Published in final edited form as:

Cat Classif Q. 2018 ; 56(2-3): 262–272. doi:10.1080/01639374.2017.1382642.

Preparing for the Future: National Library of Medicine's® Project to Add MeSH® RDF URIs to its Bibliographic and Authority Records

Diane Boehr and Barbara Bushman

Cataloging and Metadata Management Section, National Library of Medicine, Rockville Pike
MSC, Bethesda

Background

In May 2011, the Library of Congress (LC) issued a statement entitled *Transforming our Bibliographic Framework* where they introduced an initiative to investigate a replacement for MARC21 as the main encoding standard for library bibliographic data.¹ Although MARC's replacement is still being developed, the community generally agrees that library data will need to be transformed to facilitate its move into a linked data environment. The National Library of Medicine (NLM) has been active in this movement to transform bibliographic data to linked data since being asked to participate in the Bibliographic Framework Initiative (BIBFRAME) Early Experimenters in 2012.² NLM has participated in the efforts to review and revise the BIBFRAME vocabulary and models, test BIBFRAME editor tools, and investigate other linked data models. Simultaneous to these efforts, NLM published the beta version of Medical Subject Headings (MeSH)³ in Resource Description Framework (RDF) in November 2014, with the latest update released November 2016.⁴

Since 1999, subjects in bibliographic records in NLM's catalog have been recorded in the same manner as indexing records in PubMed. Topical terms are recorded independently from geographical terms and publication types (aka form/genre terms). However, when distributing subjects to subscribers, NLM artificially created subject strings that more closely matched traditional Library of Congress Subject Headings (LCSH). In preparing for a linked data environment, NLM recognized that multifaceted subject strings do not work well in a linked data environment because there is not a one-to-one correspondence between the entire subject string and a single unique identifier. Therefore, NLM ceased distributing its MARC bibliographic records with artificially reconstructed subject strings as of December 2015. NLM MARC records continue to have MeSH combined with the appropriate topical subheadings (in MARC 21 field 650 \$a \$x) but geographic locations and publication types appear in separate fields (MARC 21 651, 655) rather than as subfields of the MARC 21 650 field as shown in Figure 1.⁵

As part of the library community effort to prepare current bibliographic data for the eventual transfer to a linked data environment, NLM has been participating in the Program for Cooperative Cataloging (PCC) Task Group on URIs in MARC. This task group is charged to: “1) Identify and address any immediate policy issues surrounding the use of identifiers in MARC records that should be resolved before implementation proceeds on a large scale; 2) ... develop guidelines for including identifiers in MARC bibliographic and authority records; 3) Develop a work plan for the implementation of identifiers in \$0 and other fields/subfields in member catalogs and in PCC-affiliated utilities; and 4) ... identify and prioritize any remaining issues concerning support for identifiers in the MARC format, and initiate MARC proposals as appropriate.”⁶ As part of the this task group, NLM has collaborated with program developers for tools such as MarcEdit⁷ and Authority Toolkit⁸ to test programmatically adding Uniform Resource Identifiers (URIs) from ontologies and vocabularies modeled in RDF to MARC records.

Having successfully tested the addition of MeSH RDF URIs to various fields in NLM MARC records using the MarcNext component of MarcEdit, NLM decided to add RDF URIs to all the MeSH in the library’s bibliographic records in its production environment. These records with MeSH RDF URIs will be distributed via NLM’s normal methods to OCLC, the Library of Congress, and direct subscribers.⁹ The distribution of these augmented records will provide other libraries with a large data set to examine and use for further experimentation.

