

Issues in Vocabulary Management

*A Technical Report of the
National Information Standards Organization*

Approved September 25, 2017

About NISO Technical Reports

NISO Technical Reports describe the work and conclusions of a working group chartered by the organization to study a particular problem or issue that is related to NISO's standardization interests. However, unlike NISO standards or recommended practices, they do not have requirements for compliance or recommend "best practices" – although their conclusions may contain some recommendations. Additionally, no consensus process of the NISO voting membership was used in their creation or approval. They are strictly informative in nature.

Published by

National Information Standards Organization (NISO)
3600 Clipper Mill Road
Suite 302
Baltimore, MD 21211
www.niso.org

Copyright © 2017 by the National Information Standards Organization

All rights reserved under International and Pan-American Copyright Conventions. For noncommercial purposes only, this publication may be reproduced or transmitted in any form or by any means without prior permission in writing from the publisher, provided it is reproduced accurately, the source of the material is identified, and the NISO copyright status is acknowledged. All inquiries regarding translations into other languages or commercial reproduction or distribution should be addressed to: NISO, 3600 Clipper Mill Road, Suite 302, Baltimore, MD 21211.

ISBN: 978-1-937522-79-7

Contents

Section 1: Introduction: The Problem with Vocabularies	1
Section 2: Discovery and Evaluation for Use or Reuse	5
2.1 Communication Between Owners and Users: The Value Proposition	6
2.2 Evolution of Business Models.....	7
2.3 Use Case: Vocabularies in Transition	8
2.3.1 FAST	8
2.3.2 The Getty Vocabularies	9
2.3.3 From English-only to Multilingual (RDA).....	10
2.4 The Technical Environment.....	11
2.5 The Policy Environment and its Implications	14
2.6 Licenses	15
2.7 Application Profiles	16
Section 3: Maintenance: Status and Versioning	18
Section 4: Using, Reusing, Extending, Mapping	21
4.1 Extension.....	22
4.2 Mapping.....	23
Section 5: Vocabulary Preservation	25
Section 6: Recommendations	27
Appendix A: Terms and Definitions	28
Appendix B: Vocabularies and Tools	36
Appendix C: Additional Resources	43
Bibliography	46

Foreword

About this Recommended Practice

As interest in the new environment for sharing bibliographic information grows, questions about appropriate policies and the supporting infrastructure come increasingly into focus. NISO began exploring these questions in 2013 as part of its NISO Bibliographic Roadmap Development Project, which examined requirements for usability and adoption of advanced bibliographic exchange in a global networked environment. As one of the outcomes from this initiative, late in 2014 NISO's voting members approved a project to address issues described in the work item Development of Standards to Support Bibliographic Data Exchange.

As described in the work item, three working groups were established as well as a steering committee, consisting of the working group co-chairs and a liaison to the NISO Content and Collections Management Topic Committee, to coordinate and oversee the groups' work. The Use/Reuse working group looked at policy and social considerations, including appropriate licenses and permissions, maintenance expectations, and versioning. The Documentation working group examined standards for documentation of vocabulary properties, particularly as they relate to discovery and usage, as well as governance and sustainability issues. The Preservation working group examined the landscape issue of "orphan vocabularies," where organizations abandon vocabularies for lack of funding or when the vocabularies cannot make the transition between print and digital.

The intent of this Technical Report is to provide a background on vocabulary management for those operating in this transitional environment, where experience with policies, social constructs, and the practical aspects of moving forward using a common infrastructure tends to be scattered or missing. Other agencies—the World Wide Web Consortium (W3C) in particular—have more specialized expertise in technical areas, and though there are overlaps with NISO's effort, the W3C is better positioned to build and maintain the technical standards in this space.

The audiences for this technical report start with the communities NISO has brought together: libraries, publishers, and service providers. But beyond these communities, we hope this document may help the many individuals and groups building and sharing bibliographic and other descriptive data, as well as knowledge managers within a variety of organizations using vocabularies to solve problems. And as governments at all levels move their services to the web, joining non-profits and profit-making enterprises, new needs arise for appropriate vocabulary development and maintenance policies as well as practical information on building quality and sustainability strategies. It is for these groups that this document is intended.

To avoid overwhelming readers—who may be new to this area of work—with technical detail, we have used the bibliography to link to resources discussed in the text, and moved technical documentation to a separate section (Section 2.4, "The Technical Environment"), providing annotations to assist those wishing to delve further into those areas. There will be areas where we will quote from the W3C documents most relevant to vocabulary management, but we have avoided going into technical details beyond the basics.

NISO Content and Collection Development Topic Committee

This recommended practice is part of the portfolio of the NISO Content and Collection Development Topic Committee. At the time the Topic Committee approved this recommended practice for publication, the following individuals were committee members:

Eric Childress

OCLC Online Computer Library Center

Amy Kirchhoff

ITHAKA/JSTOR/Portico

Jody DeRidder (co-chair)

University of Alabama / University Libraries

Betty Landesman

Independent

Marti Heyman (co-chair)

OCLC Online Computer Library Center

Peter McCracken

Cornell University Library

NISO Issues in Vocabulary Management Working Group Members

Three working groups are responsible for the development of this technical report.

The following individuals served on the Documentation group.

Alexis Adkins

Cal Poly Pomona University Library

Teressa Keenan

University of Montana

Natalie Bulick (Co-chair)

Indiana State University

Michael Kim

University of California, Santa Barbara

Gordon Dunsire

Independent Consultant

Margaret Kruesi

Library of Congress

Sharon Garewal

ITHAKA/JSTOR/Portico

Dana Miller

University of Nevada at Reno

Antoine Isaac

Europeana

Susan Presley

ProQuest

Sean Glover (Co-chair)

YBP Library Services

Lars Svensson

Deutsche Nationalbibliothek

The following individuals served on the Preservation group.

Sherle Abramson-Bluhm (Chair)
University of Michigan Library

Scott Carlson
Rice University Fondren Library

Mike Doane
Independent Consultant

Christine Connors
ServiceNow

Sharon Farnel
University of Alberta

Marti Heyman
OCLC Online Computer Library
Center

Amy Kirchhoff
ITHAKA/JSTOR/Portico

Maureen McClarnon
Cengage Learning

Kim Mumbower
The Library Corporation (TLC)

Anne Washington
University of Houston

The following individuals served on the Use/Reuse group.

Joseph Busch
Taxonomy Strategies

Valentine Charles
Europeana

Dave Clarke
Synaptica

Heather Hedden
Cengage Learning

Diane Hillmann (Co-chair)
Metadata Management Associates

Daniel Lovins (Co-chair)
New York University (NYU), Division of
Libraries

Nicole Vasilevsky
Oregon Health & Science University
(OHSU)

Elizabeth Wolf
Copyright Clearance Center (CCC)

Acknowledgements

The Working Group would like to acknowledge the contributions of Alistair Morrison, formerly of Elsevier and now of Johns Hopkins University, who participated on the Documentation working group through 2016.

Trademarks, Services Marks

Wherever used in this standard, all terms that are trademarks or service marks are and remain the property of their respective owners.

Section 1: Introduction: The Problem with Vocabularies

The issues that bedevil those with interests in vocabulary development, usage, and maintenance are not new, and the recent upswing of interest in linked data has fed new interest in those issues. Linked data is a method of publishing structured data built upon standard Web technologies such as HTTP, the Resource Description Framework (RDF), and Uniform Resource Identifiers (URIs). Linked data allows computer systems to share information on the Semantic Web, thus enabling data from different sources to be connected and queried. In 2011, the Dublin Core Metadata Initiative (DCMI) organized a day-long Vocabulary Special Session at its conference in The Hague, The Netherlands. The attendees were primarily practitioners struggling to find solid footing around the issues under discussion. The report of that session¹ remains an excellent source of information on vocabulary issues in information communities today, and was the basis for developing the work items for NISO's Vocabulary Development Projects.

The discussions begun in The Hague continue to reverberate:

*"The wide ranging conversations at the DCMI special session in The Hague remind us that interoperability and the efficiencies of common approaches require guiding principles and best practices around decisions for reuse, extension of existing vocabularies, as well as development of new vocabularies. Without cooperative efforts to develop those supportive pieces, good decisions are difficult to make, much less implement."*²

There are a number of ways to look at vocabularies and their functions, depending on the community in question and how the terms of use are defined. A vocabulary defines terms. A taxonomy classifies terms, usually (but not always) in a hierarchical organization. An ontology defines terms, either asserts or infers a hierarchical organization, and defines the relationships between the terms. This view is somewhat complicated as many RDFS/OWL ontologies don't require a hierarchy for classes and properties: they can be expressed as a flat list. That isn't to say that relationships between terms aren't important; particularly when mapping is considered, relationships significantly clarify the context of terms, an aspect essential for useful mapping.

The figure below describes a somewhat simplified view of the differences between a taxonomy and an ontology:

¹ Dublin Core Metadata Initiative. "DC-2011 Vocabulary Special Session/Meeting Report" (2012), http://wiki.dublincore.org/index.php/DC-2011_Vocabulary_Special_Session/Meeting_Report

² Dunsire, Gordon, Diane Hillmann, and Jon Phipps, "Reconsidering Universal Bibliographic Control in Light of the Semantic Web", *Journal of Library Metadata*, 12, no. 2-3 (2012): 164-176. <http://hdl.handle.net/1813/36288>

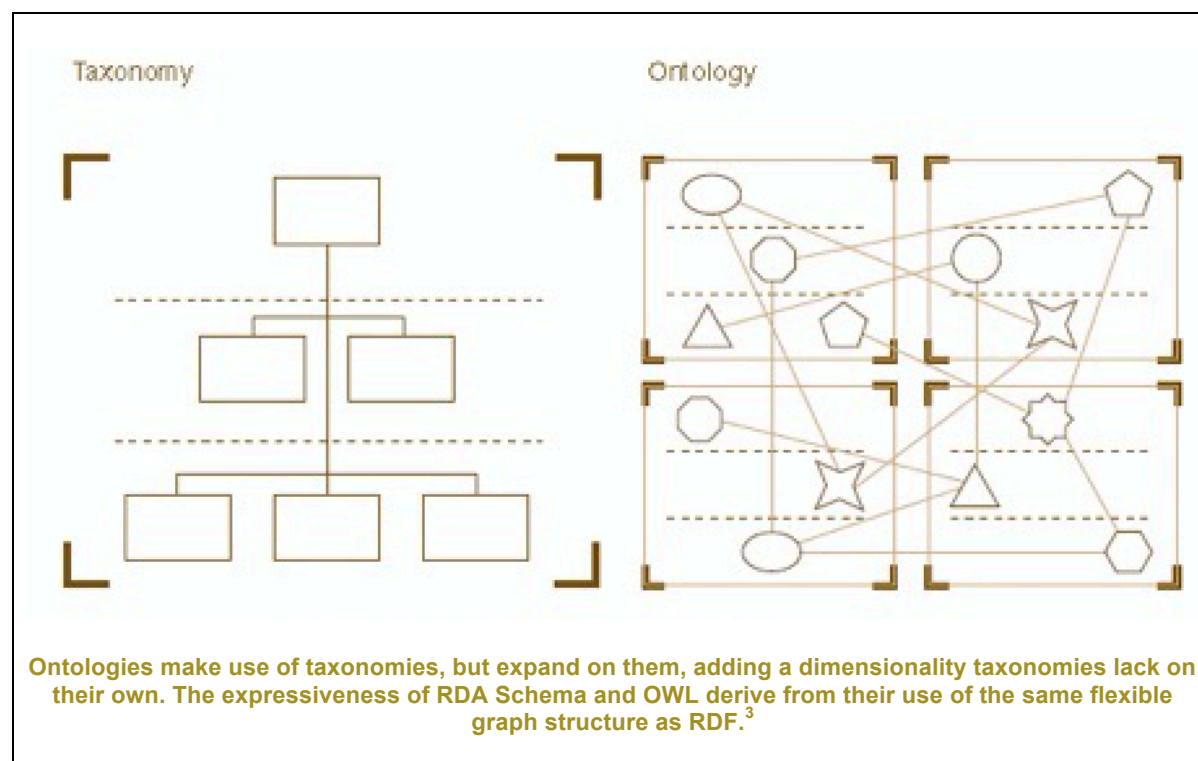


Figure 1: Taxonomies versus ontologies.

The kinds of vocabularies of interest to the identified audiences fall primarily into two basic categories: 1) element sets that support resource description in general and for particular communities, and 2) value vocabularies that provide descriptive terms available for use within those sets by creators or users of descriptive data. The element sets are sometimes called structured vocabularies, property or attribute sets, formats, metadata schema, or ontologies, but they are all intended to provide similar functionality. Most modern element vocabularies are expressed using the Resource Description Framework Schema (RDFS)⁴ or the Web Ontology Language (OWL).⁵ Value vocabularies, sometimes called concept vocabularies or controlled vocabularies, provide the terms (and URIs) that allow meaning to be understood across applications and communities. They are usually expressed using the thesaural standards Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*)⁶ or the Simple Knowledge Organization System (SKOS),⁷ which includes some of the simple thesaural relationships from ANSI/NISO Z39.19-2005, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* but extends them significantly. Common examples of structured vocabularies are MARC 21; Dublin Core

³ From PricewaterhouseCoopers. "Spinning a Data Web." *Technology Forecast*, Spring 2009, p. 4 15.

<https://www.pwc.com/cl/es/publicaciones/assets/pronostico-de-tecnologia.pdf>

⁴ W3C. "RDF Schema 1.1, W3C Recommendation." (February 25, 2014). <https://www.w3.org/TR/rdf-schema/>

⁵ Web Ontology Language (OWL). (December 11, 2012). <https://www.w3.org/OWL/>

⁶ National Information Standards Organization. "ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*" last modified January 6, 2014. http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

⁷ W3C. "SKOS Simple Knowledge Organization System." (December 13, 2012). <https://www.w3.org/2004/02/skos/>

Terms; RDA: Resource Description and Access; and BIBFRAME. Value vocabularies are often associated with a particular structure, for instance, the Library of Congress Subject Headings (LCSH) is often used to provide subject access in MARC 21 applications, the DCMIType vocabulary (DCMIType) was designed for use with Dublin Core elements, and the Getty Thesaurus of Geographic Names (TGN) was designed to be used with the Getty's cultural resources. Most value vocabularies intended for use in linked data are sufficiently generalizable to be used outside their original intended structure.

Vocabularies, at least in the traditional library and information science context, have generally been tied to a physical or virtual collection—e.g., a library, and often a specific library with a narrow subject focus. Thesauri, or concept vocabularies, are often called subject headings in the library context, and generally follow the standards for thesaurus construction using broader (BT), narrower (NT), and related (RT) relationships as defined in ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* and SKOS [Simple Knowledge Organization System]. *LCSH*, often cited as the quintessential example of subject headings, were initially developed at the tail end of the 19th century before the advent of thesaural standards, but many BT, NT, and RT relationships have been added since.⁸

Concept vocabularies, e.g., *Thesaurus of ERIC Descriptors (ERIC)*,⁹ have sometimes been published independent of a specific collection, but these vocabularies are often used to support searching of online indexes rather than in library catalogs or other discovery mechanisms.

A key problem in this area is “orphan vocabularies,” meaning vocabularies whose maintenance organizations have lost funding or cannot make the transition from print to digital platforms and there is no longer an authoritative source of terminology, policy, documentation, or licensing information. Guidance is needed for preventing orphan vocabularies and for deciding whether (and under what circumstances) the maintenance of an orphan vocabulary may be “adopted” by other organizations.

