

Task Group on URIs in MARC Year One Report

Date: October 6, 2016

Members: Robert Bremer, Steven Folsom, Paul Frank, Jean Godby, Les Hawkins, Reinhold Heuvelmann, Chew Chiat Naun, Adam Schiff, Jackie Shieh, Gary Strawn

Contributing consultants: Nancy Fallgren, Nancy Lorimer, Melanie Wacker, Terry Reese, Corine Deliot, Thurstan Young

OUTLINE:

- I. CHARGE
 - II. EXECUTIVE SUMMARY
 - III. PROCESSES
 - III.1. Define and understand HTTP URI
 - III.2. Identify issues/problems with adding URIs whether it was actually doable
 - IV. PROBLEM STATEMENT
 - IV.1. Where to place URIs in the MARC structure (\$0, \$4)?
 - IV.2. What difficulties are evidenced?
 - IV.3. What did we learn?
 - IV.4. Outcomes
 - IV.5. Next steps and in-depths analyses in year 2
 - V. RECOMMENDATIONS TO STAKEHOLDERS
 - VI. REFERENCES
-

I. 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.*

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

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*

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 Task Group should feel free to form subgroups and call on additional expertise as needed.

II. EXECUTIVE SUMMARY

The first year since the inception of the URI in MARC Task Group (TG) began, despite the extremely challenging schedules and demands, all members and most consultants devoted a great deal of their time in working together through many issues. It has been a great privilege to be part of the team in which everyone has his/her eyes on the goals.

The deliberations and recommended solutions were based on two driving principles. Firstly, the recommended solution will be across-the-board and straightforward. The implementation must have the most and broad impact, but with the least disruption to workflow and in MARC environment. Secondly, an important fact that the TG has observed and kept insight throughout various discussions. Though a lot of libraries have been anxious and in position to move forward with linked data experimentation and implementation. Majority of libraries remain ambivalent and hesitant. In such dual environments, the TG's recommendations must accommodate dual operations for a period of time. The TG must provide ways for library to decide their pace and needs when transitioning from MARC to linked data.

Early on everyone was keenly aware of the syntax and semantic complexity of identifiers in the form of dereferenceable uniform resource identifier (HTTP URI). After months' discussions, the TG firmly believed following the agile principles, specifically the scrum approaches would give the process most success in addressing URI in MARC issue:

- Figure out how to do the work
- Do the work
- Identify what's getting in its way
- Take responsibility to resolve all the difficulties within its scope
- Work with other parts of the organization to resolve concerns outside their control

1) Recognized that there were not possible across-the-board simple solutions for MARC fields concerning \$0. Therefore, the TG pushed forward the fields that could benefit \$0 without complications. See MAC Paper 2016-DP19 in REFERENCE Section.

2) Agreed upon the universal definition of \$0 for URI that describes THING (URI/Concept). Keeping in line with the overall principle of least disruption and most coverage, the TG recommended the use of HTTP URI in \$ as default URI for libraries which opt to adopt URI in \$0. Alternatively, text string identifier in \$0 to remain in force for libraries which are not ready to move forward. See MAC Paper 2016-DP18 in REFERENCE Section.

3) Agreed that the relationship entity of an RDF statement be represented in MARC. Potential candidates for expressing relationship were \$4, \$i, \$j, \$e. The consensus was to focus on \$ due to the existing subfield having been defined in all those fields where relationships can currently be expressed in MARC. The rescoping of \$4 to hold URI/property (predicate) does not prevent the library community's continued application of 3-letter relator codes. It provides an opportunity for libraries which are ready to deploy HTTP URI for relationship (property/predicate). Consensus was that \$4 alone should be redefined to carry relationship URIs: this was considered a consistent and across the board solution which would not require further amendments to the MARC formats by rescoping or defining additional subfields.

4) Identified a need to host identifier for real world object. The TG hoped to propose setting aside \$1 for identifier that points to THING (URI resource/RWO).

5) TG Members who work closely with other standards communities, such as ISNI/VIAF, have vested interests in the 024 in Authority. The 024 field appears to possess the potential of addressing relationship of an entity across vocabularies/ontologies. [1]

The TG hopes to address above items, no. 4-6 in discussion papers for MARC Advisory Committee (MAC) to consider.

The Pilot Test that the TG conducted in February-March 2016 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. Additionally, the TG is working hard further identifying potential field and/or indicator/subfield to record identifier representing a Work, a resource object. These are described in sections below in greater details.

III. PROCESSES

The TG had in mind processes that would be the least disruptive yet with the most promising results. In order to ensure cohesive and broad approaches, the TG set forth the tasks: a) define and understand uniform resource identifier and the deployment of the Web-service protocol scheme, HTTP; b) identify

issues and problems with adding URI in MARC. Is it actually doable in current system that hosts MARC data?

III.1. Define and understand HTTP URI [Charge 1.1, 1.2; Charge 4.3]

According to a MARBI position paper published in 2009,

The use of URI instead of plain text is particularly applicable to situations where the value of the...element comes from a controlled vocabulary, which could be an authority list or formal thesaurus (e.g. a name from the LC Name Authority File or a topic for an LCSH heading) or any other list of controlled codes or terms (e.g. the MARC Code List for Languages).

However, the goal of facilitating the transition from MARC to linked data now requires a more precise machine understanding of the data accessible from the URIs that have been added to MARC records.

The issue can be illustrated with an excerpt from the Library of Congress Name Authority record for Hillary Clinton, accessible at <https://lcn.loc.gov/n93010903>. Of particular interest is the list of 024 fields, which identify “standard number[s] or code[s] associated with the entity named in the 1xx field which cannot be accommodated in another field,” according to the MARC Authority definition. All of the 024 fields copied below contain URIs pertaining to Hillary Clinton.

