First Report of the Task Group on URIs in MARC
15 October 2015

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

Membership:
Jackie Shieh (GW, chair), Robert Bremer (OCLC), Reinhold Heuvelmann (DNB, liaison to MAC), Chew Chiat Naun (Cornell, liaison to SCS), Gary Strawn (Northwestern, liaison to SCA), Paul Frank (LC, liaison to NACO), Jean Godby (OCLC Research), Les Hawkins (LC, liaison to CONSER), Adam Schiff (UWa., liaison to PoCo)

Consultants:
Nancy Fallgren (NLM), John Chapman (OCLC Product Services), Steven Folsom (Cornell, VIVO), Terry Reese (Ohio State, MarcEdit), Melanie Wacker (Columbia, MODS), Nancy Lorimer (Stanford, LD4L), Galen Charlton (Evergreen), Jodi Williamschen (SkyRiver, III), Shana McDanold (Georgetown, PCC Automation)

TG Communication & Process Tracking:
The TG met virtually via WebEx bi-weekly beginning on the 3 September 2015. Documents related to the team's discussions and deliberation are shared with interested parties listed above on Google via a GW-affiliated organizational account, pccuri2015@gwu.edu.

Membership and consultants were contacted by emails, WebEx sessions and/or phone calls. Between the scheduled TG meetings, offline conversations were conducted for more focus topics.

Charge:
There are four goals that the TG was given to address concerning the policy, the implementation, the tools (existing and/or in development) and the guidelines surrounding identifier in MARC.[1] In particular the identifier in the form of dereferenceable uniform resource identifier, commonly known as HTTP URI. In light of several subfields having been designated in the MARC encoding standard for identifier, or control number in both bibliographic and authority records, the TG will work with several stakeholders, e.g. PCC Standing Committees, the MARC community (Network Development and MARC Standards Office, MAC, etc.) , etc. to develop plans and provide guidelines as the TG moves along.

Summary:
As it currently stands, the TG has come to share the vision of embedding identifiers in MARC data. The membership is keenly aware of the complexity and layers of issues as well as the enormity of impact to both metadata professionals and users who will be using the data. Any textual data contained in subfield $0 is visible to the end-user. However, with a dereferenceable (HTTP) URI in subfield $0, the user will be provided additional means to follow "the trails" to discover additional resources inside and outside his/her immediate environment.

The task group is also aware of the semantic complexity of URIs that identify real world objects versus URIs that point to an authority record for the same object. It is unclear what impact this complexity will have on the task group's work at this point.

It is the TG's hope that the progress report depicted in this document will inform PoCo and associated PCC standing committees in designing and recommending incremental changes to MARC data. Any implementation can be carried forth with little disruption to existing data providing services

Approaches:
In response to the PCC Strategic Directions (SD.3) and its timeline as the Chair of the TG, I employed the following agile principles to test existing assumptions by collocating findings and garnering support for timebox reviews and outputs. This organization of given goals is an attempt to tackle issues more from a circular process than from a traditional linear approach. However, this break from tradition may yet to prove fruitful.

Tasks were organized in separate Google documents, and TG members were asked to claim "ownership" over each task; task "owners" are responsible for forming subgroups to begin collecting, examining, preparing, and presenting findings.

The TG worked together on Task #1, *Identify and address any immediate policy issues surrounding the use of identifiers in MARC records…* to orient and familiarize with the agile approach.

Environmental Scans:
The TG reviewed the guidelines and instructions in the MARC Standards and in the OCLC Bibliographic and Formats regarding the syntax and practices of providing identifiers or control numbers. [2]

The TG identified the following issues and data problems while reviewing three full MARC records from OCLC Worldcat.org, DNB and GW:

1. Syntax and semantics of subfield zero ($0)
2. Defining repeatability, type, and associated usage
3. Identifiers from controlled and uncontrolled identity services
4. Identifiers, control numbers, etc. found in more than one subfield

The semantics of current subfield $0 is defined to encapsulate "authority record control number or standard number" and is repeatable.

The parenthetical data in subfield $0, following the model for MARC field 035 and subfield $w in MARC fields 76X-78X and 800-830, were put in place prior to the advent of uniform resource identifiers (HTTP URI). Some services have been developed based on the existing syntax.

Do multiple instances of subfield $0s refer to the same reference in the same field? Do multiple subfield $0s point to the same entity, a real world object (RWO) personal or geographic name?

In many linked data communities, an identifier is neither an authority record or a standard number. A uniform/universal resource identifier can be coined "locally," then deployed within or outside the library community (e.g. Wikidata, GeoNames, Getty, etc.) Some identifiers may point to resources that are not modeled as linked data (e.g. IMDB or MusicBrainz).