Since NLM is a national library and a PCC participant we feel a responsibility to follow existing standards. Before distributing NLM records containing MeSH RDF URIs we need to be sure our data conforms with PCC guidelines if they are going to be coded as CONSER or BIBCO. The PCC group investigating linked data and making recommendations on how and where to record that data in MARC is still in the process of submitting proposals to the MARC Advisory Committee (MAC) for new and redefined fields and subfields and recommending best practices for PCC. The PCC Task Group on URIs in MARC is discussing how to define within a MARC field the subfields that singly or combined comprise the RDF object of that field. If only a portion of a MARC field is represented by the RDF URI applied to that field, then in MARC to RDF conversion both data and understanding may be lost and the resulting triple statement may be false. For subject headings, this means that every possible subfield combination must be represented in its entirety by an RDF URI. The MeSH RDF has been developed in a manner so that the URIs fully represent the data in MARC 21 650 \$a and \$x (if present), the 651 \$a and \$x (if present), and the 655 \$a. Although MAC has approved the redefinition and use of \$0 in MARC for recording machine actionable RDF URIs,¹⁰ the PCC has yet to approve this practice for its member institutions. Until that time, PCC members cannot make use of \$0 for RDF URIs and still code their data as CONSER or BIBCO. Because the MeSH RDF URIs meet the requirement for representing the entire RDF object, they are an ideal set of test data for adding RDF URIs to MARC. Based on this, the PCC granted NLM permission to include these URIs in BIBCO records at the November 2016 Policy Committee meeting¹¹.

Starting with MeSH

Much of the community discussion with regards to adding RDF URIs to MARC data has focused on adding them for names of people and corporate bodies associated with a resource. However, NLM decided to start with adding RDF URIs only for the MeSH in the MARC 21 650, 651, and 655 fields. Limiting the scope of this initial effort has allowed us to break down the larger goal of preparing for linked data into manageable smaller tasks. We hope that sharing lessons learned from this process will inform future efforts across the community.

Because MeSH is NLM's own medical subject thesaurus for cataloging and indexing, we control its accuracy, stability, and currency. MeSH can be viewed in the MeSH Browser¹² or downloaded in XML, ASCII, MARC, and RDF formats.¹³ While MeSH contains several different categories of data, catalogers are only concerned with two types of MeSH records, Descriptors and Qualifiers. Descriptors, also referred to as main headings, are used to indicate the subject of a cataloged item. Qualifiers, also referred to as subheadings, are used in conjunction with a Descriptor to express a particular aspect of a subject (e.g., Liver—drug effects).¹⁴

For each MeSH Descriptor, there is a list of allowable qualifiers which can be combined to create a coordinated subject heading. The allowable qualifiers listed for each Descriptor are the only qualifiers permitted with that Descriptor. By recording geographical headings and publication types in separate MARC fields rather than stringing them with the topical headings and qualifiers, the possible number of valid subject headings that can be constructed with MeSH is limited, as opposed to the limitless possible combinations of main heading and free-floating subdivision combinations in LCSH. The finite number of subject headings allows assignment of unique RDF URIs for every allowable concept.

The beta version of MeSH RDF was first released in November 2014 and is current as of the 2017 MeSH vocabulary. In MeSH RDF, a dereferenceable URI was minted for each Descriptor and Qualifier by appending its unique identifier (starting with either D or Q) to the base URI of <https://id.nlm.nih.gov/mesh>. For Descriptor-Qualifier combinations (known as Descriptor-Qualifier Pairs in MeSH RDF), the unique identifier is represented as the identifier for the Descriptor followed by the identifier for the Qualifier. This new unique identifier is then appended to the base URI of <https://id.nlm.nih.gov/mesh/>. The unique identifiers for the Descriptor-Qualifier pairs are unique to MeSH MARC and to MeSH RDF and do not exist in MeSH produced in other formats. The MeSH subjects in Figure 1 are shown with their corresponding URIs in Table 1.

Method for Updating the Bibliographic Records

After making the decision to add \$0 to the fields noted above, NLM considered how to add the MESH RDF URIs, and when to perform this process. NLM recognized it is critical to add \$0 to bibliographic records programmatically as it is important that this machine-readable information be scrupulously accurate. Adding these data manually, even using copy

and paste, can lead to errors. When selecting a programmatic option to add this data, NLM considered the following issues:

- Examining the costs/benefits of using existing software versus custom programming;
- Utilizing the same program for handling the legacy data and ongoing cataloging;
- Knowing the number of existing records in the catalog and the average amount of cataloging output on a regular basis.

NLM began by experimenting with MarcEdit's MARCNext tool, which allows users to embed HTTP actionable URIs for RDF vocabularies into the MARC 21 1XX, 3XX, 6XX, and 7XX fields of MARC records by exact matching against the text strings in those fields. NLM collaborated with the MarcEdit developer to include MeSH RDF in MARCNext. Subsequent testing of MARCNext on MeSH in NLM bibliographic records indicated a 100% accuracy on matching text strings to RDF URIs.