In reality, the goal of discussing all of these activities is to highlight the importance of stability in the vocabulary environment, particularly regarding the need for interoperability as descriptive information moves into the Linked Open Data environment (‘Open’ added to the Linked Data name reflects the understanding that closed data, proprietary data, and data that are not identified globally will not be part of this environment). The activities defined in this report are inextricably tied together by that overall goal of stability.

⁸ Chan, Lois Mai. “Still Robust at 100: A Century of LC Subject Headings.” *Library of Congress Information Bulletin* (August 1998), <https://www.loc.gov/loc/lcib/9808/lcsh-100.html>

⁹ Institute of Education Resources. <http://www.eric.ed.gov>

Table 1: The relationships between the activities involved in the vocabulary environment are illustrated below.

Activity	Requires	Enables
Use/Reuse	Adequate public documentation	Preservation Stability
Documentation	Responsible long-term governance	Use/reuse Preservation, Stability
Preservation	Adequate public documentation Responsible long-term governance	Use/reuse Stability

Section 2: Discovery and Evaluation for Use or Reuse

The DC-2011 Vocabulary Special Session report (ibid.) discusses the problems of discovery and evaluation as intrinsically related concerns. The report mentions a research tool called Swoogle,¹⁰ developed as a grant-funded project between 2004 and 2007 and intended to be “... a search engine for the Semantic Web.” During that period, Swoogle provided search services across Semantic Web ontologies, instance data, and class and property terms, but although the site still exists it is no longer a ‘live’ service. Since the demise of Swoogle there has been no specific service providing search across all kinds of vocabularies, although certainly the need exists. Other related services include Datahub.io, a service of the Open Knowledge Foundation for publishing and discovering data collections, and Purl.org, a service for establishing and managing URLs that was originally developed by OCLC and lately transferred to the management of Archive.org, known for its “Wayback Machine.”

Despite the dearth of comprehensive discovery services for vocabularies, evaluation criteria for reuse of existing vocabularies are well covered in the Vocabulary Special Session report and continue to be relevant. Issues of openness, richness, and maintenance are specified, as are related services, owner trustworthiness, and persistence policies. None of these criteria are easily discernable without solid documentation about the vocabularies as a whole.

Determining which available vocabularies to use in data assumes a number of precursor steps, including identifying relevant vocabularies already in use by the target communities, assessing those vocabularies within the context of particular needs, and determining whether missing information that is essential to that assessment can be found. Important information could include the identities of the parties responsible for developing and managing the vocabulary, relevant licenses, available maintenance policies, and whether the vocabulary is open to suggestions by groups interested in using the vocabulary but outside the original community. The ability to work through those steps from discovery to evaluation for a specific purpose depends upon the availability of background documentation that can support the analysis of appropriateness for use.

Finally, there is the issue of sustainability of the vocabularies selected for use in data. A sustainable vocabulary is protected by organizational or institutional commitments, policies that make clear who makes those commitments and what they mean, as well as a record of responsible maintenance and growth. A vocabulary without those commitments may not be sustainable over time and may be a questionable investment for organizations seeking to use the vocabulary in their data. Over time, unsustainable vocabularies tend to become ‘orphans’ or ‘zombies.’ Both of these terms refer to a kind of abandonment, the difference being that ‘orphans’ tend to disappear from the open Web entirely (usually because of domain disappearance or transfer), while ‘zombies’ don’t disappear but cease to be actively maintained. In theory, at least, both varieties of resource could be embraced for current uses, or resurrected and repurposed, but guidance is needed regarding the ability and perhaps the appropriateness of using these tools. Building sustainable structures around vocabularies before they lose funding and support makes more sense than allowing them to be cast aside, and this document provides strategies to accomplish that goal.

¹⁰ U.S. Open Data Institute. “Simple Open Data.” (n.d.). <http://simpleopendata.com/>

In some respects, the problem of vocabulary discovery and availability for general use is not easily separated from issues around the loss of funding for projects building vocabulary development or management tools, almost all of which were initially developed in time-limited circumstances. This report cites several projects that were initially funded in whole or part to address issues around vocabulary provision in particular research or practice communities, but that have not received funding to extend or maintain their tools or their vocabularies. This is a significant problem, as without the ability to support the structural or conceptual vocabularies required to describe resources being aggregated or distributed, there is no such thing as useful distribution or maintenance of any data instances using those vocabularies. The many projects now being funded to consider the problems of ‘big data,’ particularly scientific research data, all depend on stable metadata vocabularies.

This concern is not that different from what we hear regularly about the crumbling physical infrastructure supporting our transportation systems (although without the scary safety implications). Many on the academic side of these questions take it for granted that funding comes and goes, and may not consider the longer-term implications of these ebbs and flows. But the reason to talk about vocabulary sustainability is that there are long-term implications of depending on funded projects to build and maintain the infrastructure around vocabularies used for linked open data, not to mention the vocabularies themselves.

2.1 Communication Between Owners and Users: The Value Proposition

The advent of the World Wide Web has changed the perspective on vocabularies, primarily because of the need to exchange data for Web usage. In this context, vocabularies are increasingly used to describe the ‘things’ of the real world, in collections other than publications and in settings besides libraries. An example is the Ontology for Biomedical Investigations,¹¹ which represents knowledge about biomedical research resources.

Each of the activities under discussion here is ideally part of a two-way conversation between the originator and consumer.

Table 2: General activities around vocabulary management and the responsible parties.

Activity	Originator	Consumer
Discovery	Owner	User
Terms of use	Owner	User
Feedback	User	Owner
Updates	Owner	User

¹¹ Bandrowski A, R. Brinkman, M. Brochhausen, M.H. Brush, B. Bug, M.C. Chibucos, Kevin Clancy, et al. “The Ontology for Biomedical Investigations.” PLoS ONE 11, no. 4 (2016): e0154556, <https://doi.org/10.1371/journal.pone.0154556>

2.2 Evolution of Business Models

Until the advent of the Web, most vocabularies were published as printed volumes (with a few available later as proprietary computer files for sale). Producing printed volumes of vocabularies was not generally a very lucrative activity, but the products were intended to financially support the building and management of the vocabulary, and generally did. The Web upset that entire model, and most vocabularies supported by print sales found themselves in an uncomfortable position with few viable choices. At first many set up commercial-style websites where vocabulary terms could be looked up one at a time and used as text labels in data. Some websites were based on a subscription model, meaning that only subscribers could access the site's functions. Some of these sites, backed by databases, also sold files that could be used inside specific applications (this option was generally only feasible for users with good IT support). Revenues plummeted, and neither owners nor users of the vocabularies were satisfied with the early alternatives to print distribution. But revenue-based models were often costly to manage as well, so vocabulary owners gradually began shifting to more openly accessible, Web-friendly models.

Users quickly became accustomed to information being freely available, which pushed vocabulary owners to provide services without obvious revenue sources to support them. A useful case study of this transition is provided by the Getty Research Institute, which moved four of its widely used vocabularies to a Linked Open Data model in 2014. The Getty experience is well documented and provides a useful case study for other organizations.¹²

One interesting aspect of the transition from the print business model to the Web linked data model is the change in how users and owners interact. In the print model, users are primarily seen as consumers, and changes in the vocabularies themselves, though perhaps filtered through advisory boards of some kind, are essentially managed top-down with little input from users. As the 'consumer' model changes to a more community-based, bottom-up culture, users are perceived as having a larger role to play, with suggestions and feedback openly solicited from user communities. Both the Getty and the Library of Congress (LC) see their user communities in this light, though LC has a longer history of those interactions in its NACO and SACO programs, part of its Program for Cooperative Cataloging (PCC).¹³ To some extent, this crowdsourcing approach partially addressed the loss of income supporting the top-down vocabulary development, but the change in perspectives can provide other benefits as well.

As users in the vocabulary context, data managers are perhaps more intermediaries than consumers, although the consumer model is useful in some aspects. These intermediate users need a context for decision making, which requires knowledge about the vocabulary, its content, scope, structure, and limitations (based on expansion decisions already made, or policies binding the growth and boundaries already in place). The history of the vocabulary's development can be useful, although not as a substitute for specific maintenance and versioning policies.

Active user communities don't grow by themselves, and require owners to think more broadly about their mission and priorities. Particularly if owners seek outside funding for

¹² Cobb, Joan. "The Journey to Linked Open Data: The Getty Vocabularies" *Journal of Library Metadata* 15, no. 3-4 (2015): 142-156, <http://www.tandfonline.com/doi/full/10.1080/19386389.2015.1103081>

¹³ Library of Congress. Program for Cooperative Cataloging (PCC). (n.d.). <https://www.loc.gov/aba/pcc/>

expansion or technical improvements, knowing who their users are and providing evidence of usage and user interaction can be crucial.

2.3 Use Case: Vocabularies in Transition

2.3.1 FAST

The Faceted Application of Subject Terminology, or FAST, began as a collaboration between the Library of Congress and OCLC in 1998.¹⁴ Its terms are programmatically generated from Library of Congress Subject Headings (LCSH). Designed to deal with the proliferation of digital resources collected by libraries and the difficulty for catalogers in keeping up with changes, FAST headings retain some of the bibliographic control of LCSH but are easier for non-experts to apply and superior to simple keywords for information retrieval. Moreover, unlike LCSH, which include “free floating subdivisions” and other aspects that defy persistent identifiers, every FAST heading is backed-up by a namespaced URL, making FAST headings more suitable than LCSH ones for use and reuse in Semantic Web applications.¹⁵ While the assignment of FAST subject headings is easier than for LCSH, FAST headings are generated from LCSH strings, which are still created one at a time by librarians, forming a bottleneck. There is no mechanism for creating FAST headings that do not derive from LCSH. Moreover, FAST is still treated more like a research project than a production service. It will be interesting to see if FAST can evolve to include contributions from non-experts or from automated natural language processing, perhaps as provisional headings that can later be curated by an editorial staff.¹⁶

¹⁴ OCLC Research. “FAST (Faceted Application of Subject Terminology).” Accessed July 22, 2015.

<http://www.oclc.org/research/themes/data-science/fast.html>

¹⁵ Library of Congress. “*Library of Congress Subject Headings*” Pre- vs. Post-Coordination and Related Issues.” Report for Beacher Wiggins, Director, Acquisitions & Bibliographic Access Directorate, Library Services, Library of Congress. Prepared by the Cataloging Policy and Support Office. March 15, 2007.

https://www.loc.gov/catdir/cpso/pre_vs_post.pdf

¹⁶ FAST is now an eight-facet vocabulary, covering Personal names, Corporate names, Geographic names, Events, Titles, Time periods, Topics, and Form/Genre. While LCSH has complex rules and syntax for stringing together (pre-coordinating) terms from these various categories, as well as complex rules for assigning these subject strings to individual publications, FAST simplifies the process by ensuring that a string only includes terms from a single facet type.

2.3.2 The Getty Vocabularies

The screenshot shows the Getty Research Institute website. The main heading is "The Getty Research Institute". Below it is a navigation bar with links: Exhibitions & Events, Special Collections, Library, Search Tools & Databases, Scholars & Projects, Publications, and About the GRI. The "Search Tools & Databases" section is active, showing a sidebar with a list of search tools and databases. The main content area is titled "Getty Vocabularies as Linked Open Data". It contains a paragraph explaining the project, a "Linked Open Data" logo, and a list of resources. The sidebar on the right contains links for "Inside Perspective", "Unlocking hidden resources for scholars", "Have a Question?", and "Contact the Vocabulary Program".

Figure 2. Starting in 2014, The Getty Research Institute made its vocabularies available as Linked Open Data.

The Getty Vocabularies began as part of a set of standards and research database projects that were supported by the Getty Art History Information Program (AHIP) beginning in 1983. The goal was to build tools and resources that could become best practices for describing art, architecture, and material culture by museums, libraries, archives, and researchers. The Art & Architecture Thesaurus (AAT) was developed independently from any library, archive, or museum collection. It was inspired by faceted indexing schemes and aspired to be a new model for information retrieval online. The other Getty vocabularies—the Union List of Artist Names (ULAN) and the Thesaurus of Geographical Names (TGN)—were developed in a more traditional way by compiling existing lists of proper names and consolidating them around named-entity authority records. The Getty Vocabularies are more or less authority

files derived from the AHIP research databases (*Répertoire international de la littérature de l'art*—RILA, Bibliography of the History of Art—BHA, the Avery Index of Architectural Periodicals, Provenance Index, Witt Library Index, and Census of Antique Art and Architecture Known to the Renaissance) as well as a widening set of collaborating museums and archives. The AAT was published in print and on CD ROM in 1990 and 1994 by Oxford University Press and the ULAN in hardcopy and on CD ROM in 1994 by G.K. Hall. The TGN was first published in 1997 in machine-readable files. From 1997 to 2014 the Getty Vocabularies were published as a freely accessible Web resource and in data files available for licensing. Beginning in 2014, the Vocabularies were made available via SPARQL endpoints as linked open data.¹⁷

Perhaps the most significant contribution of the Getty AHIP was the development of a community of librarians, archivists, museum professionals, and researchers who were beginning to think about how to work in the emerging online ecosystem. This community has become the de facto editorial board for the Getty Vocabularies and spawned projects to translate them into other languages. The Vocabularies turn up in museum collections management system software such as PastPerfect and as part of cultural heritage projects such as Europeana, which enriches its data with the AAT.¹⁸

2.3.3 From English-only to Multilingual (RDA)

RDA: Resource Description and Access is a standard for descriptive cataloging initially released in the RDA Toolkit in June 2010. Intended for use by libraries and related cultural organizations such as museums and archives, the RDA's instructions and guidelines on formulating bibliographic data succeed *Anglo-American Cataloging Rules, Second Edition* (AACR2),¹⁹ the prevailing standard for English language libraries since 1978.

RDA emerged from the International Conference on the Principles & Future Development of AACR held in Toronto in 1997.²⁰ It is published jointly by the American Library Association, the Chartered Institute of Library and Information Professionals (CILIP) in the UK, and previously by the Canadian Library Association, which has now disbanded. Maintenance of RDA is the responsibility of the RDA Steering Committee (RSC), composed initially of representatives from the three RDA publishers and the Australian Committee on Cataloguing, the British Library, the Canadian Committee on Cataloguing, and the Library of Congress. As RDA began to be considered less an Anglo-American and more of an international standard, a shift in focus occurred, which has manifested itself in the new, more international management of the standard.

The RDA Vocabularies, although available in draft form since 2008, were officially published in the RDA Registry²¹ in January 2014. The vocabularies (both the element sets and value vocabularies) are managed within a Git/GitHub environment and made available in the RDA

¹⁷ Cobb, Joan. "The Journey to Linked Open Data: The Getty Vocabularies" *Journal of Library Metadata* 15, no. 3-4 (2015): 142-156, <http://www.tandfonline.com/doi/full/10.1080/19386389.2015.1103081>

¹⁸ Charles, Valentine, and Cécile Devarenne. "Europeana Enriches its Data with the AAT." *EuropeanaPro*. (n.d.) <http://pro.europeana.eu/page/europeana-aat>

¹⁹ American Library Association, Canadian Library Association, and the Chartered Institute of Library and Information Professionals. *Anglo-American Cataloging Rules, Second Edition* (AACR2). <http://www.aacr2.org/>

²⁰ Joint Steering Committee for the Development of RDA. International Conference on the Principles and Future Development of AACR. (July 1, 2009). <http://www.rda-jsc.org/archivedsite/intlconf1.html>

²¹ RDA. RDA Registry. <http://www.rdaregistry.info/>

Registry—along with extensive documentation, examples, maps, events, and translations of the vocabularies. The Toolkit instructions also include translations, in both cases built and managed by the language communities affected. The Open Metadata Registry (OMR) is the management hub of the RDA vocabularies, including language versions, linked together with the canonical IDs (numeric) and sets of lexical aliases built from the language labels in the translated vocabularies. This structure allows for more stability for all the language versions, since changes in labels do not require changes in URIs. The introduction of translations sparked the development of spreadsheet-based export and import. Thus changes to specific language updates are automatically linked to the canonical URIs and are also linked to a specific numbered version of the original English RDA.

