Community vocabularies in RDA Toolkit

Gordon Dunsire, RSC Technical Team Liaison Officer, March 2020

Abstract

This is a briefing paper about ongoing work to review and rationalize the accommodation in RDA Toolkit of controlled terminologies for the construction of access points. The work covers several specific areas of development approved by the RSC, including: the consolidation of instructions on the use of string encoding schemes in the construction of access points; implementation of a 'community' area in the Toolkit to support policy statements, user-created documentation, and application profiles; revision of original Toolkit appendices; review of the Toolkit Resources menu; and removal of the Anglo-American focus of the original Toolkit to accommodate international perspectives.

Background

Appendices and tools

The first phase of the 3R Project involved the reformatting ('shredding') of the content of the original Toolkit to fit the new structure and design. The 3R Core Team decided that it would be better to wait until later in the project to process some of the original Toolkit appendices:

- Appendix A: Capitalization
- Appendix B: Abbreviations and Symbols
- Appendix C: Initial articles
- Appendix F: Additional Instructions on Names of Persons
- Appendix G: Titles of Nobility, Terms of Rank, etc.

The main reasons for waiting included the development of primary and equal focus of the Toolkit on individual entities and elements, and editorial policies on generalization, duplication, and consistency in the presentation of guidance and instructions.

The same approach was taken with some of the original Toolkit tools:

- Books of the Bible
- Medium of Performance

These appendices and tools were moved to the 3R Toolkit Resources menu with minimal reformatting.

Resources menu

The RSC agreed at its meeting in Santiago, Chile in October 2019 that it was now appropriate to review the drop-down menu and contents of the Resources tab in the beta Toolkit.

Sections 2 and 3 of the current beta Toolkit Resources tab menu are:

- ---
- Abbreviations and symbols
- Additional instructions on names of persons
- Capitalization
- Initial articles

- Terms of rank
- ---
- Revision history
- Books of the Bible
- Terms for medium of performance
- ---

(Section 1 is sourced from RDA Reference; section 4 is for legacy resources. These sections are under review for other reasons, including the location of Revision history.)

Anglo-American focus

The RSC agreed that the prominent display of "Books of the Bible" in the Resources menu reinforces perceptions outlined in the paper on "Western and Christian Bias in the 3R Toolkit" discussed at the Santiago meeting in 2019.

Local vocabularies

The 3R Core Team identified two other vocabularies of 'terms' that could be incorporated into the new Toolkit:

- Collective titles: this vocabulary is embedded in the original Toolkit instructions at 6.2.2.10.2. There is no element for 'collective title' and the focus on Anglo-American and literary contexts is too narrow for international applications.
- Gender terms: this vocabulary was removed from the original Toolkit before the 3R Project. RSC agreed to investigate how it might be retained as a 'local community' vocabulary for continuity with legacy metadata.

Both sets of terms have utility for local RDA communities and applications, but do not warrant treatment as 'global' RDA vocabulary encoding schemes.

Well before the 3R Project, the RDA Development Team identified a requirement for developing, maintaining, and publishing local vocabularies as part of the Toolkit. This is considered to be a significant tool for 'customizing' the Toolkit for communities for special materials such as music and rare materials.

Abbreviations for names of places

The RSC discussed the labelling of the current Toolkit Appendix B.11 for "Names of Certain Countries, States, Provinces, Territories, Etc." during discussion of the Fast Track proposal FT2020-03: Abbreviations in Place: access point for place. The RSC agreed to amend the label to improve accuracy in the context of perceptions of Anglo-American focus; the appendix covers the names of constituent parts of Australia, Canada, the United Kingdom, and the United States only, together with the names of the Union of Soviet Socialist Republics and Russian Soviet Federated Socialist Republic.

SES Project

Following discussion at the Santiago meeting, the RSC agreed to test the relocation of instructions for constructing an access point using a string encoding scheme (SES) into a more central place, nominally under the Resources menu, to improve clarity and reduce redundancy. The results of the test were given in a paper on "String encoding schemes in RDA Toolkit" discussed at the January 2020 online meeting of the RSC. The RSC decided to take Option 1, to completely remove the SES

boxes, together with associated Condition boxes and Examples boxes, from the element instructions.