024	7_	a	http://www.wikidata.org/entity/Q6294	2	uri
024	7_	a	http://dbpedia.org/resource/Hillary_Rodham_Clinton	2	uri
024	7_	a	http://viaf.org/viaf/54950123	2	uri
024	7_	a	http://isni.org/isni/0000000122802598	2	uri
024	7_	a	http://d-nb.info/gnd/119082101	2	uri
024	7_	a	http://id.ndl.go.jp/auth/ndlna/00552567	2	uri
024	7_	a	http://aut.nkp.cz/jn20000700317	2	uri
024	7_	a	http://catalogue.bnf.fr/ark:/12148/cb12543158f	2	uri
024	7_	a	http://www.idref.fr/034705171	2	uri
024	7_	a	http://datos.bne.es/resource/XX1725857	2	uri
024	7_	a	http://id.sbn.it/af/IT%5CICCU%5CUBOV%5C804461	2	uri
024	7_	a	http://cantic.bnc.cat/registres/CUCId/a11695705	2	uri
024	7_	a	https://musicbrainz.org/artist/858a3d95-e1b2-4aac-8427-a99e391ce8c5	2	uri
024	7_	a	http://www.imdb.com/name/nm0166921	2	uri
024	7_	a	http://bioguide.congress.gov/scripts/biodisplay.pl?index=C001041	2	uri
024	7_	a	http://www.nndb.com/people/022/000025944	2	uri
024	7_	a	https://ballotpedia.org/Hillary_Clinton	2	uri
024	7_	a	https://www.freebase.com/m/0d06m5	2	uri

The rows in the table can be partitioned into three categories:

- Near the bottom, the 024 fields with the peach-colored background are human-readable documents about Hillary Clinton. These are pages from popular resources maintained outside the library community, such as IMDB and BioGuide, which have been deemed

authoritative by library catalogers and authority experts. In shorthand, these URIs are standard URLs for Web pages.

- The rows with the blue background are records derived from library authority files and more modern registries designed for similar purposes. They may be pages from library authority files published on the Web, human-readable views of machine-understandable RDF data, or raw RDF. But in one form or another, all of the URIs resolve to library authorities (or simply 'Authorities') that are about Hillary Clinton. The TG refer to these URIs as Authority URIs.
- The rows with the green background contain URIs that refer to Hillary Clinton directly in a way that is technically distinct from documents about her. These URIs conform to linked data conventions described in standard Web documents such as "Cool URIs for the Semantic Web" [<https://www.w3.org/TR/cooloris/>]. The data accessible from these URIs has been published by third parties as well as the library community and encodes a rich domain model designed expressly for machine understanding. The TG refer to these URIs as Entity, or Thing URIs.

According to linked data conventions, machine processes designed to construct meaningful statements, and inferences from them, require Thing URIs. When Thing URIs are defined for people and creative works, one desirable outcome would be a machine-understandable statement such as 'Hillary Clinton is the author of the book *It Takes a Village*.' With technology available in 2016, data accessible from Web page URIs may not be machine-understandable at all, and Authority URIs may only be partially understandable. The ambiguity of URIs illustrated by the 024 fields in the MARC Authority records is also present in MARC bibliographic records.

III.2. Identify issues/problems with adding URIs whether it was actually doable [*Charge 1; Charge 3*]

pilot test of inserting HTTP URI in \$0 in bibliographic and authority data emerged as one logical first step for the TG. It helped the TG understand issues that could easily resolve in the near term and the do-ability of inserting URI in \$ in MARC environment.

The Pilot Test began in February, 2016. Members prepared sets of input data and worked with tool creators (MarcEdit and Authority Toolkit) to refine lookup algorithms for URI insertion in \$0.

The enhanced data with HTTP URIs embedded were to be ingested to several integrated library systems for evaluation. This exercise assisted the TG gaining a cohesive understanding of the role of an identifier in the form of dereferenceable URI deployed in \$0 in MARC environment.

Throughout the process, the TG began to frame the questions that might assist in the effort in transitioning MARC data to linked data. Including reached possible resolutions where potential problems may reside. Such as planning for MAC proposals in its first year.

Issues that were more long-term and may require in-depth discussions from broader community involvements, for instance, subfields such as \$4 which have been defined might have the potential to

hold HTTP URI. The repeatability and ambiguity, and significance of the ordinal sequence are less trivial and complex.

In regards to bulk processing of insertion, system performance and scalability, the Pilot Test also helped address SPARQL query adjustment on the server side. Though URIs added by hand was the least desirable exercise which could be inevitable, the TG also began documenting resources that would assist such endeavor.

The overall strategies that the TG adopted were carefully thought-out in order to achieve iterative success that will build confidence throughout phases of implementation.

IV. PROBLEM STATEMENT

To encode data suitable for transformation into RDF triples, it is necessary to be able to identify in MARC the data elements corresponding to the subject, predicate and object in each statement and/or to provide URIs for them. It quickly became apparent that the task is not simply to add subfields to allow URIs to be given - itself a non-trivial problem given the limited number of unused subfield still available in MARC - but also to negotiate the often ambiguous semantics of MARC. The TG has sought to do this through a judicious combination of redefinition proposals, clarification of existing semantics, and best practice recommendations.

Best practices for incorporating HTTP URIs in MARC BIB and Authority records without making major renovations to MARC format (taking into consideration cost/benefit analysis for an 'end of life' technology)

IV. 1. Where to place URIs in the MARC structure (\$0, \$4)? [Charge 3]

The TG developed a pilot to examine the issues surrounding the issues of adding identifiers to MARC 21 data. The work included the identification of actionable source vocabularies and creating test record sets with dereferenceable URIs embedded in the data. A variety of formats were represented in the test data and ILS vendors, programmers, system engineers, and discovery designers were consulted throughout the pilot to comment on the retrieval of actionable URIs and the appropriate policies ensuring the data are actionable in MARC 2 data.

The TG also inventoried the MARC bibliographic and authority formats to identify MARC 2 fields that contain subfields capable of accommodating URIs. In the bibliographic formats subfields \$0 and \$4 were identified as existing candidates for containing URIs, subfields \$0 and \$4 were candidates in the authority format. MARC 2 fields that might usefully contain subfield for URI, but which do not have one defined were also noted.

The TG focused on subfield \$0 and \$4 for its first three MARC Discussion papers submitted in to MAC at ALA Annual 2016.

IV.2. What difficulties are evidenced?

IV.2.1. *Adding multiple \$0? [Charge 1.2]*

The nature and use of subfield \$ has evolved in MARC since the subfield was first implemented in 2007. In 2010, it was redefined and came to include standard numbers, including URIs, in addition to its original use for authority record control numbers.

However, MARC is not specific as to which parts of a controlled heading string correspond to the \$0. Nothing in the MARC specification rules out one \$0 subfield applying to one set of subfields in a heading while a different \$0 applies to others. (To ameliorate this problem, we formed a MARC object/URI reconciliation subgroup to enumerate the subfields naming the object in each MARC field - see IV.2.2 below.) An because \$0 is repeatable, it is possible to find multiple \$0 values corresponding to the same heading subfields naming the same entity. Indeed, the latter practice is adopted by design in some implementations, notably that at the German National Library.

