had to be included in the authority records for the author in 400 fields, as these are forms that can be identified automatically. Subfields other than subfield a were not included as part of the author name attribute, to ease the matching constraints. This attribute, composed of authorized and non-authorized forms of name is referred to as "***Name***" in the rest of the report.

Uniform titles identified work records, as well as many alternate titles in English and many other languages. It was necessary, on occasion, to verify that specific titles were, in fact, variant title forms of the works studied. We did this mainly by cross checking with titles proper that were discovered in the record assembly process described in section 3.2 above. For the remaining unclear non-English titles, we verified that they were, in fact, variations of the work titles in question by asking catalogers who had the necessary language expertise. For the remaining unclear English titles, we searched in reference works and on the web for complete English-language bibliographies of the authors that would include all known title variants. Any title that remained unclear was not treated as a variant title and not included in the set of work identifier titles. This attribute, composed of authorized and non-authorized forms of title is referred to as "***Title***" in the rest of the report.

3.4 Creating Work Identifiers

The next step in the research was to create ***work identifiers*** – the data in bibliographic records that would mark those records as belonging to a particular work set. Work identifiers were composed of MARC fields and the work attributes (field content) described above in section 3.3. Work identifiers automatically identified and classified bibliographic records as belonging to a particular work set. Work identifiers are listed below in Table 1.

WORK IDENTIFIERS	DEFINITIONS
Name + Title	Combination of ***Name*** and ***Title*** attributes derived from two separate fields: 1) ***Name*** in a 100 field subfield a and ***Title*** in a 240, 245, 246, or 740 field subfield a OR 2) ***Name*** in a 100 field, subfield a and ***Title*** in a 500 field following the text "Translation of " or "Trans. of" in the beginning of the field.
Name-Title Added Entry	***Name*** and ***Title*** attributes in a 700 field, subfields a and t respectively, with a second indicator of 2.
LCC	A Library of Congress Classification number in an 050 or 090 field.

Table 1. Work Identifiers

Name never occurred in a field other than 100, subfield a. ***Title***, however, could be discovered in several different fields and still identify the record as a work record. One would expect that ***Title*** would occur in a 240 or 245 field, as those are the fields used to name a work. However, when a work is published with another work in a single physical volume, it can be cataloged using a 246 or 740 field with a uniform or variant title instead of using a name-title added entry for a contained work. This would, perhaps, not be recommended as good cataloging practice, but it occurs frequently, nonetheless.

It was necessary to restrict the name-title added entry fields included to those with a second indicator of ***2*** because only a 2 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. Unfortunately, cataloging practice has not been consistent in assigning a second indicator of 2, so some manifestation records that were excluded (see section 3.5 below) undoubtedly should have been included in the work set. Which records belonged in the work set and which represented related works would not be ascertainable without examining original items, which was not possible for this study.

The LCC number is a work attribute that is also a work identifier when discovered in a particular MARC field. Thus, by itself, LCC number in an 050 or 090 field identifies a record as belonging to a particular work set. The LC classification number could be particularly useful for identifying records such as translations, which can contain non-standard author or title information.

3.5 Creating Work Identifiers Automatically

The work identifiers above were determined by manually examining records. However, one object of the research was to determine how effectively work identification could occur when using automatic means. In other words, can work attributes (section 3.3) and work identifiers (section 3.4) be discovered and created using automatic means alone? Thus, the next step in the research was to determine how effectively work identifiers could be discovered using automatic means. Two means of automatically discovering work attributes were found:

1) using field content harvested from name authority records in the National Authority File (NAF);
2) using work identifiers harvested directly from bibliographic records.

One of the original goals of the research was to make work record identification more effective by using information about authors and works present in authority records. NAF records may be used to discover both name and title attributes. Works that appear in many editions are frequently represented in work authority records in the NAF. Once an initial work-clustering program gathers an initial set of bibliographic records representing works, the author name appearing in majority of 100 field subfield a of those records can be searched automatically in the NAF. Names appearing in the a subfields of 100 and 400 fields of the authority records can then be automatically collected, searched, and matched against 100 and 700 fields in the bibliographic records. The collection of all forms of name found would comprise the ***Name*** attribute. The OLUC program tested by O’Neill also includes mechanisms for detecting misspelling that can be used to identify records with errors.

The most common 240 field title in the initial set of work records can be searched to match authority records that contain the same titles in t subfields of 100 in the authority records. Then, all forms of title present in 400 t subfields can be collected. Next, these titles can be automatically searched and matched against 240, 245, 246, and 740 fields in bibliographic records containing the ***Name*** attribute.

Next, titles proper (245 field, subfield a) in bibliographic records containing 100 and 240 fields matching the work can be collected. These titles can then be searched in all records containing a matching name in a 100 field. This procedure works particularly well to identify translated titles that do not appear in a work authority record.

The next step would be to collect titles proper in bibliographic records containing the **Name** attribute in a 100 field and the titles discovered above following the phrase “Translation of” or “Trans. of” in a 500 field. Again, this step will increase the number of titles in additional languages.

Finally, as stated above, an LCC# associated with a work can be harvested from 050 and 090 fields in bibliographic records. Since not all works are represented by single LCC#, an algorithm would have to be developed that would: (1) identify commonly occurring LCC# in 050 and 090 fields in the initial work record set; (2) search that LCC# in the catalog; and (3) accept an LCC# as a work attribute only if the majority of records retrieved contain the name and title attributes, or variants of those attributes, of the work sought. Although unique class numbers do not exist for all highly-manifested works, it should normally be relatively easy to determine automatically if a work had one, as such a high percentage of records representing such a work would have the same LC classification number.

To summarize, the following steps are suggested to create the work identifiers representing a particular work:

1) Assemble initial set of work records (For example, OCLC's FRBR algorithm, available at: http://www.oclc.org/research/projects/frbr/algorithm.htm could be used to assemble an initial set).

2) Determine ***Name*** by searching authority records in the NAF using the predominant 100 form of name derived from the preliminary work set and collect 400 forms of name;

3) Determine ***Title*** by:

> a) searching authority records in the NAF using the predominant 100 form of name and the predominant 240 form of title derived from the preliminary work set to collect $t forms of name from all matching authority records, and
>
> b) collecting remaining variant titles from 245 fields by searching all bibliographic records containing ***Name*** and titles collected in step 3a that match a 240 field, and collecting previously undiscovered titles from the 245 field.

