Submitted on: 14.11.2017

RDA in the Wider World
Date: 11 August 2016
Location: OCLC, Dublin, OH, USA
Co-Sponsor: Committee of Principals/Joint Steering Committee for Development of RDA

The RDA Registry: supporting RDA in a multilingual world

Jon Phipps
Metadata Management Associates LLC, Jacksonville, New York, USA.
jonphipps@gmail.com

Diane Hillmann
Metadata Management Associates LLC, Jacksonville, New York, USA.
metadata.maven@gmail.com

Abstract:

Resource Description and Access (RDA) has taken the lead among international bibliographic standards in focusing on meeting the needs of a wide variety of language communities. The RDA Registry has been developed with those needs in mind, and is currently making it possible for speakers of French, Spanish, and other RDA Toolkit languages to do their descriptive work and build their user interfaces in the language of their users and staff.

Many people think of RDA as the instructions in the Toolkit and not the Resource Description Framework (RDF) vocabularies, but in the RDA context, the RDA Vocabularies perform the function of what at one time was thought of as a 'data dictionary'. We use the terms "element set" and "value vocabulary" in the context of RDF metadata vocabularies as defined by the Library Linked Data Incubator Group [1]. We use the term "ontology" for an RDF graph using properties that relate the components of element sets, and we use other terms such as "schema" and "standard" loosely, in the context of professional bibliographic metadata communities.

Keywords: RDA Registry, vocabulary management, translation management.

Introduction

Living metadata vocabularies should be expected to change over time, to evolve. A vocabulary that is allowed to change is often volatile at best, particularly when in active development, and in combination with our continuously evolving and globalized languages and cultures, provides multiple management challenges. Real world usage and the passage of time have a tendency to challenge assumptions and decisions, making stability dependent on constant evaluation of processes and requirements. And as the size and complexity of metadata vocabularies increase, maintaining system-level stability changes as well for vendors, developers, and users of the systems that require those vocabularies. In particular, metadata vocabulary development supporting multiple languages has often been limited by concerns about the complexities of reconciliation and maintenance as vocabularies build out from an original language, as well as an attitude from many vocabulary creators that English, being the exclusive language of development and developers, is sufficient.

But as the technology for managing large vocabularies matures and the functional requirements become more global in scope—thanks in no small part by the growth of the Internet—new strategies are being created to streamline development and maintenance for multilingual vocabularies. Bibliographic vocabularies, particularly those not limited by traditional Anglo-American boundaries, provide a clutch of interesting use cases, as standards like RDA expands its focus beyond English. Since 2011, the RDA governance bodies have made a strong commitment to internationalization of governance and data, and in the past two years have made considerable visible progress toward meeting that commitment. [2] One of the services in use is the Open Metadata Registry (OMR). [3]

Users in the RDA context come in several categories: Vocabulary managers charged with the creation and maintenance of the RDA vocabularies in English (the base language) were the original user focus for the vocabulary management functions of the OMR, and were dependent on the OMR services that provide change management support to update their data. A second category of users, primarily interested in using the vocabularies themselves in descriptive (or 'instance') data, are the focus of the RDA Registry, a distribution and documentation tool designed to build more efficient access for machines and humans. The third, and newest category of users, are the translators—those managing specific language versions of the RDA vocabularies within the OMR.

Language supporting RDA users was also an issue for the RDA Toolkit, where the instructional guidance for building RDA descriptions is centered. [4] Clearly, coordinating language usage and maintenance between the Toolkit and the RDA Registry was an important consideration as multiple languages were added to both tools, and better synchronization between them was planned.

