Second Report of the Task Group on URIs in MARC

15 April 2016

Charged on 1 September 2015, for an initial period of one year (ending 1 October 2016)

MEMBERSHIP AND CONSULTANTS:

Since the last report, there were several additions to the TG membership and consultancy [1]: Steven Folsom (Harvard University, member), Michelle Durocher (Harvard University, consultant, ISNI specialist), Corine Deliot (British Library, consultant, ISNI specialist), and Thurstan Young (British Library, MAC liaison).

ACTIVITIES SUMMARY:

Following the last report on October 15, 2015 [2], the TG has continued working through its charges. From the TG's bi-weekly discussions, it became evident that conducting a test sooner rather than later would be beneficial to understanding identifier management and the impact of adopting dereferenceable URIs in the current MARC environment. The test process also helped the TG address, at least in part, several items from its charges from the Steering Committee.

The TG members and consultants worked together as a group to carry out the following activities around the test: prepare spreadsheets identifying MARC fields that do and do not already have $0 defined (used both as a reference and to capture analysis), prepare datasets (authority and bibliographic MARC records) for use in testing, refine conversion tools [3], iteratively process the data sets using the conversion tools, and ingest the revised data sets in library systems (by the PCC-affiliated utility representatives, Casalini and others). In order to test the process in a short period of time, the TG focused on adding URIs to $0 only when there was an exact match between the authority heading in the MARC record and an authority source available in RDF format.

In addition, a subgroup was created to examine the status of Real World Objects (RWO) in the library community and beyond.

ACTIVITIES BY CHARGE:

Charge 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. These issues may include:

1. ***whether to use alphanumerical identifiers or URIs;***
2. ***the use of multiple identifiers for the same entity;***
3. ***where to put work and expression identifiers.***

Currently, the value in $0 contains the system control number of the related authority record, or a standard identifier such as an International Standard Name Identifier (ISNI). The control number or identifier is preceded by the appropriate MARC Organization code (for a related authority record) or the Standard Identifier source code (for a standard identifier scheme), enclosed in parentheses. The TG concluded that when $0 contains an HTTP format URI, the parenthetical (uri) is redundant because the URI is self-identifying.

The TG has begun reviewing bullet points 2 and 3 in conjunction with feedback from the pilot test. See 'Next Steps' for further information.

Charge 2: In collaboration with the PCC Standing Committees, develop guidelines for including identifiers in MARC bibliographic and authority records.

Members of the TG who lead the PCC standing committees on policy, standards and training can use lessons learned from the test as a basis for establishing guidelines for the PCC library community to apply URIs to MARC records. The PCC URI group has been working to identify and document the key issues that such guidelines will need to address, such as best practice for multiple identifiers and preferred URIs. At a later date the URI group will work with PCC leadership to hand off the development of these guidelines to relevant PCC committees.

Charge 3: Develop a work plan for the implementation of identifiers in $0 and other fields/subfields in member catalogs and in PCC-affiliated utilities. Tasks may include:

1. *determine the entities for which identifiers should be provided in an initial implementation;*
2. *identify source vocabularies that will need to be accommodated;*
3. *identify automated methods for populating and maintaining new and existing records with identifiers;*
4. *develop requirements for tools that will allow catalogers to work accurately and efficiently with linked data vocabularies;*
5. *identify functionality that will be required for library systems (including ILSs and utilities) to exchange, control, protect and update data based on identifiers;*
6. *develop a pilot project and identify partners*

Adam Schiff and Steven Folsom have taken charge of a document on formulating and obtaining HTTP URIs for RDF resources [4]. This document serves as a reference for authority resources with RDF URIs, and will help to guide both manual and automated provision of URIs.

The TG devised a test for adding URIs for RDF resources to MARC records, including the following goals:

- Focus specifically on MARC, including both authority and bibliographic records
- Focus on fields where URIs are already provided for in MARC
- Address workflow, including machine reconciliation for data up- and down-stream between utilities and local systems
- Specify preferred and/or allowed syntax for URIs and address semantic issues that may arise
- Identify community approved vocabularies, and specify preferred and acceptable RDF sources, without being unduly restrictive

After a face-to-face meeting of the TG during the ALA Midwinter Meetings 2016 in January, a subset of the TG began preparing for the test by identifying sets of input data, and working with Terry Reese and Gary Strawn toward refining the lookup algorithms of their URI insertion tools. The datasets were iteratively processed with the appropriate tools as those tools were tweaked and the results were analyzed by the TG. The resulting converted datasets were ingested to several different library systems, whose comments on that ingest are currently under review [5].