4) Determine whether an LCC number unique to the work exists by:

> a) setting a threshold, for example 95 percent of works that have 050/090 fields contain the same LCC number and
> b) performing a search of that LCC number to see if many records with work identifiers from other works are present (indicates that a single LCC number is used to identify more than one work).

3.6 Excluding Non-Work Records

The success of a work identification program is contingent upon two things: its ability to identify records that belong in the work set (recall) and its ability to exclude records that do not belong in it (precision). This project simulated automatic work identification by analyzing the bibliographic records representing editions of the works studied. As described above, each record was analyzed manually, and attributes present in records were identified to create work identifiers. However, records representing works related to the work sought have the potential to decrease precision. Many records that represent related works were incidentally retrieved in the course of assembling records for the study. Some of these records were retrieved by the O'Neill program and were included in the initial set of records sent to the researchers, and others were retrieved in the search for relevant records in the OLUC. These records posed a problem for automatic work identification because they frequently contained the same attribute combinations that work records contained, and could be misidentified as belonging to the work set.

Thus, the researchers created a set of exclusion criteria to exclude non-work records. The criteria for exclusion were developed to be as extensive as possible, with the understanding that they could be more than a program could be reasonably expected to do in a large database and still perform efficiently. In addition, as discussed further below, some of the criteria may exclude too many relevant records. However, the researchers wished to provide a list of exclusion possibilities for those interested in doing further research in this area.

Because records for related works were not systematically searched for in the OLUC nor included in the pool of records assembled for the study, the number of incorrectly excluded records may be more than appears in the tallies below, and other tallies could change. However, attempts were made to discover automatic methods by which they could be excluded by determining criteria for exclusion. The following criteria were used to exclude records from the work set:

- Presence of a name-title added entry with a second indicator other than 2 and the absence of author name in 100 field. Obviously, because of the inconsistent use of the second indicator, this step could exclude relevant records.
- Presence of the word stems “dramat*”, “adapt*”, or the words “paraphrase” or "rewritten" in a 245, 500, or 520 field (to exclude dramatizations, adaptations, and paraphrases).
- Presence of the stem “simpl*” or the words “retold”, “retelling”, or "revised" in a 245 field (to exclude simplifications, retellings, and revisions).
- 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 (to exclude comic strip versions).
- Presence of the term “kit” in a 245 subfield h (to exclude 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 is not present.

Other phrases existed in records (mostly in the 245 field) that were unclear, including:

- edited for school use
- edited for college use
- abridged for modern reading
- scenes from

Records containing these phrases were tallied as "unclear" because although it is highly likely that they indicate adaptation, it would impossible to determine without examining actual items.

In addition, criteria that only *sometimes* indicate adaptation were not used here, for example: "based on" or "edited by" in a 245 field. These phrases are ambiguous, and only sometimes indicate adaptation. Because of this ambiguity, they were ignored in this study, and all records containing them were included in the work sets, even though they are likely to indicate adaptation in some cases.

Some records were discovered that were highly likely to represent abridgements, adaptations, or parts. These include records for sound recordings that had only 1 or 2 cassettes/sound discs. The works represented in this study were all of significant length, and records for sound recordings that gave no indication of adaptation or abridgement that had 1 or 2 cassettes/sound discs only, were tallied as "Misidentified as the work" unless the record specifically indicated abridgement or parts. Likewise, the works *Little Women*, *Three Musketeers*, and *Bleak House* are seldom published in editions of fewer than 300 or 400 pages or more. For this study, records for these works that had 200 pages or fewer were also tallied as "Misidentified as the work" if they were not explicitly identified as abridgments or condensations.

Results of the exclusion process are reported in tables 2-5 below as “incorrectly excluded.” Incorrectly excluded records are records that were excluded – incorrectly – from the work set based on the automatic exclusion process described above. For this set of records, the automatic exclusion process worked extremely well – only one record in the *Little Women* work set, was excluded incorrectly (see Table 4 below).

However, it must be noted that although the exclusion criteria worked well for the four works studied here, it is far from certain that they would be effective for all other works. Specifically, it is possible that, for some works, they would exclude too many records. For example, the identification procedure would perform badly for any work that contained one of the word stems in a ***Title*** attribute. It would also perform badly, for example, for a work assigned the subject heading “Comic literature” because it would exclude relevant records containing that subject heading in a 650 field. Another weakness of the exclusion criteria is that the stems, words, and phrases used here work only for English and a few other languages; they would not work to exclude records in many other languages. Further research on exclusion criteria is needed both to identify non-English language criteria and to test the effectiveness of the exclusion criteria identified here.

3.7 Work Record Identification Results

The last step in the research was to determine how effective the work identification process described above would actually work in assembling bibliographic records related to a voluminous fiction work. All of the assembled bibliographic records that were not excluded because they contained criteria for exclusion listed in section 3.6 were examined to determine whether they contained the work identifiers created as a result of the process described in sections 3.3, 3.4, and 3.5.

The matching process was always truncated. For example, the 245 title "Little women, a condensation" was treated as a match to the standard title "Little women." Note, however, that truncated matching has the potential to include records that are not members of a work set, although it did not do so in this sample of four works. In matching field content, capitalization, diacritics, and spacing differences were disregarded. Initial articles were disregarded in matching, regardless of MARC tagging. In matching MARC fields, all indicators, subfield delimiters, other subfields and other subfield content were disregarded.

As seen in Tables 2-5, the majority of records for all four works can be identified automatically using the process described above in sections 3.2 to 3.5, and using the exclusion criteria to exclude records representing related works. Seventy-seven to 95 percent of all records for a work were correctly identified using the methods described here. Not counting *Little Women*, only two to three percent of records were unidentifiable. The reason that a large number of records representing editions of Little Women were unidentifiable is that its second part has been published separately under the title *Good Wives*; this is discussed in section 3.8, below.

In an actual catalog, with full information about work records present, the percentage of unidentifiable records would be likely to be somewhat higher, given how difficult it is to actually identify all relevant records. Between one to twelve percent of records were misidentified as members of a particular work set; that is, they would be incorrectly included in a work set. This percentage also is likely to be somewhat higher in a real catalog setting. In this sample, all of the misidentified records represented related works.

