Building a Platform to Manage RDA Vocabularies and Data for an International Linked Data World

By Jon Phipps, Gordon Dunsire and Diane Hillmann

Author note

      Jon Phipps and Diane Hillmann are principals in the firm of Metadata Management Associates, LLC, based in Jacksonville, NY. Gordon Dunsire is an independent consultant based in Edinburgh, Scotland. Correspondence should be directed to Metadata Management Associates, LLC, P.O. Box 282, Jacksonville, NY 14854 or to Diane Hillmann at metadata.maven@gmail.com.

ABSTRACT:

The management of vocabularies in the evolving linked data environment requires different tools and processes than libraries and other memory institutions have used in the past. The RDA Resource Description and Access standard has taken the lead in building tools and providing services as part of its RDA Registry development. The evolution of the current RDA Registry and the Open Metadata Registry (OMR), on which it is built, are described, including the rationale for directions, decisions and ongoing development.

KEYWORDS:

versioning, change management, vocabulary management, linked data, RDA, multilingual

Building a Platform to Manage RDA Vocabularies and Data for an International Linked Data World

Introduction

RDA: Resource Description and Access is an international cataloging standard managed and maintained by the Committee of Principals for RDA (CoP) and the Joint Steering Committee for Development of RDA (JSC). The standard is a successor to the Anglo-American Cataloging Rules and until recently its development has been undertaken by, and largely for, English-speaking communities. The addition of representation from the German-speaking communities in 2011 signalled an intention to expand the utility of RDA as a global standard. RDA is currently used or being considered for use in many non-anglophone communities around the world. [RDA around world] The CoP has made a clear statement about the internationalization of the standard:

> "RDA is a package of data elements, guidelines, and instructions for creating library and cultural heritage resource metadata that are well-formed according to international models for user-focussed linked data applications." [CoP announcement]

RDA guidelines and instructions are manifested as the RDA Toolkit. A separate RDA Registry [RDA Registry] was created in 2014 to provide documentation and support for the RDA Vocabularies — the data elements (Element Sets) and terminologies (Value Vocabularies) represented in Resource Description Framework (RDF).

RDA therefore needs an infrastructure to support the maintenance of the relatively large, complex, version-controlled, and multilingual RDF vocabularies for data elements and data values specified by the instructions. These vocabularies are necessary components for developing linked data applications using RDA data. The JSC intends well-formed data to conform to the semantics of the RDA guidelines and instructions, based on underlying international standards maintained by the International Federation of Library Associations and Institutions (IFLA), and requires the vocabularies to reflect those semantics accordingly.

The JSC is preparing for significant changes in those underlying standards. [JSC announcement] These are likely to have an impact on the RDA vocabularies, with the introduction of new entities as RDF classes, and associated attributes and relationships as RDF properties. The JSC also anticipates the deprecation of some existing entities and properties. It is uncertain when changes to the underlying standards will be completed, but there is pressure to continue to develop RDA rather than wait. The reviews of IFLA standards have taken longer than anticipated, leaving important areas of coverage under-developed, such as the treatment of

aggregation resources. This suggests that the RDF vocabularies will develop incrementally, and with varying effects on the semantic cohesion of previous versions.

The CoP expects the number of translations of the RDA guidelines and instructions to increase as they are developed for international communities. Many language communities have decided to translate only the glossary, containing the entity, element, and value labels and definitions, evidence that translations of the RDF vocabularies will occur at a greater rate than for the full text of RDA. The current workflows for translations mean that they tend to lag behind changes to the English text, and they may be incomplete because of interpretation issues with specific areas of text.

The JSC requires a vocabulary management system that can support the development and utility of RDA's RDF vocabularies in these demanding circumstances. The system needs to provide maintenance interfaces that are easy to use for a wide variety of RDA communities, facilities for version control, multilingual and multiscript capabilities, support for linked data application developers and communities, and support for standard RDF semantics. The system must meet the good-practice expectations of the Semantic Web communities. The barriers to use of the system must be low, as open as possible, to encourage widespread use of RDA.