The existence of different use cases and practices for relating headings to \$ has emerged as an issue that will need to be considered as the TG's work proceeds. In the case of OCLC's heading control functionality related to LC names and LCSH, subfield \$0 data is included as an XML tag attribute in each subfield XML tag covered by a particular authority record and is repeated as many times as needed depending on the number of subfields used to represent the name or subject. In the subsequent development of controlling for other authority files, the same approach has been taken, but instead retaining the same or different authority record control numbers in multiple \$0 subfields. [See examples at end of document]

This repeated use of \$ subfields containing the same authority record number or different authority record numbers for different parts of a heading runs contrary to the need that exists in an OCLC context of a single URI corresponding to the entire named entity given in the field. Extraneous \$0 subfields are automatically deleted in WorldCat records in fields that are otherwise controlled to particular authority file. However, this leaves unresolved the question of controlling via multiple source vocabularies within the same language of cataloging which many see as a desirable medium-to-long-term objective. Given the investment in its development and the number of controlled headings in WorldCat, completely changing the heading control functionality within WorldCat is not feasible, so the TG and OCLC staff have considered other alternatives allowing for output of needed URIs in the format which libraries would prefer in the future.

IV.2.2. *How to identify a RDF object in a MARC datafield? [Charge 4.3]*

This emerged as an important need because the ability to identify URI with its corresponding label is necessary to support both reconciliation of existing data and updates to those labels based on their association with an identifier. The only realistic way to make this identification was to document the correspondences on a field-by-field basis. Fortunately, this was very achievable for the majority of fields in widespread use. [Link to [recommendations](#). The investigation revealed a number of issues relating to

the identification of single entities vs larger sets (series, conferences) and alignment of MARC and RDA vocabularies.

IV.2.3. *What did we find in identifying relationships/multiple relationships? [Charge 4.1]*

IV.2.3.1. Relationships are expressed in MARC by a variety of means, including:

IV.2.3.1.1. Field tagging, either alone, e.g. 830, or in combination with indicators, e.g. 780/785

IV.2.3.1.2. Subfield codes, e.g. 041

IV.2.3.1.3. Codes given in subfields, e.g. 700 \$4

IV.2.3.1.4. Controlled or natural language text given in subfields, e.g. 700 \$i

IV.2.3.2. Some of these fields are very tightly bound to legacy MARC definitions, structures, and data. Redesigning 041, for example, to be hospitable to URIs would require a complete reconception of that field.

IV.2.3.3. There is the greatest value in provisioning for URIs following a 7XX \$4/\$0/\$1 pattern, with \$ repurposed to house URIs much as \$ now does. This approach seems to present a relatively low barrier to implementation while having widespread application in MARC.

IV.2.3.4. Multiple relationships can cause ambiguity where they are associated with multiple objects or multiple labels. In such cases, we recommend the expedient of simply repeating the field in order to make the associations unambiguous.

IV.2.4. *How one obtains URIs for various data sources depends on the linked data source (different data sources avail their URIs differently) and interoperability between the data source and the cataloging tool/s being used.*

To help support obtaining the right URIs for their purposes in MARC, the TG has begun a document, currently referred to as [Formulating and Obtaining URIs: A Guide to Commonly Used Vocabularies and Reference Sources](#). For commonly used vocabularies in MARC, we want to document where in the data source UI one can find the canonical URIs, that when dereferenced provides data. Going forward, for each entry in the document, we want to explain whether a data source publishes their data as Authorities, Real World Objects, or both. Also, we want to document methods available for machine access to the data. Is the data published as Linked Data available through http, available through a SPARQL endpoint, data dumps, etc.?

IV.2.4.1. MarcEdit [Charge 3]

In the summer of 2014, MarcEdit introduced a suite of tools designed to begin testing the feasibility of embedding linked data concepts into MARC records. Initially, the scope of the suite was limited to embedding HTTP URI in the \$0 in MARC fields 1xx, 6xx, 7xx in bibliographic records. This initial work focused on integration with the U.S. Library of Congress's id.loc.gov service, as well as OCLC's VIAF services for resolution. However, over the past 2 years, and in response to many of the questions and issues surfaced through the TG, the Linking services have been expanded and revised to potentially support all use-cases identified by this Task Force, as well as providing support for non-MARC21 users to configure the Linking tool for use with other MARC formats.

The MarcEdit Linking toolkit currently supports the generation of URIs for all identified fields by this Task Force for authority and bibliographic records. The application utilizes a rules file that documents field processing and service configuration values. This allows MarcEdit to quickly make changes to the rules governing field processing, as well as adding support for new collections and linked data endpoints. As of this report (9/21/2016), the MarcEdit Linked Data tool support resolution against the following linked data services:

1. U.S. Library of Congress NAF
2. U.S. Library of Congress LCSH
3. U.S. Library of Congress Children's Subject Headings
4. U.S. Library of Congress Demographic Group Terms
5. Thesaurus for Graphic Materials
6. U.S. Library of Congress Genre/Form Terms
7. U.S. Library of Congress Medium of Performance Thesaurus for Music
8. RDA Carrier Types
9. RDA Media Types
10. RDA Content Types
11. Getty Arts and Architecture Thesaurus
12. Getty ULAN
13. National Library of Medicine MESH
14. OCLC FAST Headings
15. OCLC VIAF
16. German National Library (GND)
17. [15 national library name indexes via VIAF]
18. Japanese Diet Library

Additionally, users have the ability to configure their own linked data endpoints for use with MarcEdit, so long as the service in question supports SPARQL and json. There is presently a knowledge-base article at: <http://marcedit.reeset.net/editing-marcedits-linked-data-rules-file> documenting how users can both add new collections or modify the rules used when processing a particular field.

Essentially, MarcEdit utilizes its rules file to configure MarcEdit's linked data platform to identify the proper index/service, normalization (for data query purposes), and subfields to utilize as part of any look up process. Additionally, each rule's block identifies when a field should be processed (i.e., only when used in a bibliographic record, used in an authority record, or both). For example, here is the definition for the 650 field.

```
<field type="bibliographic">
  <tag>650</tag>
  <subfields>abvxyz</subfields>
  <ind2 value="0" vocab="lcsch"/>
  <ind2 value="1" vocab="lcsnac"/>
```

```
<ind2 value="2" vocab="mesh"/>
<ind2 value="7" vocab="none"/>
<index>2</index>
<uri>0</uri>
<special_instructions>subject</special_instructions>
</field>
```

Each MarcEdit rules block is a small segment of XML that profiles field usage within a record. This is why MarcEdit's linking tool can be used with other flavors of MARC (like UNIMARC); the Linking service has no concept of MARC21 -- just for ISO2701 format -- the rules file provides that context.

