Maps and Gaps: Strategies for Vocabulary Design and Development

Diane Hillmann
Metadata Management Associates, USA
metadata.maven@gmail.com

Gordon Dunsire
Independent Consultant, UK
Gordon@gordondunsire.com

Jon Phipps
Metadata Management Associates, USA
jonphipps@gmail.com

Abstract

In this paper we discuss changes in the vocabulary development landscape, their origins, and future implications, via analysis of several existing standards. We examine the role of semantics and mapping in future development, as well as some newer vocabulary building activities and their strategies.

Keywords: vocabularies; vocabulary design; semantics; mapping; simplicity; complexity.

1. Introduction

Vocabulary development has traditionally been a group activity, bringing together communities of practice to agree on common definitions and usages of descriptive elements and vocabularies. This made sense in a world where sharing of vocabularies was limited by the high costs of publishing, and later, by the seductive influence of efficiencies inherent in building silos of information based on common vocabularies, used primarily by similar applications and institutions.

The Internet, cause of much well-documented disruption, has also agitated traditional thinking about vocabularies, leading to possibilities in vocabulary development and sharing that are only beginning to be understood. The potential of social networks to participate in vocabulary development, via tagging services or data mining, shows much promise, and has been the basis for interesting experiments, but has not yet begun to drive the important shifts from standardized vocabularies resulting from formal gathering of community expertise to more distributed and informal building of vocabularies in a variety of venues and contexts.[1]

Continuing discussions about the pros and cons of simplicity and complexity, endemic to the linked data environment at least since the inception of Dublin Core in 1995, have long had the flavor of religious argument, lacking much analysis of experience with either extreme. Shifts in thinking about the value of re-using existing vocabularies vs. re-purposing them in locally managed namespaces, as well as the best practices for extension and mapping, have surfaced in some discussions, for example IFLA's ISBD and UNIMARC bibliographic standards (Dunsire & Willer, 2011), but have yet to be integrated into standard practice. In fact, as vocabulary mapping becomes mainstream, the value of having a plethora of vocabularies to include in maps becomes much clearer. This is illustrated with an example later in this paper.

In this paper we will explore the experience of several communities involved in vocabulary development over the past twenty years, pointing out how shifts in technology and the application of vocabularies in a group of associated communities have challenged long held ideas about processes, expectations, and outcomes.

A note about terminology: As we shall see, casual use of labels believed to be commonly understood can lead to problems. We use the terms "element set" and "value vocabulary" in the

[1] Links to discussion pieces [on social tagging]: http://dublincore.org/moinmoin-wiki-archive/taggingwiki/pages/DiscussionIssues.html

context of RDF vocabularies as defined by the Library Linked Data Incubator Group (Issac et al., 2011). 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.

1.1. Top-down vs. bottom-up

We can characterize the basic design strategy behind the creation and development of a particular vocabulary as being top-down or bottom-up. The top-down approach expects the vocabulary to include relevant refinements, so vocabulary elements—classes, properties, concepts—tend to have wide scope, coarse-grained semantics and be few in number; this is exemplified by the original DC elements.[2] A bottom-up vocabulary has narrower focus with finer granularity, and often many elements. The context of a bottom-up vocabulary is usually localized to a fairly narrowly defined and often specialized knowledge domain, and in an open-world-facing linked data environment there is an expectation that its elements will need to be "dumbed down" to become a useful part of the Semantic Web. For example, this was accommodated in the early design of the RDA element sets (Hillmann et al., 2010) by the development of a parallel set of properties unconstrained by domains and ranges of Functional Requirements for Bibliographic Records (FRBR).[3] The idea of "intelligent dumb-down" in the 2003 draft of the DCMI Abstract Model, transmuted into the current Dublin Core Metadata Initiative (DCMI) level 2 of interoperability, uses sub-property relationships which require the linked element to have broader semantics than the local, source element.[4] The distinctive characteristics are summarized in Table 1.

TABLE 1: Basic characteristics of vocabulary design strategies

Characteristic	Top-down	Bottom-up
Focus	*Wide*	*Narrow*
Granularity	*Coarse*	*Fine*
Interoperability	*Refine*	*Dumb-down*
Usage	*Global*	*Local*
Size	*Small*	*Large*

These characteristics are relative, and it is natural that vocabularies which refine other top-down vocabularies exhibit a mix of top-down and bottom-up features. RDA has wide focus, global usage, and may be refined or dumbed-down for interoperability, but it has large size and relatively fine granularity because it is an application of FRBR.

