A Candid Look at Collected Works: Challenges of Clustering Aggregates in GLIMIR and FRBR

Gail Thornburg

ABSTRACT

Creating descriptions of collected works in ways consistent with clear and precise retrieval has long challenged information professionals. This paper describes problems of creating record clusters for collected works and distinguishing them from single works: design pitfalls, successes, failures, and future research.

OVERVIEW AND DEFINITIONS

The Functional Requirements for Bibliographic Records (FRBR) was developed by the International Federation of Library Associations (IFLA) as a conceptual model of the bibliographic universe. FRBR is intended to provide a more holistic approach to retrieval and access of information than any specific cataloging code. FRBR defines a work as a distinct intellectual or artistic creation. Put very simply, an expression of that work might be published as a book. In FRBR terms, this book is a manifestation of that work.[1]

A collected work can be defined as "a group of individual works, selected by a common element such as author, subject or theme, brought together for the purposes of distribution as a new work."[2] In FRBR, this type of work is termed an aggregate or "manifestation embodying multiple distinct expressions ."[3] Zumer describes aggregate as "a bibliographic entity formed by combing distinct bibliographic units together."[4] Here the terms are used interchangeably.

In FRBR, the definition of aggregates applies only to group 1 entities, i.e., not to groups of persons or corporate bodies. The IFLA Working Group on Aggregates has defined three distinct types of aggregates: (1) collections of expressions, (2) aggregates resulting from augmentation or supplementing of a work with additional material, and (3) aggregates of parallel expressions of one work in multiple languages.[5] While noting the relationships between the categories, this paper will focus on the first type.

Aggregates of the first type include selections, anthologies, series, books with independent sections by different authors, and so on. Aggregates may occur in any format, from a volume containing both of the J. D. Salinger works *Catcher in the Rye* and *Franny and Zooey* to a sound recording containing popular adagios from several composers to a video containing three John Wayne movies.

Gail Thornburg (thornbug@oclc.org) is Consulting Software Engineer and Researcher at OCLC, Dublin, Ohio.

THE ENVIRONMENT

The OCLC WorldCat database is replete with bibliographic records describing aggregates. It has been estimated that that database may contain more than 20 percent aggregates.[6] This proportion may increase as WorldCat coverage of recordings and videos tends to increase.

In the Global Library Manifestation Identifier (GLIMIR) project, automatic clustering of the records into groups of instances of the same manifestation of a work was devised. GLIMIR finds and groups similar records for a given manifestation and assigns two types of identifiers for the clusters. The first type is Manifestation ID, which identifies parallel records differing only in language of cataloging or metadata detail, some of which are probably true duplicates whose differences cannot be safely deduplicated by a machine process. The second type is a Content ID, which describes a broader clustering, for instance, physical and digital reproductions and reprints of the same title from differing publishers.

This process started with the searching and matching algorithms developed for WorldCat. The GLIMIR clustering software is a specialization of the matching software developed for the batch loading of records to WorldCat, deduplicating the database, and other search and comparison purposes.[7] This form of GLIMIRization compares an incoming record to database search results to determine what should match for GLIMIR purposes. This is a looser match in some respects than what would be done for merging duplicates. The initial challenges of tailoring matching algorithms to suit the needs of GLIMIR have been described in Thornburg and Oskins[8] and in Gatenby et al.[9]

The goals of GLIMIR are (1) to cluster together different descriptions of the same resource and to get a clearer picture of the number of actual manifestations in WorldCat so as to allow the selection of the most appropriate description, and (2) to cluster together different resources with the same content to improve discovery and delivery for end users. According to Richard Greene, "The ultimate goal of GLIMIR is to link resources in different sites with a single identifier, to cluster hits and thereby maximize the rank of library resources in the web sphere."[10]

GLIMIR is related conceptually to the FRBR model. If the goal of FRBR is to improve the grouping of similar items for one work, then GLIMIR similarly groups items within a given work. Manifestation clusters specify the closest matches. Content clusters contain reproductions and may be considered to represent elements of the expression level of the FRBR model.