This approach has allowed MarcEdit to quickly profile and examine the implication of developing URIs for linking fields, like the 880 field, which provide some unique challenges -- but can be accommodated via the current rules file format.

Utilizing the current process, MarcEdit's linking tool can accommodate a wide range of linking scenarios. For example, in an authority record:

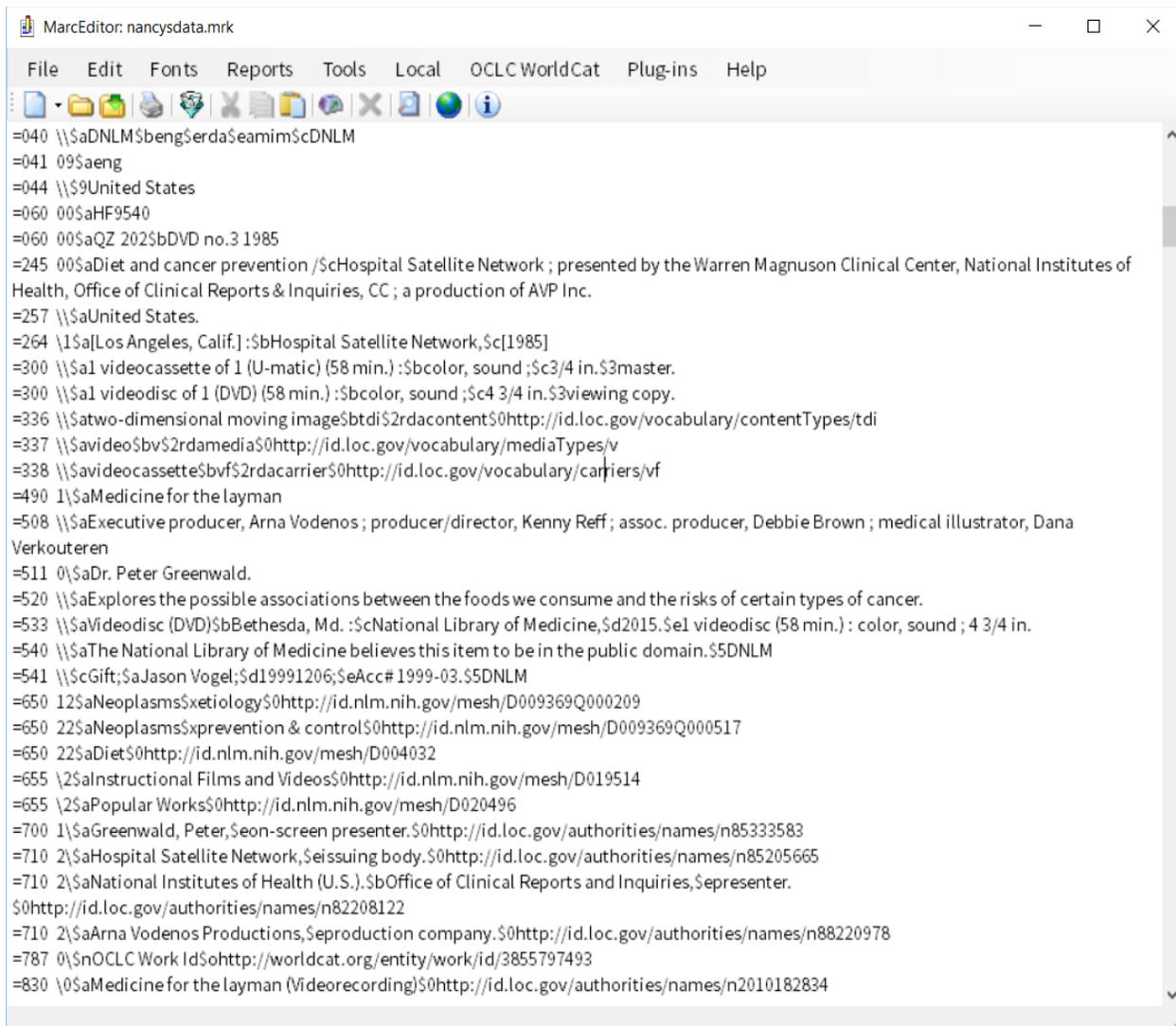
MarcEditor: adamsauthority_records.mrk

File Edit Fonts Reports Tools Local OCLC WorldCat Plug-ins Help

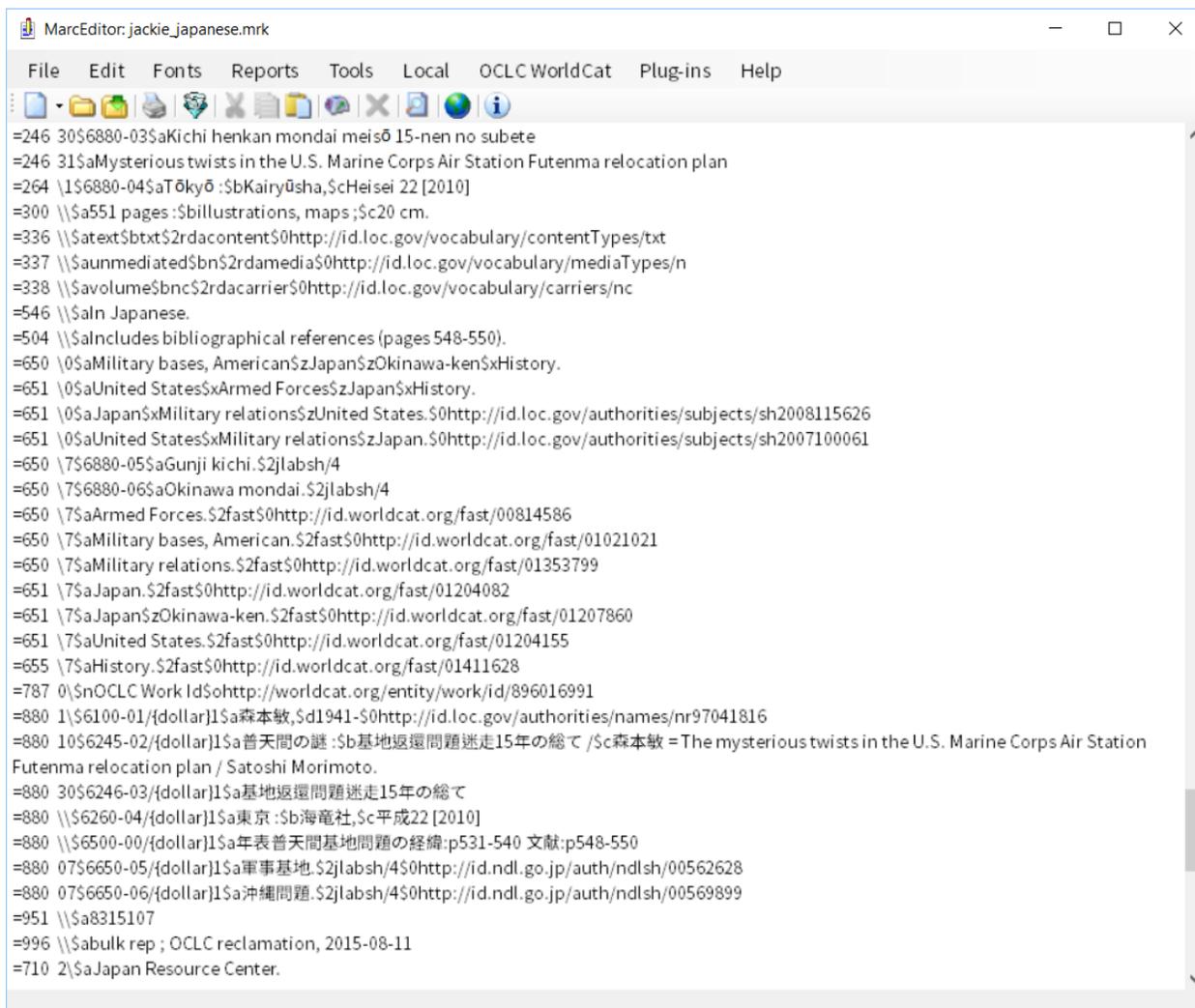
=LDR 02762cz a2200313n 45 0
 =001 oca10207248\
 =003 OCoLC
 =005 20151123044520.0
 =008 150701n\azannaabn\|||||a\aaa\|c
 =010 \\\\$ano2015086579
 =040 \\\\$aWaU\$beng\$erda\$cWaU\$dWaU
 =046 \\\\$k2008
 =100 1\|aCollins, Suzanne.\$tHunger Games (Novel)
 =370 \\\\$gUnited States\$2naf\$0http://id.loc.gov/authorities/names/n78095330
 =380 \\\\$aNovels\$2lcgft\$0http://id.loc.gov/authorities/genreForms/gf2015026020
 =380 \\\\$aDystopian fiction\$2lcgft\$0http://id.loc.gov/authorities/genreForms/gf2014026302
 =380 \\\\$aApocalyptic fiction\$2lcgft\$0http://id.loc.gov/authorities/genreForms/gf2014026226
 =380 \\\\$aScience fiction\$2lcgft\$0http://id.loc.gov/authorities/genreForms/gf2014026529
 =380 \\\\$aAction and adventure fiction\$2lcgft\$0http://id.loc.gov/authorities/genreForms/gf2014026217
 =380 \\\\$aYoung adult fiction\$2lcs\$0http://id.loc.gov/authorities/subjects/sh85149267
 =385 \\\\$nAge\$aTeenagers\$2lcdgt\$0http://id.loc.gov/authorities/demographicTerms/dg2015060011
 =386 \\\\$nna\$aAmericans\$2lcdgt\$0http://id.loc.gov/authorities/demographicTerms/dg2015060001
 =386 \\\\$ngdr\$aWomen\$2lcdgt\$0http://id.loc.gov/authorities/demographicTerms/dg2015060358
 =386 \\\\$nAge\$aWomen\$2lcdgt\$0http://id.loc.gov/authorities/demographicTerms/dg2015060358
 =530 \0\$Adapted as motion picture (work):\$aHunger Games (Motion picture)\$wr\$0http://id.loc.gov/authorities/names/no2012012602
 =500 1\|iSequel:\$aCollins, Suzanne.\$tCatching fire\$wr\$0http://id.loc.gov/authorities/names/no2015086852
 =670 \\\\$aCollins, Suzanne. The Hunger Games, 2008.
 =670 \\\\$aWikipedia, July 1, 2015\$b(The Hunger Games (novel)); The Hunger Games is a 2008 science fiction novel by the American writer Suzanne Collins. It is written in the voice of 16-year-old Katniss Everdeen, who lives in the dystopian, post-apocalyptic nation of Panem in North America; The Hunger Games is an annual event in which one boy and one girl aged 12-18 from each of the twelve districts surrounding the Capitol are selected by lottery to compete in a televised battle to the death. Country: United States. Genre: Adventure; Dystopian; Science fiction; was first

Page 1 - 2 Records per Page 100

Within a Bibliographic Record:



Across Diverse vocabularies:



