A Metadata Application Profile for KOS Vocabulary Registries

Marcia Lei Zeng
Kent State University, USA
mzeng@kent.edu

Maja Žumer
University of Ljubljana, Slovenia
Maja.Zumer@ff.uni-lj.si

1. Background

The theoretical and applied research of registries (also known as "terminology registries") of knowledge organization systems has been the focus of the NKOS (Networked Knowledge Organization Systems) community since the beginning of the Internet era. As a class, "knowledge organization system" (KOS) encompasses a wide range of types of structured vocabularies that represent concepts and terms within certain knowledge domains. Classification systems, taxonomies, and thesauri are the most well-known examples of KOS. One of the key areas has been the specification of the minimum (core) set of data elements to describe structured vocabularies in a KOS registry. A KOS registry enables the description of, and access to, a KOS resource as a whole (i.e., as a "concept scheme" as referred to by the *SKOS Simple Knowledge Organization System Reference* (Miles & Bechhofer 2009)), however there was no protocol for describing KOS in structured data, according to the *JISC Terminology Registry Scoping Study (TRSS)* (Golub & Tudhope 2008). The milestones of establishing a protocol can be seen within and beyond the NKOS community's efforts, including the production of *NKOS Registry – Draft Set of Thesaurus Attributes* of 1998 (NKOS 1998), a Dublin Core (DC)-based *NKOS Registry – Reference document for data elements* of 2001 (Vizine-Goetz 2001), *JISC Terminology Registry Scoping Study (TRSS)* report (Golub and Tudhope 2008), the initiation of a DCMI-NKOS Application Profile Task Group in 2009 (Zeng & Hodge 2011) and the activities of this Group in developing a Dublin Core application profile for describing and accessing KOS resources (KOS-AP) since 2009 (DCMI 2009; DCMI-NKOS Task Group 2012a), and the publication of *Asset Description Metadata Schema* (ADMS) by the ISA programme of European Union in 2012 (ADMS Working Group,2012). In the following sections, we will report on the outcomes of the DCMI-NKOS Task Group, which builds on the work done by the NKOS community during the last decade.

In 2009 when the Task Group decided to develop a Dublin Core application profile for the purpose of describing and accessing the KOS resources, KOS registries were the primary focus. With the development of the Web technologies, microdata is receiving more and more attention, thus describing a KOS resource in any Webpage is recognized by the Task Group as another possible applicable area of the KOS-AP. Therefore, while we discuss the KOS-AP in the context of KOS registries, the context of microdata should be considered equally important in all aspects.

2. Developing the KOS Application Profile

We have been following the requirements set by the *Guidelines for Dublin Core Application Profiles* (DCAP) (Coyle and Baker 2009) in developing the KOS-AP. The Guidelines provide a framework for the content and structure of any DCAP. A DCAP is a document (or a set of documents) that specifies and describes the metadata used in a particular application. It: a) describes what a community wants to accomplish with its application (Functional Requirements); b) characterizes the types of things described by the metadata and their relationships (Domain Model); c) enumerates the metadata terms to be used and the rules for their use (Description Set Profile and Usage Guidelines); and d) defines the machine syntax that will be used to encode the data (Syntax Guidelines and Data Formats) (Coyle and Baker 2009).

2.1 Use Scenarios

In assessing the needs for both description and access, we generalized certain use case scenarios of three major types of users. The main user types are: (1) KOS developers (including the owner(s) and creator(s)), (2) information retrieval system developers, and (3) end-users (including all other users). It should be noted that the role of a "producer" and a "user" might switch during the whole process. The use scenarios include (and are not limited to) the following:

- The developers of a KOS would want to **publish**, **share**, and **allow reusing and mapping** of their product(s). They register and publish their systems and thus **expose** the KOS product(s) to interested parties.
- Other KOS developers may be interested in an existing KOS for **reuse** or **as an example** of good practice. They may create derivative works based on an existing KOS.
- Information retrieval system (IRS) developers may want to **reuse**, **implement**, and **evaluate** a KOS, as well as to **apply** a KOS to a collection to support searching and/or navigation.
- End users and researchers may be involved in terminology-related **research and exploration** within a subject domain. They may want to **evaluate**, **align**, or **compare** KOS resources (DCMI-NKOS Task Group 2012b).