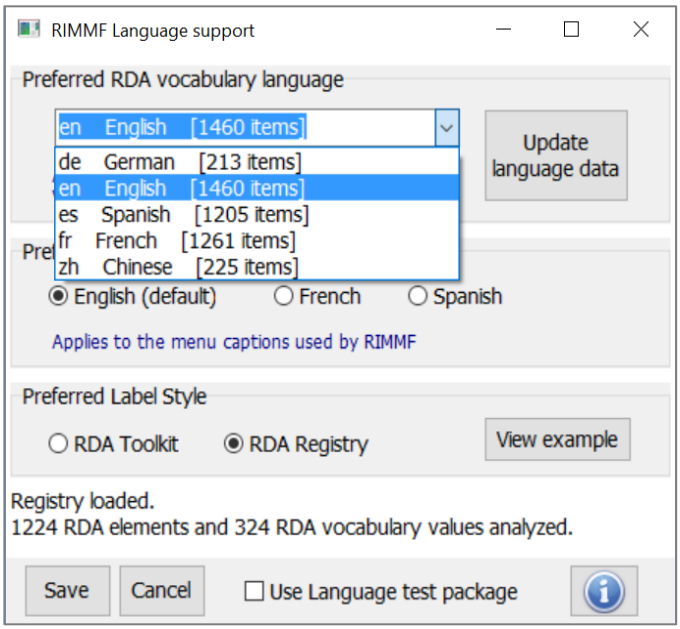

Figure 1: The developers of RDA in Multiple Metadata Formats (RIMMF) use the RDA Registry information, in various languages, as the basis for populating its cataloging tools. This capacity to match tools to users enhances acceptance of RDA, and also supports successful training and quality control of the resulting data.

Expanded Vocabulary Management Tools

Beginning with the relaunch of the RDA Vocabularies in January of 2014, development began on the RDA Registry, including a sophisticated dedicated vocabulary server, Git-based version control, GitHub-based publishing services, and integration with the OMR. [5] The OMR user interface, based on individual transactions, is increasingly showing its age, and had become particularly problematic as large vocabularies—like RDA and the IFLA vocabularies—became the norm. A maintenance strategy based on spreadsheets, introducing bulk export/import (including highly detailed transaction logging) and updating outside the OMR user interface, allowed more sophisticated versioning policies to be developed. The 'classic' OMR user interface, supporting single separate transactions for vocabularies desiring such capability, remains available although development efforts have shifted away from the OMR UI.

The original OMR workflow began with the registration of an individual user (and an organization, if relevant), to an account that gave overall administrative power to the initial registrant. That administrator could add maintainers with several levels of authorization, but language competence was not a factor in maintenance assignments, thus potentially giving access to the full array of language expressions to any maintainer, without regard to their competence in particular languages. As language-specific management capabilities were introduced, it became clear that the OMR user management controls were not adequate to support the distributed management needs of large, multi-lingual vocabularies. In consultation with IFLA vocabulary maintainers, a strategy was devised (currently under development) to allow vocabulary admins to limit maintainers to only those languages that the administrator specified, thus protecting other language versions from update by any but authorized language maintainers.

In parallel with the management upgrades supporting multiple language versions, work was progressing on the addition of language-specific versions of the RDA Toolkit, as well as building the capability to integrate the Toolkit (and its language versions) with the services of the RDA Registry. This is best described as a relationship where the Toolkit is a client of the RDA Registry vocabulary distribution services, where the RDA Registry, through its function as a global 'data dictionary' for RDA, provides the Toolkit with labels, definitions and descriptions of RDA data elements. Connection of the RDA Vocabulary URIs with internal RDA Toolkit IDs allows the RDA Toolkit to update itself as the Vocabularies for which it is providing instructions are updated, translated, and published.

With this strategy, even as the Toolkit language versions are managed separately, the relationship between Toolkit language versions and the RDA Registry language versions (managed together) can effectively provide updated information no matter where a user begins. There will necessarily be some lag time between English version changes and related language changes, and each language community associated with a translation determines the schedule and method for updating both the RDA Vocabularies and the RDA Toolkit.

One challenge as these technical developments proceed is to provide additional notification services to both the cadre of RDA Vocabulary translators and developers, using the RDA Vocabularies to provide data for user interfaces (such as the RDA Toolkit). Because English will remain the Ur language of RDA, as the English version grows and changes there will need to be notification services that will alert non-English language maintainers to those changes. Maintainers of RDA-based bibliographic data will also require notifications with a different focus, less tied to language differences, that will keep them informed of semantic changes to the RDA Vocabularies. While the current OMR provides Atom syndication feeds that provide a detailed log of additions and changes to the Vocabularies, OMR users will soon be able to subscribe to notification services providing less detail and through more distribution channels.