Current development of the tool will continue to focus on the inclusion and support of additional vocabularies, continuing to work with linked data providers around scalability issues (and ways in which MarcEdit [or services like it] can reduce impacts on their services, as well as working to profile this service to work with other flavors of MARC, like UNIMARC; to encourage further experimentation.

IV.2.4.2. Authority Toolkit [Charge 3]

The *authority toolkit* is a program for the construction and modification of authority records. One version is designed for use within OCLC's Connexion program for records in the LC/NACO authority file, but another version can work with records in files, and so with records from other sources. Both versions of the toolkit have the same capabilities. At an early stage, the toolkit acquired the ability to test terms used in authority fields such as the 370 and 372 against vocabularies available at id.loc.gov (at present: LCMPT, LCSH, LCDGT, AFSET, geographic area codes, RDA content terms, and the LC/NACO Authority File). Somewhat later, it added the ability to verify terms against the MeSH vocabulary.

(Additional vocabularies may be added in the future, based on user requests.) To perform this verification, the program needs to know which vocabularies are used to control terms in which parts of which authority fields; how to query the source to determine whether or not it is defined; and how to react to the information returned by the source. The toolkit's actions are controlled above all by the subfield \$2 code appearing in the same subfield as the term; but in the absence of a subfield \$2 code, operator preferences come into play as well. (For example, an operator may prefer that an unlabeled term be tested against MeSH first, and if not found tested against LCSH; or perhaps tested only against LCDGT.) A detailed description of the toolkit's process for verifying the content of authority fields can be found in the program's documentation at:

<http://files.library.northwestern.edu/public/oclc/documentation/#verifymenu>

If the toolkit's search for an entire term is successful, the toolkit could easily supply the corresponding URI and add it to the authority record in subfield \$0. This URI may be contained in the data provided by the source, or it could be constructed mechanically once the toolkit has extracted the appropriate identifier. As part of experimentation encouraged by the TG, on January 15, 2016, the toolkit acquired an option to add subfield \$ to fields which could be verified. (This option is described at <http://files.library.northwestern.edu/public/oclc/documentation/#optionsverification0> If a field contains more than one term, the toolkit must divide the field into multiple fields (one for each term) before it can add subfield \$0.