RDA and OMR

The RDA standard has been in development for more than a decade, and has, in the timing of its efforts during the era of Linked Open Data (LOD), significantly changed the conversation about bibliographic data.

Since 2008, when the RDA Vocabularies (by which general term we cover both element sets and value vocabularies)[1] began to be developed in the Open Metadata Registry (OMR), several goals dating from the inception of the original NSF-funded National Science Digital Library (NSDL) Registry continue to be in the forefront: provision of a simple-to-use tool to develop and manage vocabularies; making conformant XML and RDF export possible for a variety of purposes; and providing vocabulary owners and users ways to collaborate in building and managing vocabularies. The OMR, still under active development though now itself a decade old, remains the lynchpin for all the advanced services built for RDA.

As planning finally proceeded to publish the RDA Vocabularies (long in gestation in the OMR as "new-proposed", waiting for JSC review) it was clear that new thinking and new services needed to be considered, in part because, during that time, the instructions in the RDA Toolkit had continued to be maintained separately and the RDA Vocabularies were now significantly out of synchronization with the Toolkit.

The technical team behind the OMR began to look seriously at Git and GitHub, not to replace the OMR but as the basis for better and more useful services built around the OMR, to be implemented first for RDA. Git, a distributed revision control system, and GitHub, a Web-based

---

[1] The terms 'element set' and 'value vocabulary' originated with the Dublin Core Metadata Initiative, and still provide a useful distinction for the functionality of different vocabularies.

Git repository hosting service, offering all of the distributed revision control and source code management functionality of Git as well as adding its own features, was a particularly good fit for RDA vocabulary distribution needs. Some of the new services, most importantly the RDA Registry, went public in January 2014, while others--particularly documentation and issues management--have continued to evolve to this day. The continuing use of the OMR (still the place where the vocabularies are managed) in combination with the Git and GitHub-based RDA Registry, have proven to be a useful set of tools that has moved RDA firmly into a prominent position in linked data compliant publishing and management of vocabularies of all kinds. There have been a number of challenges in this development effort, and for the most part they remain challenging even as new services are built and stabilized.

Technical challenges

Multilingual vocabulary maintenance

One significant way that RDA has influenced the conversation can be found in its evolution from an English-only standard to one that is multilingual and multicultural. In 2010 the OMR developers began implementing improved multilingual functionality for RDA in the OMR, adding support for German versions of selected elements, definitions and scope notes, in collaboration with the Deutsche Nationalbibliothek. This work demonstrated that the goal of providing true multilingual services did not require totally separate vocabularies and maintenance processes, until then a basic assumption of the plan for multilingual RDA.

The trajectory of change was complicated by the early prevailing notion that RDA was confined to the guidance rules embodied in RDA Toolkit. When the RDA Vocabularies were first built in the OMR following the historic agreement between the JSC and DCMI in April 2007 [London Meeting], there was little common understanding about the relationship between the Toolkit and the Vocabularies, much less how the early plans to translate the Toolkit could be accommodated in the Vocabularies. For instance, a translation of the Toolkit would involve translating terms in the Toolkit Glossary that were the same or similar to those used for the RDA Vocabularies elements and concepts, but there were no policies or processes in place to reconcile those resources for the user community.

The Toolkit, as a textual document, albeit in digital format with linking capabilities, clearly could only be made available to other language communities via a full translation. This was assumed to be the case for the vocabularies as well. Indeed, most of the functioning models for multilingual vocabularies have chosen to build and maintain them separately, with mapping between.