The FRBR and GLIMIR algorithms this paper discusses have evolved significantly over the past three years. In addition, it should be recognized that the FRBR algorithms use a map/reduce keyed approach to cluster FRBR works and some GLIMIR content while the full GLIMIR algorithms use a more detailed and computationally expensive record comparison approach.

The FRBR batch process starts with WorldCat enhanced with additional authority links, including the production GLIMIR clusters. It makes several passes through WorldCat, each pass constructing keys that pull similar records together for comparison and evaluation. As described by Toves, "Successive passes progressively build up knowledge about the groups allowing us to refine and

expand clusters, ending up with the work, content and manifestation clusters to feed into production."[11]

Each approach to clustering has its limits of feasibility, but the FRBR and GLIMIR combined teams have endeavored to synchronize changes to the algorithms and to share insights. Some materials are easier to cluster using one approach, and some in the other.

Clustering meets Aggregates

In the initial implementation of GLIMIR, the issue of handling collected works was considered out of scope for the project. With experience, the team realized there can be no effective automatic GLIMIR clustering if collected works are not identified and handled in some way.

Why is this? Suppose a record exists for a text volume containing work A. This matches to a record containing work A, but actually also containing work B. This matches to a work containing B and also containing works C, D, and E. The effect is a snowballing of cluster members that serves no one.

How could this happen? In a bibliographic database such as WorldCat, items representing collected works can be catalogued in several ways. Efforts to relax matching criteria in just the right degree to cluster records for the same work are difficult to devise and apply.

The GLIMIR and FRBR teams consulted several times to discuss clustering strategies for works, content, and manifestation clusters. Practical experience with GLIMIR led to rounds of enhancements and distinctions to improve the software's decisions. While GLIMIR clusters can and have been undone and redone on more than one occasion, it took experience from the team to realize that the clues to a collected work must be recognized.

Bible and *Beowulf*

As are many initial production startups, the output of GLIMIR processing was monitored. Reports for changes in any clusters of more than fifty were reviewed by quality control catalogers for suspicious combinations. And occasionally a library using a GLIMIR- or FRBR-organized display would report a strange cluster.

This was the case with a huge malformed cluster of records for the Bible. Such a work set tends to be large and unmanageable by nature; there are a huge number of records for the Bible in WorldCat. However, it was noticed the set had grown suddenly over the previous two months. User interface applications stalled when attempting to present a view organized by such a set.

One day, a local institution reported that a record for *Beowulf* had turned up in this same work set. This started the team on an investigation. After much searching and analysis of the members of this cluster, the index case was uncovered.

In many cases bibliographic records are allowed to cluster based on a uniform title. What the team found connecting these disparate records was a totally unexpected use of the uniform title, a field

240 subfield a, contents: "B.". That's right, "B.". Once the first case was located, it was not hard to figure out that there were numerous uniform "titles" with other single letters of the alphabet.

So in this odd usage, Bible and *Beowulf* could come together, if insufficient data were present in two records to discriminate by other comparisons. Or potentially, other titles which started with "B."

Seeing this unanticipated use of uniform title field, the FRBR and GLIMIR algorithms were promptly modified to beware. The FRBR and GLIMIR clusters were then unclustered and redone.

This was a data issue, and unanticipated uses of fields in a record will crop up, if usually with less drama. Further experience showed more.

In the examination of another ill-formed cluster, a reviewer realized that one record had the uniform title stated as "Illiad" but the item title was Homer's "Odyssey." Of course these have the same author, and may easily have the same publisher. Even the same translator (e.g., Richard Lattimore) is not improbable for a work like this. This was a case of bad data, but it imploded two very large clusters.

Music and Identification of Collected Works

As music catalogers know, musical works are very frequently presented in items that are collections of works. The rules for creating bibliographic records for music, whether scores or recordings or other, are intricate. The challenges to software to distinguish minor differences in wording from critical differences seem to be endless.

