Alignment of RDA vocabulary IRIs with the MARC 21 encoding standard

Gordon Dunsire, Chair of RSC Technical Working Group. 25 March 2020

Summary

The MARC 21 manual for bibliographic description has two 'control' subfields for identifiers: $0 and $1 (see https://www.loc.gov/marc/bibliographic/ecbdcntf.html):

$0 - Authority record control number or standard number

$1 - Real World Object URI

Recent proposals to the MARC Advisory Committee (MAC) include examples using these subfields (e.g. http://www.loc.gov/marc/mac/2020/2020-dp04.html). They follow best practice guidelines from the Program for Cooperative Cataloging (PCC).

The RSC is concerned that the PCC guidelines and examples indicate the use of subfield $0 rather than subfield $1 for the IRIs of RDA vocabulary encoding schemes (VES), as per:

- Final report of the PCC Task Group on Linked Data Best Practices
- Formulating and obtaining URIs: a guide to commonly used vocabulary and reference sources
- PCC's FAQ 13.

Recommendation

Based on the following analysis from the RSC Technical Working Group, the RSC asks the PCC to reconsider the discussion leading to the statements in the FAQ and to consider changing its advice on the use of subfield $0 for the IRIs of RDA vocabulary encoding schemes.

Background

RDA: Resource Description and Access is described by its governors as "a package of data elements, guidelines, and instructions for creating library and cultural heritage metadata that are well-formed according to international models".[1] The package is published as RDA Toolkit.

The RDA Steering Committee (RSC) has been engaged for the past five years in the RDA Redesign and Restructure (3R) Project to align the RDA elements with the IFLA Library Reference Model (LRM) while providing "greater flexibility in the display of instructions,

---

[1] Committee of Principals Affirms Commitment to the Internationalisation of RDA. Date posted: 5 June 2015. Available at: http://www.rda-rsc.org/node/235

improved functionality for editors and translators, and … compliance with current accessibility standards …".[2] The redesigned RDA Toolkit is scheduled for release at the end of 2020. A beta version has been available for feedback and comment for the past 18 months.

RDA Toolkit includes a set of 'mappings' between RDA elements and the MARC 21 encoding standard.[3] These alignments are being reviewed and updated during the 3R Project to account for changes in RDA elements and in the MARC 21 bibliographic and authority encoding format. The task is being undertaken by experts from the British Library, responsible for alignments with the MARC 21 Bibliographic format, and Library and Archives Canada, responsible for alignments with the MARC 21 Authority format, with the RSC Technical Teams Liaison Officer and other members of RSC providing advice and support.

The new RDA instructions accommodate linked open data as one of its implementation scenarios. The scenario characteristics in the current beta version are:

"Metadata description sets are expressed in Resource Description Framework using IRIs taken from the RDA Registry.

Descriptions of the resource entities that comprise a single information resource are recorded in a separate metadata description set for each entity.

Descriptions of other entities that are associated with an information resource are recorded in a separate metadata description set for each entity.

The IRI recording method is used for values taken from a vocabulary encoding scheme, when available.

A metadata description set for an entity is linked to a metadata description set of a related entity using an IRI of the related entity."[4]

The RDA Registry includes a complete representation of RDA Toolkit entities, elements, and values in RDF. Values are represented using the Simple Knowledge Organization System (SKOS) application of RDF. The RDA value vocabularies are conformant with the SKOS Primer[5] and, as with the whole RDA Registry, 'cool URIs'[6] and so-called 'five star linked data'[7].

An example of an RDA value vocabulary (a type of vocabulary encoding system or VES) is RDA Media Type, which is intended for use with the RDA element "media type". Each distinct

---

[2] Outcomes of the 2016 RSC Meeting. Available at: http://www.rda-rsc.org/sites/all/files/RSC-Outcomes-2016.pdf

[3] Although these are published as 'mappings', to avoid confusion with semantic mappings in RDF and the Semantic Web, this paper will use the term 'alignment' for a semantic correspondence between two things, and 'mapping' for a semantic relationship between two things expressed in RDF. A mapping therefore requires that each thing and the relationship are represented by uniform resource identifiers.

[4] Note: "IRI" is Internationalized Resource Identifier; "resource entity" is one of the RDA/LRM entities Work, Expression, Manifestation, or Item.