NLM uses Ex Libris' Voyager for its cataloging. In Voyager, whenever more than 10,000 records are updated, the indexes must be regenerated. This is a time-consuming activity, which requires system downtime. NLM's catalog contains approximately 1.5 million bibliographic records. Using MARCNext to update that many records in batches of 10,000 records or less to avoid reindexing would take several months, and would require separate processes to ensure that URIs are added to new records and maintained when records are updated. Given the complexities of performing this work in small batches, NLM decided to develop its own programming to add \$0 to the MARC 21 65X fields for all the legacy data at once. We have not eliminated the possibility of eventually using MARCNext to add \$0 RDF URIs to other authority controlled fields such as personal and corporate names.

Currently, MeSH MARC authority records are loaded into Voyager and are used for validation of the bibliographic records. Ideally, the MeSH RDF URI would already be available in the MeSH-MARC file and not have to be constructed, but the file is created from an ASCII version of the MeSH vocabulary and does not contain the MeSH RDF URIs. Therefore, we needed to programmatically construct the dereferenceable URI from MeSH Unique Identifier which is present in the record and to store the RDF URI in the Voyager authority record. The RDF URI could then be pulled directly into the bibliographic record from the authority record. This approach would allow us to update the legacy records in about 58 hours (over a long weekend). The same process could easily be used to update new records on a regular basis, which NLM has chosen to do weekly.

Once we chose to store the MeSH RDF URIs in the MARC authority record and create a custom program to add the RDF URIs to the subjects in the bibliographic record, we needed to develop two programs: one for constructing and adding the RDF URIs to the MARC authority record and a second program to populate the bibliographic records. The program to populate bibliographic records with MeSH RDF URIs can be used to process all the legacy records as well as records for any newly cataloged resources.

Updating Legacy Data

Before we could write the initial program (Program 1) to add the RDF URIs to the authority records, we had to resolve several technical issues. Currently, there is no approved place in the MARC authority format to store the RDF URI representing the authority record itself. NLM decided to store the RDF URIs in a locally defined field 029 with the RDF URI in \$0, which required the addition of the 029 \$0 to our Voyager tag table as a local field. Because RDF URIs are not currently stored in the MeSH MARC data, the program would need to both construct the RDF URIs and then input them into the authority record. Since there are 850,000 authority records to edit, the records must be extracted from Voyager, updated, and then reloaded. Here are the steps for program 1.

Step 1. Extract All MeSH Authority Records from Voyager

Step 2. Construct the RDF URI by appending the MeSH ID taken from the MARC 21 035 field to <https://id.nlm.nih.gov/mesh/>

Step 3. Create field 029 \$0 containing the constructed MeSH RDF URI

Step 4. Reload authority records into Voyager

Program 2 to add the MeSH RDF URIs to the bibliographic records will use the MeSH RDF URIs generated from Program 1. Here are the steps for adding the MeSH RDF URIs to the bibliographic records.

Step 1. Extract All MeSH Authority Records from Voyager. All now have RDF URIs in the 029 field

Step 2. Extract Bibliographic Records from Voyager which have a local status of completed and contain MARC 21 65X fields

Step 3. Match each of 650, 651, and/or 655 fields in every extracted bibliographic record with 150, 151, or 155 field in the extracted MeSH authority records by \$a (if neither has \$x) or \$a and \$x combined.

Step 4. Generate a report listing any bibliographic records where one or more 65X fields failed to match an authority 15X field. These records contain invalid MeSH and need to be reviewed and revised.

Step 5. For each 65X in a bibliographic record with a matching MeSH authority 15X field, add \$0 as the final subfield with the MeSH RDF URI from the 029 in the MeSH authority record.

Step 6. Update the local NLM tag in the bibliographic record to indicate that this record has been updated. This will trigger the record to be included in the set of modified records distributed to subscribers.

Step 7. Reload bibliographic records into Voyager

Figure 2 illustrates the program workflow. Table 2 provides an example of the outputs of the programs.