A URI in the form of a URL is recorded in a subfield $u, and not subfield $0. Other subfields in MARC that have been defined for control numbers may potentially contain URLs, for example subfields $o and $w. Such practices will likely prevent successful exploitation of the data.

Existing OCLC Worldcat practices are due to system configuration and constraints over time. When a heading or authorized access point is controlled to the LC/NACO Authority File, the identifier embedded in subfield $0 is removed, except for the ones coming from the national libraries of France, Germany and the Netherlands. However, this will not inhibit OCLC from adding identifiers to bibliographic descriptions containing controlled headings or authorized access points if there are directives from the PCC and library community.

The MarcEdit suite continues to be the tool that serves technical services operations well. Over the past few months, Terry Reese (MarcEdit) has made major leaps and bounds to improve the MarcNext tool set, one of which is the identifier lookup.[3] This tool has enabled GW's goal of enhancing its bibliographic records as part of its Reclamation project this summer. As of 1 September 2015, GW has embedded ca. 4 million HTTP URI in subfield $0s for ca. 1 million bibliographic records. [4]

<u>Membership reflections so far</u>

It is vital that an identifier pointing to the resource is resolvable, unambiguous and unchanging over time. If there is not a cost-effective resolution in the short term, components of an identifier ought to be in place from which a URI can be automatically constructed.

It is highly desirable that an actionable identifier, such as a dereferenceable uniform resource identifier (HTTP URI), is deployed. Most ideally, an internationalized resource identifier, IRI, would be the identifier.[5]

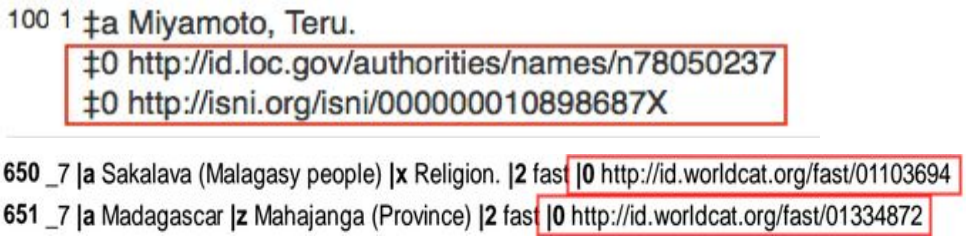
Respectfully submitted
Jackie Shieh
on behalf of the Task Group on URIs in MARC

Endnotes

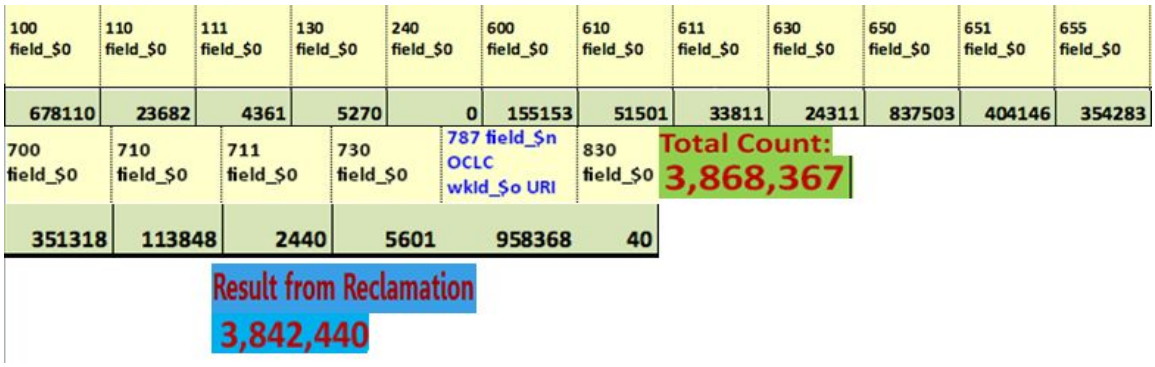
1. PCC Task Group Charge: Document
2. Links to MARC documentations concerning subfield zero ($0): http://www.loc.gov/marc/authority/ecadcntf.html ; http://www.loc.gov/marc/bibliographic/ecbdcntf.html ; http://www.loc.gov/standards/sourcelist/standard-identifier.html ; OCLC http://www.oclc.org/bibformats/en/1xx/100.html

3. MarcEdit Suite's MarcNext: http://blog.reeset.net/archives/1359

4. As of 1 September 2015, GW Voyager contains non-HTTP URIs (135,379) and over 3.8 million dereferenceable URIs.



The breakdown below of where $0 HTTP URI is stored:



5. Display of an HTTP identifier as a URI and an IRI.

URI: http://xn--rksmrgs-5wao1o.josefsson.org, http://xn--99zt52a.w3.mag.keio.ac.jp

IRI: http://räksmörgås.josefsson.org, http://納豆.w3.mag.keio.ac.jp