The difficulty of reconciling terms in the vocabularies with those used in the Toolkit instructions has been addressed by relying entirely on the OMR to manage vocabularies where they appear in the Toolkit and regularly populating the instructions (including the glossary) from the OMR data. This keeps all the data synchronized and lessens the reliance on coordinating updates of the tools using expensive human effort. As the automated workflows for these efforts are completed, users of RDA, no matter their preferred language, can be assured that instructional data and vocabularies are in sync and up to date, and when used in descriptive data can be automatically and efficiently updated.

2.4 The Technical Environment

Technical standards and tools for vocabulary development and management range across a very wide spectrum. Some standards for vocabularies, for instance ANSI/NISO Z39.19-2005 (R2010) *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, focus on the general issues of the pre-Web environment. Since the Semantic Web emerged, the W3C has taken the lead in developing technical standards for many aspects of vocabulary development.

It is beyond the scope of this working group to establish technical standards for reuse, reservation, and documentation of bibliographic vocabularies. Instead, we refer those interested in vocabulary development and management to external technical documents, primarily from the W3C, that represent broad community consensus and deep technical expertise. The documents listed below offer best practices for working with Linked Data including use of persistent URIs, multilingual representations of ontology terms, version control metadata, and application profiles.

- *Cool URIs for the Semantic Web*²² is the ongoing technical reference for minting and serving URIs for linked data resources. Inspired by Tim Berners-Lee's 1998 article "Cool URIs Don't Change,"²³ this Interest Group Note (not a W3C recommendation) discusses conventions for content negotiation (e.g., serving an HTML representation vs. RDF/XML), server redirects, and distinguishing Web documents from real-world objects.
- *Best Practice Recipes for Publishing RDF Vocabularies*²⁴ is the ongoing reference for technical guidelines on how to serve RDF vocabularies (and their elements) on the Web,

²² W3C. "Cool URIs for the Semantic Web." (December 3, 2008). <https://www.w3.org/TR/cooluris/>

²³ Berners-Lee, Tim. "Cool URIs Don't Change." (1998). <https://www.w3.org/Provider/Style/URI>

²⁴ W3C. "Best Practice Recipes for Publishing RDF Vocabularies." (August 28, 2008). <https://www.w3.org/TR/swbp-vocab-pub/>

expanding on "Cool URI" best practices.²⁵ The recipes vary depending on one's use of hash vs. slash²⁶ namespaces, the level of content negotiation required, and other factors.

- *Data on the Web Best Practices* (2017)²⁷ is a W3C Recommendation that goes beyond vocabularies to cover data of all types on the Web, mostly those considered to be in the "Linked Open Data cloud", but also restricted-access data, where privacy or intellectual property must be taken into consideration. Most relevant to the Bibliographic Roadmap is Best Practice 8.9, which defines various types of vocabularies, recommends reuse of pre-existing vocabularies, and cites related recommendations. The Data Catalog Vocabulary (DCAT)²⁸ is highlighted for its reuse and incorporation of Dublin Core, FOAF, and SKOS vocabularies. Other important factors include the reputation of the sponsoring agency, the quality of documentation, and the provision of unique Web-based identifiers that resolve to term definitions and other documentation. Also noteworthy is Best Practice 16, which advises the reader to "Choose the right formalization level," that is, to opt for a level of formal semantics that fits both the instance data and likely target applications. The recommendation is to "design for wide use," accommodating the widest range of applications and data sources.
- *W3C Best Practices for Publishing Linked Data*²⁹ has in some ways been superseded by the *W3C Data On the Web Practices* document but remains a helpful resource.³⁰ While focused on government-sourced data, the guidelines apply equally well to linked data from other publishers. Chapter 6, "Standard Vocabularies," explains the value of using well-established (i.e., "standard") vocabularies before attempting to create new ones. A checklist is provided to help developers appropriately reuse or extend pre-existing vocabularies. Note that some best practices have been considered, reprised, and/or generalized in the later work of the W3C Data On the Web Practices Working Group, above.
- *W3C Library Linked Data Incubator Group Final Report*³¹ reviews library data sets available on the Web (as of 2011) and how they can be reused, extended, and combined when conforming to Linked Data standards. Chartered by the W3C from May 2010 through August 2011, the group's charge was "to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web activities—focusing on Linked Data—in the library community and beyond, building on existing initiatives, and identifying collaboration tracks for the future." Recommendations include: (a) identifying vocabularies for early conversion to RDF; (b) engaging in Semantic Web standardization work; (c) having data and system designers enhance user services based on linked data design principles; and (d) having librarians

²⁵ W3C. "Best Practice Recipes for Publishing RDF Vocabularies." (August 28, 2008).

<https://www.w3.org/TR/swbp-vocab-pub/>

²⁶ W3C. HashVsSlash. (March 22, 2016). <https://www.w3.org/wiki/HashVsSlash>

²⁷ Lóscio, Bernadette Farias, Caroline Burle, Newton Calegari (eds.). W3C Recommendation 31 January 2017. Accessed June 16, 2017. <https://www.w3.org/TR/dwbp/>

²⁸ W3C. Data Catalog Vocabulary (DCAT). (January 16, 2014). <https://www.w3.org/TR/vocab-dcat/>

²⁹ Hyland, Bernadette, Ghislain Atemezang, Boris Villazón-Terrazas (eds.) W3C Working Group Note 09 January 2014. Accessed September 10, 2016. <https://www.w3.org/TR/2014/NOTE-ld-bp-20140109/>

³⁰ Lóscio, Bernadette Farias, Caroline Burle, Newton Calegari (eds.). "Data on the Web Best Practices," W3C Candidate Recommendation. (August 30, 2016). Accessed September 8, 2016. <https://www.w3.org/TR/dwbp/>

³¹ Baker, Thomas, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, et al. "W3C Library Linked Data Incubator Group: Final Report, 25 October 2011." (Accessed Sept. 10, 2016). <https://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>

and archivists contribute to the long-term preservation of element sets and value vocabularies.³²

- W3C Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets (2011),³³ supplements the main Incubator report, above. It distinguishes among three types of library data: instance data (datasets), value vocabularies, and metadata element sets, providing definitions and examples for each. Its stated goal is to “identify a set of useful resources for creating or consuming Linked Data in the library domain.”
- Library Linked Data Resources,³⁴ authored by the Library Linked Data Incubator Group, covers library-oriented element sets and value vocabularies, some of which have been modeled (or re-modeled) in RDF. Examples include ISBD, FRBR, RDA, and MARC.³⁵
- W3C Vocabulary Management White Paper³⁶ outlines a policy for versioning vocabularies, modeled on policies previously established for SKOS and DCMI. In some cases, it has been superseded by other recommendations found on this list. A more recent W3C plan for vocabulary management can be found at the Vocabulary Services³⁷ webpage, but it remains in the planning stages.
- W3C Vocabulary Market³⁸ advises on how RDF-based vocabularies can be discovered or publicized on the Web. It also provides examples of vocabularies that are well-suited to bibliographic data (as well as e-commerce, provenance, and other domains).³⁹ As of 2016, however, the list is no longer maintained. For a more formal reference to W3C-maintained vocabularies, see the “RDF Vocabularies Current Status” page.⁴⁰ For a list of RDF vocabularies that are reused across W3C specifications, see the 2013 “core initial context” for RDFa,⁴¹ which is based on a 2010 usage study of RDF vocabulary⁴² over the (Semantic) Web.⁴³
- *Generating HTML documentation of OWL*⁴⁴ is dated, but includes a list of tools for generating HTML documentation of OWL ontologies, including references to other lists; examples of vocabularies that use the tools; and comments on their suitability.⁴⁵

³² Ibid.

³³ Isaac, Antoine, William Waites, Jeff Young, Marcia Zeng. 25 October 2011. Accessed Sept. 10, 2016 <https://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/>

³⁴ Library Linked Data Community Wiki, W3C Library Linked Data Incubator Group. https://www.w3.org/2001/sw/wiki/LLD/Library_Data_Resources

³⁵ Ibid.

³⁶ W3C. “W3C VocabMgtWhitePaper: White Paper on Vocabulary Management.” (Last updated September 4, 2009). <https://www.w3.org/wiki/VocabMgtWhitePaper>

³⁷ W3C. W3C Vocabulary Services. (June 6, 2013). <https://www.w3.org/2013/04/vocabs/>

³⁸ W3C VocabularyMarket (Last updated in 2011). (Last updated April 1, 2013). <https://www.w3.org/wiki/VocabularyMarket>

³⁹ Ibid.

⁴⁰ W3C. RDF Vocabularies Current Status. (2017). https://www.w3.org/standards/techs/rdfvocab#w3c_all

⁴¹ W3C. RDFa Core Initial Context. (June 8, 2017). <https://www.w3.org/2011/rdfa-context/rdfa-1.1>

⁴² W3C. Vocabulary Search on the Semantic Web for RDFa Default Profiles. (March 1, 2013). <https://www.w3.org/2010/02/rdfa/profile/data/>

⁴³ Ibid.

⁴⁴ W3C. “Generating HTML Documentation of OWL.” (Last updated June 2, 2012). https://www.w3.org/2011/prov/wiki/Generating_HTML_documentation_of_OWL

⁴⁵ Ibid.

2.5 The Policy Environment and its Implications

It was clear to the Working Groups from the outset that perhaps the most significant gap in the vocabulary development environment was the dearth (or complete lack) of documentation around so many critical public vocabularies. Documentation underlies the provision of discovery services (both within and across vocabularies) as well as other services essential to support content providers, vocabulary developers, and maintainers in their efforts to provide reliable data to the Semantic Web. For instance, the right kind of documentation can be used by content owners to justify the investment in vocabularies and maintenance. Sustainability and preservation services are also dependent on appropriate documentation to function effectively when long-term viability is at risk.

Potential decision makers also need background information on the organization developing and managing a vocabulary, as well as any policies developed to support the effort. Policies and practices around versioning and change notification are critical for potential users intending to maintain their content over time using automated processes. Even for those updating using human effort, which is more expensive, knowing where change has occurred—instead of having to hunt through files—is critical. Documentation also includes information on the institutional policies around maintenance, mission, and governance—all categories that require effort and long-term commitment by owners.

The W3C Data on the Web Best Practices advocate that a number of different aspects of policy be documented, including information about access, usage, versioning, licenses, and quality measures. The report highlights three metadata archetypes: structural, descriptive, and localized. For example, structural vocabularies document how the vocabulary is organized internally, and this structure can be used to display the hierarchy of connections on a webpage or other interface for an end user.

Descriptive information available for human and machine consumption is often provided both within the vocabulary or via a dedicated means to share vocabulary data (using an API or other form of data publication designed for machine-to-machine communication). For example, through an API, the W3C Data Catalog Vocabulary (DCAT) can be reused to provide descriptive information by pointing from one linked dataset to another that is associated with a different vocabulary.⁴⁶

Perhaps the most useful, comprehensive, and authoritative example of providing descriptive documentation for structured vocabularies is the Linked Open Vocabulary (LOV) model. LOV-recommended data includes contributors (attribution data), frequency of maintenance, date of origination; descriptions; links to datasets about the vocabulary; version information; namespace; namespace prefix; number of properties and classes; and which vocabularies are referenced, extended, specialized, or generalized, when such information is known to exist about any given vocabulary in LOV's integrated hub of vocabularies. LOV's submission metadata template, usage of which is required of vocabularies that submit to its database, is often cited as a clear set of basic descriptive recommendations.

There is not a direct analog for value vocabularies similar to LOV, but a combination of ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* and SKOS provides some context for

⁴⁶ W3C. Data on the Web Best Practices. (Jan. 31, 2017.) <https://www.w3.org/TR/dwbp/>

discussion. For the top level of vocabulary description, the documentation suggested by LOV is directly applicable, and because value vocabularies tend to be more volatile than element sets, information about where responsibility for maintenance and change management is focused can be most important. Some groups, for instance the DCMI NKOS Task Group,⁴⁷ are considering some of these questions, but it's difficult to determine from the task group's webpage how actively it is working.

There is no absolutely accepted method for how to provide metadata for particular vocabularies. The information might be shared from datasets within the data itself, for example, the LOV recommends including a preferred namespace and namespace prefix within the vocabulary.⁴⁸ The W3C Best Practices for Publishing Linked Data for creating "Good URIs" could be another example.⁴⁹ More extensive information, including policies on maintenance, versioning, and further development plans could be maintained for the vocabulary on webpages, which are often supported via popular tools like GitHub and BitBucket.

2.6 Licenses

One of the challenges in determining the ownership and shareability of information resources is that terms of use are often expressed as free text or not expressed at all. Because most licensing regimes in the digital universe were developed with content and software rather than vocabularies or metadata in mind, there is continuing confusion about whether and how vocabularies can be licensed. Content aggregators such as Europeana and the Digital Public Library of America (DPLA) have collaborated to develop and publish a set of relevant statements about digital content at <http://rightsstatements.org/>. As useful as this resource is, particularly for shared cultural heritage objects, it does not seem to cover vocabularies at this point.

For the most part, currently available public vocabularies lack license statements, which is unhelpful, because many potential users will assume that the lack of a license means that the vocabulary owner intends to disallow use by others—and indeed they should not make such assumptions. Until Creative Commons recently developed its CC0 license, there were no open licenses suitable for general use by vocabularies.

"CC0 helps solve this problem by giving creators a way to waive all their copyright and related rights in their works to the fullest extent allowed by law. CC0 is a universal instrument that is not adapted to the laws of any particular legal jurisdiction, similar to many open source software licenses."⁵⁰

Prior to the availability of CC0, vocabulary owners intending not to assert rights often used other CC licenses in the 'CC BY' category, all of which required attribution of the original

⁴⁷ DCMI NKOS Task Group. "Developing a Dublin Core Application Profile for KOS Resources". Available at: http://wiki.dublincore.org/index.php/DCMI_NKOS_Task_Group. Last modified Sept. 30, 2015.

⁴⁸ Baker, Thomas, Pierre-Yves Vandenbussche, and Bernard Vatant. "Requirements for Vocabulary Preservation and Governance," *Library Hi Tech*, 31, no. 4 (November 2013): 657-668. <http://dx.doi.org/10.1108/LHT-03-2013-0027>

⁴⁹ Hyland, Bernadette, Ghislain Atemezang, and Boris Villazón-Terrazas (eds.) "Best Practices for Publishing Linked Data." W3C Working Group Note (January 9, 2014). Accessed September 10, 2016. <http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/>

⁵⁰ Creative Commons. Share Your Work: Public Domain. (n.d.). <https://creativecommons.org/share-your-work/public-domain/>

creator when reusing or deriving from a resource. Although it is fairly clear how such a license would operate for resources not subject to change, it was not clear how such attribution would be managed in the vocabulary world, where change is a given.