Moreover, musical sound recordings are largely collected works due to the nature of publication. As noted by Papakhian, personal author headings are repeated oftener in sound recording collections than in the general body of materials.[12] There are several factors that may contribute to such an observation. There are likely to be numerous recordings by the same performer of different works and numerous records of the same work by different performers. Composers are also likely to be performers. The point is, for sound recordings an author statement and title may be less effective discriminators than for printed materials.

Vellucci[13,14] and Riley[15] have written extensively on the problems of music in FRBR models. The problems of distinguishing and relating whole/part relationships is particularly tricky. Musical compositions often consist of units or segments that can be performed separately. So they are generally susceptible to extraction. These extractive relationships are seen in cases where parts are removed from the whole to exist separately, or perhaps parts for a violin or other instrument are extracted from the full score. Software must be informed with rules as to significant differences in description of varying parts and varying descriptions of instruments, and in this team's experience that is particularly difficult.

Krummel has noted that the bibliographic control of sound recordings has a dimension beyond item and work, that is, performance.[16] Different performances of the same Beethoven symphony

need to be distinguished. Cast and performer list evaluation and dates checking are done by the software. However, the comparisons the software can make are susceptible to fullness or scarcity of data provided in the bibliographic record. There is great variation observed in the numbers of cast members stated in a record. Translator and adapter information can prove useful in the same sense of roles discrimination for other types of materials.

This is close scrutiny of a record. At the same time consider that an opera can include the creative contributions of an author (plot), a librettist, and a musical composer. Yet these all come together to provide one work, not a collected work.

Tillett has categorized seven types of bibliographic relationships among bibliographic entities, including the following:

1. Equivalence, as exact copies or reproduction of a work. Photocopies, microforms are examples.

2. Derivative relationships, or, a modification such as variations, editions, translations.

3. Descriptive, as in criticism, evaluation, review of a work.

4. Whole/part, such as the relation of a selection from an anthology.

5. Accompanying, as in a supplement or concordance or augmentation to a work.

6. Sequential, or chronological relationships.

7. Shared characteristic relationships, as in items not actually related that share a common author, director, performer, or other role. [17]

While it is highly desirable for a software system to notice category 1 to cluster different records for the same work, that same software could be confused by "clues," such as in category 7. And the software needs to understand the significance of the other categories in deciding what to group and what to split.

To handle these relations in bibliographic records, Tillett discusses linking devices including, for instance, uniform titles. Yet uniform titles are used for the categories of equivalence relationships, whole/part relationships, and derivative relationships. This becomes more and more complex for a machine to figure out. Of course, uniform titles within bibliographic records are supposed to link to authority records via text string only. Consideration should ideally be given to linking via identifiers, as has been suggested elsewhere.[18]

Thematic Indexes

Review of scores and recordings GLIMIR clusters showed a case where Haydn's symphonies A and B were brought together. These were outside the traditional canon of the 104 Haydn symphonies and were referred to as "A" and "B" by the Haydn scholar H. C. Robbins Landon. This mis-clustering highlighted the need for additional checks in the software.

The original GLIMIR software was not aware of thematic indexes as a tool for discrimination.

Thematic indexes are numbering systems for the works of a composer. The Kochel Mozart catalog, as in K. 626, is a familiar example. These designations are not unique to a given composer, that is, they are intended to be unique for a given composer, but identical designators may coincidentally have been assigned to multiple composers. While "B" series numbers may be applied to works of Chambonnières, Couperin, Dvořák, Pleyel, and others, the presence of more than one B number is suggestive of collected work status. For more on the various numbering systems, see the interesting discussion by the Music Library Association.[19]

However, the software cannot merely count likely identifiers in the usual place. This could lead to falsely flagging aggregates; one work by Dvořák could have B.193, which is incidentally equivalent to opus 105. Clearly, any detection of multiple identifiers of this sort must be restricted to identifiers of the same series.

String Quartet Number 5, or Maybe 6

Cases of renumbering can cause problems in identifying collected works. An early suppressed or lost work, later discovered and added to the canon of the composer's work, can cause renumbering of the later works.