As the value and the necessity of the RDA Vocabularies to support the RDA scope and structure [RDA Scope and Structure] became clearer, issues of synchronization between the instructions and the vocabularies began to loom larger. A particularly difficult question in early synchronization discussions was the source of updated information common to the Vocabularies and the Toolkit: the vocabularies registered in the OMR feeding data to the Toolkit; or the text of the Toolkit feeding data to the OMR? The OMR had been built to provide such distribution

services, and the Toolkit capability to provide such data would have to be built from scratch, so the initial agreement for development was limited to the sharing of identifiers between the systems. These issues have been addressed again more recently following the implementation of a new infrastructure for maintaining the Toolkit, and work is proceeding to allow updating of the Toolkit from the new RDA Registry.

Multilingual URI management

The first vocabularies available in the OMR were based on the Simple Knowledge Organisation System (SKOS) [SKOS] for value vocabularies, then brand new and not yet implemented elsewhere. The developers were aware at that time that value vocabularies tended to be more volatile and required strategies for change if they were to be useful in production environments. URI stability in environments where identifiers are based on terminology in languages that evolve quickly seemed unsustainable for the OMR and for those maintaining data based on the vocabularies. This was seen as a particular issue in SKOS, where the ability to add alternative labels to assist discovery suggested a scenario where the preferred label might well, in the passage of time, flip position with the alternative label. A perfect example of this volatility is shown in the various labels used for what have come to be known as 'USB flash drives' or 'flash drives', formerly called 'thumb drives' and dozens of other names.[2]

With this in mind, default URIs in the former and current OMR for SKOS vocabularies are numeric. There has been a considerable amount of criticism of this strategy, primarily from developers who had become used to being able to 'read' URIs to determine the semantics of the resource. This desire to shortcut the need to reference the coagulant source or documentation to determine correct semantics, though understandable, meant that URI stability would require more complex (and expensive) maintenance in the OMR — 'drift' in labels away from the original label encoded in the URI would require either a new URI or deprecation of the original URI. This likely increase in the number of deprecated preferred labels and URIs would require a heavy load of redirects, particularly for vocabularies in the process of development.

English-based URIs were introduced when the OMR was extended to manage RDF element sets, partly because element vocabularies were expected to be less volatile than concepts, but this did not prove to be the whole case. The OMR helpfully builds URIs automatically from a standard base for the registered element set, plus the 'name' of the element (generally a camel-cased version of the label). This avoids a great deal of creative naming decisions and typos (although the automatic URI can be overridden during the process). Particularly for RDA, where the element names tend to be descriptive and long, many lengthy URIs were created in the draft element sets, leading to almost as much displeasure as the opaque URIs.

---

[2] The article in Wikipedia redirects from 91 other names for 'USB flash drives.'
(http://dispenser.homenet.org/~dispenser/cgi-bin/rdcheck.py?page=USB_flash_drive)

For the RDA Vocabularies, the RDA developers have chosen to use "opaque" canonical URIs based on numbers, as well as providing developer-friendly "lexical aliases", for the elements in a range of languages. Lexical aliases are true resolvable IRIs, defined as a sub property of owl:sameAs, that provide a set of language-specific, "readable" IRIs for each element. This keeps the canonical URI language-neutral and allows developers, especially those for whom English is not optimal, to work comfortably in their own language with IRIs that are more memorable than the canonical URIs. Since a lexical alias by definition is based on the label, changes to the label can and do generate a new lexical alias, but historical aliases are maintained and will continue to redirect to the canonical URI indefinitely, solving the problem of label 'drift'. Because all the data in the registry is tagged with a language, lexical aliases used in data can be redirected to the canonical URI as the data is ingested.

Thus the strategy of canonical URIs and associated lexical aliases provide the best of both approaches, and is also a significant boost to the utility of the multilingual approach desired by the JSC.

User management in a multilingual environment

Routine management of the multilingual RDA vocabularies is expected to be accomplished in the OMR by agencies responsible for the translations, either as part of a translation effort destined for the Toolkit, or separately. The original OMR management control for users was relatively light — a designated administrator assigned roles to other members of a team, which enabled them to function within the OMR at a level matching their training and capabilities. The OMR was not originally built to accommodate multiple teams with varying permissions and languages, so a new user management system was needed to permit language specialists to operate only within the languages for which they had permission, ensuring that neither error or terror could take place.