BLEAK HOUSE / Dickens	**Total**	**Percent**
Correct ID	**368**	**95%**
Name & Title Attributes	*366*	*95%*
NTAE	*0*	*0%*
LC Class number	*2*	*1%*
Not Identifiable	**7**	**2%**
Name & Title not identifiable	*0*	*0%*
Title only not identifiable	*7*	*2%*
Name in 7XX or 245 $c only	*0*	*0%*
Misidentified	**5**	**1%**
Incorrectly Excluded	**0**	**0%**
Unclear	**6**	**2%**
TOTAL	**386**	**100%**

Table 2. Bleak House

KIDNAPPED / Stevenson	**Total**	**Percent**
Correct ID	**435**	**94%**
Name & Title Attributes	*427*	*92%*
NTAE	*2*	*0%*
LC Class number	*6*	*1%*
Not Identifiable	**9**	**2%**
Name & Title not identifiable	*0*	*0%*
Title only not identifiable	*8*	*2%*
Name in 7XX or 245 $c only	*1*	*0%*
Misidentified	**9**	**2%**
Incorrectly Excluded	**0**	**0%**
Unclear	**10**	**2%**
TOTAL	**463**	**100%**

Table 3. Kidnapped

LITTLE WOMEN / Alcott	**Total**	**Percent**
Correct ID	**601**	**77%**
Name & Title Attributes	*588*	*75%*
NTAE	*3*	*0%*
LC Class number	*10*	*1%*
Not Identifiable	**74**	**9%**
Name & Title not identifiable	*6*	*1%*
Title only not identifiable	*61*	*8%*
Name in 7XX or 245 $c only	*7*	*1%*
Misidentified	**89**	**11%**
Incorrectly Excluded	**1**	**0%**
Unclear	**19**	**2%**
TOTAL	**784**	**100%**

Table 4. Little Women

THREE MUSKETEERS / Dumas	**Total**	**Percent**
Correct ID	**658**	**94%**
Name & Title Attributes	*511*	*73%*
NTAE	*0*	*0%*
LC Class number	*147*	*21%*
Not Identifiable	**14**	**2%**
Name & Title not identifiable	*1*	*0%*
Title only not identifiable	*8*	*1%*
Name only not identifiable	*5*	*1%*
Name in 7XX or 245 $c only	*0*	*0%*
Misidentified	**25**	**4%**
Incorrectly Excluded	**0**	**0%**
Unclear	**6**	**1%**
TOTAL	**703**	**100%**

Table 5. Three Musketeers

Name+Title correctly identified a record as belonging to a particular work set in most cases. For some works, it is clear that ***LCC*** would be a highly effective attribute for identifying work records that do not contain a correct ***Name+Title.*** For instance, 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 tables 2-5, only a single attribute combination was tallied, even though many records contained more than one attribute combination that allowed them to be classified as belonging to a particular work set. For example, many records contained both ***LCC*** and ***Name+Title.***

3.8 Discussion of Record Identification Study

It is clear from the results of this study that that further research and experimentation in this area could be very profitable. A small number of automatically identifiable attributes used together in attribute combinations to automatically classify work records for the four works studied varied from 86 to 95 percent, which is high, especially given the large number of records associated with each of these works. If one looks only at *Bleak House, Kidnapped*, and *Three Musketeers,* the rate of success is 94-95 percent.[1] This high of a success rate for three out of the four works studied gives hope that automatic classification techniques would frequently be successful for other works.

[1] In the previous publication of these results in the *Proceedings of the 12th ASIS&T SIG/CR Classification Research Workshop* (2001), the percentage of successful identification reported for *Three Musketeers* was 65%. At the time, the researchers decided not to consider the French title appearing in a 240 field as being classifiable automatically; that decision was subsequently changed. In addition, further refinement of the identification process and reconsideration of problem records changed the other percentages slightly.

Little Women presents an unusual challenge because it consists of two parts – *Little Women* (part 1) and *Good Wives* (part 2) – which are sometimes published separately. Although this information is relatively easy for a person to ascertain and understand from the records for *Good Wives* and *Little Women*, we could not determine a way in which an automatic identification program could discover the relationship between *Good Wives* and *Little Women*. Thus, we included records without uniform titles or alternate titles containing the title *Little Women* that represented items entitled *Good Wives* items in the "not identifiable" category. It must be assumed that some small percentage of individual works will have unusual characteristics, such as this one, that will impede record identification. However, with a minimal amount of human intervention, these problem works could have greatly increased success of being identified automatically.

Three Musketeers records could have proved to be not as easily identifiable as records for the other works because it was originally published in French and has been translated more frequently than the other works. The success of the identification process in identifying records for *Three Musketeers* relies in part on the success of automatically identifying the *LCC* associated with it. Without a unique LCC, only 73 percent of the relevant records would have been identifiable. Older works that were originally published in languages other than English, such as *Three Musketeers*, could pose significant challenges for automatic identification. First, the uniform titles that are found in 240 fields are not always correct. The uniform title for *Three Musketeers* is *Trois Mousquetaires*, but many records for French and other language versions had the English version of the title in the 240 field.

Another method of identification that is likely to increase recall for non-English language works is using the multiple step process described above to increase the number of records retrieved that have no uniform title:

1) after retrieving an initial set of records using author + title, harvest all alternate titles found in 245 fields in records that also contain 240 fields; in this research, it was clear that work authority records were missing many alternate titles;
2) re-run the search for relevant records using author + variant titles harvested in bibliographic records from the above step.

The automatic identification methodology proposed here results in a relatively 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 those works studied. They include such manifestations as children's adaptations, sequels to the works, etc. These records are, for various reasons, unable to be recognized in the part of the simulation that attempts to exclude all records representing works related to the work studied.

Clearly, manifestations of the authors' collected works contain the individual works of an author. Although here collected works that did not have the title of the work somewhere in the record were not treated as work records, they could, in future research, be included and studied, and mechanisms for identifying them automatically developed as well.