A small task force of RSC members and policy statement writers was set up in February 2020 to continue this work.

Discussion

Community vocabularies

The working term 'community vocabulary' is used for a controlled terminology used in the construction of values for RDA access point elements. Specific terms are used in SES constructors; for example, a conventional collective title or a title of a book of the Bible may be used in a value of Work: access point for work, and a gender term may be used as a distinguishing characteristic in a value of Person: access point for person.

A community vocabulary is characterized as:

- Used only by specific RDA communities.
- Covering a limited number of languages and scripts.
- Not requiring translation in every Toolkit language.
- Requiring maintenance by experts in specific languages, scripts, and cultures.
- Not necessarily conforming to a full vocabulary encoding scheme: concepts may lack definitions, scope notes, IRIs, or notations, and coverage may be incomplete.

A community vocabulary should be no more prominent in the basic Toolkit menu system than an RDA VES, to avoid perceptions of favouring some RDA communities over others.

All community vocabularies should be gathered in one place for consistency and ease of selection.

All community vocabularies should be individually accessible via distinct sub-menu items, URLs, and citation numbers. This allows re-use in policy statements, user documentation, and application profiles.

A top-level 'Community vocabularies' item should be added to the Resources menu, and all specific vocabularies should be moved or added as sub-menu items, preserving existing hierarchical structures where appropriate. This presents vocabularies with equal prominence.

Section 3 of the Resources tab becomes:

- ---
- Community vocabularies
    - Terms for medium of performance
    - Titles of books of the Bible [amended for clarity and consistency]
        - Titles of books of the Bible: Library of Congress-Program for Cooperative Cataloging [amended for consistency]
        - Livres de la Bible: Bibliothèque et Archives Canada-Bibliothèque et Archives nationales du Québec [to be re-translated]
        - Festlegungen für den deutschen Sprachraum zum Erfassen der bevorzugten Titel von biblischen Schriften [to be re-translated]
- ---

Local vocabularies

The new menu provides consistent accommodation for the local vocabularies for collective titles and gender:

- Community vocabularies
    - Terms for collective titles
    - Terms for gender

Terms of rank

The original Appendix G: Titles of Nobility, Terms of Rank, etc. was given the shorter and more consistent title “Terms of rank” and reformatted into a more consistent structure when it was moved to the beta Toolkit.

The content has the characteristics of a community vocabulary, and covers only the countries of France, Indonesia, and the United Kingdom, and the Iban language.

Terms of rank should be moved from Section 2 of the Resources tab menu to the sub-menu of Community vocabularies in Section 3. The sub-menu items for the four existing areas of coverage should be re-titled for clarity and consistency:

- Community vocabularies
    - Terms of rank
        - Terms of rank in the Iban language
        - Terms of rank used in France
        - Terms of rank used in Indonesia
        - Terms of rank used in the United Kingdom

Abbreviations and symbols

The original Toolkit Appendix B for abbreviations and symbols has two distinct parts: general guidelines and instructions for using abbreviations in the values of specific elements; and lists of abbreviations in specific scripts and languages.

Reformatting the general guidelines and instructions for consistency with the 3R content, and applying new editorial policies to reduce redundancy and duplication, results in all of the content of this part of Appendix B being removed or relocated to specific elements.

The lists of abbreviations have a structure and function that is similar to other ‘community’ vocabularies. They should be moved from Section 2 of the Resources tab menu to the new ‘Community vocabularies’ sub-menu in Section 3 of the Resources tab menu.

The lists of abbreviations of names of places and words in specific scripts and languages are labelled:

- Latin Alphabet Abbreviations
- Cyrillic Alphabet Abbreviations
- Greek Alphabet Abbreviations
- Hebrew and Yiddish Abbreviations
- Names of Certain Countries, States, Provinces, Territories, Etc.

The accuracy and clarity of the label for names of places was noted during the Fast Track discussion noted above.

The lists for specific ‘alphabets’ are confined to specific writing systems or scripts. The only remaining usage of the term “alphabet” in the beta Toolkit is in the context of the original appendices on abbreviations and names of persons. Otherwise, the term ‘script’ is used for writing systems and ‘language’ for linguistic systems, recognizing that a single script may express many languages and a single language may be expressed in multiple scripts.

The Latin alphabet abbreviations cover multiple languages.