2.7 Application Profiles

The notion of application profiles has largely been developed by the Dublin Core Metadata Initiative (DCMI) over the past decade and a half, but the idea of 'profiles' is now in use in a number of metadata communities. DCMI's definition, "A DCAP defines metadata records which meet specific application needs while providing semantic interoperability with other applications on the basis of globally defined vocabularies and models," has been used by data providers and communities to provide structured, detailed information about data decisions and usage. DCMI was probably among the first to discuss Application Profiles in a formal sense, spurred by the publication of an article by the late Rachel Heery and her colleague Manjula Patel.⁵¹ In the article, the authors discussed the concept of 'mix and match' whereby more than one schema could be used to structure a set of instance data, but also that more than one vocabulary could be used to populate the schema. The mechanism for documenting those decisions, as well as limitations (repeatability, conditions of usage and instructions, and/or a rationale for decision making) have been gathered by the DCMI in its publication "Guidelines for Dublin Core Application Profiles."⁵²

As user-facing documentation, an Application Profile (AP) documents the decisions and context of vocabularies used within a data set. This should include how the vocabulary is used, where it is used, and the conditions of use. For instance, can only one vocabulary be used in a particular location, or may more than one be employed, if applicable? An example of this might be a bronze statue with a wooden base, which would generally be described as one object made with two different materials. This might be expressed as a 'repeatable' usage of the element-set vocabulary term for materials used, and might also include an instruction about how multiple terms should be expressed (within one instance of the expression or divided into multiple expressions), as well as the source of the terms used to describe the materials. Those decisions are important, because they affect how instance data and the vocabularies they contain are maintained and shared, and they provide crucial support for reuse of the data itself in other contexts.

An interesting example of the usefulness of APs is shown in the Kinematic Models for Design Digital Library (KMODDL) project (see Figure 3), which makes available related collections of 19th-century kinematic models held by several institutions. The project faced several challenges, among them a plethora of objects starting with the three-dimensional models designed in the 19th century for demonstration and/or teaching purposes, stereolithography files for creating 3-D replicas of the models, various images of the models (still, moving, and interactive), as well as tutorials and related textual resources including books and articles. Among the most innovative parts of the KMODDL AP is the use of published classifications of the models themselves as the basis for organizing the collections and developing several local vocabularies. This strategy ensured that materials about each specific model would be linked together, and the models associated with other

⁵¹ Heery, Rachel and Manjula Patel. "Application Profiles: Mixing and Matching Metadata Schemas," *Ariadne* 25 (September 24, 2000). <http://www.ariadne.ac.uk/issue25/app-profiles>

⁵² Dublin Core Metadata Initiative. Guidelines for Dublin Core Application Profiles. (May 18, 2009). <http://dublincore.org/documents/profile-guidelines/>

models in the collection that illustrate the same principles of motion. This extensive categorization allowed the collections to be extremely browseable.

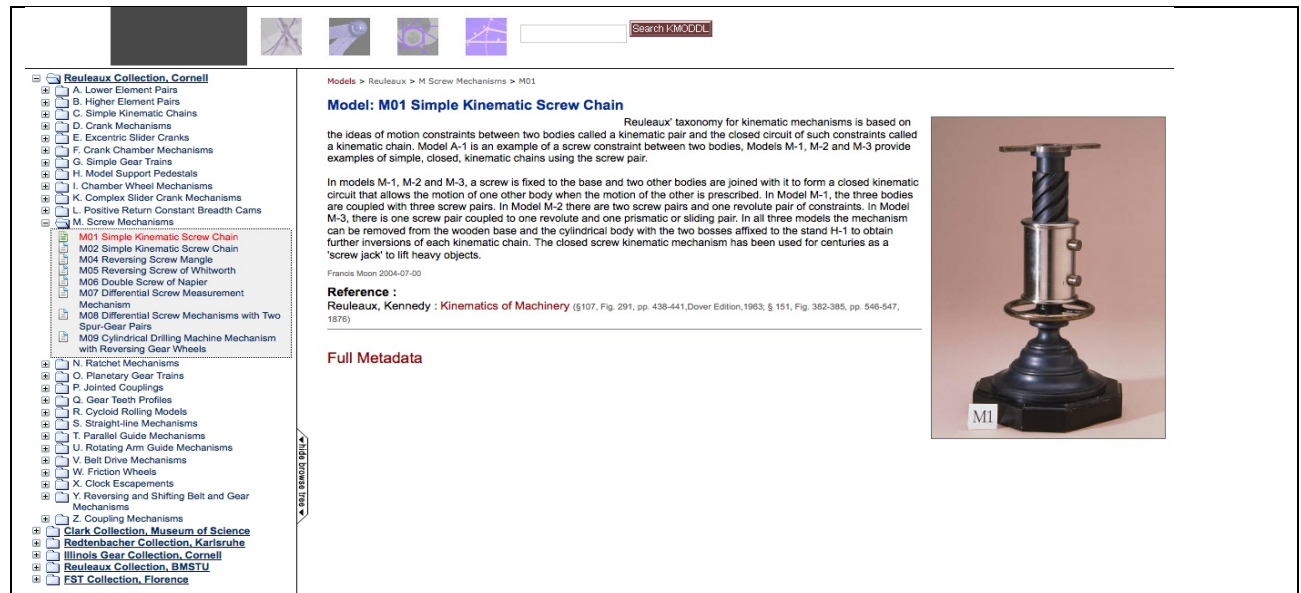


Figure 3: A screenshot of a KMODDL page, illustrating a particular model, its description, and, on the left panel, the overall categorizations relating to the illustrated model, designed to provide navigation using the available vocabularies.

Many APs are only represented by documentation—when an AP uses only one existing vocabulary and gives loose guidelines for using it in a specific context, this may be sufficient. For the future, and for more complex APs, there's a desire move a step further, and add a machine-readable enforcement mechanism, perhaps a schema, to validate the data or some other technology to accomplish a similar goal. Determining whether data going out or coming in (and referencing an AP) can be trusted is an important measure of quality and predictability, relevant for all kinds of data users. One important aspect of any AP is that it makes a connection between a particular element set with value vocabularies intended to be used by a particular community.

Machine readable APs have been requested and discussed for some years, with various groups engaged in discussions about how machine readability can be accomplished to meet important goals. One recent group, convened by DCMI, has identified a series of requirements for validation of APs. One validation language considered by the DCMI group is SHACL (Shape Constraints Language), being developed by the W3C's RDF Data Shapes WG as a language “for describing structural constraints and validate RDF instance data against those.”⁵³

⁵³ W3C. RDF Data Shapes Working Group. (June 12, 2017). https://www.w3.org/2014/data-shapes/wiki/Main_Page

Section 3: Maintenance: Status and Versioning

All vocabularies have a lifecycle, from initial development to retirement, affecting how maintenance and sustainability concerns play out. Characterization of a point in the lifecycle of a ‘collection’ of terms (the top level of a vocabulary) and the terms within the collection are functionally different. For instance, a status of “Active, currently maintained” might be appropriate for a vocabulary as a whole, but not appropriate for an individual concept or term. In a practical sense, without knowing when the determination of the application of a status was applied or changed, the status designation for the top level of the vocabulary itself loses meaning; without indication of maintenance activity within the vocabulary, the status is of questionable utility. The same problem occurs when determining the usefulness of a Web resource—a ‘last modified’ date that is years in the past might lead to concerns about the usefulness of the site and the information it contains.

The status of ‘deprecated’ is usually reserved to flag an individual term within a vocabulary as no longer used (for whatever reason) and is only rarely applied to a top-level vocabulary description. Deprecating is usually recommended instead of deleting, because the deletion of terms creates a problem for communities still using the term as well as those systems doing regular updating of the caches they keep for internal use. In some respects, at least in the short term, doing nothing but applying a flag (and a date) to the individual terms so that a service importing the vocabulary can filter deprecated terms makes more sense than deleting information that will create maintenance problems for users of the vocabulary. The status of deprecated applied to an individual term conveys valuable information for a service provider. Any routine developed to remove a deprecated term from a cached vocabulary or data that uses the deprecated term must start by identifying the URI and possibly the text label. Deleted information may leave a hole, but it’s not an identified hole, and requires more sophisticated matching of files to identify which terms are no longer present.

When an entire vocabulary is abandoned by its owner or community, it’s very unlikely that further work will be attempted on the description of the vocabulary as a collection, or indeed on the individual terms, that would be helpful for a potential user. In most cases, neither the vocabulary description nor the terms can be updated or modified by anyone other than the ‘owner’ (who presumably has permissions to do so, conferred by whomever manages the server where the vocabulary resource is stored). Presumably such permission could be transferred to another person or organization, either temporarily in order to allow the addition of information for users (or potential users) or to allow a new owner to maintain the vocabulary; a vocabulary without information about governance (active or inactive) can only be ignored, supplanted, or ‘taken over’ by copying the data to a new domain and attaching new URIs or other persistent identifiers.

This implies that recommending that a ‘status’ be applied at the point when the vocabulary is abandoned is unrealistic, even if there were general agreement on what status terms to use. Descriptive data at the top level that includes ownership (or contact) information is best supplied early in the development of the vocabulary, in the hope that any subsequent change in status could be inferred and validated. The date of the last update and the party responsible for it should be supplied automatically by the tool or management system in place—and not left to humans.

It’s in this context that the notion of versioning—the application of a version designation or number—becomes an important part of maintenance activity. It is most useful to think about

the rationale for and practice of versioning in the context of the familiar software used with our personal computers and mobile devices.

“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 (not to mention “apps”) coped with the necessity of updating their products as operating systems changed, competition for users grew, and functionality sought by customers became more sophisticated. Operating systems, software applications, and open standards such as HTML and JavaScript are increasingly interdependent and a seemingly minor change can have a devastating ripple effect. 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.

Software updates in general use version numbering to identify for users, and updating software, the version of individual software packages on a computer. 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. The software development community has recently begun to move toward a formal specification of version management known as “Semantic Versioning”.⁵⁴

From the point of view of a vocabulary owner, adopting a versioning strategy comes with a number of benefits, most importantly making the updating of content using the vocabulary easier and more amenable to automated routines. Vocabularies that use appropriate version strategies should be more attractive to those searching for well-designed and well-maintained vocabularies to employ within their data. In a rational world, good vocabularies that are extensively relied upon create value for data aggregators as well as the community at large that is presumably looking for good data to reuse.

For existing vocabularies, transitioning from a closed, file-based vocabulary to one expected to support linked data can be rocky. A shift from seeing vocabularies as primarily files distributed through bibliographic utilities to viewing them as more open resources that exploit the distribution services of the Web has not often been smooth.

“If we accept the premise that vocabulary semantics will change, there are very few methods to create stable systems that can rely on linked data. One option (preferred) is to use vocabularies from systems that provide stable URIs for past, present, and future versions of the vocabulary or (not preferred) to create a local, stable shadow vocabulary and map the local vocabulary to the public vocabulary over which you have little or no control. Mapping vocabularies in this way gives you the opportunity to maintain the semantic stability of your own system, your own ‘knowledge base’, while still providing the ability to maintain semantic integration with the global pool of linked data. Clearly, this is not an inexpensive proposition.

There are a number of related issues here that would also benefit from broader discussion. Large public vocabularies have tended to make an incomplete transition from print to online, getting stuck ... attempting to use the file management

⁵⁴ Hillmann, Diane, Gordon Dunsire, and Jon Phipps. “Versioning Vocabularies in a Linked Data World,” (Paris: IFLA, 2014). <http://hdl.handle.net/1813/40559>

processes of the print era to manage change behind a 'service' front end that isn't really designed to do the job it's being asked to do. What needs to be examined, soon and in public, is what the relationship is between these files and the legacy data which hangs over our heads like a boulder of Damocles. Clearly, we're not just in need of access to files (whether one at a time or in batches) but require more of the kinds of services that support libraries in managing and improving their data. These needs are especially critical to those organizations engaged in the important work of integrating legacy and project data, and trying to figure out a workflow that allows them to make full use of the legacy public vocabularies.⁵⁵

⁵⁵ Metadata Matters. Versions and Services, Part 1. (May 15, 2013).
<http://managemetadata.com/blog/2013/07/23/versions-and-services-pt-1/>

Section 4: Using, Reusing, Extending, Mapping

The popular notion of using or reusing existing vocabularies is based primarily on efficiency—why assume the burden of responsibility for a vocabulary you didn't develop when the URIs allow you to use the parts relevant to you? Heath and Bizer, in their book *Linked Data: Evolving the Web into a Global Data Space* are firm proponents of this strategy:

“If suitable terms can be found in existing vocabularies, these should be reused to describe data wherever possible, rather than reinvented. Reuse of existing terms is highly desirable as it maximises the probability that data can be consumed by applications that may be tuned to well-known vocabularies, without requiring further pre-processing of the data or modification of the application.”

The authors suggest criteria for reuse of existing vocabularies based on this preference:

“In selecting vocabularies for reuse the following criteria should be applied: [a.] Usage and uptake – is the vocabulary in widespread usage? Will using this vocabulary make a data set more or less accessible to existing Linked Data applications? [b.] Maintenance and governance – is the vocabulary actively maintained according to a clear governance process? When, and on what basis, are updates made? [c.] Coverage – does the vocabulary cover enough of the data set to justify adopting its terms and ontological commitments? [d.] Expressivity – is the degree of expressivity in the vocabulary appropriate to the data set and application scenario? Is it too expressive, or not expressive enough?”⁵⁶

There are important, but widely misunderstood, rules for reusing an extant vocabulary, perhaps the most important (and most ignored) being the rule that one should not change the meaning of terms within a vocabulary when reusing it in another context.

“When re-using terms from other vocabularies, respect the formal definitions and constraints declared by their maintainers. In a Semantic Web context there is a strong social convention by which it is the owner of a vocabulary – or to be more precise, the owner of the URI domain under which its terms are coined – who declares the meaning of a given term, and anyone re-using that term should respect its declared meaning. Of course, terms are subject to redefinition through actual use – for example, the term owl:sameAs has been misused on such a massive scale that its original meaning has arguably been compromised – but URIs can only serve to anchor meaning if this principle is followed.”⁵⁷

This rule is largely ignored in a number of widely used tools, and only some of the problems created are obvious. Some early DC adopters used what is called in DCMi 'dotty syntax,' for instance 'dc.format.extent', and others creatively added subproperties when needed, with too little concern for assertions of the original vocabulary.

⁵⁶ Heath, Tom and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. <http://linkeddatabook.com/editions/1.0/>

⁵⁷ Baker, Thomas, Pierre-Yves Vandenbussche, and Bernard Vatant. "Requirements for Vocabulary Preservation and Governance," *Library Hi Tech*, 31, no. 4 (November 2013): 657-668. <http://dx.doi.org/10.1108/LHT-03-2013-0027>

But there are other risks with the efficient reuse of existing vocabularies:

“There are two schools of thought on vocabulary design. The first says you should always reuse terms from existing vocabularies if you have them. The second says you should always create your own terms when given the chance.

The problem with the first is you are beholden to someone else's sensibilities should they change the meaning of terms from under you (if you think the meaning of terms are fixed, there are safer games for you to play than vocabulary design). The problem with the second is term proliferation, which leads to a requirement for data integration between systems (if you think defining the meaning of terms is not coveted, there are again safer games for you to play than vocabulary design).