4. Part 2: Record Clustering

4.1 Methodology

Records assembled for this study were also examined to discover attributes that could be used to cluster them into intelligible groupings automatically.[2] Operational definitions for relationship-based clusters were suggested by two research projects: Carlyle (1997) and Carlyle (1999 & 2001). Clusters used here were formulated from an analysis of filing rule codes reported in the 1997 article, and augmented by cluster possibilities gleaned from Tillett's taxonomy of bibliographic relationships (1991) and Smiraglia's work on augmentations (1992). Additional clusters were based on the 1999 and 2001 research projects, in which users were asked to cluster manifestations of a particular work, and attributes used for clustering were analyzed. Relationship-based clusters used in this study include:

- illustrated editions
- editions in collections of 2 or more works
- editions with amplifications, e.g., introductions
- large print, Braille, etc. editions
- non-English language editions
- nonbook format editions
- abridgements
- parts, selections, only
- English language editions without amplifications, illustrations, etc.

Operational definitions for each cluster were comprised of record attributes discovered through an initial analysis of each record. Individual record attributes are referred to as "cluster indicators" in the discussion that follows. Operational definitions for each cluster were created in order to formulate an automatic clustering program. Record attributes were comprised of MARC fields and subfields and the content of those fields and subfields. See Appendix 3 for complete operational definitions. When specific subfields are not delineated within a field, the presence of a cluster indicator may appear in any subfield.

- ***Illustrated editions*** were identified using a combination of fixed field information and information from the following fields: 250, 300, 440, 490, 500, and 700 subfield e.
- ***Editions with amplifications*** were identified using the presence of words, sometimes truncated, such as “afterword”, “annotated”, “introduction”, etc. in 245 subfield c, 250, and 500 fields.
- ***Large print, Braille, or other orthographic variation editions*** were identified using a combination of fixed field information, and the presence of words such as “large type” and “Braille” in 245 subfields c and h, 250, 440, 490, 500, 553, 650, and 830 fields.
- ***Editions in collections*** were identified using author, uniform title, and other information in 100, 500, 501, 505, and 700 fields. As noted above in section 3.8,

[2] The total number of records associated with each work in this part of the research do not exactly match the total number of records in the identification part of the research because of the status of problem records, which were not considered at times during the analysis process.

editions in collections that did not contain the uniform title or alternate title somewhere in the record were not studied.

- ***Editions composed of parts, selections only*** were identified using information present in uniform title fields, LC call numbers, and subfields in related work added entries.
- ***Abridgements*** were identified using the presence of "abridge*" (for abridgement, abridged, etc.) or "condens*" (for condensed, condensation, etc.) in a variety of fields.
- ***Non-English editions*** were identified using a combination of fixed field information and information from the 041, 240, 245, 500, and 520 fields.
- ***Nonbook format editions*** were identified using a combination of information in fixed fields, 245 subfield h, and a variety of fields associated with specific formats.
- ***English language editions without illustrations, amplifications, etc.*** This cluster was composed of records without any of the cluster indicators identified above.

Each of the MARC records assembled for *Bleak House*, *Kidnapped*, *Little Women*, and *Three Musketeers* were reviewed to discover cluster indicators present. Operational definitions for each cluster were created and refined throughout the analysis. While the operational definitions include extensive lists of cluster indicators, it is likely that analysis of records for other works, and other types of works such as nontextual or nonfiction works, will reveal other cluster indicators. Cluster indicators were formulated in as extensive and inclusive a way as possible, to make them widely applicable. We recognize that in an actual application of an automatic clustering program, refinements and simplifications might need to be made.

Because the works selected have a relatively long history, a number of records that were incorrectly cataloged were included in the study. A question we encountered was whether to include cluster indicators that represented incorrect cataloging. For instance, in one record for a microform, the term "microform" appeared not in a $h of a 245 field, but in a $b. In another, records included illustration cluster indicators such as "ill." in the correct field 300, but in $a and not subfield $b. Each occurrence of incorrect cataloging was considered, and the extent to which it would have the potential to incorrectly cluster items if used as a cluster indicator was used to determine whether it should be included as a cluster indicator. For example, we decided not to include the presence of "microform" in a 245 $b because of its potential to cluster records incorrectly. Fortunately, most microform records contain more than one cluster indicator. In the "ill." example, we included the presence of "ill." in a 300 $a because we did not believe it would incorrectly cluster records as illustrated editions.

When the operational definitions were finalized, records were analyzed again to tally the number of records that would be represented in each cluster. Records that could not be identified and clustered automatically were analyzed to determine the record enhancements necessary to include them in an automatic clustering program.

4.2 Clustering Results

One notable result of the study is that records frequently contain more than one indicator of a single cluster type; in other words, cluster indicators are often redundant. For example, a record for an illustrated edition may contain cluster indicators in the illustrations fixed field, statement of responsibility (245 subfield c), and physical description areas (300 field). This is especially true for nonbook format indicators, such as microforms and sound recordings. For the study, each record is counted only once in a single cluster. However, a single record can belong to more than one cluster, that is, items are frequently appear with amplifications and illustrations. Here, each record is tallied with each of its applicable clusters; thus, a record can be counted once in the illustrations cluster and once in the amplifications cluster. As a result, the percentages in the tables below (Tables 7-10) do not add up to 100 percent.

Results provided below have been separated into English language and non-English language items to highlight differences in record structure, content, and quality based on language of original text. Anecdotal evidence suggests that records for editions in languages other than English are frequently shorter than English-language records, in part because minimal level cataloging procedures are frequently applied to them. Also, wide variations in quality have been noted in records of non-English language items.

Table 6 indicates the distribution of records analyzed by language and by work, and indicates the percent of the total each work represents. Note that *Three Musketeers*, because it was written originally in French, has many more non-English records than the other works.

Works	**English**	**Non-English**	**Total Records**	**% of Total**
Bleak House	359	27	386	17%
Kidnapped	424	39	463	20%
Little Women	521	212	733	32%
Three Musketeers	336	367	703	31%
Totals:	**1640**	**645**	**2285**	**100%**
Percent:	**72%**	**28%**	**100%**	

Table 6. Distribution of Records Analyzed by Work and Language

Most of the *Bleak House* records, 293, or 76 percent of the total, would cluster by illustration (see Table 7). Amplifications (editions that contain introductions, prefaces, afterwords, commentaries, etc.) comprise eighteen percent of the *Bleak House* records. No orthographically variant records were found in the *Bleak House* record set, and few records represent collections, parts, abridgements, and nonbook editions. Ten percent of the records represent translations. Nonbook editions representing sound recordings comprise 67 percent of the total number of nonbook editions (fourteen records). The remainder represent microforms (33 percent, or seven records). The remaining ten percent of records represent English language editions without special features.