Because textual labels are used for various purposes, some users would like to distinguish between the 'official' Registry labels, and labels designed for specific uses. For RDA, the most obvious use case is to be able to store and maintain specific labels for the Toolkit which have not been verbalized (e.g., do

not begin with ‘has’, ‘is’ or other addition which implies directionality). The RDF expressions of the vocabularies use the ‘canonical’ URI (canonical because it identifies all the base element and variants, regardless of language or style), leaving display options to users of the data. As usage of these kinds of use-case-specific labels would be driven by a specific application profile there are potentially many use cases not yet defined. Also under discussion are possibilities for embedding a direct link from a vocabulary element to the usage and guidance available to Toolkit subscribers.

Element Sets: RDA Manifestation properties
Elements: has title proper

Detail | Statements | History

Profile property	Object	Language
name	titleProper	English
label	has title proper	English
description	Relates a manifestation to the chief name of a resource (i.e., the title normally used when citing the resource).	English
type	property	
domain	http://rdaregistry.info/Elements/c/C10007	
uri	http://rdaregistry.info/Elements/m/P30156	
status	Published	
subPropertyOf	http://rdaregistry.info/Elements/m/P30134	
lexicalAlias	http://rdaregistry.info/Elements/m/titleProper.en	English
hasUnconstrained	http://rdaregistry.info/Elements/u/P60515	
instructionNumber	02/03/2002	
ToolkitLabel	title proper	English
ToolkitDefinition	The chief name of a resource (i.e., the title normally used when citing the resource).	English
label	a pour titre propre	French
description	Met une manifestation en relation avec le titre principal d'une ressource (c'est-à-dire le titre normalement utilisé pour citer la ressource).	French
ToolkitLabel	titre propre	French
ToolkitDefinition	Le titre principal d'une ressource (c'est-à-dire le titre normalement utilisé pour citer la ressource).	French
ToolkitLabel	título propiamente dicho	Spanish
ToolkitDefinition	El nombre principal de un recurso (i.e., el título normalmente usado para citar el recurso).	Spanish

Figure 2: An OMR display for a specific element, including several language translations. Some properties shown here, identified as ‘Toolkit’ properties, are managed in the OMR though not currently integrated into the OMR editorial display.

Translation workflow

Work continues to expand and improve the import and export capability to support translators. The updated export process under development will store in the export log the version of English exported, and place an export identifier in the file name. When the translation is imported back into the OMR it will reference the export identifier and be able to tell from the history which version of English vocabulary any particular entry is translating.

The very first time a translation is published its version number will match the English version that was translated, even if the English version has moved on. The translation can continue to be updated, reimported, and published, incrementing its own published version number as needed, remaining tied to a specific English version, until a new export is created, translated, and imported which will then reference the English version that was exported for the updated translation work.

A	B	C	D	E	F	G
reg_id	uri	preferred_label_en	preferred_label_fr	RDA_Toolkit_Label_en	RDA_Toolkit_Label_fr	definition[0]_en
533	RDACarrierType:1001	Audio carriers		Audio carriers		
534	RDACarrierType:1002	audio cartridge	cartouche audio	audio cartridge	cartouche audio	A cartridge containing an audio tape.
535	RDACarrierType:1003	audio cylinder	cylindre audio	audio cylinder	cylindre audio	A roller-shaped object on which sound waves are incised or indented in a continuous circular groove.
536	RDACarrierType:1004	audio disc	disque audio	audio disc	disque audio	A disc on which sound waves, recorded as modulations, pulses, etc., are incised or indented in a continuous spiral groove.
537	RDACarrierType:1005	sound-track reel	bobine de piste sonore	sound-track reel	bobine de piste sonore	An open reel holding a length of film on which sound is recorded.
538	RDACarrierType:1006	audio roll	rouleau audio	audio roll	rouleau audio	A roll of paper on which musical notes are represented by perforations, designed to mechanically reproduce the music when used in a player piano, player organ, etc.
539	RDACarrierType:1007	audiocassette	cassette audio	audiocassette	cassette audio	A cassette containing an audio tape.
540	RDACarrierType:1008	audiotape reel	bobine de bande audio	audiotape reel	bobine de bande audio	An open reel holding a length of audio tape to be used with reel-to-reel audio equipment.