[5] SKOS Simple Knowledge Organization System Primer. 18 August 2009. Available at: https://www.w3.org/TR/skos-primer/

[6] Cool URIs for the Semantic Web. 03 December 2008. Available at: https://www.w3.org/TR/cooluris/

[7] Linked data < Design issues. Last change: 2009/06/18. Available at: https://www.w3.org/DesignIssues/LinkedData.html

value is represented as an instance of SKOS Concept and assigned a URI. The URI is the subject of RDF triples giving more information about the concept, such as its preferred label, alternate labels, notation, definition, and scope note in various languages:

rdamt:1001 skos:prefLabel "audio"@en .[8]
rdamt:1001 skos:definition "A media type used to store recorded sound, designed for use with a playback device such as a turntable, audiocassette player, CD player, or MP3 player."@en .
rdamt:1001 skos:scopeNote "Media used to store digitally encoded as well as analog sound are included."@en .
rdamt:1001 <http://metadataregistry.org/uri/profile/rdakit/toolkitLabel> "audio"@nl .
rdamt:1001 skos:definition "Media gebruikt voor de opslag van geluidsopnames, ontworpen om te gebruiken met een afspeelapparaat zoals een draaitafel, cassettespeler, cd-speler of mp3-speler."@nl .
rdamt:1001 skos:notation "1001@en"

Other triples relate the concept to its value vocabulary (knowledge organization system) and status within the vocabulary:

rdamt:1001 skos:inScheme <http://rdaregistry.info/termList/RDAMediaType> .
rdamt:1001 omr:status <http://metadataregistry.org/uri/RegStatus/1001> .

The MARC 21 Bibliographic format encodes values of media types with tag 337.[9] Specific subfields relevant to the RDA vocabulary include $a (Media type term), $b (Media type code), and $1 (Real World Object URI). These correspond respectively to the three recording methods for the same 'value' that RDA specifies:
$a = structured description (a human-readable string label from a VES)
$b = identifier (a machine-readable string label from a VES)
$1 – IRI (an IRI/URI)

The subfields also correspond to the SKOS ontology vocabulary:

$a = preferred label
$b = notation
$1 = concept (instance URI)

This suggests that the correct alignment between the RDA element and MARC 21 encoding for 'media type' is:

RDA	MARC 21 Bibliographic	SKOS property
media type	337	
Recorded as structured description	$a	prefLabel
Recorded as identifier	$b	notation
Recorded as IRI	$1	[concept URI]

[8] Note: "rdamt:1001" is a shortened string referent to the URI <http://rdaregistry.info/termList/RDAMediaType/1001>
[9] 337 - Media Type (R). December 2019. Available at: https://www.loc.gov/marc/bibliographic/bd337.html

Thus a full encoding of the RDA media type 'audio' in MARC 21 is:

337 ##$aaudio$b1001$1http://rdaregistry.info/termList/RDAMediaType/1001

There is a one-to-one correspondence between the values recorded in each subfield, so only one subfield needs to be used in practice and the choice of subfield is application dependent.

The consistency of RDA element value vocabularies and of the MARC 21 encoding standard allow this alignment to be applied to all the RDA elements with a value vocabulary that have a corresponding MARC 21 tag.

However, recent proposals to the MARC Advisory Committee (MAC) include examples that encode the concept URI in subfield $0 rather than subfield $1. For example, the paper on "Renaming Field 345 and Defining New Subfields for Aspect Ratio and Motion Technique in the MARC 21 Bibliographic Format"[10] includes the example:

345 ##$dmixed aspect ratio$0http://rdaregistry.info/termList/AspectRatio/1003$2rdaar

The value recorded in subfield $d is the English preferred label taken from the RDA Aspect Ratio Designation value vocabulary ($a is used for a different purpose in this tag), and the value recorded in subfield $0 is the concept URI.

Discussion

The RSC understands that the MAC examples are following "Formulating and obtaining URIs: a guide to commonly used vocabularies and reference sources"[11], prepared by the Program for Cooperative Cataloging Task Group on URIs in MARC. This document includes the RDA value vocabularies and correctly states "RDA Value Vocabularies are modeled as skos:Concepts. RDA Elements are properties". It then states "MARC Subfield for Capturing URIs: Values, $0. Elements, $4".

The MARC 21 Bibliographic format (Appendix A) states:

"Subfield $0 contains the system control number of the related authority or classification record, or a standard identifier such as an International Standard Name Identifier (ISNI). These identifiers may be in the form of text or a Uniform Resource Identifier (URI).

Subfield $1 contains a URI that identifies an entity, sometimes referred to as a Thing, a Real World Object or RWO, whether actual or conceptual. When dereferenced, the URI points to a description of that entity. A URI that identifies a name or label for an entity is contained in $0."