Top-down vocabulary development is often based on a conviction that the development of the vocabulary is driven by an assessment of user needs. This conviction tends to hold even when, unlike FRBR, there has not been a formal (or informal) assessment of user needs with which to drive decisions. While FRBR focuses on user tasks as part of its core methodology, the Bibliographic Framework Transition Initiative (BIBFRAME) is looking for user requirements in the legacy data of MARC 21 records, and the schema.org initiative seeks the wisdom of crowds for developing refinements. Given that the future role of traditional catalogs (not to mention libraries) continues to be unclear, assumption of user needs based on historical precedent seems a poor substitute for a real examination, though the current level of uncertainty about the future is clearly an impediment.

[2] http://dublincore.org/documents/dces/
[3] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[4] http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/2003-12-18/

1.2. Generality and Specificity

A defining belief in those developing new schema vocabularies over the past twenty years is that it is better to start simple, with a common use case, and from that beginning to add more detail to address more complex use cases. MARC 21 began that way, working from a simple notion of book cataloging to become, for half a century, the behemoth of library metadata. The MARC development trajectory, though it went from books, to serials, then maps, multi-media, digital materials and other special materials, eventually was pulled back together, starting in the mid-1980's, in a painful and expensive multi-year process called Format Integration (Glennan, 1995). In the era of the MARC silo, this upheaval was probably necessary, but in an era of constant change, it provides a potent lesson for current vocabulary managers.

In the bibliographic community, this consolidation effort was paralleled (and influenced) by the development of International Standard Bibliographic Descriptions (ISBDs) from 1971 to 1992, when the problems of resource categorization caused by digital carrier formats were forcing a re-examination of the principles and purposes of cataloging within the worldwide professional community. This work changed the focus of development from specialization to generalization, resulting in the 2011 publication of a single consolidated ISBD with an element set based on common attributes of information resources and a basic value vocabulary for distinguishing content and media categories. "The consolidated ISBD is intended to serve as a standard for description of all types of published materials up to the present date, and to make it easier to describe resources that share characteristics of more than one format. In addition, it facilitates the work of keeping the ISBD updated and consistent for the future".[5]

2. Defining Experiences: Dublin Core

But the best example of the dangers of starting simple is Dublin Core (DC), which debuted in 1995 with 13 elements (which rapidly expanded to 15) and then attempted to build upon that small base two years later after much contention about the wisdom of doing so. The tension between the camps for and against expansion was noted in the report for DC-4 in Canberra (Weibel, Renato & Cathro, 1997):

> *The Minimalist point of view reflects a strong commitment to the notion that DC's primary motivating characteristic is its simplicity. This simplicity is important both for creation of metadata (for example, by authors unschooled in the cataloging arts) and for the use of metadata by tools (for example, indexing harvesters, which will probably not make use of detailed qualifiers or encoding schemes). The goal of semantic interoperability across communities can only be achieved if there is a simple core of elements that are understood to mean the same thing in every case. Additional qualifiers support specifying, modifying, and particularizing the meaning of an element. Since this will probably be done in different ways by different groups at different times, it will potentially lead to semantic drift in the elements, and consequent loss of semantic interoperability.*
>
> *The Structuralists as a group accept the danger of this semantic drift in exchange for the greater flexibility of a formal means of extending or qualifying elements such that they can be made more useful for the needs of a particular community*.

Simplicity in DC is expressed in the plain language of labels and other annotations, the small number of elements, a low-barrier access and maintenance infrastructure, and one-size-fits-all core appeal, as well as the coarse grain of its semantics; all the hallmarks of a top-down strategy. The strong market penetration of DC in its early years marked the success of this strategy. There was nothing else like it in terms of ease of implementation for casual authors of digital documents and it remains far, far better than nothing.

[5] http://www.ifla.org/publications/international-standard-bibliographic-description

This first pass at more specificity, built and ratified in Canberra, approved a heterogeneous group of "qualifiers" proposed by a variety of specialist groups. But continuing this model of extension as a large group process proved to be contentious and time consuming, so the DC community chose a model not that different from MARC's Advisory Committee (known as MARBI) but called the Usage Board (DCUB). The UB was intended to respond to the many complaints that the 15 elements were insufficient, not particularly useful as a simple set, and as designed, unusable for any but the sparest applications. As the first task of the UB, a process for communities working under the DC umbrella to submit proposals for additional qualifiers was developed, and the group began to evaluate those proposals.