The label “Hebrew and Yiddish Abbreviations” mixes a script (Hebrew) with two languages (Hebrew and Yiddish), and is inconsistent with the “alphabet” labels for other scripts.

An alphabet is a category of script; the other categories are ‘syllabary’ and ‘logographic script’. Non-alphabet scripts may support standard abbreviations, so the use of ‘alphabet’ is too narrow in an international context.

The ‘working’ labels for the lists of script abbreviations are:

- Abbreviations for countries and states
- Abbreviations in Cyrillic script
- Abbreviations in Greek script
- Abbreviations in Latin script
- Abbreviations in Hebrew script

There are no symbols in any of the lists; it is not possible to abbreviate a symbol, so symbols are unlikely to be added. The title of the super-menu and menu page can be shortened to “Abbreviations”:

- Community vocabularies
    - Abbreviations
        - Abbreviations for countries and states
        - Abbreviations in Cyrillic script
        - Abbreviations in Greek script
        - Abbreviations in Latin script
        - Abbreviations in Hebrew script

Amended Resources tab menu

Consolidating these developments results in amended Sections 2 and 3 of the Resources tab menu:

- ---
- Additional instructions on names of persons
- Capitalization
- Initial articles
- ---
- Community vocabularies
    - Abbreviations
        - Abbreviations for countries and states
        - Abbreviations in Cyrillic script
        - Abbreviations in Greek script
        - Abbreviations in Latin script
        - Abbreviations in Hebrew script
    - Terms for collective titles

    - Terms for gender
    - Terms for medium of performance
    - Terms of rank
        - Terms of rank in the Iban language
        - Terms of rank used in France
        - Terms of rank used in Indonesia
        - Terms of rank used in the United Kingdom
    - Titles of books of the Bible
        - Titles of books of the Bible: Library of Congress-Program for Cooperative Cataloging
        - Livres de la Bible: Bibliothèque et Archives Canada-Bibliothèque et Archives nationales du Québec [to be re-translated]
        - Festlegungen für den deutschen Sprachraum zum Erfassen der bevorzugten Titel von biblischen Schriften [to be re-translated]
- ---

Section 2 of the Resources tab menu

Expanding the remainder of section 2 of the new Resources tab menu gives:

- Additional instructions on names of persons
    - Names in the Arabic alphabet
    - Burmese and Karen names
    - Chinese names containing a non-Chinese given name
    - Icelandic names
    - Indic names
    - Indonesian names
    - Malay names
    - Roman names
    - Romanian names containing a patronymic
    - Thai names
    - Recording surnames that include an article and/or proposition
        - [Latent sub-items based on a mix of countries and languages]
- Capitalization
    - General guidelines for capitalization
    - Names of agents and places
    - Title of work
    - Titles of manifestation
    - Other elements
        - Edition statement
        - Numbering of serials
        - Numbering within sequence
        - Notes
        - Details of elements
    - General guidelines for English language capitalization
    - Personal names and terms of rank, etc.
    - Names of people, etc.
    - Place names
        - Geographic features, regions, etc.

        - Political divisions
        - Popular names
    - Names of structures, streets, etc.
    - Names of corporate bodies
    - Religious names and terms
        - [Various sub-items]
    - Names of documents
    - Names of historical and cultural events and periods
    - Decorations, medals, etc.
    - Names of calendar divisions
    - Names of holidays
    - Scientific names and terms
        - [Various sub-items]
    - Trade names
    - Single and multiple letters used as words or parts of compounds
    - Other languages
        - [23 languages and language groups]
- Initial articles
    - Initial articles listed by language
    - Initial articles listed by word or words

There are multiple issues that will be resolved when these guidance and instructions are shredded into the new structure:

- Terminology (e.g. documents have titles, not names).
- Duplication or mis-location of instructions for specific elements (e.g. historical periods, etc. are timespans).
- Inconsistent presentation of general guidelines and instructions for all appropriate elements, instructions for specific elements, community vocabularies, etc.

Shredding will be carried out in stages similar to the processing of the content of the abbreviations appendix:

1. Extraction of community vocabularies and relocation to the Community vocabularies menu.
2. De-duplication and relocation of content to specific elements.
3. De-duplication and relocation of content to guidelines and general instructions for processing names, titles, and access points.
4. Improvement of consistency and clarity of remaining content. Location of remaining content in the Resources menu will be considered at this final stage.