Timing of the Process

After deciding to add the \$0 and writing the various programs, we considered when to run the programs to add the RDF URIs to our records. We had to weigh several factors in determining the timing: other demands on our system and system development staff, and the potential impact on our users, subscribers, and other NLM products that use our MARC data. Loading and re-indexing our data requires system downtime for our Voyager ILS. This project had to be carefully coordinated with other NLM projects impacting the ILS and our programming staff. The project could not conflict with the NLM annual Year End Processing, when the MeSH vocabulary is updated; the NACO Phase 3 update, where all NACO nodes (including NLM) must update their authority files with the mass changes done by LC to upgrade records to Resource Description and Access (RDA); and an upgrade to the latest release of the Voyager ILS. NLM prefers to schedule large update projects over a long weekend which lessens staff and user down-time and provides additional time to process large files or recover from any identified problems. Therefore, NLM selected the weekend leading up to July 4 to update the legacy data.

Ongoing Processing of New Bibliographic Records

Simultaneous to developing the program to handle the legacy data, we developed another automated mechanism for adding RDF URIs to new and/or changed records. While updating individual records could be done using MarcNext, it would not fit in well with our existing cataloging workflow. Since Program 2 contained the needed functionality to add RDF URIs to ongoing cataloging, we decided it was most efficient to use it with some minor changes. The steps for adding RDF URIs to new and updated records (Program 3) are:

- Step 1. Extract all bibliographic records that have been added or updated within the last week
- Step 2. Match each of MARC 21 650, 651, and/or 655 fields in every extracted bibliographic record with 150, 151, or 155 field in the extracted MeSH authority records by \$a (if neither has \$x) or \$a and \$x combined
- Step 3. Generate a report listing any bibliographic records where one or more 65X fields failed to match an authority 15X field. These records contain invalid MeSH and need to be reviewed and revised
- Step 4. For each 65X in a bibliographic record with a matching MeSH authority 15X field, add \$0 as the final subfield with the MeSH RDF URI from the 029 in the MeSH authority record
- Step 5. For each 65X in a bibliographic record already containing a \$0, verify the MeSH RDF URI matches the RDF URI in the 029 of the corresponding authority record. If they do not match, replace the existing MeSH RDF URI with the correct one. Note each 65X must have only one \$0 RDF URI
- Step 6. Update the local NLM tag in the bibliographic record to indicate that this record has been updated. This will trigger the record to be included in the set of modified records distributed to subscribers.

Step 7. Reload bibliographic records into Voyager

Testing of the Programs

Before actually updating records in the production environment, NLM ran several tests of the programs in our Voyager test environment. To minimize processing time, our initial goal was to process 100,000 records at a time, but that proved to overwhelm the system, so the records were loaded in batches of 10,000 records as recommended by Voyager documentation.

The first test run of Program 1 demonstrated that our authority data was very clean but a few MeSH authority records had either never been loaded or accidentally deleted from the authority file. These were easily identified and restored. The first test run of Program 2 found ca. 4100 invalid MeSH in our bibliographic records. While 4100 errors was a somewhat daunting number, when considering that our catalog contains over 1.5 million records with an estimated average of 3 MeSH each, that is actually only a .001% error rate. Most of the errors were spacing problems, capitalization, punctuation, or incorrect MARC tagging which cannot be identified in Voyager's built-in validation program. But when matching on text strings it is critical that the strings be 100% accurate and these 4100 errors had to be corrected before proceeding. By sorting the errors into categories 80% of the incorrect MeSH were corrected in an automated fashion using MarcEdit. The remaining 800 records were distributed to staff for cleanup, which was completed in two days.

One of the bonuses of this project is that by running Program 2 on weekly basis we are ensuring the subjects in our catalog remain pristine, as it is impossible to eliminate all human errors.

Impact on Other NLM Processes

NLM will require new procedures to ensure that URIs remain current in its authority and bibliographic files during the annual update of authority and bibliographic records to accommodate updated MeSH vocabulary. A new annual programming task will be needed to add a MeSH RDF URI in the 029 field to the new and revised authority records, just as is being done for our legacy data in Program 1. For all bibliographic records where the MARC 21 65X fields have changed due to changes in the MeSH vocabulary, the MeSH RDF URI will need to be added or updated in \$0 as happens in Program 3.

