**RDA phase 3: Interim report**

January 12, 2015

**Introduction**

In March 2014 the Program for Cooperative Cataloging appointed the PCC RDA Authorities Phase 3 Task Group to consider RDA-related changes to the LC/NACO Authority File (NAF) beyond those accomplished by an earlier Task Group.[[1]](#footnote-1) There are two major tasks to be achieved in this work:[[2]](#footnote-2)

1. All AACR2 records whose 1XX field contains no elements contrary to RDA are to be re-coded as RDA.
2. 024 fields for ISNIs and possibly other identifiers should be added, subject to various limiting factors.

The group was also to consider other changes to authority records. For example, certain 678 fields should be re-formulated as 670 fields, and fuller consideration could be given to subfield $c in access fields for personal names. Some of these changes will affect the 1XX field, and some would not.

Because the Phase 3 work requires the updating of a majority of the records in NAF, the group was charged to work closely with the NACO nodes to ensure that they were aware of the upcoming replacement of a large number of records, and to help negotiate a schedule for the work.

The Task Group has been busy in the intervening months. It has prepared a number of documents detailing various aspects of this project. The Task Group has already begun to develop, and test, elements of the computer program that will be used to make the conversion. The Task Group has also received requests to add other manipulations to its work; these changes are chiefly based on revisions to RDA made in April 2014.

The Task Group has been in close contact with staff at the Library of Congress, and through LC staff with the NACO nodes. All parties are coming to an understanding of how this massive project can be accomplished. The process and schedule that appear to fit all needs are sufficiently different from the original concept that the Task Group believes it best to present this intermediate report to the larger community.

This interim report describes the state at which these complicated matters stand at the current time. The report describes in general terms the work to be performed (with reference to more specific documents), and proposes a schedule for performing the work.

**Scheduling considerations**

Phases 1 and 2 of work on the LC/NACO Authority File were performed by a program that was limited to 30,000 updates per day. The program started with the first record in the file (the record with the lowest system-assigned control number); if that record required a change the program updated it, and then moved on to the next record. When the program had updated the 30,000th record, it stopped. The next day, it took up the work where it left off, and continued until it had again found 30,000 records to change. The program accomplished each phase of this work in about a month's time.

Taken together, Phases 1 and 2 updated about 500,000 authority records. By way of contrast, because most AACR2 authority records contain 1XX fields that are acceptable for use under RDA, millions of authority records will be updated for Phase 3. Because a rate of 30,000 changes per day means that it would take a program about one year to complete the work (Monday-Friday, with additional allowance for holidays, vacations and other interruptions), the NACO nodes were asked if it would be possible instead to perform all of the work at one go: the NACO nodes would receive all of the changes as one large update, and would load the entire set. The NACO nodes were receptive to this revised plan for doing the work, but with one important proviso: they asked that the changed records be divided into two groups: records in which the 1XX field is being changed must be issued separately from records in which the 1XX field is not being changed.

There are yet other factors that affect the scheduling of this work. In the case of the Voyager system at the Library of Congress, the easiest way to re-load such a large number of records is to perform what is called a "slam" load, in which the entire file is replaced.[[3]](#footnote-3) Performing this work involves ExLibris staff (ExLibris are the providers of the Voyager system); and paying ExLibris to do the work requires a purchase order. Internal considerations dictate that such a purchase order can only be issued to cover work that begins on or after April 1, 2015. In the case of the British Library, upgrades to their Aleph system (also provided by ExLibris) will occur in the first quarter of 2015. Although the British Library system might be able to absorb a slightly increased number of daily updates (in the range of 5,000 per day) during the first quarter of 2015, it will not be possible to absorb a larger number, and certainly not 30,000 per day.

Calls for changes to authority records have arisen since the Task Group was formed. The April 2014 updates to RDA contained two groups of changes that can be performed by program:

* Changes to the medium of performance for music (these can occur in the 1XX, 4XX and 5XX fields)
* Changes to relator terms (in authority records, these are limited to the 5XX block of fields)

Especially in the case of the changes to the medium of performance for music, it would be convenient if the work could be performed as quickly as possible. This would not only save catalogers from having to make changes to individual records (with the accompanying possibility for error), but would help enforce consistency within NAF.

These factors lead us to the inevitable conclusion that the work originally termed "Phase 3" must be divided into two distinct projects. For convenience and future reference, these will be identified by uppercase alphabetic characters "A" and "B".