Figure 3: Snippet of a spreadsheet including English and French for the RDA Carrier Type value vocabulary. The next column to the right, Column H, includes the French version of the definition represented in English in Column G.

The flexibility of the import/export spreadsheet strategy allows both defaults for the normal case and additional options where necessary or useful. For new translations, the request for a spreadsheet would normally include the English version with an empty column for the translated values to be supplied. If the particular translation is already in the OMR and an update is the goal, the request would usually be for English and the existing translation to be updated in contiguous columns on the spreadsheet. When the translated values are added or modified, the completed spreadsheet can be imported back to the OMR by a translator with the appropriate permissions.

Figure 4: Although implementation of multilingual capabilities in the RDA value vocabularies is in process, this display of value de-referencing is very close to the planned final version. Note that the alternative label is expressed only in English, because there is not a translation of that label available in the Registry.

For RDA, language translations can be identified as necessary, down to the level of regional or cultural differences, and described using standard language codes. For instance, a French translation can specify that it is based on French as used in France, or in French-speaking Canada. Translators for either can export spreadsheets containing both variants, so that they can be easily compared. Each of

these variants can be treated as separate translations, and versioned separately. Requests for a published version of the RDA Vocabularies in a specific regional variant would contain the specific regional translation (fr-CA), falling back to the more general variant (fr) when no distinct region-specific translation is available and finally the omnipresent English when no other translation is available.

Versioning and Extension in a Multilingual Environment

The OMR was built to manage change at the most granular level, with every transaction on any level of vocabularies recorded and visible. This strategy in theory allowed OMR users to specify and name as versions snapshots (called 'timeslices') as a method of establishing versions of published vocabularies that could support regular and rational methods of updating vocabularies. [6] As useful as this feature was, it was largely ignored by the OMR's user community, despite its obvious similarity to other version-control system's (Git in particular) point-in-time commit snapshots.

As part of the integration of Git and GitHub into the OMR's functions, the versioning issues were rethought. Git, a distributed version control system, is often used to support versioning, and given the complications of versioning translations as well, a strategy is being developed to strengthen the approach to versioning within the OMR and the RDA Registry. [7] Based on the notion of Semantic Versioning in which a 3-level numbering system provides an intrinsic understanding of the degree of change represented by which level of the version number is incremented, the RDA Registry maintainers attempt to assign a specific version number to each published release of the RDA Vocabularies. GitHub supports tagging a specific commit of a repository with a version number and identifying it as a 'release'. This works well when each individual component of a complex system has its own version number, or when there is a single repository that is monolithically updated. When the repository has many variations (translations) of the same elements, each existing at multiple version numbers, there is less support for the current single repository publishing scheme.

Although the current updating process links the language translation to the appropriate English version, more experience suggests that separate versioning information beyond the semantic version numbering might be needed. One option is to assign translations their own minor version with the formal semantic 'translationOfEnglishVersion' relationship to the 'source' version. In addition, Semantic Vocabulary Versioning is only intended to cover semantic and structural changes in the RDA Vocabularies and doesn't include a concept of technical changes to the vocabularies, such as changing the structure of the JSON-LD representation in a way that doesn't alter the semantics or the resulting RDF, or adding new descriptive metadata for the vocabulary. One potential solution is to include a fourth level to the convention of Semantic Vocabulary Versioning, strictly to cover technical changes.