When analysing the tasks of different users, we are using the user tasks defined by the IFLA *Functional Requirements for Bibliographic Records* (FRBR) (1998), with the extension of the user tasks defined by the *Functional Requirements for Subject Authority Data* (FRSAD) (2011):

- using the metadata to find a KOS that corresponds to the user's stated search criteria (e.g., in the context of a search for all KOS on a given subject, or a search for a KOS issued under a particular title);
- using the metadata retrieved to identify a KOS (e.g., to confirm that the KOS described in a record corresponds to the document sought by the user, or to distinguish between two KOS products or two editions that have the same title);
- using the metadata to select a KOS that is appropriate to the user's needs (e.g., to select a KOS in a particular language, or to choose a release of a KOS that is compatible with the hardware and operating system available to the user);
- using the metadata to acquire or obtain access to the KOS described (e.g., to place a purchase order or to access online an electronic KOS product stored on a remote server);
- using the data to explore the different KOS resources that are available in a registry (e.g., get acquainted with the subject coverage of a KOS or discover available KOS resources in a specific domain) (DCMI-NKOS Task Group 2012b).

2.2 The Conceptual Model

2.2.1 A model built based on the characteristics of the KOS resources

KOS resources, regardless of their different structures, share a number of characteristics that distinguish them from other creative works:

- The **continuity** of KOS works. Almost all KOS resources need to be continuously developed, following immediately the changes in the real world. A KOS scheme or system would lose its value and credibility if not constantly and timely updated.
- The **diversity** of the 'family' members. New versions of a scheme may be regularly released. From the same 'root' system (e.g., *Dewey Decimal Classification* (DDC)) versions may 'grow' and further extend the original (e.g., abridged and extended versions, translations).
- The **shared authorship**. In addition to the dynamic derivatives within the 'family', KOS works are usually not developed or used as stand-alone resources. Reuse, mapping, re-alignment, and derivation outside of the 'family' are common use cases. Even within the same 'family', later versions of a KOS scheme are usually based on previous ones, fully or partially, while the authorship also changes along with the new versions/editions. It is important to know the relationships among the different KOS resources to enable implementation and interoperability.
- The **complexity** of relations among KOS resources. Taking DDC as an example, all editions, versions, and the derivations of them are complex. The print version of DDC 22 was published in 2003 (in English); a web version was made available nearly simultaneously in WebDewey. At the time of initial publication, the underlying database was represented in a proprietary markup language (ESS), and distributed to translators after being transformed into ESS XML (an XML version of the same markup language). The German translation was published in print and web versions in 2005; the German web version, MelvilClass, presents the DDC in a different end user format from that used in WebDewey. (Žumer, Zeng, and Mitchell 2012). Nevertheless, often the situation is not as straightforward as it seems on the surface. For example, translations of a KOS can be symmetrical, locally tailored, or selective. Furthermore, a translation, extraction, and reuse can be at different levels or limited to a subset of the original work.
- **Tendency towards micro-level** management. Recently we are seeing continuous micro-level updating and translation of individual elements, so the same scheme available on a website today might be different from that of yesterday and no macro-level 'versions' are released. For example, the *Art and Architecture Thesaurus* (AAT) printed versions were published in 1990 and 1994. *AAT Online* has been the same "edition" since 1998 (Getty Vocabularies Program 2013), while the thesaurus has been extended to several languages and to over 34800 concepts and 245500 terms (Harpring, 2013). The intellectual property rights and provenance data also may be assigned and managed at the individual concept and term level (for an example, check this URI for the concept "smartphone" from the *Library of Congress Subject Headings* (LCSH) at: http://id.loc.gov/authorities/subjects/sh2007006251.html).