Bleak House	**English**	**Non-English**	**Totals**	**% of BH recs.**
Illustrations	280	13	293	76%
Amplifications	67	1	68	18%
Large print, etc.	0	0	0	0%
Collections	3	1	4	1%
Parts	5	2	7	2%
Abridgements	4	0	4	1%
Non-English Lang.	NA	NA	39	10%
Nonbook	21	0	21	5%
English eds. only	37	NA	37	10%

Table 7. Distribution of *Bleak House* Record Clusters

The *Kidnapped* record set contains fewer illustrated editions than *Bleak House*, although these editions nonetheless make up a cluster that includes over half of the records (265 or 57 percent, see Table 8). Amplifications account for eleven percent of the *Bleak House* records, and a similarly small percentage (three) represent some kind of orthographically variant edition. Of these, one record is for a Braille edition, and the rest are large print editions. Three percent of the records represent editions that contain the work in a collection; only a single translation contains a part or parts of the work. Abridgements comprise five percent of the set, and non-English language versions (six percent) make up the rest. The *Kidnapped* record set contains almost double the number of non-book editions that the other two works do, at ten percent, which is likely due to its status as a work for children. The records remaining, English language editions, account for 20 percent of the total.

Kidnapped	**English**	**Non-English**	**Totals**	**% of Kidn. recs.**
Illustrations	242	23	265	57%
Amplifications	49	0	49	11%
Large print, etc.	13	0	13	3%
Collections	15	0	15	3%
Parts	0	1	1	0.2%
Abridgements	21	0	21	5%
Non-English Lang.	NA	NA	27	6%
Nonbook	48	0	48	10%
English eds. only	92	NA	92	20%

Table 8. Distribution of *Kidnapped* Record Clusters

Like the other works, *Little Women* (Table 9) has been published frequently in illustrated manifestations (61 percent of records), and seldom in amplified, orthographic-variation, or collected manifestations (6%, 2%, and 0 percent respectively). Five percent of the manifestations represent parts. However, as noted above, because the second half was published from time to time under the title *Good Wives*, and it was not possible to identify those items. Thus, this percentage should actually be much higher (74 records for *Little Women* were not identifiable; 61 of those were for *Good Wives*). *Little Women* has been abridged and translated regularly; eight percent of the manifestations are abridgements, and 29 percent are translations. Nine percent of records in the study represent nonbook manifestations, and twelve percent represent English language editions without special characteristics.

Little Women	**English**	**Non-English**	**Totals**	**% of LW recs.**
Illustrations	305	136	441	61%
Amplifications	40	0	40	6%
Large print, etc.	12	0	12	2%
Collections	1	0	1	0%
Parts	30	3	33	5%
Abridgements	55	6	61	8%
Non-English Lang.	NA	NA	210	29%
Nonbook	65	1	66	9%
English eds. only	84	NA	84	12%

Table 9. Distribution of *Little Women* Record Clusters

The analysis of *Three Musketeers* reveals patterns similar to the other works (see Table 10). Nearly half of the records represent items that are illustrated. Twelve percent represent amplifications, and few records are contained in large print (0%), collections (2%), parts (1%), and abridgements (6%) clusters. Non-English items represent a large proportion (28%) of this work, much larger than the other works. This may be a result of the fact that it was originally published in French. Six percent of the records cluster as non-book. Finally, English language editions without illustrations or other special features account for fourteen percent of the total items.

Three Musketeers	**English**	**Non-English**	**Totals**	**% of 3M recs.**
Illustrations	167	165	332	47%
Amplifications	40	47	87	12%
Large print, etc.	0	0	0	0%
Collections	3	8	11	2%
Parts	5	2	7	1%
Abridgements	29	10	39	6%
Non-English Lang.	NA	NA	364	52%
Nonbook	37	2	39	6%
English eds. only	98	NA	98	14%

Table 10. Distribution of *Three Musketeers* Record Clusters

As expected, some records were found that contained incorrect cluster indicators, or contain cluster indicators that cannot be identified automatically (Table 11). Six percent of the records analyzed (90 records) were identified as not belonging to a cluster to which they should, in fact, belong. The most problematic records analyzed represented works published in collections. Forty-two records could not be identified automatically as containing the work analyzed, which is 47 percent of the total number incorrectly clustered records. Amplifications comprised 26 percent of incorrectly clustered records, while records for items representing parts or selections of works came in third (14 percent). Few materials published in non-English language editions, illustrated editions, and abridgements clustered incorrectly, and none of the large print, Braille, or non-book materials clustered incorrectly.

Records Unable to be Clustered	**Number**	**% Unclusterable**
Illustrations	5	4%
Amplifications	26	20%
Large print, etc.	0	0%
Collections	46	36%
Parts	41	32%
Abridgements	6	5%
Non-English Lang.	5	4%
Nonbook	0	0%
English eds. only	NA	NA
TOTAL UNCLUSTERABLE	129	100%

Table 11. Distribution of Records Unable to be Clustered

4.3 Discussion of Clustering Study

Effective displays that clarify the nature of items retrieved and present them in a brief, summary display would require clustering a large number of records. In each work analyzed, illustrated editions formed much larger clusters than the other cluster types, with a range for each individual work from 47 to 76 percent of the records for the different works. All of the other clusters were smaller, except non-English editions of the *Three Musketeers.*

The widely varying sizes of the clusters, and especially the large size of the illustrations clusters, point to the desirability of developing alternative clustering methods. One alternative would be cluster all illustrated editions that also cluster into other groups with those groups only, and not in the illustrations clusters, leaving the illustrated editions cluster to be composed only of illustrated editions that have no other cluster attributes. Another possibility would be sub-clustering the largest clusters. For example, the illustrations cluster could be sub-clustered into abridged illustrated editions, illustrated editions with introductions, etc.