The process of evaluating the data input and output contributed to a more cohesive understanding of the role of an identifier, in particular, the dereferenceable URI, deployed in a MARC environment. Conducting a pilot test was a prudent approach and evidently the necessary first step for the TG as a whole. It helped define focal points for TG discussion, such as the syntax of URI in $0 and elsewhere. This exercise also helped TG members tasked with formulating PCC policies, guidelines and best practices surrounding the recording of URIs in $0 and other subfields.

Discussions of URIs in MARC were not exclusively focused on $0, but also extended to other subfields, such as $w, $4, $i, $e/$j and $o (Oh), etc., subfields which have the potential to hold HTTP URI. Testing revealed that provisioning for URIs in MARC presents additional layers of complexity that require further consideration, i.e., repeatability, pairing, ambiguous relationships, and significance of the ordinal sequence.

As a result of the test and following many good-spirited discussions, the TG has formed the view that initial implementations should focus on elements that can be defined clearly and unambiguously from a machine processing perspective. While further work can be done, and in some cases may be worth doing, to make problematical areas of MARC more hospitable to URIs, the TG believes that return on investment should be taken into account and that beyond a certain point MARC development will reach a point of diminishing returns.

Charge 4: In consultation with the MARC Advisory Committee, technologists versed in linked data best practices, and other stakeholders, identify and prioritize any remaining issues concerning support for identifiers in the MARC format, and initiate MARC proposals

as appropriate. Prioritization of issues should take into account impact, feasibility, and the late stage of MARC's life cycle. Issues may include:

1. ***accommodating entities and relationships not currently well provisioned for identifiers in MARC;***
2. ***consistency of provisions across MARC fields;***
3. ***addressing distinction of URIs pointing to real world objects vs URIs pointing to documents/authorities.***

The Task Group should give priority to actions that will lead to tangible results during the lifetime of the PCC Strategic Directions, 2015-2017. The group should feel free to form subgroups and call on additional expertise as needed.

In March/April, the TG focused on authoring two discussion papers and collaborating with the British Library on editing the $4 and $w discussion papers that were introduced at Midwinter. As a result, the TG authored and submitted the papers below to MAC:

- Redefining $0 in the authority, bibliographic, and holdings formats to allow dereferenceable URIs to be implied in the absence of standard identifier source code prefix: (uri)
- Adding Subfield $0 to Fields in the MARC 21 Bibliographic and Authority Formats

In addition, the TG provided input on two papers planned for submission to MAC by the British Library:

- Redefining subfield $4 to encompass URIs for relationships in the MARC 21 Authority and Bibliographic Formats
- Expanding the Definition of Subfield $w to Encompass Standard Numbers

These papers address, to some degree, the desire to consistently use $0 in MARC for RDF object URIs and/or $4 for RDF relationship URIs. The lack of adequate provision in MARC for unambiguous identification of relationships has been a serious shortcoming in MARC.

After consideration the British Library will be deferring the paper which recommends expansion of subfield $w. This is to allow time for a consensus to be reached on identifiers for RWOs, identifiers for documents about RWOs and where best to record them in MARC. Deferral will also allow the British Library time to establish whether there is still a use case for expanding subfield $w.

The TG also created a small subgroup to address the question of distinguishing an identifier that represents the RWO from one that describes the RWO. The TG has found that making this distinction is not trivial. The RWO subgroup prepared findings that it hopes will dispel misunderstanding about RWOs among the membership and perhaps position another discussion

paper that the TG will submit to MAC for ALA Midwinter 2017. In sum, the RWO subgroup submitted the following initial findings:

- We acknowledge that the distinction between RWOs and documents about RWOs is fundamentally important and should eventually be formally discoverable in a bibliographic record. But until this distinction can be realized, the URIs in $0 ambiguously refer to both RWOs and documents about RWOs.
- But a recommendation at this time for descriptions for MARC data format is premature because many library authority files are in flux.
    - They may not be published as RDF.
    - If published as RDF, they may not be modeled as strictly interpreted RWOs.
    - A broader community discussion is required to resolve the issue of library Authorities and RWOs.
- In a future recommendation, RWOs should be identifiable in the bibliographic record, perhaps labeled with the keyword 'RWO' or recorded separately in a $1 subfield.