Extension

There are well-known issues with straight translation between languages, and the RDA strategy for dealing with these issues is useful for filling in those gaps as well as making RDA more useful in other cultural heritage communities. The key to solving these is the notion of extension, which extends well beyond the usual practices used for vocabulary development. For language communities, extension enables the addition of terms where straight translation is insufficient. A simple use case occurs where a term of local significance is narrower than an existing RDA term. The vocabulary for RDA Base Material provides a hypothetical example: the term 'bronze' is not currently on the list, and adding that term, and perhaps terms for other specific alloys might make sense for the museum community in particular. Another use case arises if the translation of Term A in English cannot be simply accommodated with a single Term A in another language, an extension can be built to include more terms or term relationships in the additional language.

It is possible to envision an environment where distributed catalogers working in close connection with institutions and materials being described could provide additional terms and relationships within communities that have built the technical infrastructure to support localized vocabulary development.

Such an infrastructure could include a shared domain and tools to provide URIs, definitions and relationships as well as a community based review process to publish such additions for the use and reuse of community members and beyond. In fact, such support could be provided within the OMR, should communities organize themselves to use it.

These extended terms would not necessarily be incorporated in the more generalized or the English version of RDA, but could be published and maintained by that specific language or practice community, with the infrastructure and mapping relationships that could allow the work done to be used by everyone.

Continuing challenges

The history of bibliographic vocabulary development has been largely top down, with little or no accommodation of specialized needs of communities desiring to work within a generalized consensus model. There were costs to that model, paid largely by the specialized cataloging communities for music, law, maps, etc., and those costs were considered to be intrinsic to the effective sharing of metadata in the bibliographic universe. The larger RDA development effort continues in part in that mode of generalized approach and consensus change, but in addition recognizes that current technology requires neither the straightjacket of total compliance nor the overhead inherent in the process of broad approval by the RDA organization to co-exist and thrive. This indeed represents the potential of bottom up vocabulary development meeting specialist needs, yet linked to the general vocabularies providing the important secure basis for sharing data.

The RDA Vocabularies are designed to be used at the system integration level in multilingual environments to present RDA-based bibliographic information to both machines, for transport and indexing and ultimately understanding as AI systems become more intelligent, and to humans, for display and understanding of the data in any context. Much effort has been put into clarifying our understanding of the individual FRBR-based entities and identifying and describing those entities. To date very few (if any) library system vendors have embraced the opportunities afforded by the rich, extensible, multilingual descriptive metadata environment of RDA.

RIMMF has begun to show the possibilities for metadata creation and management, making good use of the RDA vocabularies to drive a rich, inherently multilingual cataloging interface. Library system vendors sitting on the sidelines and waiting for the correct amount of ‘user demand’ will be left scrambling to catch up with whichever system vendor finally recognizes and capitalizes on the possibility to create, index, and present metadata that can be understood by both systems and users in a localized language environment anywhere in the world, *without the additional cost of having to maintain that interface themselves.*

Bibliographic metadata now lives in a world that does not consist only of speakers of English, nor is it constrained by the need to know, and thoroughly understand, the English language. Realizing the full potential represented by the expansion of RDA into language and specialized communities is still largely lacking the participation of the library system vendors, but that could (and should) change.

REFERENCES

[1] Isaac, Antoine, William Waites, Jeff Young, & Marcia Zeng. Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report. Available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/.

[2] Dunsire, Gordon. “Towards an Internationalization of RDA Management and Development”, Italian Journal of Library, Archives and Information Science, v. 7, no. 2, 2016. Available at: http://leo.cineca.it/index.php/jlis/article/view/11708

[3] Open Metadata Registry (OMR). Available at: http://metadataregistry.org

[4] RDA Toolkit. Available at: http://www.rdatoolkit.org/

[5] RDA Registry. Available at: http://www.rdaregistry.info/

[6] Phipps, Jon. "Timeslices and Versions." The Registry Blog, March 26, 2008. Available at: http://metadataregistry.org/blog/2008/03/26/timeslices-and-versions/

[7] Hillmann, Diane, Gordon Dunsire and Jon Phipps. "Versioning Vocabularies in a Linked Data World", paper presented at the Paris Satellite of the IFLA 2014 conference in Lyon, France. Available at: http://hdl.handle.net/1813/40559