As the teams and administrative users scale up and the languages increase, the teams will need to determine acceptable limitations on the scope of their ability to change information they lack authorization to make, perhaps in languages in which they are not competent. To do this requires a much more complex and layered approach to user management, with a thorough understanding of how administrators will work to maintain quality. The current limited list of maintainer roles is being expanded, and maintainers will be assigned the languages in which they can work, as well as be linked via notifications and change feeds to the administrator making those assignments.

Workflow, distribution

One missing facility in the OMR that was identified early on was the lack of upload capability — in essence use of the OMR was limited to, and therefore by, the human user interface. Once decisions were made on the processes used to publish the RDA Element Set, this gap became more problematic. Currently in active development is the capability to upload

vocabulary data to the OMR using tabular data (CSV files), and to push the changes made to the RDA Registry. Download of CSV templates for initial use by translators or others carrying out maintenance in bulk is also part of this intended facility.

The chief difficulty in enabling spreadsheet upload has been, in some respects, the OMR's unique capability for creating and storing a detailed history of changes at the individual 'statement' level, capturing dates, specifics of the change (adds, changes and deletes) and the vocabulary maintainer who made the change. This data drives the automatic notification feeds, and also allows administrators to determine where gaps in understanding or training have occurred.

Figure 1: Screenshot of OMR History page

To view the complete element set, see:
http://metadataregistry.org/schemahistory/list/schema_property_id/13634.html

GitHub and similar services such as Bitbucket, have become increasingly vital to software developers in the management and distribution of software code (with the Git software managing version control), primarily for open source and community supported projects. Since the OMR generates vocabularies in a variety of RDF flavors, including such popular serialization formats as turtle and JSON-LD, managing publishing and distribution has often been a challenge. The decision to use Git and GitHub for the RDA Registry has been an important breakthrough, because it has allowed a shift in focus to vocabulary distribution and workflow using automated services to publish vocabularies and update distribution servers.

More recently, the version control capability of Git has been used for writing projects, including documentation of GitHub supported functionality and code. The attractiveness of this model is that the documentation is managed and versioned using the same tools that manage the vocabularies. The version control aspects are especially useful, since documentation has traditionally not made effective use of versioning.[3]