The following illustration shows an authority record as verified by the authority toolkit, with the option to add subfield \$0 during verification turned on. (For this experiment, subfield \$0 was locally defined for some fields.)

```

Rec stat: n Entered: 160909 Char: a
Type: z Upd status: a Enc lvl: n Source:
Roman: ? Ref status: n Mod rec: Name use: a
Govt agn: ? Auth status: a Subj: a Subj use: a
Series: n Auth/ref: a Geo subd: n Ser use: b
Ser num: n Name: a Subdiv tp: n Rules: z

001 1104760
005 19930217061745.8
010: : $a n 84020843
040: : $a DLC $b eng $e rda $c DLC $d NJP
046: : $f 1950 $2 edtf
100:1 : $a Strawn, John, $d 1950-
370: : $a Lima (Ohio) $2 naf $0 id.loc.gov/authorities/names/n81129565
370: : $c United States $2 naf $0 id.loc.gov/authorities/names/n78095330
370: : $e Larkspur (Calif.) $2 naf $0 id.loc.gov/authorities/names/n79132166
375: : $a Males $2 lcdgt $0 id.loc.gov/authorities/demographicTerms/dg2015060003
670: : $a Foundations of computer music, c1985: $b CIP t.p. (John Strawn)
670: : $a Phone call to author, 4-12-84 $b (John Michael Strawn; b. 1-22-50)
670: : $a His Modeling musical transitions, 1985: $b t.p. (John Michael Strawn)

```

Although the toolkit can often discover information about compound terms (such as some corporate bodies with subordinate units, and some LCSH headings) for which an authority record exists for some parts but not all, the toolkit cannot supply subfield \$0. (There is no authority record, and so no URI, that represents the entire term.) The toolkit also cannot add subfield \$0 to fields that contain multiple terms, if the field contains an aggregation of terms, rather than a collection of independent items. (Example: the toolkit cannot add subfield \$0 to the 382 field.)

The task of discovering that a term given in an authority record is defined in an external vocabulary is made more difficult because the searching mechanisms available do not always compensate

appropriately for operator variations in punctuation, capitalization and the use of combining diacritics. In addition, the response time experienced by the toolkit can vary widely, even for the same term searched repeatedly within a brief time; and some services are unavailable over the weekend. If the potential of linked data is to be enjoyed, services providing data must ensure that their entry mechanisms are robust and flexible, and available at all times.

IV.2.4.3. Lookup online (e.g., VIAF, Getty ULAN, Geonames, Wikidata)

Online lookup requires manual operation. Users must be well versed in SPARQL queries that individual services provide. Getty ULAN works differently to Geonames and Wikidata. The URI returns from a query may not be a RDF URI but one that may land user onto a Web page or document.

IV.3. *What did we learn? [Charge 1.3; Charge 3]*

IV.3.1. Tackle low hanging fruit/what can we do in Year 1

The TG's activities during Year 1 were designed to position the MARC community to take tangible steps toward incorporating linked data URIs into its processes within an achievable timeframe. Therefore, the TG put aside some tasks, such as overhaul of certain legacy MARC data elements, that would have delayed progress with the TG's practical objectives. The tool development undertaken by Terry Reese and Gary Strawn was designed to advance these objectives; but so were the Formulating URIs document and the MARC object/URI reconciliation work, both of which document information that will be needed by other stakeholders, and the work IDs in MARC proposal which seeks to remove one of the main barriers to routine incorporation of work identifiers in MARC records.

IV.3.2. Add \$1 where it's not defined (not simple)

One of the TG's goals was also to identify and add \$1 to fields that currently do not have one defined. The TG found the following MARC fields that needed \$1 defined:

bibliographic: 046, 257, 260/264, 375, 753;
authority: 046, 360, 375, 377, 663, 680, 681

These fields do not render an easy resolution when considering \$1 which reflects the resource object for an entity described. The TG conducted thorough analyses and concluded that only 25 and 37 could contain a URI that is unambiguous between the field and the object it represents, leaving out more complicated cases, e.g. fields 264 "Production, Publication, Distribution, Manufacture, and Copyright Notice", and 382 "Medium of Performance."

One of the issues confronted with drafting discussion paper 2016-DP19 was the extent of effort needed to individually propose subfield \$1 for MARC 21 fields that do not contain it. MAC accepted the paper as

proposal and there was agreement “that similar changes such as those recommended this paper might in the future be considered as part of a MARC Fast-Track process.” Being able to fast-track proposals for defining subfield \$0 in field which do not contain it will considerably streamline the process in the future.

IV.3.3. Strategies in lieu of limited life cycle of MARC environment

Though many may see MARC is “dead,” the system remains a viable tool that delivers metadata for data discovery. It is also, however, a legacy format that reflects in its somewhat baroque structure a long history of accretion to meet varied and changing needs. In pursuing its goals, the TG has adopted a strategy of pursuing changes that can be applied coherently across MARC and maximize return on the library community’s investment of effort. There are economical and sensible approaches in determining what to do. The TG always kept in mind of recommendations must cause the least disruption for data transition from MARC to linked data. There is unlikely to have a wholesale possibility of inserting HTTP URI, though possibly most, but not all of MARC fields and/or subfields.

The TG is committed to work through a list of tasks and identify viable solutions. While \$0, after one year’s deliberation, seemed a straightforward solution for URI representing resource object, more discussions needed with regards to predicate that denotes relationship. MARC data have not been consistent in expressing relationship. Combination field, indicators, and subfields raises complexity for the process.

IV.3.4. ILS analysis results

Some ILSs would not load the processed records because of the presence of \$0. Others loaded, but did nothing with the data.

The TG members mocked up files of bibliographic and authority data adding various URIs in subfield \$0 wherever subfield \$0 is currently defined in MARC. These files were uploaded into a number of ILS systems to see if the addition of subfield \$0 with URIs caused problems. No significant problems were found. These files included URIs in subfield \$ which were not prefixed with the (uri) identifier.

In OCLC, the same \$0 subfields were also not problematic. OCLC’s validation of subfield \$0 does not check the structure of subfield \$0 in the same way as it does for control numbers in 760-787 subfield \$w or URLs in \$u subfields. Use of URIs in subfield \$4 to express relationship information would require a change to OCLC’s validation of \$4 subfields, but that may be readily changed without extensive effort.

IV.3.5. Tools needed: MarcNext, Authority Toolkit

Currently, the TG has tested and continued to work with MarcNext and Authority Toolkit. The TG members continues collecting and recording additional tools and resources that facilitate practitioners in identifying and validating an RDF URI.