Completely different clustering strategies should also be investigated, for example, using record sets types defined by Elaine Svenonius (1988). Her record clusters included work, text, typesetting (or edition), sub-edition, imprint, and reprint (p. 7). These equivalence sets, excluding work sets, can also be used to determine the degree to which clustered displays assist users in navigating large retrieval sets. OCLC has done a significant amount of work automatically identifying work records toward creating FRBR-based displays (IFLA Study Group, 1998). More information about these projects is available on their FRBR site, available at: http://www.oclc.org/research/projects/frbr/default.htm.

A further consideration is the presence of non-English language records in other clusters. In the results presented above, the non-English language items were included in other clusters when they contained those indicators. One reason supporting the exclusion of non-English language items in other clusters is that cluster indicators for non-English items are sometimes difficult, or impossible, to identify automatically. For example, it would be a long process requiring language expertise in many languages to identify non-English terms for amplifications in all of the languages represented even in this small work sample. However, it would be useful to have user research confirming the effectiveness of segregating non-English language materials into a single cluster.

Several problems with automatic clustering become apparent during the record analysis. First is the extensive list of cluster indicators generated by this project. The list is so long that any software developed to accomplish clusterings would be difficult to program, and very slow to run, particularly in large databases. One possible solution would be to identify those most prevalent cluster indictators, and restrict the automatic clustering. Based on cursory examination of the data, the distribution across the different indicators is wide, which would mitigate against selection of a smaller number of indicators for programming. However, many FRBR discussions are recommending restricting FRBRization to a limited number of works. These are the same works that the clustering indicators were developed for, and limiting the works that a program operated on would make such a program, though large, more efficient.

Another significant problem exists because of errors in the MARC tagging of records representing works published in collections. A name-title added entry (700 field, second indicator 2) for analytical entries is required to identify a work in a collection correctly. Unfortunately, many of the incorrectly clustered records contained a 700 field with a blank second indicator. A blank second indicator also means that the item described is a related work, making it impossible to identify collections that have incorrectly tagged 700 fields. It should be noted that a blank second indicator was once correct tagging, hence the large number of records in this category Another problem is the use of title-added entry fields, for example, the 740 field or the 246 field, to identify the works present in the item. Again, it would be impossible to determine when these fields indicate the presence of an additional work.

One limitation of the amplification indicators used here is that when an item indicates that it is an amplification, this information often appears on a title page. When it appears on a title page, it is usually recorded in the 245 $c (statement of responsibility) or in a note (500) field. The ramifications of this are that identifying amplifications requires text word identification in these fields, for example, the words: “introduction,” “preface,” "afterword," and other words that can indicate the presence of an amplification. These words seem to indicate the presence of an amplification relatively well, although grouping could be hindered when these words, although present, do not indicate an amplification.

A final consideration is incorrect or missing information in bibliographic records that also has an impact on record identification. These situations would produce incorrect clustering regardless of whether the clustering is done automatically or manually. An example of incorrect information in records was found in this project in records for *Bleak House* and *Three Musketeers.* Most of the books representing these works are between 400 and 800 pages long, but a number of records were found with 200 or fewer pages without any indication of abridgement, condensation, or adaptation. It is likely that a number of these records actually represent items that are abridgements or adaptations of some kind. Also, sound recordings for the unabridged editions of these works usually contain ten or more cassettes; some of the sound recording records analyzed contained one or two cassettes, again, without any indication that they were abridgements or adaptations.

5. Future Research

This study lays the groundwork for research that may proceed in a variety of directions. First, this project takes fiction works as its focus. Subsequent studies might build on the results of this project by examining a selection of non-fiction works to determine whether sets of records representing fiction and non-fiction works differ or are similar. Such an analysis would allow for the creation of more general automatic record identification and clustering techniques. Concurrently, automatic record identification and clustering techniques could be tested and refined by applying them to a broader range of works represented by large numbers of bibliographic records.

Another possibility for future research would be to include identification and clustering of records representing works *related* to a particular work. It is likely, based on the findings of a previous research project (Carlyle 1996), that records representing works related to a work are far more numerous than records representing manifestations of a work. These records are often retrieved in a user's search for a particular work, and so should be considered for their potential to be identified and clustered for effective display.

Sub-clustering of records is another avenue for further research. Records in each cluster could be analyzed further to determine the extent to which they may be sub-clustered into equivalence sets identified by Svenonius (1988) and Svenonius and O'Neill (1988). As stated above, these include typesetting (edition), sub-edition, imprint, and reprint. Operational definitions of a somewhat different sort could be derived for sub-cluster types to include identification of parts of records that could be used for matching. These operational definitions could be analyzed using records identified in the study to determine the extent to which they would successfully sub-cluster records automatically.

An important extension of this research might be to devise automatic identification and clustering strategies in a prototype system featuring organized work displays. The effectiveness of such a system could then be compared to systems featuring more conventional work displays or to systems featuring other types of organized or clustered work displays. This research might incorporate studies of user and system effectiveness.

6. Conclusion and Recommendations

This research was conducted to deepen our understanding of voluminous works. These works are represented in online catalogs by many bibliographic records, and variations in the records can mean that they are not all retrieved in a search for that work, or that the retrieved records are displayed in an incomprehensible or unhelpful manner. It is likely that these works are among the most frequently sought works in online catalogs, thus, it makes sense to conduct research to alleviate these problems. The research conducted in this paper proposes two methods to address these problems. First, it proposes a sophisticated algorithm to identify records that represent works. Second, it proposes a method of clustering records related to a work into meaningful clusters.

Recommendations based on the results of the identification study include:

- Augment the current work identification algorithms being used by OCLC, including the FRBR Work Set algorithm, with methods suggested here, especially:
    - use of information contained in authority records, i.e., alternate forms of author name and work title,
    - use of LCC number (especially for works published originally in a language other than English), and
    - use of titles proper harvested from bibliographic records containing uniform titles.
- Do further research of this type on works of non-fiction.
- Do further research to determine the types of situations in which minimal human intervention in a mostly automated process would be fruitful.
- Do further research on including collected works (that is, editions entitled “Collected Works,” etc.) in automatic identification algorithms.
- Do further research on identification of records representing related works.
- Do further research to investigate the nature of works – are difficulties related to identification similar from work to work? if not, how are they different?
- Refine exclusion criteria by testing on a wide variety of works.