**Phase 3A:** *Work that involves changes to the authority 1XX field.* Because of the various scheduling constraints, this phase must be completed **before January, 2015**.[[4]](#footnote-4) This keeps the file internally consistent, and gets changes to 1XX fields out of the way before the British Library begins its system work. This means in turn that the elements to be included in this part of the project must be fully defined, and the programming completed and tested, as soon as possible. This work can proceed at a rate of up to 30,000 updates per day. The work was performed in the same manner as Phase 2: as a phased sweep through the entire NAF.

**Phase 3B:** *The re-coding of large numbers of AACR2 records as RDA, with no change to the 1XX field.* This work cannot take place until **after April 1, 2015.** The definition and programming for this work is simpler than the work for the first part, but it all needs to be in place, and tested to the extent possible, before April 1, 2015. (Further testing will probably occur after April 1, 2015.) The ease with which this project can be defined and programmed is more than offset by the complexity of the work required by the various NACO nodes to handle the large number of updated records.

**Phase 3A: Definition**

Phase 3A involves work that makes changes to the 1XX field of authority records. More specifically, it involves the following categories of change:

* Changes involving the music medium of performance statement (subfield $m)[[5]](#footnote-5)
* Changes involving subfield $c of personal names[[6]](#footnote-6)
* Make sure that authority records for undifferentiated personal names are appropriately labeld[[7]](#footnote-7)
* Re-formulate instances of field 678 that are not suitable for public display, as 670 fields.[[8]](#footnote-8)
* All changes to the 1XX previously defined for Phases 1 and 2[[9]](#footnote-9)
* Evaluate the contents of the 370 field[[10]](#footnote-10)

It was originally part of the Phase 3A scheme to replace obsolete indicators in the 1XX field with blanks, and to remove unnecessary terminal punctuation. It was not possible to generate an estimate of the scope of this work before the production run of Phase 3A began. After the first few days of the production part of Phase 3A, the repair of obsolete 1XX indicators was found to create too many updates, and this part of the conversion program was turned off. This means that many 1XX fields in older records will continue to feature obsolete indicators and unnecessary terminal punctuation.

As noted above, all records changed during either Phase 3A or 3B will be subject to additional changes. Most of these were already defined for Phases 1 and 2, but at least one change is new, and one has been expanded.[[11]](#footnote-11)

* Generate the 046 field for personal names[[12]](#footnote-12)
* Generate the 382 field from certain music medium of performance statements[[13]](#footnote-13)
* Generate the 368 $c, 368 $d, 370, 371, 372, 373, 374 and/or 375 fields from information derived from subfield $c of personal names[[14]](#footnote-14)

If the program makes any change to a record during Phase 3B, the program will reset the 005 field in the finished record, to reflect the current date and time. (Because of its different working method, this is not necessary during Phase 3A.)

The Task Group has considered, but has chosen not to pursue, the following work:

* Generate the 375 field from "His" or "Her" at the beginning of authority 670 fields. A sampling of records indicated that pronouns were used far too carelessly to serve as the basis for derived information.
* Remove ordinal numbers at the beginning of titles of nobility in subfield $c (for example, change" 2d Earl of Brompton" to just "Earl of Brompton")
* Delete authority records that represent multiple language expressions (subfield $l contains "&", or "Polyglot"). Although these records may be deleted, and the Task Group hopes that they will be, this work will be independent of the Task Group's operations.

**Phase 3B: Definition**

Phase 3B involves everything not yet done that can be done to a record, as long as the change does not involve the an authority record's 1XX field. More specifically, it involves all of the categories of change defined for Phase 3A (most of which should now have no effect), plus the following:

* Replace obsolete indicators in 4XX and 5XX fields with blanks[[15]](#footnote-15)
* Supply qualifiers for initialisms and acronyms in 410 fields, when that can be done reliably.[[16]](#footnote-16)

**Appendix A: Slam load at the Library of Congress for Phase 3B**

It will be more efficient for the Library of Congress to replace its entire authority file, rather than to make replacements of millions of individual authority records. In Voyager parlance, the replacement of the entire contents of a file in this manner is called a *slam* load.[[17]](#footnote-17) To make matters more complicated, there are actually two separate authority files in the LC system to be updated, and the contents of the two are not identical.

* LCDB: This is LC's main production database. The authority file contains the "official" copy of the LC/NACO authority file, plus an unofficial copy of the LCSH authority file.
* CDS: This database is used to generate various products that are distributed to other organizations. The authority file contains a second "official" copy of the LC/NACO Authority File, plus the "official" representation in MARC format of the LCSH authority file. This database also contains a copy of each record that has been *deleted* from either of these files.