Problems with this strategy began surfacing almost immediately. The UB was beginning the process of reviewing proposals at the same time as they were developing their process and criteria for approval. After several rounds of proposals, which saw additions for educational materials and collections (prepared by two of the most active DC communities), the process was essentially stalled, and DCMI has determined that it will not expand the vocabularies further.

3. Defining Experiences: RDA

One top-down effort that did not start simple was the Resource Description and Access (RDA) standard for bibliographic description, originally called AACR3, referencing its direct antecedent, the Anglo-American Cataloging Rules, second edition (AACR2). The Joint Steering Committee for the Development of RDA (JSC) explicitly discarded the structure of AACR2, which had started as rules for book cataloging, to which were added special rules as multimedia and digital materials started flooding into libraries—essentially in parallel with the development continuum of MARC and ISBD standards. FRBR, developed in the final years of the twentieth century, became the underlying model instead. Far from being a simple start, FRBR is focused on the requirements of catalog users and is intended to accommodate the complexity of a multi-format, multi-version world, where relationships both abstract and concrete can be described and specified.

But even the RDA developers succumbed to the siren song of simplicity, and identified a number of elements as "core", without a clear plan for how that classification would work in practice. Because some of the elements defined as core are used only for specific materials ("Key" is defined as core, for instance, but intended for music), the core elements are not particularly useful for simple record validation.

An example of the problems inherent in traditional top-down development occurred in the first years of development of RDA. The creation of RDA from AACR2 necessarily used the existing management infrastructure to provide continuity and resources, and focus was directed to retaining the utility of the standard and its vocabularies in the new environment rather than engaging the community to be found already there. After all, the New World's inhabitants speak unknown languages and have unfamiliar customs and ways of doing things. Criticism from the community over the process and results was harsh, and the doors gradually opened to let in opinions and review processes involving constituent groups. This turmoil led to the unlikely participation by another community entirely, DCMI, resulting in the historic agreement of April 2007 and the subsequent development of the RDA vocabularies in parallel with the completion of the guidance text (British Library, 2007). At each stage, shifts driven by internal and external participation and review arguably improved the final product, but with the penalty of drawn-out timetables.

Although at first blush the RDA effort has also been classic top-down (although at least based on solid past experience), the stated intention of appealing to Semantic Web communities and supporting extension for other specialist communities opened the possibility of a more distributed effort. As already noted, the design of the RDA vocabularies became much more bottom-up. RDA and ISBD are collaborating on mappings that necessarily require coarser-grained properties with common semantics (Willer, 2012). Each of these collaborating communities is asking itself

the same questions about who is responsible for what, what it will cost, and whether formal, expensive protocols are really necessary for improving interoperability. This, of course, is a classic bottom-up scenario - so where does the tipping point occur? Will an approach that mixes bottom-up and top-down work? The design strategy of schema.org vocabularies is to crowd-source development so that "extensions that gain significant adoption on the web may be moved into the core".[6] This is similar to the fate of qualifiers to the original DC elements, and can easily fall prey to political and economic forces; will it work for search engines better than it works for human users?

4. Why maps?

The acknowledged growth of new bibliographic schemas over the past few years has been called "chaos", "anarchy" or, less pejoratively, "proliferation". This point of view is understandable, but not very useful. If nothing else, the continuing proliferation confirms that what exists is not meeting all the needs to be found out in the world. Perhaps a better direction is to assert that more metadata vocabularies makes for a richer metadata environment, able to meet a broader array of needs. One analogy that seems to apply is that the traditional notion of top-down "we know what you need" approaches provide only limited choices, which may be insufficient outside existing silos. A more chaotic metadata environment provides lots of choices, but those choices may be difficult to navigate for many practitioners.

Once we get beyond the idea that semantic mapping is just a substitute for crosswalking, we can see that mapping can be used to address the vocabulary chaos as well as to bridge existing and emergent gaps, thus providing a path to the semantic web we've all been anticipating (Dunsire et al., 2011). Crosswalking is too often a reductive process that dilutes the power of semantics by substitution—it is top-down, deductive, and analytic. Mapping can more usefully be seen as an additive process, in which the map adds otherwise missing semantics through semantic relationships; it is bottom-up, inductive, and synthetic.