This work is expected to take several months.

Language tagging

The language of Toolkit content must be indicated for use by screen-reader software.

Individual terms and phrases that are not in the base language (English or a full translation) must also be indicated to translators. This content needs to remain in its own language, without translation.

There is a standard technique for tagging content to indicate language. The mark-up is applied to contiguous blocks of content in the same language. For example, an entire element or guidance page

has a single tag for its language at the start of the content. If there is content in a different language embedded with the page, it is surrounded by a language tag. When a screen-reader reaches the end of the tag, it reverts back to the language of the page.

The number of tags required is determined by the number of contiguous portions of content in a language that is different from the base.

The community vocabularies that have content in specific languages and scripts will be reformatted to make contiguous portions larger and reduce the amount of language tagging required. This will also improve the re-usability of content for language-based RDA communities.

String manipulation

All of the content of the Community vocabularies area of RDA Toolkit is associated with the processing of string values of RDA elements.

Most of the content is associated with the processing of strings in string encoding schemes, including the omission of initial articles, application or expansion of abbreviations, normalization of capitalization, restructuring of name and title strings, and provision of controlled strings for specific constituent elements or boilerplate. SESs are used for the construction of access points, which are structured description values.

Some of the content is associated with controlling the terminology of values of attribute elements which may or may not be used in SESs. These elements accommodate the use of a local vocabulary encoding scheme for structured description values.

SES Project

Further development of the structure and location of the content of sections 2 and 3 of the Resources tab menu will be carried out within the SES Project.

The Project so far has:

- Removed specific value selectors and punctuation patterns from access point elements.
- Modularized and normalized the structure of value selector and punctuation pattern components of a string encoding scheme.
- Assigned a CMS identifier that is the basis of a URL for each SES and its components.
- Tested the use of SES URLs in policy statements.

Future activities include:

- Remove condition blocks that are exclusively associated with SESs from access point elements.
- Modularize and normalize the structure of condition blocks associated with SESs so that they can be incorporated in policy statements, user documentation, and application profiles. This will explore the presentation of conditions associated only with access point construction. Individual conditions can be re-assembled in the CMS into different blocks for presentation and use. Many individual conditions are CMS boilerplate supporting one-stop updating. Conditions have been structured to enable future development of material-specific views of global RDA based on work category, content type, carrier type, etc. that can be developed for local RDA.
- Shred section 2 of the Resources tab menu to relocate appropriate content to Community vocabularies and string encoding schemes.

- Review and amend the organization of community vocabularies for use in policy statements, etc.
- Normalize, etc. the use of introductory text in SESs and community vocabularies to improve 'localization' support in the CMS for translators and policy statement writers.
- Review and develop the use of conditions to drive the selection of SESs and community vocabularies, in the context of wider localization for material-specific communities.

The SES Project task force will continue to liaise with the RDA Policy Statement Writers Group, the RSC Application Profiles Working Group, and the RSC Translations Working Group and report to RSC at appropriate intervals.

Impact

This work will improve the clarity, consistency, usability, and re-usability of Toolkit content.

Changes to Toolkit content will have no impact on current policies and practice or on applications of RDA metadata.

The presentation of content used by RDA communities that are focused on language, religious, geopolitical, or other 'local' context is more equitable.

It will be easier to accommodate the addition of content for RDA communities based on language, script, or local culture. Existing gaps in content will be easier to identify.

This work will inform future decisions by the RSC on governance and maintenance of content for global and local RDA communities. The new labels and menu locations of the 'community' content only indicate a potential boundary. The underlying structure of the RDA content management system and application of the DITA standard treat all content on an equal basis; the Toolkit menu structure is controlled by a virtual directory that can be easily and quickly updated.

The amount of content to be processed in full translations of the Toolkit is reduced. The 'full translation boundary' that delineates what must be translated is controlled by a physical directory in the CMS. The directory includes only section 2 of the Resources menu tab. Relocation of content to section 3 moves it over the boundary and reduces the 'noise' or non-translatable content included in the 'must translate' section 2.

All of the work is scheduled for completion in beta versions of the new RDA Toolkit.