Special modifications are needed to the conversion program to make the updating of LC's database possible. These features, coupled with steps to be taken at the Library of Congress, will ensure that at the end of Phase 3B, both databases contain the same set of records as they did at the beginning. The following is an outline of the complex steps that will be needed in order for this to happen as planned.

* All record-sharing between the NACO nodes will be suspended, and all pending updates will be propagated as appropriate.
* The entire LCDB authority file will be extracted to one or more files in MARC communication format. (Segmenting the work will not only make recovery in case of a problem simpler; it will also make it possible for the Phase 3B work to be performed on multiple workstations simultaneously.)
* Certain parts of the CDS authority file will be extracted to files in the MARC communication format: the active LCSH records, the deleted LCSH records, and the deleted LC/NACO records.
* The conversion program will process the records exported from LCDB, and will produce two output streams.[[18]](#footnote-18) One output stream will contain all of the records in the input file, whether changed or not. The second output stream will contain just the LC/NACO authority records, whether changed or not.
* The first output stream will be used to replace the LCDB authority file in its entirety. The second output stream will be merged with the records previously extracted from the CDS authority file (active LCSH records, and deleted LCSH and LC/NACO records), to replace the CDS authority file in its entirety.
* Record-sharing between the NACO nodes will be re-started. The NACO nodes will receive as part of their next request for updated records, and load according to their own best lights, the millions of authority records updated during Phase 3B.
1. The group's charge is available here: <http://files.library.northwestern.edu/public/rdaphase3/docs/> [↑](#footnote-ref-1)
2. As was the case for Phases 1 and 2, all records updated for any reason during Phase 3 will be considered for additional changes; this secondary work should be assumed through this document. For example, an authority record for a personal name containing subfield $d will be considered for the automated generation of an 046 field. When Phase 3 is complete, all records in NAF will have been reissued for one RDA-related reason or another. [↑](#footnote-ref-2)
3. Appendix A gives an idea of just how complex this process will be for the Library of Congress. At the time of writing the planning document outlining each step runs to three full pages, with the expectation that the document will only grow as the date approaches. [↑](#footnote-ref-3)
4. LC asked that this work be completed by December 15, 2014. In reality, the work was completed on January 6, 2015. [↑](#footnote-ref-4)
5. Described in *RDA phase 3. Music performance medium.docx* [↑](#footnote-ref-5)
6. Described in *RDA phase 3. Subfield c.docx*. Some subfield $c texts previously being declared to be not acceptable for use under RDA will be re-categorized as acceptable under RDA. [↑](#footnote-ref-6)
7. Described in *RDA phase 3. Undifferentiated personal names.docx* [↑](#footnote-ref-7)
8. Described in *RDA phase 3. Authority 678 field.docx* [↑](#footnote-ref-8)
9. Although there has been substantial effort made to prevent new and updated authority records from containing elements not in accordance with RDA, it is inevitable that some will have slipped through. For example, authority records continue to be created that use a formulation such as "b. 1902" in subfield $d of personal names, even though this is not the accepted structure. [↑](#footnote-ref-9)
10. Described in *RDA phase 3. Re-examining the 370 field.docx* [↑](#footnote-ref-10)
11. These changes are described in more detail in *RDA phase 3. When a record is otherwise updated.docx* [↑](#footnote-ref-11)
12. During Phases 1 and 2, this work was limited to cases in which the authority 100 field contained subfield $d. This work has been expanded to include elements buried in the 670 field. This additional work is described in *RDA phase 3. Mining authority 670 fields for birth and death dates.docx* [↑](#footnote-ref-12)
13. This is limited to cases of standard combinations of instruments, and is described in *RDA phase 3. Music 382 field.docx* [↑](#footnote-ref-13)
14. This work is described in the document referenced earlier that also describes the enhanced treatment of subfield $c of personal names: *RDA phase 3. Subfield c.docx* [↑](#footnote-ref-14)
15. Described in *RDA phase 3. Obsolete indicators and punctuation.docx* [↑](#footnote-ref-15)
16. Described in *RDA phase 3. Qualification of 410 fields.docx* [↑](#footnote-ref-16)
17. During a slam load, some or all indexing features are disabled, which vastly speeds up the process. Indexes are regenerated as a separate step once the database has been rebuilt. [↑](#footnote-ref-17)
18. Each stream may consist of multiple files. [↑](#footnote-ref-18)