Figure 2: Screenshot of RDA Registry versions page
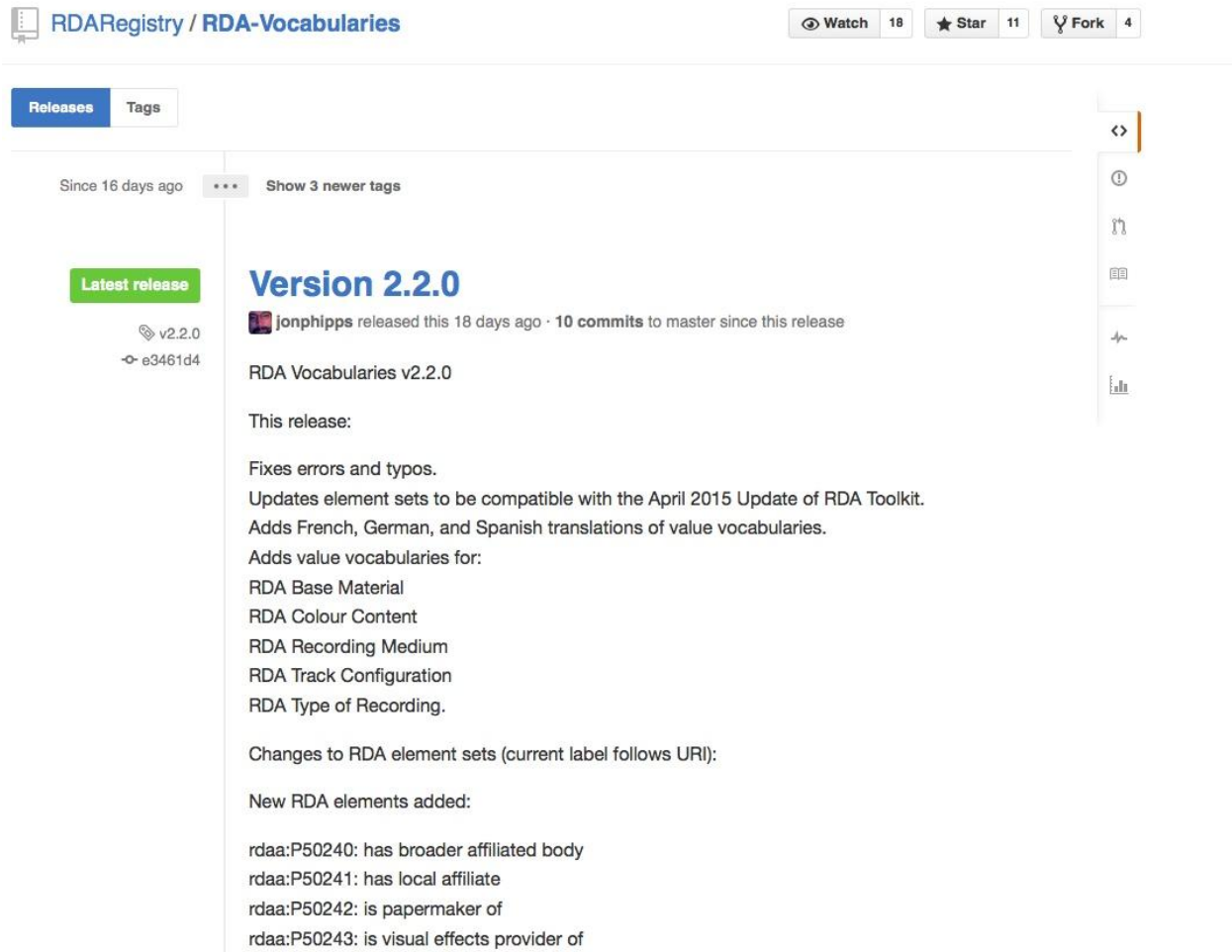
For the complete page, see: https://github.com/RDARegistry/RDA-Vocabularies/releases
]

Stability and versioning

The evolution, still very incomplete, from an entirely central data distribution model to a more widely distributed approach has been a continuing struggle for traditional memory institutions. These institutions have, over time, developed complex sharing policies and practices

---

[3] For information on using github for writing projects, see http://readwrite.com/2013/11/08/seven-ways-to-use-github-that-arent-coding

designed for local catalogs or consortial services. Many of these institutions moved onto the web late, in fits and starts, and retained an emphasis on individual look-up, or, at best, distribution of entire files.

Maintaining data quality over time is one of the ongoing challenges facing the linked data community, especially since the (linked) vocabularies used to describe the data rarely travel with the data. For instance, if published linked data (regardless of schema) contain vocabularies managed by a range of institutions and services, how are the changes in those vocabularies to be communicated to both publishers and consumers of the data? Capturing entire vocabulary files on a regular basis and looking for changes (not all of which will be obvious or relevant) seems a primitive solution. Vocabularies can't reasonably be resolved every time the data is accessed, and vocabularies in active use by the linked data community may suddenly cease to be maintained or resolve, so a local cache must usually be maintained. How do consumers and publishers of linked data know when to refresh the cache?

Adding modification dates or notes to web pages, without specifying what those changes represent, assigns significant effort and costs to the instance data managers, hardly an incentive to maintain their own data responsibly. Those imagining that there will be generalized agreement about now such notifications will be built and distributed are not looking past the old boundaries: not all these vocabularies will be managed by librarians or by library institutions for libraries, and those that are will evidence a full range of function and dysfunction, as they do now.