What's good about the first approach is macroscopic—there are less terms on the whole. What's good about the second approach is microscopic - terms have local stability and coherency. Both of these approaches are wrong insofar as neither represents a complete solution. They also transcend technology issues, such as arguments over RDF versus XML. And at differing rates, they will produce a need to integrate vocabularies.”⁵⁸

4.1 Extension

For projects or users needing more than a general vocabulary, extending the general vocabulary to support specialized needs is one way to build out without starting from scratch. The figure below shows an extension to RDA by a fictional group using the KidLit domain to build out RDA to better support children's literature, using the pattern RDA uses for building sub-properties into its main element set (Figure 4).

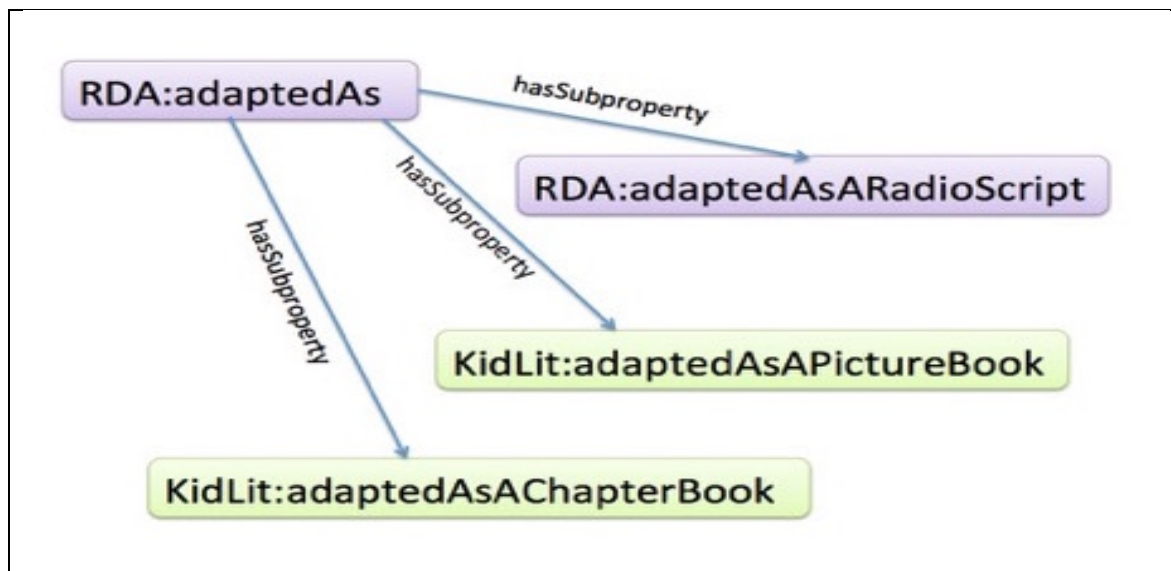


Figure 4: An illustration of an extension to RDA.

⁵⁸ de hÓra, Bill. “Vocabulary Design and Integration.” (April 28, 2007). http://www.dehora.net/journal/2007/04/data_integration.html

The extension strategy illustrated above has the added advantage of building a map at the same time as it expands the vocabulary, using the relationships with the existing vocabulary as a way to ensure that the extensions can always be ‘dumbed down’ for users of the data who prefer not to use extensions. In other words, if builders of a system have no real interest in picture books or chapter books, they are free to ignore any or all extensions from that community when designing their data usage.

Although RDA vocabularies often have domains and ranges, there is an unconstrained set that contains no domains or ranges and thus is particularly useful for extensions and mapping.

Most important from an interoperability standpoint when extensions are developed is that care must be taken to document those extensions as a *separate vocabulary*, making sure that domain and range assertions match the original vocabulary and appropriate separate namespaces conventions are used.

4.2 Mapping

In the MARC era, mapping meant crosswalks, defined as a “...specification for mapping one metadata standard to another.”⁵⁹ Crosswalks were generally thought to be one to one between two schemas, usually developed by large institutional agents who needed to build and use crosswalks during their regular activities. Crosswalks were sometimes available beyond the institutional setting, e.g., the Library of Congress made its crosswalk from MARC to Dublin Core (and Dublin Core to MARC) available on a webpage, but most were rarely distributed beyond the creating institution. In addition, crosswalks tended to be rather blunt instruments, assuming only one relationship, ‘same as,’ between term 1 in the source schema and a target term in another schema.

To some extent, the concern about ‘too many’ vocabularies stems from a fear of a world more out of control, even as moves towards URIs and public schemas seem to argue in the opposite direction.

*“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.”*⁶⁰

⁵⁹ St. Pierre, Margaret and William P. LaPlant, Jr. “Issues in Crosswalking Content Metadata Standards.” (Oct. 15, 1998). http://www.niso.org/publications/white_papers/crosswalk/

⁶⁰ Hillmann, Diane, Gordon Dunsire, and Jon Phipps. “Maps and Gaps: Strategies for Vocabulary Design and Development,” in *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2013*. <http://hdl.handle.net/1813/42443>

But in reality, the new environment has many other advantages, ones that are not always obvious when concerns about rapid growth continue to focus the conversation.

“The meaning of “mapping” changes radically on moving from a database and record based approach to an open, multi-domain, global, shared environment based on linked data technologies—where anybody can say anything about any topic, validity constraints are not acknowledged, a nearly infinite number of properties can be defined to describe an infinite number of entities, and authority is multi-dimensional and often ephemeral. The classic approach to such apparent chaos is to attempt increased control, increased filtering, increased restrictions, and limited access. This approach hinders appreciation of the broad diversity of perspective that comes with a world of open data.”⁶¹

One aspect of the important shift from crosswalks to maps is that who’s doing the mapping and what their goals might be begins to turn on its head. For crosswalks, there was no real need for more than one such tool, or at least, no such needs were expressed. In the MARC environment, where there were limited sources of data and a few large institutions dominating distribution, a generalized strategy seemed sensible. In an open linked data world with many distributors and consumers of data with no central ‘hub’ and myriad goals, maps present a different sort of value. It seems but a matter of time before maps are created and shared as vocabularies are now: as separate resources, each with its own rationale and flavor.

⁶¹ Dunsire, Gordon, Diane Hillmann, Jon Phipps, and Karen Coyle. “A Reconsideration of Mapping in a Semantic World,” in *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2011. <http://dcpapers.dublincore.org/pubs/article/view/3622/1848>

Section 5: Vocabulary Preservation

In order to understand the issues of long-term stability and preservation, it's important to consider description and maintenance requirements within a larger, longer-term context. Sustainable discovery and access of vocabularies over time requires adequate documentation, appropriate maintenance policies, and active governance, as does any activity focused on current use and re-use.

"As a foundation for data sources meant to be usable in the long term, the value of any given vocabulary depends on the perceived certainty that the vocabulary—in both its machine-readable and human-readable forms—will remain reliably accessible over time and that its URIs will not be sold, re-purposed, or simply forgotten. Vocabulary maintainers move on to other projects or retire. Resources owned by institutions may be neglected or become unavailable. As the givers of meaning to datasets, vocabularies are of vital importance to the scholarly record and cultural memory. However, their preservation will not happen automatically; it must be planned. The requirements for long-term preservation must consider a timeframe that stretches beyond the planning horizon of any institution that exists today."⁶²

In the digital preservation community built around content, the use of a community model—such as LOCKSS ("Lots of Copies Keeps Stuff Safe"⁶³)—has been enthusiastically embraced, but there are different issues when preserving resources whose content changes significantly over time. Baker et al.⁶⁴ argue that LOCKSS can be used effectively by mirroring information caches among multiple repositories, and they provide some detail about how such activity could be technically accomplished. Attractive as that appears, it seems a stretch in a world that has only begun to come to grips with the complexities of building and maintaining useful vocabularies for those able to discover them, especially in an environment of sparse resources.

In the future, we can imagine a broadly distributed ecosystem for vocabulary creation, maintenance, and use based on a commonly agreed URI infrastructure, built to support distribution of terms to consumers based on their explicit preferences. The Food and Agriculture Organization (FAO) implements such a model for AGROVOC⁶⁵ and it is instructive to review its features. The 32,000-concept agricultural vocabulary available in 27 languages is envisioned as a "quarry" where owners of local information systems can select relevant parts of the vocabulary and add their own concepts, depending on the needs of a specific country or domain.⁶⁶

⁶² Baker, Thomas, Pierre-Yves Vandenbussche, and Bernard Vatant. "Requirements for Vocabulary Preservation and Governance," *Library Hi Tech*, 31, no. 4 (November 2013): 657-668. <http://dx.doi.org/10.1108/LHT-03-2013-0027>

⁶³ LOCKSS. Lots of Copies Keeps Stuff Safe. (n.d.). <https://www.lockss.org/>

⁶⁴ Baker, Thomas, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, et al. "W3C Library Linked Data Incubator Group: Final Report, 25 October 2011." (Accessed Sept. 10, 2016). <https://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>

⁶⁵ Food and Agriculture Organization of the United Nations. "AGROVOC Multilingual Agricultural Thesaurus." (Last modified July 15, 2016). <http://aims.fao.org/registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>

⁶⁶ Baker, Thomas and Johannes Keizer. "Linked Data for Fighting Global Hunger: Experiences in Setting Standards for Agricultural Information Management," in *Linking Enterprise Data* (Springer, 2010): 177-201. Accessed January 2017. <http://www.fao.org/docrep/014/am324e/am324e.pdf>

The maintenance of AGROVOC is decentralized, based on the VocBench Web service for editing, publishing, and maintaining vocabularies. Partners from the AGROVOC community can add their own translations to existing AGROVOC concepts, and add new, specialized concepts that fit local needs not yet addressed by the concepts from the 'official' core vocabulary. FAO recognizes and welcomes many contributors to its community of collaborators. New terms can be suggested using the centralized Web service. These terms can be submitted for integration in the core vocabulary and approved by the relevant editors. But they can also be used in local applications by those who created and need them, as soon as they are created in VocBench. In a truly Linked Data approach, thanks to mappings to other vocabularies, AGROVOC can also be used as a hub from which to access many other vocabularies available on the Web. AGROVOC is a large and well-supported effort, with a robust community that illustrates how flexibility and good management can build stability over time.

Section 6: Recommendations

Given the important gaps in current practices around vocabulary development and the important technical challenges already being faced and expected in future, there is clearly a need for broader discussion and delineation of best practices to meet the goals of the vocabulary-development community going forward. This report provides important background for that discussion. The recommendations here are primarily aimed at the NISO community and are intended to provide a basis for a more specific best-practices document.

That said, these general recommendations provide signposts for such a document:

1. Appropriate metadata clearly underpins most discovery or use/reuse scenarios
2. An overall approach to funding should be developed with a broad community in mind
3. Guidance needs to be developed to cover the gap between the knowledge organization system (KOS) and library communities, which may or may not have technical support
4. The role of organizations such as NISO in providing educational support at a level useful for building a community of vocabulary 'experts' should be explored, recognizing that technical expertise is only part of the need
5. Requirements for adequate discovery of extant vocabularies should be defined
6. Best practices for metadata should include requirements for machine- and human-focused documentation

At the top level (for the vocabulary as a whole), the following aspects should be considered minimal:

1. Definition, scope and owner of the vocabulary
 - a) Vocabulary name
 - b) Vocabulary URI
 - c) Approximate (or actual, if available) number of preferred terms, classes, and properties
 - d) Detailed description of subject area covered or a list of topics included
 - e) Vocabulary 'owner' or responsible organization or individual. This may include historical information if there have been changes in responsible parties over time, as well as funding source(s) and contact information
 - f) Status (under development, active, retired, etc.)
 - g) Versioning and maintenance policies
 - h) Languages available

License and specific provisions for use and reuse

- a) If there is no license, this should be declared and any expectations defined
- b) Explicit permissions to: adapt, change/modify, distribute, create derivative works, extend, map
- c) Access policies and preferences, in particular whether a vocabulary should be cached if servers are subject to overload
- d) Delivery mechanisms and file formats available

Members of the NISO community should be lauded for wading into this complex environment. Although the job is not yet done, it is intended that this report should enable next steps with more detailed and experience-based recommendations for those in the community who wish to participate in this environment with confidence.

Appendix A: Terms and Definitions

These terms are defined in the context of vocabularies applied in bibliographic data, not in the broader context of bibliographic information instance data.

Application Profile

In the computer information sciences, an application profile consists of a set of metadata elements, policies, and guidelines defined for a particular application.

Source: [Dublin Core](http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml#A),
<http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml#A>

Application Programming Interface

An Application Programming Interface (API) is a machine endpoint designed to provide programmatic, often high-level, access to data and/or virtual machines via a standardized protocol.

Source: [Wikipedia](https://en.wikipedia.org/wiki/Application_programming_interface), https://en.wikipedia.org/wiki/Application_programming_interface

Authority file

A set of established headings and the cross-references to be made to and from each heading, often citing the authority for the preferred form or variants. Types of authority file include name authority files and subject authority files.

Also referred to as **Value vocabulary**.

Source: [ANSI/NISO Z39.19-2005 \(R2010\)](http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*,
http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

Classes

Groups of resources having certain properties in common and therefore put together as members of one concept. For example, Dublin Core Classes are the terms of the DCMI Type Vocabularies, e.g., `dcmitype:Collection` and `dcmitype:Dataset`.

Source: [Dublin Core](http://dublincore.org/documents/dcmi-terms/#H6), <http://dublincore.org/documents/dcmi-terms/#H6>

Classification scheme

A method of organization according to a set of pre-established principles, usually characterized by a notation system and a hierarchical structure of relationships among the entities. *NOTE: A classification scheme often also includes an index.*

Source: [ANSI/NISO Z39.19-2005 \(R2010\)](http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*,
http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

Commercial use

The carriage of persons or property for any fare, fee, rate, charge or other consideration, or directly or indirectly in connection with any business or other undertaking intended for profit.

Source: [Lectric Law Library](http://www.lectlaw.com/def2/u013.htm), <http://www.lectlaw.com/def2/u013.htm>

Controlled vocabulary

A list of terms that have been enumerated explicitly. This list is controlled by and is available from a controlled vocabulary registration authority. All terms in a controlled vocabulary must have an unambiguous, non-redundant definition. *NOTE: This is a design goal that may not be true in practice; it depends on how strict the controlled vocabulary registration authority is regarding registration of terms into the controlled vocabulary.*

At a minimum, the following two rules must be enforced:

1. If the same term is commonly used to mean different concepts, then its name is explicitly qualified to resolve this ambiguity. *NOTE: This rule does not apply to synonym rings.*
2. If multiple terms are used to mean the same thing, one of the terms is identified as the preferred term in the controlled vocabulary and the others are listed as synonyms or aliases.
Source: [ANSI/NISO Z39.19-2005 \(R2010\), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*](http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf),
http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

Data Preservation

Data preservation is a comprehensive topic, which includes things such as backups, archives, data conversion, reformatting, and rescue.

Source: [DataOne Best Practices: Decide What Data to Preserve](https://www.dataone.org/best-practices/decide-what-data-preserve),
<https://www.dataone.org/best-practices/decide-what-data-preserve>

Datatypes

Syntax Encoding Schemes (SES) are the rules that specify how a value has to be structured. Dublin Core defined Syntax Encoding Schemes as DCMI Metadata Terms, e.g., `dcterms:W3CDTF` for date specification. The relationship between Dublin Core datatypes and properties is as follows: certain properties—date, identifier, etc.—may be typed by a Syntax Encoding Scheme, where the Syntax Encoding Scheme dictates the syntax of the values used with that property.