The full report will be at the TG group page later in 2016,
http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

NEXT STEPS:

[Charge 1] While it is an acceptable practice in MARC to have multiple identifiers for the same entity in one field via repeating subfields, that does not translate well to RDF. Preliminary feedback from ILS vendors indicates that where possible it is preferable, when there are multiple $0s in a single MARC field, for that field to be repeated for each $0.

[Charge 2] There was confusion even among the TG regarding what a URI does and what it means in a linked data environment. When URL, URN, and permalink can all interchangeably be called URIs, grasping the essence of the intended function of a URI can be quite baffling. Additionally, relationships among URIs in a given context is also unclear to practitioners. This confusion communicated to the TG that the guidelines the TG will recommend, working closely with Standing Committee on Standards (SCS), Standing Committee on Policy (SCP) and Standing Committee on Training (SCT), need to be coupled (perhaps tripled) with example use cases throughout each workflow. While we need to state clearly guidelines which vocabularies and sources are preferred, we should not be unduly restrictive.

[Charge 3] The TG has opened discussions with PCC-affiliated utilities and authority vendors to devise deliverables in regard to programmatically inserting URIs in legacy MARC data or in newly created MARC data until we have moved from MARC to a linked data cataloging environment. The TG has reached out to OCLC Production Service for in-depth, in-person discussion at ALA Annual 2016.

[Charge 3] Review and analyze ILS feedback on test datasets. Reports from the PCC-affiliated ILS teams revealed issues with the converted test datasets. Some of the issues are clearly the result of differences in system regarding the handling of URI data in subfields such as $0, $4, $e/$j, $2, and $5, etc.; however, the reports are still being reviewed and TG discussion pending [5].

[Charge 4] Several issues that remain regarding use of URIs in $0 require more research and discussion:

- Repeating $0s when they can refer to different objects or multiple other subfields in a single MARC field. It is programmatically impossible to determine which subfield each $0 URIs referencing because sequencing and order of subfields has no meaning to programs. For example,

  ```
  382  0\$aviolin$n1$n1$s2$2lcmpt
  $0http://id.loc.gov/authorities/performanceMediums/mp2013015782
  $apiano
  $0http://id.loc.gov/authorities/performanceMediums/mp2013015550
  ```

- How to handle MARC fields, i.e. 041, where each subfield conveys a relationship for the data value it contains. In this case, there would be multiple $0s for each language but the relationship of the language to the resource is lost.

- What is possible for the many entities which still lack a perfect match URI? The TG came to the conclusion that on the first try, the result will not be 100 percent satisfactory. Experience with these tasks will nevertheless help libraries at large consider issues surrounding new tools and alternative workflows capable of addressing this challenge.

[Charge 4] RWOs. Continue to monitor W3C, library, and other communities' discussion and implementation of RWO URIs to determine best practices for the library community. Potentially prepare a discussion paper to MAC for ALA Midwinter 2017.