Most of us old enough to have witnessed the personal computer revolution and subsequent growth of mobile devices have lived through several stages of evolution as developers of applications coped with the necessity of updating their products as operating systems changed, competition for users grew, and functionality sought by customers became more sophisticated. Current practices for updating software optimize fast distribution of changes and are increasingly automatic, despite past emphasis on user control in an effort to avoid malware.

As so many other development communities have discovered, there is tremendous benefit to being able to reference specific versions of software libraries in order to maintain and ensure system stability. The average Linux server isn't going to automatically update its libraries every time it is accessed, and it is the rare systems engineer who would create a system that automatically updated to the absolute latest version of every library on which it depended. Yet this is the situation that often exists with linked data described by 'institutionally managed' vocabularies. Linked data described by distributed vocabularies should be semantically stable indefinitely and this requires the ability to reference specific, stable vocabulary versions.

The OMR initially created a crude (remember it was a decade ago) timestamp-based versioning system that allowed a direct reference to any vocabulary as it existed at any point in time. A version tag could be associated with any timestamp in order to 'name' it and dereferencing that named version would retrieve the vocabulary as it existed at that point in time. Looking at the way Git much more sensibly manages this same process using 'commits' and commit identifiers, and the way GitHub uses Git tags as published version identifiers has been

instructive, and the process of commit, then tag with a version, then commit to GitHub, has been integrated into the OMR publishing workflow. Currently the version that users of the vocabularies retrieve from the RDA servers is always the most currently published version. Publishers wishing to lock in a particular version must reference that version in the GitHub repository. This is of course suboptimal and the OMR developers are actively investigating alternatives that would allow direct dereferencing of published versions on the vocabulary server itself.

The RDA Registry and Semantic Versioning

Over time the software industry has refined their practices to be able to indicate via the version number the extent of change represented in an update. Many in the software development community have begun to embrace a formal specification of version management known as "Semantic Versioning".

"Without compliance to some sort of formal specification, version numbers are essentially useless for dependency management. By giving a name and clear definition to the above ideas, it becomes easy to communicate your intentions to the users of your software. Once these intentions are clear, flexible (but not too flexible) dependency specifications can finally be made." [Preston-Werner]

There have been some attempts to apply semantic versioning principles to ontologies, making the point that there are more similarities with the requirements for software than differences, as well as some general similarities to the management of application programming interfaces:

"OWL ontologies should be semantically versioned, which means two things:
* make the ontology's version identifier structured & meaningful, i.e., encode some meaning in the string of characters that makes up the version identifier; and
* change the version identifier according to well-understood, public, and reasonable rules.
Which suggests, of course, that a version identifier, plus a strategy for changing version identifiers, is a simple signaling mechanism intended to make multi-party coordination games cheaper and less disruptive for the participants. Consumers and producers of an ontology, no less and no more than of an API, are engaging in a multi-party coordination game in which costs should be kept as low as possible. Semantic versioning is one such cost control mechanism." [Clark, 2011]

It seems clear that in order to use a semantic versioning model to manage a similar level of complexity across the web itself requires that vocabulary managers and management systems pay better attention to the way they capture and describe change, focusing their effort at a very

granular level, not necessarily at the traditional "record" level so ingrained in current library authority control systems.

The RDA Registry, using GitHub capability for version-tagging published RDF vocabularies maintained in the OMR, provides semantic versioning based on the way software is versioned, and to support the same kinds of automated or guided updating by data publishers [Registry versioning]. The responsibility for determining the version number itself, documenting the changes made to a published vocabulary that necessitated a version increment at any semantic level, and deciding the level of version increment required, remains the responsibility of individual vocabulary administrators. The OMR, through its use of Git and GitHub as part of the publishing workflow, provides more than adequate support for effective semantic versioning.