These dynamic and complex characteristics require a multi-layered model to present the complex attributes of KOS recourses and the relationships among them. The section "Selecting or Developing a Domain model" in *Guidelines for Dublin Core Application Profiles* (Coyle and Baker 2009) presents two domain model examples: a simple model and a FRBR-based model. We decided to build the conceptual model for the KOS-AP on the FRBR family (see Figure 1) according to the characteristics of KOS resources.

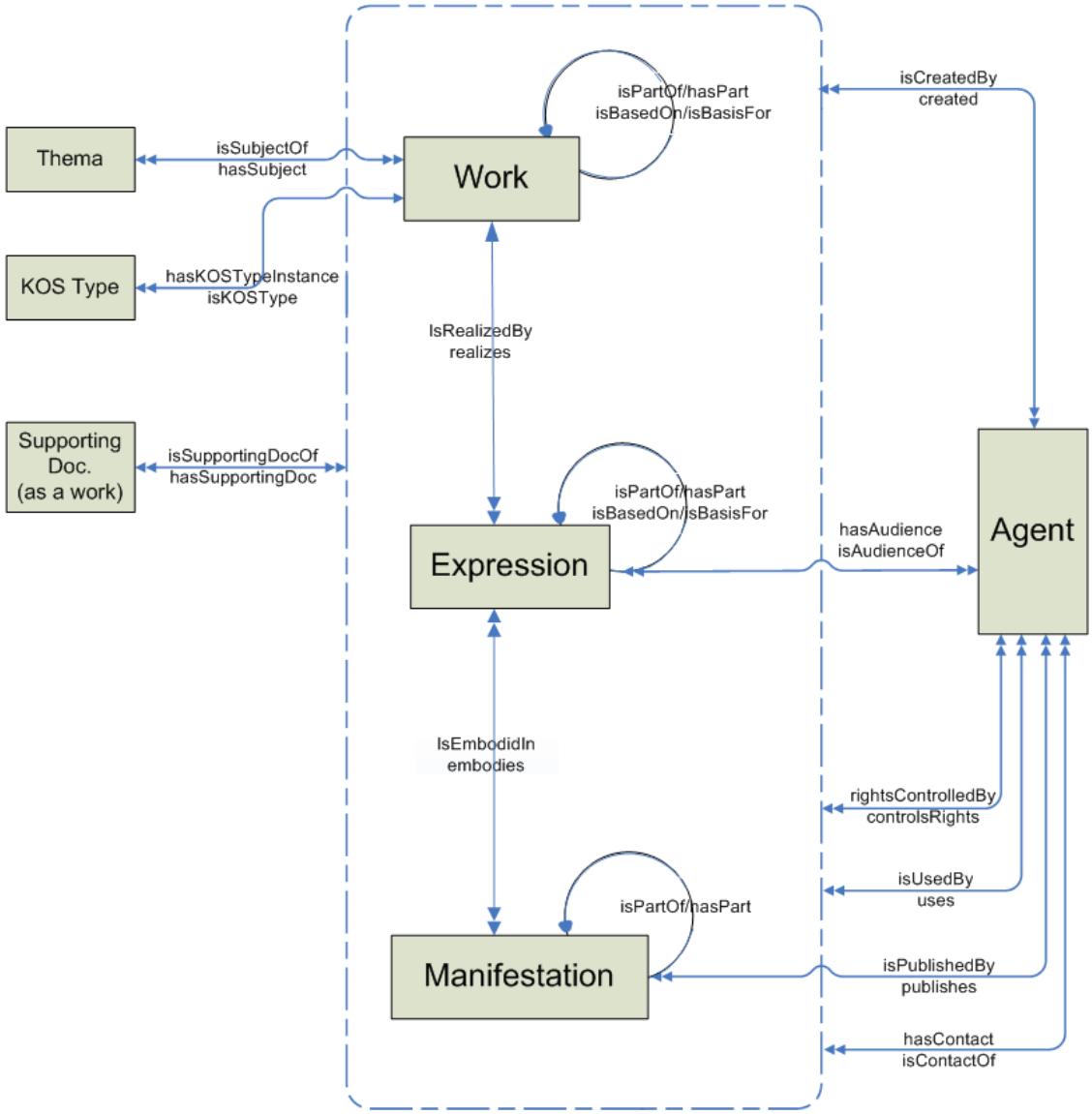

Figure 1. The Conceptual Model of KOS-AP. Source: DCMI-NKOS Task Group, 2012c.