Source: [Dublin Core User Guide](http://wiki.dublincore.org/index.php/User_Guide), http://wiki.dublincore.org/index.php/User_Guide

Elements

Synonym: Element Sets, Properties

The first version of the Dublin Core properties was called the Dublin Core Element Set (DCMES), and the term continues to be used as an alternative term for properties.

See also: Properties, metadata element sets

Also referred to as: **Structured vocabulary**.

Source: [DCMI Metadata Terms, Section 6: Classes](http://dublincore.org/documents/dcmi-terms/#H6), <http://dublincore.org/documents/dcmi-terms/#H6>

Entity

Something that has separate and distinct existence and objective or conceptual reality.

Source: [Merriam-Webster](http://www.merriam-webster.com/dictionary/entity), <http://www.merriam-webster.com/dictionary/entity>

Exchange format

A machine-readable format for representing information, which is intended to facilitate exchange of the information between applications.

Source: [ISO 25964—the international standard for thesauri and interoperability with other vocabularies](http://www.niso.org/schemas/iso25964/), <http://www.niso.org/schemas/iso25964/>

Interoperability

The ability of two or more systems or components to exchange information and to use the information that has been exchanged.

Source: [ISO 25964—the international standard for thesauri and interoperability with other vocabularies](http://www.niso.org/schemas/iso25964/), <http://www.niso.org/schemas/iso25964/>

Knowledge management

The process of capturing, developing, sharing, and effectively using organizational knowledge. It refers to a multi-disciplinary approach to achieving organizational objectives by making the best use of knowledge.

Source: [Wikipedia](https://en.wikipedia.org/wiki/Knowledge_management), https://en.wikipedia.org/wiki/Knowledge_management

Local purposes

Use by something or someone of a particular region or part, or to each of any number of these.

Source: [Oxford Dictionaries](https://en.oxforddictionaries.com/definition/us/local), <https://en.oxforddictionaries.com/definition/us/local>

Also referred to as **Local use**

Linked Open Data

Tim Berners-Lee outlined four principles of linked data in his "Linked Data" note of 2006, paraphrased along the following lines:

1. Use uniform resource identifiers (URIs) to name (identify) things.
2. Use Hypertext Transfer Protocol (HTTP) URIs so that these things can be looked up (interpreted, "dereferenced").
3. Provide useful information about what a name identifies when it's looked up, using open standards such as Resource Description Framework (RDF), SPARQL Protocol, RDF Query Language (SPARQL), etc.
4. Refer to other things using their HTTP URI-based names when publishing data on the Web.

Berners-Lee later added a fifth rule—create open content—to define linked open data.

Source: [W3.org](https://www.w3.org/DesignIssues/LinkedData.html), <https://www.w3.org/DesignIssues/LinkedData.html>

Map (verb)

To establish relationships between the concepts of one vocabulary and those of another.

Source: [ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies](https://www.iso.org/iso/catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53658) First Edition, 2013-03-15, https://www.iso.org/iso/catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53658

Mapping (gerund/verbal noun)

The process of establishing relationships between the concepts of one vocabulary and those of another.

Source: [ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies](https://www.iso.org/standard/53658.html) First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Mapping (noun)

Product of mapping process. Relationship between a concept in one vocabulary and one or more concepts in another.

NOTE 1: A mapping generally has a direction.

NOTE 2: Exceptionally, a mapping may involve a combination of two or more target vocabularies, where one or more of them act(s) as qualifier to the other.

Source: [ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies](https://www.iso.org/standard/53658.html) First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Metadata

Data that identify attributes of a document; typically used to support functions such as location, discovery, documentation, evaluation, and/or selection.

NOTE: Preferred terms or notations selected during the indexing process are commonly applied as metadata values.

Source: ISO 25964-2:2013, *Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies* First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Name authority list

Controlled vocabulary for use in naming particular entities consistently.

NOTE: The entities in question are unique individuals, such as Benjamin Disraeli, Kilimanjaro, or the Bayeux Tapestry, rather than classes such as politicians, mountains, or embroideries. A name authority list may also be known as a name authority file. A name authority list is sometimes referred to simply as an authority list.

Source: ISO 25964-2:2013, *Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies* First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Ontology

In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse. It is thus a practical application of philosophical ontology, with a taxonomy.

Source: Wikipedia, [https://en.wikipedia.org/wiki/Ontology_\(information_science\)](https://en.wikipedia.org/wiki/Ontology_(information_science))

Also referred to as: **Structured vocabulary**

Source: ISO 25964-2:2013, *Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies* First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Properties

Elements are the "core" attributes of resources, used for the uniform structured resource description. Properties like dc:title, dc:creator, etc. are defined in the Dublin Core Metadata Element Set (DCMES), which is the Set of the fifteen generic elements created in the 1990s. And properties like dcterms:title, dcterms:alternative, etc. are defined as DCMI Metadata Terms - also known as subproperties or refinements of the DCMES. In record based metadata systems properties are usually called metadata fields.

Also referred to as: **Elements, element sets**

Source: Dublin Core User Guide, http://wiki.dublincore.org/index.php/User_Guide

Proprietary vocabulary

A vocabulary that is used, produced, or marketed under exclusive legal right of the inventor or maker.

Source: Merriam-Webster, <http://www.merriam-webster.com/dictionary/proprietary>

Public vocabulary

A vocabulary that has been published or broadcast for public consumption, is available on request to the public, is accessible on-line or otherwise to the public.

Source: IT Law Wiki, http://itlaw.wikia.com/wiki/Publicly_available_information

Schedule

Terms, notations, captions, cross-references and scope notes set out to exhibit the content and structure of a structured vocabulary.

Source: ISO 25964-2:2013, *Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies* First Edition, 2013-03-15, http://itlaw.wikia.com/wiki/Publicly_available_information

Search thesaurus

Vocabulary intended to assist searching even though it has not been used to index the documents being searched.

NOTE: Search thesauri are designed to facilitate choice of terms and/or expansion of search expressions to include terms for broader, narrower or related concepts, as well as synonyms. Optionally, a thesaurus complying with part 2 of ISO 25964—the international standard for thesauri and interoperability with other vocabularies can be used as a search thesaurus.

Source: ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Semantic Web

The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF).

Source: Wikipedia, https://en.wikipedia.org/wiki/Semantic_Web

Structured vocabulary

An organized set of terms, headings or codes representing concepts and their inter-relationships, which can be used to support information retrieval.

Source: ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

*Also referred to as: **Element vocabulary (or element set), schema vocabulary, property vocabulary***

*Also referred to as: **Metadata element sets***

A metadata element set defines classes and attributes used to describe entities of interest. In Linked Data terminology, such element sets are generally made concrete through RDF Schemas or OWL Web Ontology Language⁶⁷ ontologies, the term "RDF vocabulary" often being used as an umbrella for these. Usually a metadata element set does not describe bibliographic entities, rather it provides elements to be used by others to describe such entities.

Source: W3C Library Linked Data Incubator Group, https://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/#Value_vocabularies

Subject heading scheme

Synonyms: Subject heading language, Subject heading list, SHL

Structured vocabulary comprising terms available for subject indexing, plus rules for combining them into pre-coordinated strings of terms where necessary.

Source: ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Synonym ring

A group of terms that are considered equivalent for the purposes of retrieval.

Source: ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, http://www.niso.org/apps/group_public/download.php/12059/z39-19-2005r2010.pdf

⁶⁷ W3C. OWL Web Ontology Language Overview. (February 10, 2004). <https://www.w3.org/TR/owl-features/>

Taxonomy

A collection of controlled vocabulary terms organized into a hierarchical structure. Each term in a taxonomy is in one or more parent/child (broader/narrower) relationships to other terms in the taxonomy.

Source: [ANSI/NISO Z39.19-2005 \(R2010\)](http://www.niso.org/apps/group_public/download.php/12059/z39-19-2005r2010.pdf), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, http://www.niso.org/apps/group_public/download.php/12059/z39-19-2005r2010.pdf

Terminology

Set of designations belonging to one special language.

Source: [ISO 25964-2:2013](https://www.iso.org/standard/53658.html), *Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies* First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

Thesaurus

Plural: thesauruses, thesauri

A book that lists words in groups of synonyms and related concepts.

Source: [Oxford Dictionaries](https://en.oxforddictionaries.com/definition/thesaurus), <https://en.oxforddictionaries.com/definition/thesaurus>

Value vocabulary

Vocabularies used to express concepts or values in instance data. DCMI uses the term Vocabulary Encoding Scheme for these vocabularies, which assumes more internal structure than a term list in text. (Example: Getty Art & Architecture Thesaurus). A value vocabulary defines resources (such as instances of topics, art styles, or authors) that are used as values for elements in metadata records. Typically, a value vocabulary does not define bibliographic resources such as books but rather concepts related to bibliographic resources (persons, languages, countries, etc.). The resources in a value vocabulary are "building blocks" with which metadata records can be populated. Many libraries mandate specific vocabularies for selecting values for a particular metadata element. A value vocabulary thus represents a controlled list of allowed values for an element. Examples include: thesauri, code lists, term lists, classification schemes, subject heading lists, taxonomies, authority files, digital gazetteers, concept schemes, and other types of knowledge organization systems. To be useful for linking of data, value vocabularies should have Hypertext Transfer Protocol (HTTP)⁶⁸ Uniform Resource Identifiers (URIs)⁶⁹ assigned for each value; these URIs would then appear in a metadata record instead of or in addition to the literal value.

Also referred to as: **Content vocabulary, concept vocabulary, controlled vocabulary**

Source: [W3C Library Linked Data Incubator Group](https://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/#Value_vocabularies), https://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/#Value_vocabularies

Vocabulary

Words used on a particular occasion or in a particular sphere: *the vocabulary of law; the term became part of business vocabulary*

Source: [Oxford Dictionaries](http://www.oxforddictionaries.com/us/definition/american_english/vocabulary), http://www.oxforddictionaries.com/us/definition/american_english/vocabulary

⁶⁸ W3C. HTTP: Hypertext Transfer Protocol. (June 11, 2014). <https://www.w3.org/Protocols/>

⁶⁹ Wikipedia. Uniform Resource Identifier. (May 29, 2017). https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

Vocabulary adaptation and extension

Adaptation of vocabularies is often in the form of extension of extant vocabularies to meet needs of communities that wish to use the base vocabulary but need to fill in gaps to make the vocabulary usable in their context. Extensions may begin with new properties or sub-property extensions based on an existing property. Various issues should be considered when adapting or extending an existing vocabulary: Is permission required from the original owner/publisher? Are there copyright or licensing legal positions that need to be addressed? Is it desirable, or even possible, to work within the existing vocabulary, or should a new version be generated?

Source: NISO Issues in Vocabulary Management Working Group

Vocabulary alignment and mapping

Most vocabularies are not developed or used in a vacuum, and in a rich and varied vocabulary environment there is generally a need to explore and document relationships between vocabularies in use. The process of vocabulary alignment could result in shareable maps that express those relationships in ways that allow machines to make connections, whether for discovery, translation, or migration. There are difficulties in managing this, including how to maintain the alignment as the aligned vocabularies evolve, and how to document the alignment. Key requirements for documentation include provenance and information about any tools that were used if the alignment was performed automatically.

Source: NISO Issues in Vocabulary Management Working Group

Vocabulary control

Management of a vocabulary in order to disambiguate and constrain the form of the terms and limit the number of concepts and terms available for indexing.

Source: [ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies](#) First Edition, 2013-03-15, <https://www.iso.org/standard/53658.html>

More simply, the process of organizing a list of terms (a) to indicate which of two or more synonymous terms is authorized for use; (b) to distinguish between homographs; and (c) to indicate hierarchical and associative relationships among terms.

Also referred to as: **Controlled vocabulary**.

Source: [ANSI/NISO Z39.19-2005 \(R2010\), Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies](#), http://www.niso.org/apps/group_public/download.php/12059/z39-19-2005r2010.pdf

Vocabulary Encoding Scheme (VES)

Concept Scheme identifies controlled vocabularies—such as thesauri, classifications, subject headings, taxonomies, etc.—whose terms may be used as values.

Source: [Dublin Core User Guide](#), http://wiki.dublincore.org/index.php/User_Guide

Vocabulary interoperability

The ability to exchange and use information from a vocabulary (usually in a large heterogeneous network made up of several local area networks).

Source: [Vocabulary.com](#), <https://www.vocabulary.com/dictionary/interoperability>

Vocabulary maintenance

Maintenance is the process of keeping a vocabulary in line with changes in term usage, encoding practice, and understanding as well as with the needs in the community. Good maintenance practices should be based on openly available policies that enable distributed and configurable notifications for known users of the vocabulary. Appropriate versioning

practices should be included in maintenance policies.

Source: NISO Issues in Vocabulary Management Working Group

Vocabulary ownership

The act of owning and controlling a vocabulary.

Source: Vocabulary.com, <https://www.vocabulary.com/dictionary/ownership>

Vocabulary reuse

The practice of adopting or adapting general or specialized vocabularies used by other projects or communities in order to avoid the costs of building and maintaining new vocabularies (accepting the risk that reused vocabularies may develop in non-ideal directions) and to enhance alignments with extant vocabularies, up to and including maps that can be used by others.

Source: NISO Issues in Vocabulary Management Working Group

Vocabulary use

The determination and application of appropriate vocabularies to structure or include within descriptive data. Evaluation of vocabularies in the context of the organization or project might include identifying licenses, maintenance policies, and fit.

Source: NISO Issues in Vocabulary Management Working Group

Appendix B: Vocabularies and Tools

Vocabulary Directories, Repositories or Collections

AberOWL

<http://aber-owl.net>

This ontology repository was jointly developed by researchers from King Abdullah University of Science and Technology, the Centre for Computational Biology at the University of Birmingham, and the University of Cambridge. AberOWL hosts 523 ontologies in primarily biomedical subjects, and registered users may upload ontologies to the site. The website has a synonym ring controlled vocabulary to support searching for ontologies by topics. Overview metadata for each ontology comprises homepage URL, contact email, Topics, Species, Class count, CreatedBy, OWL version, comment, and label. Ontologies can all be browsed and downloaded in the ont format, and SPARQL endpoints are provided for each. Some ontologies can be dynamically visualized.

ANDS (Australian National Data Service, Research Vocabularies Australia)

<https://vocabs.ands.org.au/>

Research Vocabularies Australia, formally launched in September 2015, is the "controlled vocabulary discovery service" of the Australian National Data Service (ANDS), supported by the Australian Government. Research Vocabularies Australia comprises 72 vocabularies, published by governmental and nongovernmental organizations, mostly in the sciences but intended to "grow to cover a broad spectrum of research fields—across sciences, social sciences, arts and humanities." Vocabularies are searchable by the following metadata: Subject (not controlled and can be multi-valued), Publisher, Language, and License. Additional metadata is provided for Format and Access. Options for access are API/SPARQL, direct download, and online.

Athena Plus, Access to Cultural Heritage Networks for Europeana

<http://www.athenaplus.eu/>

AthenaPlus is a best practice Europeana network composed of 41 partners from 21 member states with the main objectives being [1] to contribute more than 3.6 million metadata records, from both the public and private sectors, focusing mainly on museum content; [2] to improve search, retrieval and re-use of content; improve multilingual terminology management; and offer a SKOS export and publication tool/API for content providers; and [3] to experiment with enriched metadata adapted for users with different needs (e.g., tourists, schools, and scholars). ATHENA was built on the perceived success of earlier projects—LIDO and the ATHENA Ingestion Server and Mapping Tool (MINT), which were developed in order to further advance and complete the effective infrastructure and tools for making digital content available through Europeana. AthenaPlus started in March 2013 and was funded through August 2015.