Discussion

In the old centralized days, practice evolved to use computer power to make batch changes, with cleanup efforts organized to deal with outliers after the automated processes did what they could. Separate costs for these services were not generally available, and were normally not broken out from the overall costs for the regular provision of metadata, even though, in most cases, acquiring improved data after changes were made was neither easy nor cheap. Libraries with sufficient staff and expertise to help with manual cleanup efforts were not motivated to break out these costs for their institutional managers and in fact many of those efforts did not make the trip from local cataloging operations to the centralized cache, primarily because of financial disincentives to do so.

To bring general vocabulary management policies and practices forward from the traditional centralized file management model still used by many of the bibliographic institutions of the MARC era, requires much more discussion and effort. Many of the institutional services developed in the recent past to distribute traditional vocabularies (usually supported by printed book sales), built their web presences based on look-up functionality, while full files were reserved for paid subscribers. This business model is gradually fading, but appropriate management practices for vocabularies are still widely misunderstood and largely ignored.

The RDA Registry envisions a broader role for vocabulary management in the linked data environment based primarily on improved distribution practices designed to provide stability to still-evolving linked open data publishers and consumers. The bibliographic portion of the Semantic Web movement is determined to be open, eschewing the long-standing data 'ownership' struggles (not resolved, but no longer much discussed). Clearly, the development of open services to support the transition from traditional data distribution to one that does not assume centrally maintained instance records must assume open data as a basis.

RDA is a potent use case for demonstrating the difficulties of vocabulary management, maintenance and publishing in a changing world. We believe that the infrastructure of the OMR's RDF generator, publishing tools, Git, and Github make it possible to publish a sophisticated, versioned, and stable multilingual vocabulary. The difficulties of designing and publishing a

vocabulary, even without multilingual capabilities, versioning and multiple authors is significant, but with the right tools, not impossible.

References

[Clark, 2011] Clark, Kendall (2001). Semantic versioning and OWL ontologies. Clark & Parsia Blog. Retrieved from http://weblog.clarkparsia.com/2011/09/19/semantic-versioning-and-owl-ontologies/

[CoP announcement] RDA Committee of Principals. Committee of Principals affirms commitment to the internationalisation of RDA. Retrieved from http://www.rda-jsc.org/node/235

[Hillmann, Dunsire & Phipps 2014] Hillmann, Diane, Dunsire, Gordon and Phipps, Jon. Versioning vocabularies in a linked data world. Proceedings of the IFLA Satellite meeting in Paris, France, 2014. Retrieved from http://ifla2014-satdata.bnf.fr/pdf/iflalld2014_submission_Hillmann_Dunsire_Phipps.pdf

[JSC announcement] Joint Steering Committee for the Development of RDA. JSC takes pragmatic approach to period of change. Retrieved from http://www.rda-jsc.org/node/229

[London Meeting] Data Model Meeting: British Library, London 30 April - 1 May 2007. Retrieved
from: http://www.bl.uk/bibliographic/meeting.html

[Preston-Werner] Preston-Werner, Tom. Semantic Versioning 2.0.0. Retrieved from http://semver.org/

[RDA around the world] RDA around the world / edited by Marie-France Plassard and Gordon Dunsire. London: Routledge, 2015. ISBN: 978-1-138-898899

[RDA Registry] RDA Registry. Retrieved from http://www.rdaregistry.info/

[RDA Scope and Structure] RDA Scope and Structure. 1 July 2009. Retrieved from http://www.rda-jsc.org/docs/5rda-scoperev4.pdf

[Registry Versioning] RDARegistry/RDA-Vocabularies. Releases. https://github.com/RDARegistry/RDA-Vocabularies/releases

[SKOS] SKOS Simple Knowledge Organisation System home page. Retrieved from: http://www.w3.org/2004/02/skos/