There are a variety of ways that mapping can become a standard step in vocabulary development workflows. An initial map provided by a vocabulary owner is a starting point for many possible maps, fulfilling many possible needs for a host of users. A map created by a vocabulary owner also provides a basis for intelligently exporting data in a variety of formats as a matter of routine. A map created further down the line, to fulfill other needs not anticipated by the owner, can potentially be endorsed by the owner, adopted by others, or, of course, depending on the quality of the map, find no other users. If multiple vocabularies can be envisioned as part of a richer metadata environment, so too can the maps connecting them. For maps to reach the services that can consume them, they will need to be managed in the same way vocabularies should be: open, discoverable, and reusable with provenance and versioning. Discovery, too, has its top-down and bottom-up aspects. The Open Metadata Registry (OMR) enables searching of concepts and elements at a granular level to discover in what vocabularies they appear, but other discovery options like the European Commission's Joinup project[7] and the Linked Open Vocabularies project (LOV)[8] focus more on upper level vocabulary descriptions and provide discovery services primarily at that "semantic asset" level.

4.1. Using maps to identify gaps

The presence of an increasing number of element sets intended to describe the same corpora—often at different levels of specificity—has been seen as a problem, so long as the defining question has been "What format should I choose" for particular purposes or projects. But in a web-based world where connections and relationships provide the structure on which we will

[6] http://schema.org/docs/extension.html
[7] https://joinup.ec.europa.eu/
[8] http://lov.okfn.org/dataset/lov/

increasingly depend, semantic mapping creates new opportunities for interoperability without pain, no matter what decisions have been made in the past.

Stepping back from the focus on the individual element set, the ability to create maps between existing sets to identify gaps in the whole array provides a possible way ahead for all the communities operating in this bibliographic space. In Figure 1 below, the results of combining individual properties illustrate that bigger picture.

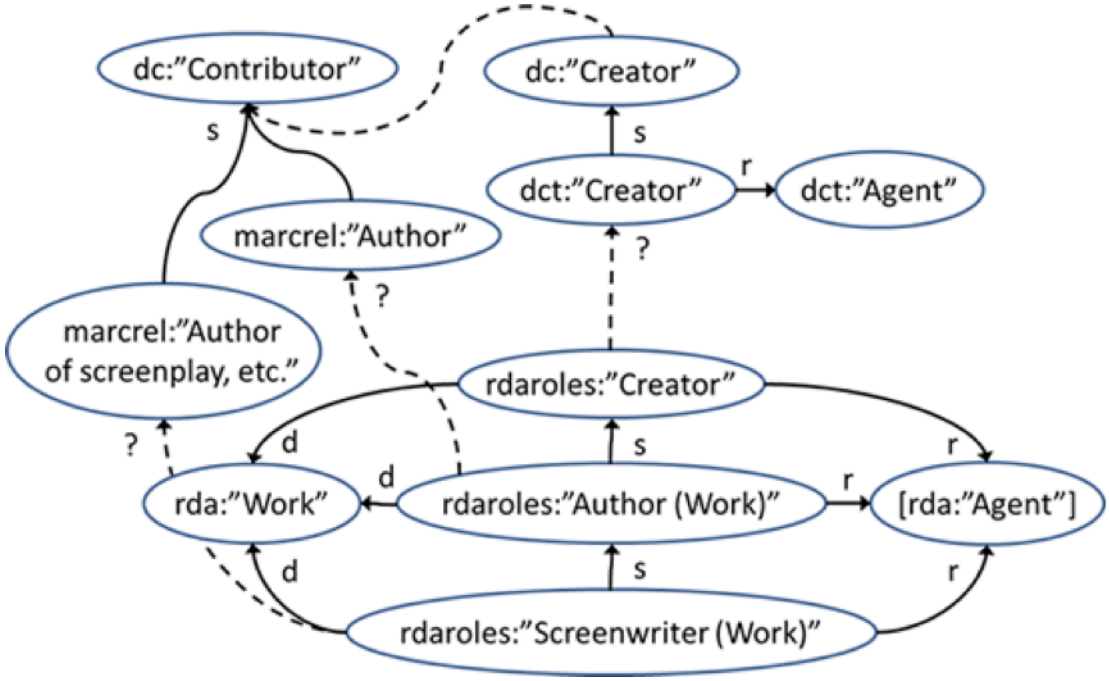

FIG. 1: RDF graph of a map between RDA, MARC 21, and DC elements for the role of screenwriter. Key: d: rdfs:domain; r: rdfs:range; s: rdfs:subPropertyOf.