Respectfully submitted,

Jackie Shieh
on behalf of the Task Group on URIs in MARC

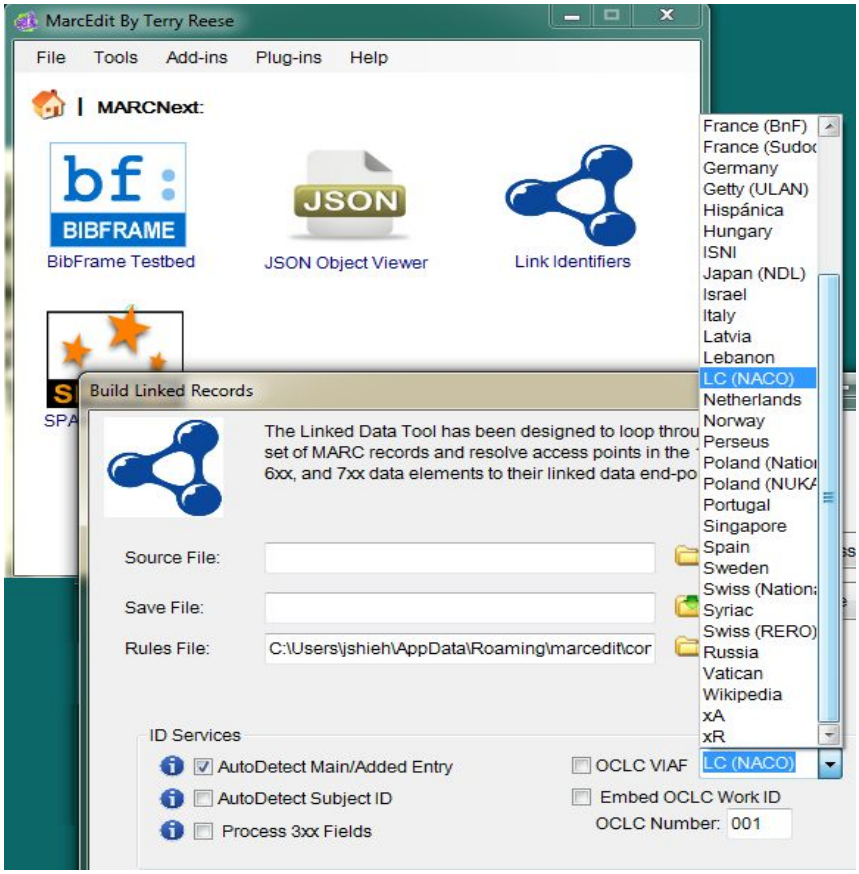
ENDNOTES:

1. Full list of names and institutions for members and consultants is located on the PCC Task Group Web site. http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

2. The first Task Group report is posted on the PCC Task Group Web site. http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

3. Tools:
   a. The pilot test was mostly conducted via Terry Reese's MarcEdit suit, *Link Identifiers function*. *RDA and beyond*, https://youtube.com/embed/B4bZkxad-FM provides the tutorial for data transformation.

      In MarcEdit interface, the *Linked Identifiers* function lists available vocabularies from drop-down as shown below:



Vocabularies may be modified by editing the rules file that invokes the MarcEdit search algorithm (located in C:\Program Files\MarcEdit 6\shadow\configs\linked_data_profile.xml).

```
Supported Vocabularies:
  Value: lcshac
  Description: LC Childrens Subjects

  Value: lcdgt
  Description: LC Demographic Terms

  Value: lcsh
  Description: LC Subjects

  Value: lctmg
  Description: TGM

  Value: aat
  Description: Getty Arts and Architecture Thesaurus

  Value: ulan
  Description: Getty ULAN

  Value: lcgft
  Description: LC Genre Forms

 Value: lcmpt
 Descirption: LC Medium Performance Thesaurus

 Value: naf
 Description: LC NACO Terms

 Value: naf_lcsh
 Description: lcsh/naf combined indexes.

 Value: mesh
 Description: MESH indexes
```

b. Gary Strawn's Authority Toolkit helped in understanding cataloger's workflow in constructing URI when catalogers and system interact directly. bit.ly/1Hl1jST
c. Steven Holloway (James Madison University) and Joseph Kiegel (University of Washington) offered invaluable insights from their respective tests.
d. SPARQL query algorithm that Columbia colleagues used while investigating AAT vocabularies

4. *Formulating URI document*, compiled by Adam Schiff and Steven Folsom. The final version will be posted on the PCC Task Group Web site, later in 2016. http://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

5. Reports from testers who graciously provided their findings:
a. Innovative/Skyriver. (Jodi Williamschen)
b. OCLC. (Robert Bremer, John Chapman, & Jean Godby)
c. Sirsi/Dynix (Horizon, Anythink Libraries, Colorado)
d. ExLibris (via TG member libraries)
e. Casalini. (Tiziana Possemato & Michele Casalini)
f. Koha (Galen Charlton)
g. Steven Holloway (James Madison Univeristy)