Impact on the Community

Data created by NLM resides in its own catalog and is distributed to OCLC and other subscribers. Before making large-scale changes to our bibliographic database, we notify subscribers so they are prepared for the change. Subscribers to NLM data must decide to either take or remove \$0 from the data they receive. All libraries that incorporate MARC 21 65X \$0 data into their catalogs need to think about how it might affect their public displays. In fact, as a Cataloging in Publication (CIP) partner with LC, NLM had to consult with LC about the effect a 65X \$0 would have on the CIP data display in a printed book. After LC determined that the data would be sent to the publishers, NLM agreed to send LC its CIP

records before \$0 is added to the 65X fields. Both libraries agreed the CIP data block in a published book should not visibly display data that is meant solely for machine processing. For the same reason, NLM has also chosen not to display the \$0 in the default views of its two public catalog interfaces, LocatorPlus and the NLM Catalog; however, the RDF URIs will be visible in the raw data views of both services.

Lessons Learned

The lessons we learned as we engaged in the process to add MeSH RDF URIs to our records have been invaluable. We furthered our understanding of the impact of the change on our systems and users. Adding RDF URIs to bibliographic data is a complex task with many moving parts, that needs to start with scrupulously clean data for optimal results. It requires understanding the data and the implications of changing that data. It requires planning and coordination with staff throughout the library to determine the impact on other systems and processes. For NLM, it also involves coordination with our cataloging partners and the larger library community. NLM's role as a national library presents unique challenges when attempting to make significant changes to its bibliographic data, but every library will face challenges. Any library that wishes to enhance its data with RDF URIs should consider the effect those changes will have on other parts of the library, end users, and/or cooperative cataloging partners. We encourage more libraries to start investigating these issues within their own institutions to better prepare themselves and their data for a linked data environment.

Next Steps

Now that we have added MeSH RDF URIs to our bibliographic and authority records, we will begin to look at other areas where we might enhance our data in preparation for a move to a linked data environment. We will be collecting feedback from consumers of our enhanced data. Pending the outcome of distributing our records with RDF URIs, as well as any PCC recommendations regarding adding RDF URIs to bibliographic and/or authority records, we will investigate adding RDF URIs to other MARC fields including MARC 21 6XX (i.e. 600, 610, 611 and 630) fields not included in this effort, 1XX, and/or 7XX fields. We remain involved with the PCC Task Group on URIs in MARC and will utilize our experience to help inform future recommendations and best practices. We will continue to work with developers of MarcEdit, the Authority Toolkit, and possibly other tools, which can help us along the way. We will work within NLM to ensure that MeSH RDF URIs are distributed in MeSH data. We will investigate adding terms and RDF URIs from external vocabularies to our bibliographic records, such as the Getty Thesaurus of Geographic Names. We also want to look at other NLM products and services and coordinate linked data efforts to further connect more of our data together.

Conclusion

To better position ourselves to participate in a future in which library data are exposed to the web and can interact with data from other sources, we need to prepare our MARC data for the transition to a linked data environment. Adding MeSH RDF URIs to NLM's

bibliographic and authority records does not turn these records into linked data, as they are still siloed in our catalog. However, it is the first step needed before we can plan to publish our data on the Web. Having completed this process and reporting back to the community on our experiences, NLM will be in position to assist the community in future implementation.

It is our hope that by outlining NLM's challenges and decision-making process, we can assist other libraries in understanding the complexities of this work. Some of NLM's challenges were unique, but we've endeavored to provide a path for other libraries to follow. It will take the combined effort of the entire community to move library data into a linked data environment. It will not be a quick and easy transition, but we can accomplish it, by taking one step at a time.

Acknowledgments

Acknowledgements: The authors wish to thank the following NLM staff members for their assistance with this effort: Tuck Arnold, Nancy Fallgren, Iris Lee, Jennifer Marill, Tina Shrader, Mark Vereb This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Reference