Figure 1 gives an example of using bottom-up maps to identify gaps in top-down ontologies. This is a possible RDF graph of mappings between properties representing the role or relator of screenwriter in three vocabularies: RDA: resource description and access[9], MARC 21[10], and Dublin Core[11]. The graph contains three sub-maps: the lower map between RDA properties is declared in the RDA element set; the upper map between DC properties is declared in the DC Terms element set; and the upper-left map between MARC and DC properties is declared in the MARC Relator Codes element set. Note also that the RDA class Agent is in the process of being assigned as the range of the RDA properties. The mappings indicated by dashed links are potential links between individual elements to create the super-map. In this map, the suggested linking property is rdfs:subPropertyOf, and the direction of each relationship ensures semantic coherency when the map is used to generate entailments from data triples. In general, the object property is always less constrained by domain and range than the subject property in each related pair. Thus the RDA property for the specific role of screenwriter, constrained by a domain and a range, must be a sub-property of the MARC property for the equivalent role of screenwriter or author of screenplay, while the RDA property for author must be a sub-property of the equivalent MARC property.

[9] http://rdvocab.info/roles
[10] http://id.loc.gov/vocabulary/relators/
[11] http://dublincore.org/documents/dcmi-terms/

The MARC Relator Codes were mapped as sub-properties of the DC contributor property in 2005.[12] If this relationship is correct, the implication is that the DC creator property must be a sub-property of the DC contributor property if the super-map is to remain coherent. This relationship, however, is only declared explicitly in the DC terms element set. To improve interoperability between MARC and RDA data triples, the MARC sub-map requires extension to the DC 1.1 creator property, rather than, say, the more efficient replacement of the DC 1.1. contributor property with the creator property. This represents a known gap in the Dublin Core ontology, which effectively requires external maps to use the DC terms rather than the 1.1 element set. The issue is more complicated than first appears. The definition of the creator property in RDA is circular, and conveys no intrinsic meaning. The creator property in DC equates creating with making, and is easily misinterpreted as meaning "manufacturer"; in RDA, creator is a role associated with a FRBR Work, and a manufacturer with a FRBR Manifestation, which in turn is related semantically to publisher in ISBD and RDA. Publisher is a separate and semantically isolated property in DC. Both "creator" and "contributor" are terms engendering significant discussion in professional cataloguing circles, much of it centered on the meaning and context of primary responsibility, which is explicit in the DC definition. Is it responsibility for the primary part of the resource, or is it primary responsibility for the resource as a whole? This example does not seem to meet the DC goal of "a simple core of elements that are understood to mean the same thing in every case", and illustrates the downside of simplicity and a weakness of the top-down strategy.

5. Recent shifts of sand

The MARC 21 standard, long the gold standard for rich library metadata world-wide, has been metaphorically shoved aside, perhaps prematurely (and certainly without useful discussion of the value of its semantics). Its intended 'replacement', BIBFRAME, is still in its early stages. BIBFRAME started out in fall of 2011 as the Bibliographic Framework Transition Initiative, a major Library of Congress project, and has recently set up a separate website—http://BibFrame.org—for information on the initiative, accompanied by a discussion list.[13]

The proposed BIBFRAME semantic model occupies a gap between DC and RDA, so the initiative is to be welcomed. But rather than putting all bibliographic eggs in a BIBFRAME basket, it is probably better to consider it, in a mapping context, as just another player, another map to fit into communal mapping targets. Interoperability requirements are not the same thing as usage requirements; they are based on different functional contexts. The ontology you use to maintain your metadata never has to be the ontology you use to publish your metadata, and the ontology used to describe the metadata you harvest never has to be the ontology you use to manage your metadata.