Clustering software needs must be very attentive to discrete numbers in music, but can it be clever enough? Paul Hindemith (1895–1963) works offer an example. His first string quartet was written in 1915, but long suppressed. His publisher was generally Schott. Long after Hindemith's death, this first quartet was unearthed, and then was published by Schott. The publisher then renumbered all the quartets. So quartets previously 1 through 6 became 2 through 7. The rediscovered work was then called "No. 1," though sometimes called "No. 0" to keep the older numbering intact. Further, the last two quartets did not even have opus numbers assigned and were both in the same key.[20] This presents a challenge.

Anything Musical

Another problem case emerged when reviewers noticed a cluster contained both the unrelated songs "Old Black Joe" and "When You and I were Young Maggie." On investigation, the cluster held a number of unrelated pieces. Here the use of alternate titles in a 246 field had led to overclustering, and the rules for use of 246 fields were tightened in FRBR and GLIMIR. As in the other problem cases, cycles of testing were necessary to estimate sufficient yet not excessive restrictions. Rules too strict split good clusters and defeat the purpose of FRBR and GLIMIR.

At this point the GLIMIR/FRBR team recognized that rules changes were necessary but not sufficient. That is, a concerted effort to handle collected works was essential.

Strategies for Identifying Collected Works

The greatest problem, and most immediate need, was to stop the snowballing of clusters. Clusters containing some member records that are collected works can suddenly mushroom out of control.

Rule 1 was that a record for a collected work must never be grouped with a record for a single work. If all in a group are collected works, that is closer to tolerable (more on that later).

With time and experimentation, a set of checks were devised to allow collected works to be flagged. These clues were categorized as types: (1) considered conclusive evidence, or (2) partial evidence. Type 2 needed another piece of evidence in the record.

Finding the best clues was a team effort. It was acknowledged that to prevent overclustering, overidentification of aggregates was preferable to failure to identify them. Several cycles of tests were conducted and reviewed, assessing whether the software guessed right.

Table 1 illustrates the types of checks done for a given bibliographic record. Here the "$" is used as abbreviation for subfield, and "ind" equals indicator.

Area	Field	Rule	Notes
Uniform Title	240 $a and no $m, $n, $p, or $r	Title in $ a on list of terms, without the other subfields listed, IS collected work	This is a long list of terms such as "symphonies," "plays," "concertos," and so on.
Title	245	Contains "selections," IS collected	
	245	245 with multiple semi colons and doc type "rec"	
	246	If four or more v246 fields with ind2 = 2, 3, or 4, IS collected.	If more than 1 246, consider partial evidence
Extent	300	If 300$a has "pagination multiple" or "multiple pagings," IS collected	
Contents Notes	505$a and $t	1. Check $a for first and last occurrences of "movement". If Not multiple movement occurrences and does have	IF all / any the above produce more than one pattern instance or more

		multiple " / " pattern. 2. If the above doesn't find multiple patterns, also look for " ; " patterns. 3. If the above checks don't produce more than 1 pattern, look for multiple " – " patterns. 4. Count 505s $t cases. 5. Count $r cases.	than one $t, or more than one $r, IS collected.
Various fields for Thematic Index clues	505a	If any v505 $a, check for differing Opuses. (This also checks for thematic index cases too.) If found, IS collected.	For types Score and Recording
Related work	740	If 1 or more 740 and 1 has indicator 2 = 2", IS collected .	If only multiple 740s, partial evidence
Author	700/710/711/730	Check for $t and $n. And check 730 ind 2 value of "2." If 730 with ind2 = 2 or multiple $t is found, IS collected.	If only 1 $t, partial evidence
	100/110/111, 700/710 730	If format recording, and both records are collected work, require cast list match to cluster anything but manifestation matches.	That is, do not cluster at content level without verifying by cast.

Table 1. Checks on Bibliographic Records.

Frailties of Collected Works Identification in Well-Cataloged Records

The above table illustrates many areas in a bibliographic record that can be mined for evidence of aggregates. The problem is that cataloging practice offers no one rule mandatory to catalog a collected work correctly. Moreover, as WorldCat membership grows, the use of multiple schemes of cataloging rules for different eras and geographic areas adds to the complexity, even assuming that all the bibliographic records are cataloged "correctly." Correct cataloging is not assumed by the team.