Recommendations based on the results of the cluster study:

- Investigate different clustering or sub-clustering algorithms that would create more evenly sized clusters, for example, using record-set-types based on Svenonius (1988) or FRBR-based displays of expression and manifestation sets.
- Do further research on clustering of records representing related works.
- Do further research to investigate the nature of works – are types of clusters similar from work to work? If not, how are they different?
- Include records representing related works in future research.

7. Acknowledgements

I would like to acknowledge the contributions of the following in the completion of this project, in alphabetical order: Harry Bruce, Collette Davis, Karen Fisher, Lisa M. Fusco, Maurice Green, Joe Janes, Elizabeth S. Knight, Sara Ranger, and Joel Summerlin.

8. Publications Based on this Research

Allyson Carlyle and Joel Summerlin. (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. Clare Beghtol, Lynne C. Howarth, and Nancy J. Williamson. Wurzburg: ERGON Verlag: 320-326.

Allyson Carlyle and Joel Summerlin. (2002) “Transforming Catalog Displays: Record Clustering for Works of Fiction.” *Cataloging & Classification Quarterly*. v. 33, no. ¾ : 13-25. With Joel Summerlin. [Republication with slight revision of conference paper originally published in 2000.]
Also available at:
http://projects.ischool.washington.edu/acarlyle/Papers/transforming_displays.htm

Allyson Carlyle and Sara Ranger. (2001) “Facilitating Retrieval of Fiction Works in Online Catalogs.” *Proceedings of the 12th ASIS&T SIG/CR Classification Research Workshop, November 4, 2001, Held at the 64th ASIS&T Annual Meeting, November 2-8, 2001, Washington, D.C.* Efthimis N. Efthimiadis, ed. Silver Spring, MD: American Society for Information Science and Technology: 1-11.
Also available at:
http://projects.ischool.washington.edu/acarlyle/Papers/fiction_works_sigcr02.htm

9. Works Cited

Ayres, F.H., L.P.S. Nielsen, M.J. Ridley, and I.S. Torsun. (1995). *The Bradford OPAC: A New Concept in Bibliographic Control.* British Library R & D Report 6183. Wetherby, West Yorkshire: British Library Research and Development Department.

Ayres, F.H., L.P.S. Nielsen, and M.J. Ridley. (1998). *The Bradford OPAC2 (BOPAC2): Managing and Displaying Retrievals from a Distributed Search in Z39.50*. British Library Research and Innovation Report 103. Wetherby, West Yorkshire: British Library Research and Innovation Centre. Also available at: http://www.bopac2.comp.brad.ac.uk/~bopac2/report

Carlyle, Allyson. (1996). "Ordering author and work records: An evaluation of collocation in online catalog displays." *Journal of the American Society for Information Science*. 47: 538-554.

Carlyle, Allyson. (1997). "Fulfilling the Second Objective in the Online Catalog: Schemes for Organizing Author and Work Records into Usable Displays." *Library Resources & Technical Services*. 41 (2): 79-100.

Carlyle, Allyson. (1999). "User Categorisation of Works: Toward Improved Organisation of Online Catalogue Displays." *Journal of Documentation.* 55 (2): 184-208.

Carlyle, Allyson. (2001). "Developing Organized Information Displays for Complex Works: A Study of User Clustering Behavior." *Information Processing & Management*. 37 (5): 677-699.

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998) Functional Requirements of Bibliographic Records. München: K.G. Saur. Also available at: www.ifla.org/VII/s13/frbr/frbr.pdf.

Matthews, J., G.S. Lawrence, and D.K. Ferguson. (1983). *Using Online Catalogs.* New York, NY: Neal Schuman.

O'Neill, Edward T. (1994). "Manifestations of Fiction Works." *Annual Review of OCLC Research 1994.* Dublin, OH: 11-15.

Smiraglia, Richard P. (1992). *Authority Control and the Extent of Derivative Bibliographic Relationships*. Ph. D Dissertation, University of Chicago, Chicago, Ill.

Svenonius, Elaine. (1988). "Clustering Equivalent Bibliographic Records." *Annual Review of OCLC Research, July 1987-June 1988.* Dublin, OH: 6-7.

Svenonius, Elaine and Edward T. O'Neill. (1988). "Custering Equivalent Bibliographic Records." *OCLC Research Review.* Dublin, OH.

Tillett, Barbara Ann Barnett. (1987). *Bibliographic Relationships: Toward a Conceptual Structure of Bibliographic Information Used in Cataloging.* Ph.D. Dissertation, University of California, Los Angeles.

Tillett, Barbara B. (1991). A taxonomy of bibliographic relationships. *Library Resources & Technical Services*. 35 (2): 150-158.

Wiberley, Stephen E. Jr., Robert Allen Daugherty, and James A. Danowski. (1995). "User Persistence in Displaying Online Catalog Postings: LUIS." *Library Resources & Technical Services*. 39 (3): 247-264.

Appendix 1. Fields and Field Content Used In Work Identification Coding

Fields and Field Content

Note: Not all of the codes identified here were actually used in the final work identification process.
Key to table below:
EST=English standard title, which is usually uniform title except for Three Musketeers. For Three Musketeers, both French and English titles were counted equally as the "English Standard Title."
The EST is always truncated – that is, any text following the EST is ignored.
Alt. Title=Alternate title: all other acceptable (that is, automatically identifiable) titles, including individual titles of parts, translated titles and translated titles of parts. The Alt. Title is always truncated – that is, any text following the Alt. Title is ignored.