This suggests that it would be useful if all managers of schemas and other standards were to develop element sets and value vocabulary representations that match the source semantics at the finest granularity and make them available along with maps of the internal ontologies. There is as yet no official version of MARC 21 suitable for linked data, but there is an unofficial one at http://marc21rdf.info created to help research into vocabulary issues [Full disclosure: the authors of this paper are behind that effort]. The Library of Congress is still promising an official version, but this has not yet happened; without it, BIBFRAME can only plug gaps in the coarse-grained environment of simple bibliographic description.

6. Conclusion

Although the re-use of existing vocabularies has long been a stated goal of the new resurgence of vocabulary development over the last 10 years, public opinion has shifted to advocating the re-

[12] http://lcweb2.loc.gov/diglib/loc.terms/relators/dc-contributor.html
[13] http://www.loc.gov/marc/transition/

purposing of existing vocabularies, instead of relying on the good will and good practices of the original vocabulary owner. There are a number of strategies extant for this re-purposing, most are reliant on the maturity of the mapping solutions to come. Certainly strategies that provide for positive and open pathways for extension and communication within and among communities of practice are much more likely to survive.

There are new tools which provide technical solutions to this conundrum: GitHub, and similar services providing open source code sharing, allow users to use code directly or to fork and/or change it as well. Users who fork or change can recommend to the vocabulary owner that they pull in the changes or additions, but either way, the original and the changes are visible to others, and both carry information on versions and change history. Github works on a somewhat top-down model, given that most changing of extant vocabularies is likely to go in a more specialized direction, but it allows authorized groups to interact easily and transparently, with supportive structures and tools.

In the absence of general operational applications of linked data in bibliographic communities, the main benefits of developing vocabularies one-link-beyond the formal scope of standard schemas and terminologies lies in representing alignments between different standards for future data interoperability. For example, the alignment of RDA with ISBD is supported by the set of unconstrained RDA elements maintained by JSC.[14] But "it will be necessary to develop unconstrained ISBD elements and map RDA elements to them" because of overlaps in definitions and granularity. The maintainers of ISBD must therefore continue to work "for the future".

Finally, of course, there is no top or bottom, core or periphery, or other spatial coordinate system behind the Web or an RDF map. We have shown that design strategies eventually succumb to the real world. But we conclude by favoring the bottom-up approach, not least because it preserves the local, without loss of data or other unhappy compromises. When embarking on a journey it's usually better to start from home.

References

British Library (2007, May/June) Data Model Meeting, British Library, London, 30 April-1 May 2007. Retrieved July 1, 2013 from http://www.bl.uk/bibliographic/meeting.html.

Dunsire, Gordon, Diane Ileana Hillmann, Jon Phipps & Karen Coyle. (2011, Sept.). A Reconsideration of Mapping in a Semantic World,. Proceedings of the International Conference on Dublin Core and Metadata Applications. Retreived July 1, 2013 from http://dcpapers.dublincore.org/pubs/article/view/3622.

Dunsire, Gordon & Mirna Willer. (2011, Dec.). UNIMARC and Linked Data. IFLA Journal, 37 (4). December 2011. Retrieved July 1, 2013 from http://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-37-4_2011.pdf.

Glennan, Kathryn P. (1995, Fall). The Final Phase. MC Journal: The Journal of Academic Media Librarianship, 3(2), 1-31. Retrieved July 1, 2013 from http://wings.buffalo.edu/publications/mcjrnl/v3n2/glennan.html.

Hillmann, Diane, Karen Coyle, Jon Phipps & Gordon Dunsire. (2010, Jan./Feb.). RDA Vocabularies: Process, Outcome, Use. D-Lib Magazine, 16(1/2). Retreived July 1, 2013 from http://dlib.org/dlib/january10/hillmann/01hillmann.html.

Isaac, Antoine, William Waites, Jeff Young, & Marcia Zeng. (2011, Oct. 25). Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report. Retrieved July 1, 2013 from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset/.

Weibel, Stuart, Renato Iannella & Warwick Cathro. (June, 1997). The 4th Dublin Core Metadata Workshop Report. D-Lib Magazine, 3(6). Retrieved July 1, 2013 from http://www.dlib.org/dlib/june97/metadata/06weibel.html.

Willer, Mirna. (2012). Alignment of the ISBD element set with RDA element set – RDA, Appendix D.1. Retrieved July 1, 2013 from http://www.rda-jsc.org/docs/6JSC-ISBD-Discussion-1.pdf.

[14] http://www.rda-jsc.org/working2.html#community-isbd-discussion-1