Software Confounded

With all the checks outlined in the table, the team still found cases of collected works that seemed to defy machine detection.

One record had the two separate works, *Tom Sawyer* and *Huckleberry Finn*, in the same title field, with no other clues to the aggregate nature of the item.

The work *Brustbild* was another case. For this electronic resource set, *Brustbild* appeared to be the collection set title, but the specific title for each picture was given in the publisher field.

A cluster for the work *Gedichte von Eduard Morike* (score) showed problems with the uniform title which was for the larger work, but the cluster records each actually represented parts of the work.

The bad cluster for *Si ku quan shu zhen ben bie ji*, an electronic resource, contained records which each *appeared* to represent the entire collection of 400 volumes, but the link in each 856 field pointed only to one volume in the set.

Limitations of the Present Approach

The current processing rules for collected works adopt a strategy of containment. The problem may be handled in the near term by avoiding the mixing of collected works with noncollected works, but the clusters containing collected works need further analysis to produce optimal results.

For example, it is one thing to notice scores "arrangements" as a clue to the presence of an aggregate. The requirement also exists that an arrangement should not cluster with the original score. The rules for clustering and distinguishing different sets of arrangements present another level of complexity. Checks to compare and equate the instruments involved in an arrangement are quite difficult; in this team's experience, they fail more often than they succeed. Without initial explication of the rules for separating arrangements, reviewers quickly found clusters such as Haydn's *Schopfung*, which included records for the full score, vocal score, and an arrangement for two flutes.

An implementation that expects one manifestation to have the identifier of only one work is a conceptual problem for aggregates. A simple case: if the description of a recording of Bernstein's *Mass* has an obscurely placed note indicating the second side contains the work *Candide*, *Mass* is likely to be dominant in the clustering effect, with the second work effectively "hidden." This manifestation would seem to need three work IDs, one for the combination, one for *Mass*, and one for *Candide*. This does not easily translate to an implementation of the FRBR model but could perhaps be achieved via links. Several layers of links would seem necessary. A manifestation needs to link to its collected work. A collected work needs links to records for the individual works that it contains, and vice versa, individual works need to link to collective works. This can be important for translations, for example, into Russian, where collective works are common even where they do not exist in the original language.

Lessons Learned

First and foremost, plan to deal with collected works. For clustering efforts this must be addressed in some way for any large body of records.

Secondly, formats will gain the focus. The initial implementation of the GLIMIR algorithms used test sets mainly composed of a specific work. After all, GLIMIR clusters should all be formed within one work. These sets were carefully selected to represent as many different types of work sets as possible, whether clear or difficult examples of work set members. Plenty of attention was given to the compatibility of differing formats, given the looser content clustering.

These were good tests of the software's ability to cluster effectively and correctly within a set that contained numerous types of materials. Random sets of records were also tested to cross check for unexpected side effects. What in retrospect the team would have expanded was sets that were focused on specific formats. Recordings, scrutinized as a group, can show different problems than scores or books. The distinctions to be made are probably not complete.

Another lesson learned in GLIMIR concerned the risks of clustering. The deliberate effort to relax the very conservative nature of the matching algorithms used in GLIMIR was critical to success in clustering anything. Singleton clusters don't improve anyone's view. In the efforts to decide what should and should not be clustered, it was initially hard to discern the larger scale risks of overclustering. Risks from sparse records were probably handled fairly well in this initial effort, but risks from complex records needed more work. Collected works is only one illustration of risks of overclustering.

FUTURE RESEARCH

The current research suggests a number of areas for possible further exploration:

- The option for human intervention to rearrange clusters not easily clustered automatically would seem to be a valuable enhancement.
- There is next the general question, what sort of processing is needed, and feasible, to distinguish the members of clusters flagged as collected works?
- Part versus whole relationships can be difficult to distinguish from the information in bibliographic records. Further investigation of these descriptions is needed.
- Arrangements of works in music are so complex as to suggest an entire study by themselves. Work on this area is in progress, but it needs rules investigation.
- Other derivative relationships among works: Do these need consideration in a clustering effort? Can and should they be brought together while avoiding overclustering of aggregates?
- How much clustering of collected works may actually be helpful to persons or processes searching the database? How can clusters express relationships to other clusters?