IV.3.6. Need to be able to easily report duplicates found in VIAF, etc., and need a way to know which URI to use when duplicates are found

Throughout the first year of investigation and deliberation, the TG learned though vocabularies and ontologies are structured per standards and published for adoption, some are more domain specific than others. Often there are more than one methods to structure a body of data. Duplications can be expected across various datasets. The reconciliation of URI is one of the tasks that the TG has recognized yet not in a position to recommend solution in the near term.

IV.4. Outcomes

IV.4.1. MAC Discussion Papers [*Charge 4*]

The TG was aware that some aspects of its intended goals were not yet accommodated by the MARC format. Following the defined workflows of MARC governance and standardization, the TG submitted several discussion papers to the MARC Advisory Committee (MAC). As an initial preparation, an informal discussion paper entitled "URIs in MARC: Call for Best Practices", by Steven Folsom, had been discussed during the June 2015 MAC meeting. It focused on subfield \$0, "Authority record control number or standard number", its current usage, its capability for URIs, and addressed some aspects of best practice. The paper generated extensive discussion, and there was broad agreement that the time was right for the library community to begin using URIs consistently. Steven Folsom was asked to cooperate with the PCC to develop a formal MAC Discussion Paper.

In fall 2015, the British Library (BL) submitted two papers to MAC for the January 2016 meeting, independently of the TG, covering title to title relationships via subfield \$w, and specific relationship information, then discussed using subfield \$0. The approaches taken by the BL in its papers, coupled with the approach taken by the TG, resulted in MAC suggesting that the British Library and the PCC should collaborate on submitting a paper for June 2016.

During the MAC meetings at the ALA Annual Conference in Orlando in June 2016, three papers were presented by or in cooperation with the TG: Discussion Paper No. 2016-DP18, entitled "Redefining Subfield \$ to Remove the Use of Parenthetical Prefix '(uri)' in the MARC 2 Authority, Bibliographic, and Holdings Formats" described the syntactical improvement that a subfield \$0 containing a URI without the parenthetical prefix "(uri)" would allow, so that automated processes could use the content of these \$0's without having to strip away prefix. The discussion paper was discussed at the MAC meeting, and the recommendation was made that the discussion paper be upgraded to proposal status; it was approved at the meeting as proposal. From now on \$ containing an identifier in the form of web retrieval protocol, e.g. HTTP URI, should not be given a parenthetical prefix.

second paper was presented to the MAC, Discussion Paper No. 2016-DP19 "Adding Subfield \$0 to Fields 25 and 37 in the MARC 2 Bibliographic Format and Field 37 in the MARC 2 Authority Format." It resulted from extensive analyses of the MARC Bibliographic and Authority formats by the TG, selecting fields which are to be controlled by an identifier. Only those fields where an identifier can be applied with clear correspondence between the field and one entity were included in the paper. The discussion paper was discussed at the MAC meeting, and the recommendation was made that the discussion paper be upgraded to proposal status; it was also approved at the meeting as a proposal. Both changes will be included into the update 23 to the MARC documentation, to be expected in fall 2016.

The third paper, Discussion Paper No. 2016-DP17 "Redefining Subfield \$4 to Encompass URIs for Relationships in the MARC 21 Authority and Bibliographic Formats" was presented by the British Library in consultation with the TG. This paper generated vivid discussions. It was acknowledged that the approach to recording URIs for relationships using subfield \$ was preferable to any of the other alternatives outlined by the paper. The distinction between relator codes and relationship codes in the MARC format was questioned. As of now, an across-the-board solution for recording URIs for any data element in MARC, subfield or field, seems to be preferred by NDMSO over what it regards as an ad hoc solution for single elements. This discussion will be continued; this paper should not be considered in isolation, but rather in the context of the other papers which the TG is in the process of submitting. Taken as a whole, it is hoped that they will achieve the comprehensive solution which is sought throughout the MARC formats.

IV.4.2. Formulating & Obtaining URI document [Charge 3.2]

A draft document was for commonly used sources for authorities and identifiers. For each source, screen captures were made showing where a URI could be found for a particular entity, or how to formulate a URI once the identifier for the entity is known. Before making this document available widely, it must be determined how best to organize it. Some resources provide URIs that directly represent a *thing* and others provide URIs that reference an authority (e.g., controlled or standard vocabularies, which may or may not have underlying metadata about the thing) or a resource describing a thing. The document needs to be able to distinguish this and inform catalogers which URIs are for real world objects and which are not. In order to be helpful to developers building tools, the document intends to also include descriptions of how data sources provide machine access to the data. Is the data published as Linked Data available through http, available through a SPARQL endpoint, data dumps, etc.? Another issue that must be determined is where to put the final document, and how it will be maintained. Should it be cooperatively maintained by the community (such as on a wiki), or should some group within PCC take responsibility for keeping it up to date and adding to it?

IV.4.3. Revisions to OCLC handling of HTTP URIs [Charge 3.1]

The question arises as to whether it would be better for catalogers to enter all needed URIs directly into the shared bibliographic record in WorldCat or whether OCLC should provide options for output of URIs based on data present in particular MARC fields and profiled library preferences. Clearly, some libraries will embrace use of URIs for their web-based catalogs while others may find them problematic in local displays of bibliographic information. OCLC staff have looked into the issue and believe that the use of output options would likely produce more consistent results as well as meet the varying needs of libraries.

The TG members are drafting a spreadsheet outlining the subfields that together name an entity for which a corresponding URI could be added in subfield \$0. That spreadsheet will be useful as the basis for future specifications for use by OCLC system developers. It will allow for a comparison of what is desired by the PCC cataloging community in terms of URIs corresponding to the entire named entity versus the existing use of subfield \$0 and subfield-\$0-like information used in OCLC heading controlling functionality. That heading control functionality allows for control numbers in multiple \$ subfields corresponding to different parts of a named entity, i.e., corporate name hierarchies, names and titles,

subjects and separately controlled subdivisions, etc. These are cases where output of multiple URIs corresponding only to part of the named entity would not be preferred.

OCLC cataloging policies in this area are expected to evolve as this TG makes recommendations and OCLC development work moves ahead on the proposed output options for URIs.

IV.5. *Next steps and in-depths analyses in year 2 [Charge 3; Charge 4]*

In 2016-2017, the TG will continue an agenda focused on practical outcomes. Work is already well advanced on several of the following items.