BARTOC (Basel Register of Thesauri, Ontologies & Classifications)

<http://bartoc.org/>

BARTOC is a comprehensive database/terminology registry of all kinds of knowledge organization systems, created and managed by the Basel University Library. The database lists more than 2,700 vocabularies of all kinds, in all languages, in all subject areas, in any publication format, and in any form of accessibility. The database of knowledge organization systems is hosted on Drupal and the metadata for the vocabularies is standardized and made available as Linked Open Data, comprising the fields of: URI, Title, Alternate and/or

English Title, Author, Abstract, Coverage, Type, Format, Size, Licenses, Access, Dewey Decimal Classification subject, Dewey Decimal Classification Main Class, Wikidata identifier, URL link, Language, Topic, Year of Creation, and publisher location. Bartoc is also distinguished by its use of several controlled vocabularies to index vocabularies, including Dewey Decimal Classification, multilingual EuroVoc descriptors, and Dublin Core. Controlled vocabularies are also used to support searching by format, license, and language. Mappings to KOS records in Wikidata provide links to Wikipedia articles. BARTOC coordinates its development with the International Society for Knowledge Organization (ISKO) and Project coli-conc (<https://coli-conc.gbv.de/>).

Finto

<http://finto.fi/en>

Finnish thesaurus and ontology service Finto enables publication and browsing of vocabularies and offers interfaces for integrating the thesauri and ontologies into other applications and systems. It contains 33 Finnish ontologies listed under the categories of: General, Society, Geography, Science and Medicine, Art and Culture, and Languages and Literature. Ontologies are either in Finnish only or are bilingual/multilingual in Finnish, English, and/or Swedish. Finto provides a REST interface and a SPARQL endpoint, so the ontologies can be integrated into other applications. All of the software is open source.

Heritage Data

<http://www.heritagedata.org/>

Heritage Data was developed by the Forum on Information Standards in Heritage (FISH), a group comprised of 13 UK cultural heritage organizations and agencies and the Historic Environment Information Resources Network (HEIRNET). The British governmental organization Arts & Humanities Research Council (AHRC) funded Semantic ENrichment Enabling Sustainability of arCHAEological Links (SENESCHAL), "a Knowledge Exchange project based on enhanced vocabulary services that aims to make it significantly easier for vocabulary providers to make their vocabularies available as Linked Data."

All of the vocabularies are available to download as SKOS RDF and also as PDFs of the alphabetical thesaurus and hierarchies of terms for viewing. Complete Linked Data URL/URI information can be downloaded in RDF formats of N-Triples, Turtle, JSON, or XML.

NCBO Bioportal

<http://bioportal.bioontology.org/>

Bioportal is a biomedical ontology repository service of the NIH Funded National Center for Biomedical Ontology (NCBO). There are hundreds of ontologies, many of which can be downloaded directly from the site. The vocabularies can be searched by Category, Group, and Format. Metadata filters (for sorting) the list of ontologies are by Popular[ity], Size, Projects, Notes, and Upload date. A great deal of metadata and summary information is provided for each vocabulary, including history of uploads of versions and a graph of downloads over time. The following metadata is provided in varying degrees of detail: Acronym, Visibility (e.g., "public"), Bioportal PURL, Description, Status, Format, Contact, Home page, Publication page, Documentation page, Categories, Groups, and License information.

ONKI - Finnish Ontology Library Service

<http://seco.cs.aalto.fi/services/onki/>

The ONKI repository is managed by the Semantic Computing Research Group in cooperation with the University of Helsinki and Aalto University. It contains Finnish and international ontologies and vocabularies and thesauri suitable for publishing content on the

Semantic Web. ONKI currently lists 87 vocabularies in Finnish and/or English, each of which can be searched on the site, and 68 of which can be downloaded from the site in RDF/XML.

Ontobee

<http://www.ontobee.org>

Designed for ontologies, Ontobee is a linked data server hosted by Dr. Yongqun "Oliver" He's laboratory research group at the University of Michigan. Ontobee aims to facilitate ontology data sharing, visualization, query, integration, and analysis. It dynamically dereferences and presents individual ontology term URIs to HTML webpages for user-friendly Web browsing and navigation, and to RDF source code for Semantic Web applications. Ontobee is the default linked data server for most [OBO Foundry library ontologies](#). It provides a tabular list of 181 biomedical ontologies, which can each be individually searched and browsed for terms on the Ontobee website. Ontologies (with OWL extension) can be downloaded, and lists of terms can be downloaded as Excel spreadsheets or as text. For each ontology, URLs are provided for IRI, downloading, documentation, and publisher homepage, along with an email contact.

Ontology Lookup Service

<http://www.ebi.ac.uk/ols>

Ontology Lookup Service (OLS) is a repository for biomedical ontologies that was developed and is maintained by the Samples, Phenotypes and Ontologies Team at European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) with EU grant funding. OLS hosts 523 ontologies in primarily biomedical subjects. The ontologies are listed in a table that can be sorted by column headers for ontology name, acronym, description, or load date. Overview information for each ontology may include ontology IRI, Version IRI, Ontology id, version date, and last load date, with variable information on number of terms, license, editor's note, default language, contributor, comments, and creator names. All ontologies can be browsed, searched, and downloaded in the .owl format and include links to the ontology homepages.

Taxonomy Warehouse

<http://www.taxonomywarehouse.com/>

Taxonomy Warehouse is an information website directory dedicated, as the name suggests, to taxonomy. The site includes an extensive directory of taxonomies, events, organizations, publishers, books, and other publications as well as blogs, list of individuals related to taxonomy, and a marketplace. Launched in 1999, this longstanding, respected site is free to users and vocabulary publishers. There are no defined criteria for inclusion other than the availability of sufficient information on what may make a vocabulary useful to a particular community. Current Metadata may include vocabulary name, publisher, type, description, URL for information, URL for online access, notation scheme, term count, revision cycle, languages, and subject categories.

Selected Examples of Available Tools for Vocabulary Development

This list is not intended to be exhaustive but to highlight some tools that could be explored by projects or groups beginning the task of vocabulary development.

Athena Plus tool: TMP2

<http://www.athenaplus.eu/index.php?en/212/tmp2>

The TMP2 (Thesaurus Management Platform) is a Web portal for thesauri management. It was developed in the framework of two European projects dedicated to cultural heritage and is freely accessible (read only) and utilized via user accounts. TMP2 offers creation and editing of thesauri, specification of all types of hierarchical relations, extensive concept information, management of multilingualism, integration of ISO standards, import and export in SKOS, JSON, and RDF as well as thesaurus mapping. The site includes extensive documentation on projects and uses for the tool.

Fluent Editor and Fluent Editor 2015

<http://www.cognitum.eu/Semantics/FluentEditor/>

This ontology editor, which uses Controlled Natural Language, is a comprehensive tool for editing and manipulating complex ontologies. Fluent Editor provides an alternative to XML-based OWL editors and is considered more suitable for human-machine interaction. Its main feature is the usage of Controlled English as a knowledge-modeling language. Supported via Predictive Editor, it prohibits the entering of a sentence that is grammatically or morphologically incorrect and actively helps the user during sentence writing. Facilitating collaborative editing and instant testing and allows custom plugins, Fluent Editor is compatible with OWL 2, OWL-DL, OWL-RL, SWRL, SPARQL, RDF, and OCNL. The editor is free for individual developers, open source projects, academic research, education, and small professional teams.

Git/Github

<https://github.com/>

GitHub is a Web-based [Git](#) repository⁷⁰ hosting service⁷¹, mostly used for source code though increasingly for documentation, projects, and vocabularies. It offers all of the distributed version control⁷² and source code management⁷³ functionality of [Git](#) (a version control system) as well as adding its own features. It provides access control⁷⁴ and several social networking-like functions, such as issue tracking, intended to support communities and projects. GitHub also allows registered and non-registered users to browse public repositories on the site, although a user must create an account to contribute to the site. Multiple desktop clients and Git plugins that integrate with the platform have been created by GitHub and third parties.

Knoodl

<http://www.knoodl.com/>

Knoodl is a Distributed Information Management System (DIMS)TM that leverages existing Web infrastructure. Knoodl contains tools for creating, managing, analyzing, and visualizing RDF/OWL descriptions. It includes features that facilitate collaboration in all stages of these activities. Knoodl's key component is a semantic software architecture that supports

⁷⁰ Wikipedia. Repository (Version Control). (September 17, 2016).

[https://en.wikipedia.org/wiki/Repository_\(version_control\)](https://en.wikipedia.org/wiki/Repository_(version_control))

⁷¹ Wikipedia. Internet Hosting Service. March 27, 2017. https://en.wikipedia.org/wiki/Internet_hosting_service

⁷² Wikipedia. Repository (Version Control). (September 17, 2016).

[https://en.wikipedia.org/wiki/Repository_\(version_control\)](https://en.wikipedia.org/wiki/Repository_(version_control))

⁷³ Ibid.

⁷⁴ Wikipedia. Access Control. (June 2, 2017). https://en.wikipedia.org/wiki/Access_control

Emergent Analytics. Knoodl is hosted in the Amazon EC2 cloud and can be used for free. It may also be licensed (agreement to Terms of Use) for private activity behind a Firewall as MyKnoodl, currently with no fees.

NeOn Toolkit

<http://neon-toolkit.org>

The NeOn Toolkit is the ontology engineering environment originally developed as part of the [NeOn Project](#) and now supported by the NeOn Foundation. This toolkit is a state-of-the-art, open source multi-platform ontology engineering environment that provides comprehensive support for the ontology engineering life-cycle. The toolkit is based on the Eclipse platform, a leading development environment, and provides an extensive set of plug-ins covering a variety of ontology engineering activities.

Open Metadata Registry (OMR)

<http://metadataregistry.org>

OMR is a simple free tool for building vocabularies (structural in RDFS and concept in SKOS). It is currently being expanded to support the more complex requirements for projects, translated vocabularies, and spreadsheet based updating. OMR also provides hosting services for vocabularies under its own URI and can assist those who wish to use their own namespace to provide redirects and resolution services through the OMR.

OpenSKOS

<http://openskos.org/>

OpenSKOS, part of the [CATCHPlus](#) project, which developed products to promote cooperation and consistency in the information infrastructure of the heritage sector, is an Exchange and Publication Platform for SKOS Vocabularies. OpenSKOS software has an editor that can be utilized by login as a "tenant"; as SaaS (Software as a Service); or, by downloading the source code, to manage SKOS vocabularies. The OpenSKOS software also supports OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) with two [metadataPrefixes](#): *oai_dc* (Open Archives Initiative Dublin Core) and *oai_rdf* (Open Archives Initiative Resource Description Framework). It provides API's on SKOS (Simple Knowledge Organization System)⁷⁵ including "Find Concepts," "Create/Retrieve/Update/Delete Concepts," "Institutions," and "Collections."

OWLGrEd

<http://owlgred.lumii.lv/>

This free UML style graphical editor for [OWL](#) ontologies has additional features for graphical ontology exploration and development. It is associated with the Institute of Mathematics and Computer Science, University of Latvia.

Protégé

<http://protege.stanford.edu/>

Protégé is a free, Java-based, open-source plug-in architecture that can be adapted to build both simple and complex ontology-based applications. Developers can integrate the output of Protégé with rule systems or other "problem solvers" to construct a wide range of intelligent systems. Developed by the Stanford Center for Biomedical Informatics Research, a desktop version and Web version are currently available. Protégé supports the latest OWL 2 Web Ontology Language and RDF.

⁷⁵ W3C. "SKOS Simple Knowledge Organization System." (December 13, 2012).
<https://www.w3.org/2004/02/skos/>

SCENT for GLAM (Galleries, Libraries, Archives and Museums)

<http://www.scent-glam.eu/#/h>

SCENT (Semantic and Collaborative Environment for a Network of Terminologies) is a tool that allows users to create or import, SKOSify, and edit a terminology; map concepts from one terminology to another; and export terminology and mappings in SKOS. SCENT is developed in accordance with Semantic Web standards.

SKOS Shuttle

<https://skosshuttle.ch>

SKOS Shuttle calls itself “Thesaurus Management as a (collaborative) service,” and is managed by Semweb LLC. Although it makes available a ‘free start’ and has a list of other levels, costs are not clearly delineated. The service advertises several aspects of its services, including RDF editing (in two modes), integration of OWL Ontologies, customized relations and attributes, and an Orphan Concept Analysis for quick ‘deorphanization,’ as well as other special services.

Semantic Turkey

<http://semanticturkey.uniroma2.it/>

Semantic Turkey is a platform for Knowledge Acquisition and Management realized by the Artificial Intelligence Research Group at the University of Rome, Tor Vergata. Having adopted W3C standards for knowledge representation belonging to the RDF family, Semantic Turkey turns the popular Firefox Web browser into a rich and extensible framework. A Firefox-based user interface can keep track of relevant information from visited websites and organize collected content according to imported/newly created ontologies. Domain experts and ontology developers can thus build ontologies starting from source information they find on the Web without any need to interconnect heterogeneous tools and applications.

Skosmos

<http://skosmos.org/>

This open source Web-based SKOS browser and publishing tool is currently used by Finto, FAO / AGROVOC, Rhineland-Palatinate spatial data initiative classifications, UNESCO Thesaurus, and the University of Oslo Library thesauri. Features include search and browse vocabularies, an alphabetical index, a thematic index, structured concept display, a visualized concept hierarchy, and a multilingual user interface.

Vitro

<http://vitro.mannlib.cornell.edu/>

First developed for a research and scholarship portal at Cornell University, Vitro is a general-purpose Web-based ontology and instance editor with customizable public browsing. It is a Java Web application that allows creating or loading ontologies in OWL format, editing instances and relationships, and building a public website to display data. Data may be indexed and searched via Apache Solr. The latest source code may be downloaded from the VIVO project's GitHub repository.

Vocbench

<http://vocbench.uniroma2.it/>

VocBench is a web-based, multilingual, collaborative development platform for managing SKOS thesauri. Designed to meet the needs of semantic web and linked data environments, VocBench development has also been driven by the feedback gathered from a community of users made of public organizations, companies and independent users looking for open source solutions for maintaining their ontologies, thesauri, code lists, and authority resources.

VocBench's tools and functionalities facilitate both collaborative editing and multilingual terminology. It also includes administration and group management features that permit flexible roles for maintenance, validation, and publication.

VocBench 2 was developed as a collaboration between the Food and Agriculture Organization of the United Nations (FAO) and the ART Group of the University of Rome, Tor Vergata.