Code	Field/Subfield
Da	100 $a correct
Db	245 $a begins with EST
Db2	245 $a is Alt. Title (=translated & in work au rec)
Dc	245 $b present, not EST nor alt. Title
Dd	245 $b present, is EST (or $p)
Dd2	245 $b present, is ALT title (or $p) or begins with ALT
De	245 $c has correct author
Df1	240 $a is EST
Df2	240 $a is alt. title (translated & in work au rec)
Df3	240 $a begins with EST
Df4	240 $a begins with ? [this was not used]
Dg	700 $a has correct author
Dh	740 or 246 is EST (and Db not coded); field begins with correct title
Di	740 or 246 is alt. title; field begins with correct title
Dj	700 12 NTAE correct: $a (correct author) $t alt. title
Dk	700 12 NTAE correct: $a (correct author) $t EST
Dl	Call # is correct and is first element (disregard subfield indicators)
Dm	Call # is correct and is second element
Dn	245 $a contains EST but not as first element
Dn2	245 $a contains alt. title but not as first element
Dn3	title not recognizable
Dr	100 field is 400 form ($a, or whole field) in au record
Ds	505 contains EST or Alt title [not used in SIG-CR]
Dt	500 contains “Translation of” or “Trans. of “ + EST or Alt title (truncated – cd. be something after that)
Dx	Misidentified as an edition of the work – it is not.
Dz	Unidentifiable as an edition of the work, although it probably is an edition
UN	Unclear whether this is an edition or not an edition of the work
DO/DP	Incorrectly excluded

Other coding guidelines

If call number has extra number on the end	Do not code as Dl or Dm
Disregard punctuation (even between words) & capitalization; diacritics are treated as the character without the diacritic.	Unless in middle of words/call # (typos)
If author name or title is spelled wrong	Do not code – cannot be used as identifying attribute
If author name is abbreviated	Do not code – cannot be used as identifying attribute
If 700 has indicators 1 and 2 but $t is not EST	Code Dg
If 245 $a has EST first and more text following	Code Db
If 245 $a has incorrect article but the 2nd indicator is correct	Code Db
If 245 $b has uniform title anywhere	Code Dd
Disregard indicators, even if incorrect (except in 700 $t)	

Appendix 2. Work Identification Codes

Specific Work Identifiers

Name + Title Identifiers (see individual code keys in Appendix 1)

1. Da + Df1 [Name* + 240 EST**]
2. Da + Db [Name + 245 $a EST]
3. Da + Dd [Name + 245 $b EST]
4. Da + Dh [Name + EST in Other T******]
8. Da + Df2 [Name + 240 ALT***]
9. Da + Dn [Name + EST in 245 $a]
10. Da + Db2 [Name + 245 ALT]
14. Da + Dn2 [Name + ALT in 245 $a]
20. Dr + Db2 [ALTName*****+ 245 $a ALT]
21. Dr + Db [ALTName + 245 $a EST]
22. Dr + Df1 [ALTName + 240 EST]
24. Da + Dt [Name + 500=”Translation of” or “Trans. of” + EST or ALT; may have text following]
25. Da + Ds [Name + 505 contains EST or ALT]

Name-Title Added Entry

5. Dk [EST Name/Title Added Entry]

LCC#

6. Dl [Class no. in LC class no. field, $a]

NOT ***Used to Identify Work – All coded as DZ (Not identifiable)***

7. Dg + Dn3 [700 name/not rec name + not recognizable title]
11. Da + Dn3 [Name + not recognizable title]
12. Da + Di [Name + 246/740 ALT] (OCLC #40156204)
13. Dj [ALT Name/Title Added Entry] (found 1)
15. Dg + Db [700 Name**** + 245 $a EST]
16. Dg + Df1 [700 Name + 240 EST]
17. Dg + Dh [700 Name + Other T]
18. De + Db [245 $c Name + Tproper EST, $a]
19. De + Df1 [245 $c Name + 240 EST]
23. Da only [Name]

*Name=Correct authority record 100 form, in 100 field, subfield a only.
**EST=English standard title (see Appendix 1)
***ALT=Alternate work title (see Appendix 1)
****700 Field Name = Correct form in $a
*****ALT Name = 400 field form of name in Authority Record appears in 100 $a in Bib Record
*****Other T = a) title appearing in 740 or 246 field OR b) edition title not identifiable as related to the work (i.e., not in authority record and not appearing in a 245 in any bib records paired with a correct 240) OR c) EST appears in a 245 $c

Appendix 3. Operational Definitions for Clusters

Records belonging to a particular cluster were identified by looking for the presence of various cluster indicators in records. These indicators are listed below for each cluster type.

abridgements: presence of "abridge*" in a 245, 250 $a, 511, or 520 field; presence of "condens*" in a 245, 250 $a, 511, or 520 field.

amplifications: presence of the terms "afterword", "annot*", "comment*", "intro*", "note*, "prolog*", or "pref* in a 245 $c or 250 field; presence of "introd*", "commentary", "notes", "supplement" in a 500 field.

editions in collections: presence of uniform title of work in 505 field + 100 $a author of work; presence of 700 field with 2nd indicator "2" with correct author and uniform title for work + (100 $a is not author of work *or* 245 $a is not title of our work); presence of 700 field with 2nd indicator "2" with author and uniform title for another work; 500 $a or 501 $a beginning with "With" or "Bound with."

illustrated editions: presence of any letter in an Ills. fixed field; presence of the truncated (truncation indicated by *) terms "ill*", "port*", "plate*", or "map*) in either the 245 $c, the 300 field, or the 250 field; presence of "ill*" in a 440 or 490 field; presence of "illus* in a 500 field; presence of "ill*" in a 700 $e.

large print, Braille, or other orthographic variation editions: presence of "d" or "f" in the Form fixed field; presence of the term "large type" in a 245 $c, 250, 440, 490, 500, 650 $a, or 830 field; presence of the term "large print" in a 245 $h; presence of the term "Braille" in a 245 $h or a 553 field.

non-English editions: presence of any text other than "eng" in Lang fixed field; presence of 041 1_; presence of 240 $L if not "English"; presence of "translat*" in 245 $c, 500, or 520 field

nonbook format editions: *Sound Recordings*: presence of Type fixed field "i" or "j" presence of 007 $a "s"; presence of 245 $H "sound*"; presence of 300 field $a with ("sound*" or "cassett*") ; presence of 305 or 362 fields. *Computer files:* presence of Type fixed field "m"; presence of 245 $h "computer file"; 256, 538, or 856 fields present; *Microforms:* presence of Form fixed field "a", "b" or "c"; presence of 245 $h "microform"; presence of "micro*" in 533 field $a.

editions composed of parts, selections only: presence of 240 $a uniform title + ($k "selections" or presence of $p); presence of LC call number for work indicating parts; presence of 700 field with 2nd indicator "2" with correct author and uniform title for work + presence of ($k "selections" or presence of $p)

English language editions without illustrations, amplifications, etc.: all of the records not clustered in any of the above clusters.