As noted above, the URI of a concept taken from an RDA value vocabulary is not the URI of a label for the concept. The SKOS properties prefLabel and altLabel are used to record the various

---

[10] MARC Discussion Paper No. 2020-DP04. December 20, 2019. Available at:
http://www.loc.gov/marc/mac/2020/2020-dp04.html

[11] Formulating and obtaining URIs: a guide to commonly used vocabularies and reference sources. Version date: 2020-01-15. Available at:
https://www.loc.gov/aba/pcc/bibframe/TaskGroups/formulate_obtain_URI_guide.pdf

labels associated with a concept and do not themselves have unique URIs. RDA value vocabularies are multilingual, so there may be anything from 1 (preferred English) label to dozens of preferred and alternate labels in English and other languages. All are labels for the same concept, which is a Real World Object. The intended usage for linked open data implementations of RDA is that a label is obtained for display purposes by de-referencing the concept URI and utilizing SKOS semantics: only one preferred label is allowed for each language, so it is programmatically easy to switch language displays of the same metadata. Definitions and scope notes in multiple languages are also obtained by de-referencing.

The description of label strings, including names, titles, access points, and local identifiers is accommodated in RDA by implementing the LRM Nomen entity. If it is necessary for an application to specify a precise string value, the application can record the string as an instance of Nomen, assigning a URI for the string and recording statements about the nomen, such as language, script, relationship to other nomens (such as derivation), and administrative data such as agent and date of assignment of the string as a referent to its 'named' entity.

The RSC is concerned about the difference in the Toolkit alignment of the RDA elements and MARC 21 encoding and the advice offered by the PCC.

The conflicting alignments will confuse users of RDA Toolkit and the RDA Registry, including cataloguers and application developers.

Applying the PCC guidelines will result in less effective use of RDA metadata in applications.

An application that finds a URI for an RDA value in subfield $1 can use a generalized process for de-referencing the URI to obtain its RDF/SKOS graph and extract the value of the preferred label, definition, etc. in the language required for display or other functions. The application can also use RDA's de-referenced redirection to link to an HTML document that displays the graph in a human-readable format: an RDA Registry page.

Conversely, an application that finds the URI in subfield $0 must assume that the URI refers to a document (the Registry page) that can only be hyperlinked to the application display. The application can only extract and re-use the content of the graph with a specific program that understands the relationship between the syntax of the XHTML and the semantics of the RDF/XML serialization of the concept graph. If the RSC changes the HTML layout, etc., the application must update the 'connector' or extraction program.

The discussion on MARC 21 subfields $0 and $1 in the report of the PCC Task Group on Linked Data Best Practices[12] says:

"A URI given in $0 usually refers to an authority file or controlled vocabulary, which typically serves the primary function of providing a preferred label for an entity. Pairing the authority URI with the label helps the identifier to perform its primary function of supporting the label. By contrast, $1 URIs refer directly to entities and are drawn from linked data sources that typically do not support a single preferred label."

---

[12] PCC Task Group on Linked Data Best Practices Final Report. September 12, 2019. Available at: https://www.loc.gov/aba/pcc/taskgroup/linked-data-best-practices-final-report.pdf

This is accompanied by a reference to a proposal on "Use of Subfields $0 and $1 to Capture Uniform Resource Identifiers (URIs) in the MARC 21 Formats"[13]. Unfortunately, the RSC Technical Working Group cannot see how the reference justifies the statement.

The PCC report also references "URI FAQs"[14]; question 13 is "Why are skos:Concepts not considered Real World Objects (RWOs) with respect to $0 and $1?". The answer includes:

"skos:Concepts [sic], the central class in SKOS, are used to build entries within a particular Knowledge Organization Scheme. The concept works as a proxy for a thing in the real world, and it can have statements about it that do not apply to the RWO, e.g. versioning information for the term, or what scheme the concept is in -neither of which is true about the RWO".

The RSC Technical Working Group does not agree with the statement that a "concept works as a proxy for a thing in the real world". Rather, a URI works as a proxy for a concept in the real world, and every URI is a proxy for some thing. This is supported by the discussion in the proposal on the use of $0 and $1: "the use of $1 to hold URIs that refer directly to a Thing or RWO (Person, Place, Thing, Concept, etc.)".

---

[13] MARC Proposal No. 2017-08. May 16, 2017. Available at: https://www.loc.gov/marc/mac/2017/2017-08.html

[14] URI FAQs. September 26, 2018. Available at:
https://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI%20FAQs.pdf