Appendix C: Additional Resources

- Abelson, Hal, Ben Adida, Mike Linksvayer, and Nathan Yergler. "ccREL: The Creative Commons Rights Expression Language." (2008).
<https://www.w3.org/Submission/ccREL/>
- Allemang, Dean and James Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 2nd Ed. (Burlington, MA: MorganLib Kaufmann, 2011).
- Arp, Robert, Barry Smith and Andrew D. Spear. *Building Ontologies with Basic Formal Ontology*. (Cambridge, MA: MIT Press, 2015).
- Baker, Thomas, Emmanuelle Bermès, Karen Coyle, Gordon Dunsire, Antoine Isaac, Peter Murray, Michael Panzer, et al. "W3C Library Linked Data Incubator Group: Final Report, 25 October 2011." (Accessed Sept. 10, 2016).
<https://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>
- Berners-Lee, Tim. "Cool URIs Don't Change." (1998). <https://www.w3.org/Provider/Style/URI>
- CITEC, DERI, UPM and DFKI. "Lemon - The Lexicon Model for Ontologies." (n.d.)
<http://lemon-model.net/index.php>
- Commission for the Management and Application of Geoscience Information (CGI):
<http://www.cgi-iugs.org/>
- DCMI NKOS Task Group. "Developing a Dublin Core Application Profile for KOS Resources". Available at:
http://wiki.dublincore.org/index.php/DCMI_NKOS_Task_Group. Last modified Sept. 30, 2015.
Example of an AP, also includes use cases and a proposed list of elements.
- DCMI Vocabulary Management Options [Last updated in 2011]. Accessed Sept. 10, 2016.
http://wiki.dublincore.org/index.php/Vocabulary_Management_Options
- Dublin Core User Guide: http://wiki.dublincore.org/index.php/User_Guide. Accessed June 13, 2017.
- Dunsire, Gordon, Diane Hillmann, and Jon Phipps, "Reconsidering Universal Bibliographic Control in Light of the Semantic Web", *Journal of Library Metadata*, 12, no. 2-3 (2012): 164-176. <http://hdl.handle.net/1813/36288>
- Europeana. "White Paper on Best Practices for Multilingual Access to Digital Libraries." (n.d.) <https://docs.google.com/document/d/1pxp3JS72odvUc2PYn3-0gwXY-8WrmJhzbcKm9wnNueM>
- Geoscience Terminology Working Group (GTWG). (2016). http://www.cgi-iugs.org/tech_collaboration/geoscience_terminology_working_group.html
- Getty Research Institute. (n.d.). <http://www.getty.edu/research/>
- Halpin, Harry and Tom Baker. "Vocabulary Hosting: A Modest Proposal," 2010 AAAI Spring Symposium: Linked Data Meets Artificial Intelligence 2010.
<http://www.aaai.org/ocs/index.php/sss/sss10/paper/download/1140/1450This>
- Harvard Law Today. "Harvard's Perma.cc Receives Grant to Expand its Tools for Saving Sources on the Web." (April 14, 2016). <http://today.law.harvard.edu/harvards-perma-cc-receives-grant-expand-tools-saving-sources-web-2/>

- Hyland, Bernadette, Ghislain Atemezang, and Boris Villazón-Terrazas (eds.) "Best Practices for Publishing Linked Data." W3C Working Group Note (January 9, 2014). Accessed September 10, 2016. <http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/>
- Isaac, Antoine, William Waites, Jeff Young, and Marcia Zeng. "Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets: W3C Incubator Group Report 25 October 2011." Accessed Sept. 10, 2016. <https://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/>
- Joint Steering Committee for the Development of RDA. International Conference on the Principles and Future Development of AACR. (July 1, 2009). <http://www.rda-jsc.org/archivedsite/intlconf1.html>
- Library of Congress. "Library of Congress Subject Headings (LCSH)." Accessed Sept. 15, 2016. <http://id.loc.gov/authorities/subjects.html>
- Library of Congress. "*Library of Congress Subject Headings*" Pre- vs. Post-Coordination and Related Issues." Report for Beacher Wiggins, Director, Acquisitions & Bibliographic Access Directorate, Library Services, Library of Congress. Prepared by the Cataloging Policy and Support Office. March 15, 2007. https://www.loc.gov/catdir/cpsa/pre_vs_post.pdf
- Library of Congress. "MARC 21 Format for Bibliographic Data." Accessed Sept. 15, 2016. <https://www.loc.gov/marc/bibliographic>
- Lóscio, Bernadette Farias, Caroline Burle, Newton Calegari (eds.). "Data on the Web Best Practices," W3C Candidate Recommendation. (August 30, 2016). Accessed September 8, 2016. <https://www.w3.org/TR/dwbp/>
- National Center for Biomedical Ontology. NCBO BioPortal. (n.d.). <http://bioportal.bioontology.org/>
- Open Semantic Framework. "Ontology Tools" [last major update 2010]. http://wiki.opensemanticframework.org/index.php/Ontology_Tools. Accessed Sept. 9 2016.
A list, classification, and description of tools for ontology development, editing, mapping, visualization, analysis, etc. Note that this resource's most recent major update was in August 2010 and most recent edit in November 2014.
- Pattueli, M. Cristina, Alexandra Provo, and Hilary Thorsen. "Ontology Building for Linked Open Data: A Pragmatic Perspective," *Journal of Library Metadata*, 15 (2015): 265-294. DOI: <http://dx.doi.org/10.1080/19386389.2015.1099979>
- Research Data Alliance (RDA). RDA Registry. (February 3, 2017). <http://rdaregistry.info>
- RDA. RDA Registry. <http://www.rdaregistry.info/>
- RDA. RDA Toolkit. (n.d.). <https://access.rdatoolkit.org/>
- Sikos, Leslie F. Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data. (Apress, 2015). Accessed July 17, 2016. <https://www.safaribooksonline.com/library/view/mastering-structured-data/9781484210499/>
- St. Pierre, Margaret and William P. LaPlant, Jr. "Issues in Crosswalking Content Metadata Standards." (Oct. 15, 1998). http://www.niso.org/publications/white_papers/crosswalk/
- U.S. Open Data Institute. "Simple Open Data." (n.d.). <http://simpleopendata.com/>

- W3C. "Best Practice Recipes for Publishing RDF Vocabularies." (August 28, 2008).
<https://www.w3.org/TR/swbp-vocab-pub/>
- W3C. "Cool URIs for the Semantic Web." (December 3, 2008).
<https://www.w3.org/TR/cooluris/>
- W3C. "Generating HTML Documentation of OWL." (Last updated June 2, 2012).
https://www.w3.org/2011/prov/wiki/Generating_HTML_documentation_of_OWL
- W3C. LLD/Library Data Resources. (October 21, 2011).
https://www.w3.org/2001/sw/wiki/LLD/Library_Data_Resources
- W3C. "W3C VocabMgtWhitePaper: White Paper on Vocabulary Management." (Last updated September 4, 2009). <https://www.w3.org/wiki/VocabMgtWhitePaper>
A brief white paper that sketches a policy for versioning vocabularies, modeled after those found in W3C documents (e.g., SKOS) and DCMI.
- W3C VocabularyMarket (Last updated in 2011). (Last updated April 1, 2013).
<https://www.w3.org/wiki/VocabularyMarket>

Bibliography

- American Library Association, Canadian Library Association, and the Chartered Institute of Library and Information Professionals. *Anglo-American Cataloging Rules, Second Edition* (AACR2). <http://www.aacr2.org/>
- Baker, Thomas and Johannes Keizer. "Linked Data for Fighting Global Hunger: Experiences in Setting Standards for Agricultural Information Management," in *Linking Enterprise Data* (Springer, 2010): 177-201. Accessed January 2017. <http://www.fao.org/docrep/014/am324e/am324e.pdf>
- Baker, Thomas, Pierre-Yves Vandenbussche, and Bernard Vatant. "Requirements for Vocabulary Preservation and Governance," *Library Hi Tech*, 31, no. 4 (November 2013): 657-668. <http://dx.doi.org/10.1108/LHT-03-2013-0027>
- Bandrowski A, R. Brinkman, M. Brochhausen, M.H. Brush, B. Bug, M.C. Chibucos, Kevin Clancy, et al. "The Ontology for Biomedical Investigations." PLoS ONE 11, no. 4 (2016): e0154556, <https://doi.org/10.1371/journal.pone.0154556>
- Chan, Lois Mai. "Still Robust at 100: A Century of LC Subject Headings." *Library of Congress Information Bulletin* (August 1998), <https://www.loc.gov/loc/lcib/9808/lcsh-100.html>
- Charles, Valentine, and Cécile Devarenne. "Europeana Enriches its Data with the AAT." *EuropeanaPro*. (n.d.) <http://pro.europeana.eu/page/europeana-aat>
- Cobb, Joan. "The Journey to Linked Open Data: The Getty Vocabularies" *Journal of Library Metadata* 15, no. 3-4 (2015): 142-156, <http://www.tandfonline.com/doi/full/10.1080/19386389.2015.1103081>
- Coyle, Karen and Thomas Baker. "Guidelines for Dublin Core Application Profiles". <http://dublincore.org/documents/profile-guidelines/>
- Creative Commons. Share Your Work: Public Domain. (n.d.). <https://creativecommons.org/share-your-work/public-domain/>
- DataONE. "Decide What Data to Preserve." <https://www.dataone.org/best-practices/decide-what-data-preserve>
- de hÓra, Bill. "Vocabulary Design and Integration." (April 28, 2007). http://www.dehora.net/journal/2007/04/data_integration.html
- Dublin Core Metadata Initiative. "DC-2011 Vocabulary Special Session/Meeting Report" (2012), http://wiki.dublincore.org/index.php/DC2011_Vocabulary_Special_Session/Meeting_Report
- Dublin Core Metadata Initiative. "DCMI Metadata Terms." (Date last issued: June 14, 2012). <http://dublincore.org/documents/dcmi-terms/#H6>
- Dublin Core Metadata Initiative. "DCMIType Vocabulary." (October 11, 2010). <http://dublincore.org/documents/2010/10/11/dcmi-type-vocabulary/>
- Dublin Core Metadata Initiative. Guidelines for Dublin Core Application Profiles. (May 18, 2009). <http://dublincore.org/documents/profile-guidelines/>

- Dunsire, Gordon, Corey Harper, Diane Hillmann, and Jon Phipps. "Linked Data Vocabulary Management: Infrastructure Support, Data Integration, and Interoperability," *Information Standards Quarterly* 24, no. 2/3 (Spring/Summer 2012): 4-13.
<http://www.niso.org/publications/isq/2012/v24no2-3/dunsire/>
- Dunsire, Gordon, Diane Hillmann, Jon Phipps, and Karen Coyle. "A Reconsideration of Mapping in a Semantic World," in *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2011*.
<http://dcpapers.dublincore.org/pubs/article/view/3622/1848>
- Food and Agriculture Organization of the United Nations. "AGROVOC Multilingual Agricultural Thesaurus." (Last modified July 15, 2016). <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>
- Getty Institute. Getty Thesaurus of Geographic Names (TGN). (Last updated February 2, 2016). <http://vocab.getty.edu/tgn/>
- Gruber, Thomas R. "A translation approach to portable ontology specifications." *Knowledge Acquisition* 5, no. 2 (1993): 199-220, <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>
- Heath, Tom and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. <http://linkeddatabook.com/editions/1.0/>
- Heery, Rachel and Manjula Patel. "Application Profiles: Mixing and Matching Metadata Schemas," *Ariadne* 25 (September 24, 2000). <http://www.ariadne.ac.uk/issue25/app-profiles>
- Hillmann, Diane, Gordon Dunsire, and Jon Phipps. "Maps and Gaps: Strategies for Vocabulary Design and Development," in *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2013*.
<http://hdl.handle.net/1813/42443>
- Hillmann, Diane, Gordon Dunsire, and Jon Phipps. "Versioning Vocabularies in a Linked Data World," (Paris: IFLA, 2014). <http://hdl.handle.net/1813/40559>
- Institute of Education Resources. <http://www.eric.ed.gov>
- Library of Congress. Program for Cooperative Cataloging (PCC). (n.d.).
<https://www.loc.gov/aba/pcc/>
- Library of Congress Subject Headings (LCSH). (n.d.).
<http://id.loc.gov/authorities/subjects.html>
- Linked Open Vocabularies (LOV). Accessed January 2017. <http://lov.okfn.org/dataset/lov>
- Linkeddata.org. n.d. "Linked Data – Connect Distributed Data Across the Web."
<http://linkeddata.org/>
- LOCKKS. Lots of Copies Keeps Stuff Safe. (n.d.). <https://www.lockss.org/>
- Metadata Matters. Versions and Services, Part 1. (May 15, 2013).
<http://managemetadata.com/blog/2013/07/23/versions-and-services-pt-1/>
- National Information Standards Organization. "ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*" last modified January 6, 2014.
http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf

National Information Standards Organization. ANSI/NISO Z39.19-2005 (R2010). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* (2010).

http://www.niso.org/apps/group_public/download.php/2591/z39-19-2005r2010.pdf

National Information Standards Organization. NISO Bibliographic Roadmap Development Project. (2014). <http://www.niso.org/topics/tl/BibliographicRoadmap/>

OCLC Research. "FAST (Faceted Application of Subject Terminology)." Accessed July 22, 2015. <http://www.oclc.org/research/themes/data-science/fast.html>

Open Metadata Registry (OMR). (n.d.). <http://metadataregistry.org/>

PricewaterhouseCoopers. "Spinning a Data Web." *Technology Forecast*, Spring 2009, p. 4 15.

<https://www.pwc.com/cl/es/publicaciones/assets/pronostico-de-tecnologia.pdf>

RDF Schema 1.1 (RDFS). (February 25, 2014). <https://www.w3.org/TR/rdf-schema/>

Swoogle, Semantic Web Search. 2007. <http://swoogle.umbc.edu>

Vandenbussche, Pierre-Yves and Bernard Vatante. "LOV Metadata Recommendations for Linked Open Vocabularies," version 1.1 (August 19, 2012). Accessed January 2017. http://lov.okfn.org/Recommendations_Vocabulary_Design.pdf

W3C. Data Catalog Vocabulary (DCAT). (January 16, 2014). <https://www.w3.org/TR/vocab-dcat/>

W3C. Data on the Web Best Practices. (Jan. 31, 2017.) <https://www.w3.org/TR/dwbp/>

W3C. HashVsSlash. (March 22, 2016). <https://www.w3.org/wiki/HashVsSlash>

W3C. HTTP: Hypertext Transfer Protocol. (June 11, 2014). <https://www.w3.org/Protocols/>

W3C. OWL Web Ontology Language Overview. (February 10, 2004). <https://www.w3.org/TR/owl-features/>

W3C. RDF Schema 1.1, W3C Recommendation. (February 25, 2014). <https://www.w3.org/TR/rdf-schema/>

W3C. RDF Vocabularies Current Status. (2017). https://www.w3.org/standards/techs/rdfvocab#w3c_all

W3C. RDF Data Shapes Working Group. (June 12, 2017). https://www.w3.org/2014/data-shapes/wiki/Main_Page

W3C. RDFa Core Initial Context. (June 8, 2017). <https://www.w3.org/2011/rdfa-context/rdfa-1.1>

W3C. "SKOS Simple Knowledge Organization System." (December 13, 2012). <https://www.w3.org/2004/02/skos/>

W3C. Vocabulary Search on the Semantic Web for RDFa Default Profiles. (March 1, 2013). <https://www.w3.org/2010/02/rdfa/profile/data/>

W3C. W3C Vocabulary Services. (June 6, 2013). <https://www.w3.org/2013/04/vocabs/>

W3C. "Web Ontology Language." (December 11, 2012). <https://www.w3.org/OWL/>

Wikipedia. Uniform Resource Identifier. (May 29, 2017). https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

Wikipedia. Repository (Version Control). (September 17, 2016).
[https://en.wikipedia.org/wiki/Repository_\(version_control\)](https://en.wikipedia.org/wiki/Repository_(version_control))