CONCLUSION

Clustering bibliographic records in a database as large as WorldCat takes careful design and undaunted execution. The navigational balance between underclustering and overclustering is never easy to maintain, and course corrections will continue to challenge the navigators.

ACKNOWLEDGMENTS

REFERENCES

1. Barbara Tillett, "What is FRBR? A Conceptual Model for the Bibliographic Universe," last modified 2004, accessed November 22, 2013, http://www.loc.gov/cds/FRBR.html.

2. Janifer Gatenby, email message to the author, November 10, 2013.

3. International Federation of Library Associations (IFLA) Working Group on Aggregates, *Final Report of the Working Group on Aggregates, September 12, 2011*, http://www.ifla.org/files/assets/cataloguing/frbrrg/AggregatesFinalReport.pdf.

4. Maja Zumer and Edward T. O'Neill, "Modeling Aggregates in FRBR," *Cataloging and Classification Quarterly* 50, no. 5–7 (2012): 456–72.

5. IFLA Working Group on Aggregates, *Final Report.*

6. Zumer and O'Neill, "Modelling Aggregates in FRBR."

7. Gail Thornbug and W. Michael Oskins, "Misinformation and Bias in Metadata Processing: Matching in Large Databases," *Information Technology & Libraries* 26, no. 2 (2007): 15–22.

8. Gail Thornburg and W. Michael Oskins, "Matching Music: Clustering versus Distinguishing Records in a Large Database," *OCLC Systems and Services* 28, no. 1 (2012): 32–42.

9. Janifer Gatenby et al., "GLIMIR: Manifestation and Content Clustering within WorldCat," *Code{4}Lib Journal* 17 (June 2012),http://journal.code4lib.org/articles/6812.

10. Richard O. Greene, "Cataloging Alchemy: Making Your Data Work Harder" (slideshow presented at the American Library Association Annual Meeting, Washington, DC, June 26–29, 2010), http://vidego.multicastmedia.com/player.php?p=ntst323q.

11. Jenny Toves, email message to the author, December 17, 2013.

12. Arsen R. Papakhian, "The Frequency of Personal Name Headings in the Indiana University Music Library Card Catalogs," *Library Resources & Technical Services* 29 (1985): 273–85.

13. Sherry L. Vellucci, *Bibliographic Relationships in Music Catalogs* (Lanham, MD: Scarecrow, 1997).

14. Sherry L. Vellucci, "FRBR and Music," in *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*, ed. Arlene G. Taylor (Westport, CT: Libraries Unlimited, 2007), 131–51.

15. Jenn Riley, "Application of the Functional Requirements for Bibliographic Records (FRBR) to Music," www.dlib.indiana.edu/~jenlrile/presentations/ismir2008/riley.pdf.

16. Donald W. Krummel, "Musical Functions and Bibliographic Forms," *The Library*, 5th ser. 31 (1976): 327–50.

17. Barbara Tillett, "Bibliographic Relationships: Toward a Conceptual Structure of Bibliographic Information used in Cataloging," (PhD diss., Graduate School of Library & Information Science, University of California, Los Angeles, 1987), 22–83.

18. Program for Cooperative Cataloging (PCC) Task Group on the Creation and Function of Name Authorities in a Non MARC Environment, "Report on the PCC Task Group on the Creation and Function of Name Authorities in a Non MARC Environment," last modified 2013, http://www.loc.gov/aba/pcc/rda/RDA%20Task%20groups%20and%20charges/ReportPCCTGonNameAuthInA_NonMARC_Environ_FinalReport.pdf.

19. Music Library Association, Authorities Subcommittee of the Bibliographic Control Committee, "Thematic Indexes Used in the Library of Congress/NACO Authority File," http://bcc.musiclibraryassoc.org/BCC-Historical/BCC2011/Thematic_Indexes.htm.

20. Jay Weitz, email message to the author, May 6, 2013.