1. "Transforming our Bibliographic Framework: a statement from the Library of Congress, May 13, 2011," Library of Congress, accessed March 29, 2017, <https://www.loc.gov/bibframe/news/framework-051311.html>.
2. "Bibliographic Framework Initiative," Library of Congress, accessed March 15, 2017, <http://www.loc.gov/bibframe/>.
3. Medical Subject Headings," U.S. National Library of Medicine, accessed March 29, 2017, <https://www.nlm.nih.gov/mesh/>.
4. "Medical Subject Headings (MeSH) RDF Linked Data (beta)," U.S. National Library of Medicine, accessed March 15, 2017, <https://id.nlm.nih.gov/mesh>.
5. Boehr D, "Discontinuing Distribution of Cataloging Bibliographic Records with Artificially Reconstructed Subject Strings, NLM Technical Bulletin, No. 404 (May-Jun 2015), accessed March 15, 2017, https://www.nlm.nih.gov/pubs/techbull/mj15/mj15_cataloging_unstringing_survey.html.
6. "PCC Task Group on URIs in MARC Charge," Program for Cooperative Cataloging, accessed March 15, 2017, <http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html>.
7. "MarcEdit," MarcEdit Development, accessed March 15, 2017, <http://marcedit.reeset.net/>.
8. "Authority toolkit: create and modify authority records," Strawn Gary L., accessed March 15, 2017, <http://files.library.northwestern.edu/public/oclc/documentation/>.
9. For any libraries who do not currently subscribe to NLM's Catalog Record Data services, they are freely available at <https://www.nlm.nih.gov/databases/download/catalog.html>.
10. "MARC 21 Format for Bibliographic Data, Appendix A – control subfields," Library of Congress, November 2016, accessed March 10, 2017, <http://www.loc.gov/marc/bibliographic/ecbdcntf.html>.
11. PCC Policy Committee, PCC Policy Committee (PoCo) Meeting, November 3–4, 2016 Outcomes, p.4, accessed April 4, 2017, <https://www.loc.gov/aba/pcc/documents/PoCo-2016/PoCo-2016-Outcomes.docx>.
12. "Medical Subject Headings 2017," U.S. National Library of Medicine, accessed March 16, 2017, <https://meshb.nlm.nih.gov/#/fieldSearch>.
13. "Files Available to Download: 2017 MeSH Files," U.S. National Library of Medicine, accessed March 16, 2017, <https://www.nlm.nih.gov/mesh/filelist.html>.
14. "MeSH Record Types, U.S. National Library of Medicine," accessed March 16, 2017, https://www.nlm.nih.gov/mesh/intro_record_types.html.

In the NLM database:

650 12 \$a Acquired Immunodeficiency Syndrome \$x ethnology

650 22 \$a Cross Cultural Comparison

650 22 \$a Health Policy

651 _2 \$a Africa \$x ethnology

651 _2 \$a Caribbean Region \$x ethnology

655 _2 \$a Congresses

Subject strings created for distribution prior to 2016:

650 12 \$a Acquired Immunodeficiency Syndrome \$x ethnology \$z Africa \$v Congresses

650 12 \$a Acquired Immunodeficiency Syndrome \$x ethnology \$z Caribbean Region \$v Congresses

650 22 \$a Cross Cultural Comparison \$z Africa \$v Congresses

650 22 \$a Cross Cultural Comparison \$z Caribbean Region \$v Congresses

650 22 \$a Health Policy \$z Africa \$v Congresses

650 22 \$a Health Policy \$z Caribbean Region \$v Congresses

Figure 1.

Example of deconstructed vs. reconstructed subject strings

Figure 2.
Program Workflows

Table 1.

MeSH RDF URIs for Subject Headings in Figure 1

MARC 21 tag	Subject Heading	RDF URI
650	\$a Acquired Immunodeficiency Syndrome \$x ethnology	http://id.nlm.nih.gov/mesh/D000163Q000208
650	\$a Cross Cultural Comparison	http://id.nlm.nih.gov/mesh/D003431
650	\$a Health Policy	http://id.nlm.nih.gov/mesh/D006291
651	\$a Africa \$x ethnology	http://id.nlm.nih.gov/mesh/D000349Q000208
651	\$a Caribbean Region \$x ethnology	http://id.nlm.nih.gov/mesh/D017691Q000208
655	\$a Congresses	http://id.nlm.nih.gov/mesh/D016423

Table 2.

Example outputs

Authority Heading/Subheading	Authority Record 035\$a MeSH Unique Identifier	Authority Record 029\$0 URI	65X Field Bibliographic Record
150 __ \$a Critical Care	035 __ \$a D003422	029 __ \$0 https://id.nlm.nih.gov/mesh/D003422	650 12 \$a Critical Care \$0 https://id.nlm.nih.gov/mesh/D003422