- IV.5.1. In collaboration with OCLC, develop a specification for outputting URIs based on internal linkages present in WorldCat data.
- IV.5.2. Complete the MARC object/URI reconciliation document and seek to incorporate the information into formal MARC documentation.
- IV.5.3. Produce work ID recommendation and use it in pilot implementation.
- IV.5.4. Produce discussion paper or proposal for handling relationships in MARC.
- IV.5.5. Consider additional targeted reconciliation projects.
- IV.5.6. In consultation with stakeholders, evaluate need for additional MARC proposals or best practices
- IV.5.7. RWO recommendations
- IV.5.8. Identify “homes” in PCC or elsewhere for aspects of the TG’s work that will need further exploration or continuing upkeep.
- IV.5.9. Outreach, advocacy, training
- IV.5.10. Etc.

V. RECOMMENDATIONS TO STAKEHOLDERS

During its first year, the TG was very much focused on the needs and interests of the many different stakeholders. This is reflected both in the outcomes of the work completed so far (see *Sec. IV.4. Outcomes*) as well as in the plans laid out for year 2 (see *Sec. III. 5. Next steps and in-depths analysis in year 2*). After careful consideration, the TG proposes the implementation of URIs in MARC for the near-term. The sooner this process can begin, the sooner the data providers, e.g. libraries, can produce the data that can be more easily transformed into linked data. In order to facilitate progress towards this goal, the TG developed the recommendations already outlined in the report above, such as the spreadsheet identifying the phase 1 entities for identities, i.e. the subfields that together name an entity in each MARC field (see *Sec. IV.4.3. Revisions to OCLC handling of HTTP URIs*) and the draft document *Formulating an Obtaining URIs: A Guide to Commonly Used Vocabularies and Reference Sources*. The TG hopes that this document could be used as starting point to develop an official list of PCC sanctioned initial source vocabularies for embedding URIs.

For the sake of consistency, expediency, and accuracy, it is advisable to use automated processes for populating MARC records with URIs. Individual catalogers doing this work manually is not a desirable

practice, and could be less efficient. Several possible ways to accomplish this goal, have been outlined in this report (see *Secs IV.2.4.1 MarcEdit, IV.2 .4.2 Authority Toolkit and IV.4.3. Revisions to OCLC handling of HTTP URIs*).

Outreach, advocacy and training will be a core goal of phase 2. The TG is planning o working closely with stakeholders, such as other PCC committees, to influence cataloging policies and best practices that have been identified problematic for the implementation of URIs in MARC.

Training needs related to implementation (for example how to obtain URIs or the difference between authorities and real world objects) will be communicated to the PCC Standing Committee on Training, so that appropriate training can be either identified or developed.

Though MARC is the most prominently used schema for library metadata, it is frequently used alongside many others that may or may not allow for the inclusion of URIs. In addition to that concern, are the maintenance of identifiers recommendation, in relation to reconciliation, and possible ILS functional requirements. The TG on URIs in MARC is recommending that new TGs be formed concerning URIs for non-MARC metadata.

VI. REFERENCES

1. The subgroup, Work IDs in MARC, has identified potential fields and scenarios to accommodate a work identifier (or multiple work identifiers). Considerations have been given to legacy data, whether a work identifier (ID) already established in an authority format, or not (7XX \$t, 1XX/240). An unambiguous relationship of a work ID among various vocabularies (024), and relationships among variant of a work, etc. The subgroup will present recommendations to the community in 2017.

Links:

Meetings of the MARC Advisory Committee: Agendas and Minutes

2015-06 MAC meeting:

http://www.loc.gov/marc/mac/an2015_age.html

<http://www.loc.gov/marc/mac/minutes/an-15.html>

2016-01 MAC meeting

http://www.loc.gov/marc/mac/mw2016_age.html

<http://www.loc.gov/marc/mac/minutes/mw-16.html>

2016-06 MAC meeting

http://www.loc.gov/marc/mac/an2016_age.html

<http://www.loc.gov/marc/mac/minutes/an-16.html>

Papers:

Informal discussion paper: "URIs in MARC: A Call for Best Practices" (Steven Folsom, Discovery Metadata Librarian, Cornell University)

https://docs.google.com/document/d/1fuHvF8bXH7hldY_xJ7f_xn2rP2Dj8o-Ca9jhHghleUg/edit?pli=1

Discussion Paper No. 2016-DP04: Extending the Use of Subfield \$0 to Encompass Linking Fields in the MARC 21 Bibliographic Format (British Library)

<http://www.loc.gov/marc/mac/2016/2016-dp04.html>

Discussion Paper No. 2016-DP05: Expanding the Definition of Subfield \$w to Encompass Standard Numbers in the MARC 21 Bibliographic and Authority Formats (British Library)

<http://www.loc.gov/marc/mac/2016/2016-dp05.html>

Discussion Paper No. 2016-DP17: Redefining Subfield \$4 to Encompass URIs for Relationships in the MARC 21 Authority and Bibliographic Formats (British Library in consultation with the PCC Task Group on URIs in MARC)

<http://www.loc.gov/marc/mac/2016/2016-dp17.html>

Discussion Paper No. 2016-DP18: Redefining Subfield \$0 to Remove the Use of Parenthetical Prefix "(uri)" in the MARC 21 Authority, Bibliographic, and Holdings Formats (PCC Task Group on URI in MARC in consultation with the British Library)

<http://www.loc.gov/marc/mac/2016/2016-dp18.html>

Discussion Paper No. 2016-DP19: Adding Subfield \$0 to Fields 257 and 377 in the MARC 21 Bibliographic Format and Field 37 in the MARC 2 Authority Format (PCC URI in MARC Task Group)

<http://www.loc.gov/marc/mac/2016/2016-dp19.html>

MARC Format Overview: Status Information:

<http://www.loc.gov/marc/status.html>

Examples for Sec. IV.2.1.

This LC subject heading string is linked to three different authority records. The links are OCLC's ARNs. No single \$0 could be output for this subject access point.

```
650 #0 †a Neurologists<Link:2068890> †z New Zealand<Link:255121> †v  
Biography.<Link:4933801>
```

This medical subject string is linked to one authority record, although the controlling process links individual subfields. It is a candidate for output of a single \$0 with a URI because the links all refer to the single authority record. In the case of MeSH, unlike LCSH, the \$0 subfield displays in Connexion. See OCLC record #957132118.

```
650 12 †a Neurology<Link:(DNLM)D009462Q000266> †x history.  
<Link:(DNLM)D009462Q000266>
```

Displays as:

650 12 Neurology †x history. †0 (DNLMD009462Q000266

So, it could be output with single \$ containing the corresponding URI for the MeSH heading.