2.2.2 Main entities of the KOS-AP conceptual model

According to FRBR, the entities belong to three groups, which can briefly be labeled as products, agents, and thema (the FRBR family entities are highlighted with italics in the text):

1) The central part of the model (Figure 1) includes three entity types representing products of creative endeavour: *Work* (distinct intellectual creation), *Expression* (realization of a *work*), and *Manifestation* (embodiments of an *expression*) belong to the so-called Group 1 entities in the FRBR model. The remaining entity type, *Item* (exemplar of a *manifestation*), is not included here.
2) On the right is Agent, which is linked through relationships with Group 1 entities, such as creation, production, distribution, and ownership. This is a generalized entity of the Group 2 entities defined by FRBR (including *Person* (an individual) and *Corporate body* (organization or group of individuals)).
3) On the top-left is *Thema*, anything that can be the subject of a *work*, as defined by FRSAD. *Thema* corresponds to the aboutness of a *work*. Consequently, KOS Type corresponds to the isness of a *work*.

In addition to these three 'groups', the Group 1 entities may be described by a supporting document, which is considered a *work*. Therefore a relation between the Group 1 entities and the supporting documents is also presented in the model.

2.2.3 Basic relationships between *Work*, *Expression*, and *Manifestation*

The central part of the conceptual model diagram contains three entity types, *Work* (W), *Expression* (E), and *Manifestation* (M). The relationships among them are essential for reflecting the complexity of KOS resources.

First, there are relationships between these instances of different entity types, i.e., *Work* to *Expression*, and *Expression* to *Manifestation*. Their relationships can be summarized as:

Table 1. Relationships between entities of different types

Work (W)-to-Expression (E):	(E) realizes (W)
Expression (E)-to-Manifestation (M):	(M) embodies (E)

To better understand these relationships, we can use a real KOS, *ASIS&T Thesaurus*, as an example. *ASIS&T Thesaurus* was first published in 1994. In addition to the new versions (created by different authors), multiple translations (translated by different translators) have been published or used internally. The thesaurus has been released for different needs in the online and Web environment with various formats. When modelling it according to FRBR, we can see them in different layers (DCMI-NKOS Task Group 2012d; Žumer, Zeng, & Hlava 2012):

1) *ASIS&T Thesaurus* as a whole is a *work*;
2) different versions (such as Version 1994 in English, Version 2005 in English, and Version 2012 in French) are different *expressions* of this *work*; and
3) the printed edition of the 2010 English version and the SKOS Linked Data representation of the same version, which are examples of *manifestations.*

With this model it is easy to identify the relationships between different KOS products. For example, when a multilingual or translated KOS scheme is brought into this model, it can be understood as:

- for a language-specific thesaurus: an *expression* of a *work* in a particular language;
- for a translation of an original version: a relationship with the original *expression*;
- for a selected translation of a version which is a translation itself: an *expression* based on another *expression*; or
- for an extraction of a classification *work* that is partially released as RDF triples for Linked Open Data purpose: a specific *manifestation*.

2.2.4 Basic and extended relationships between entities of the same type

There are also relationships between instances of entities of the same type, i.e., *Work* to *Work*, *Expression* to *Expression*, and *Manifestation* to *Manifestation*. They can be summarized as:

Table 2. Relationships between entities of the same type

Work (W)-to-Work (W):	based on (W), is part of (W)
Expression (E)-to-Expression (E):	based on (E), is part of (E), other relation (E)

All of these relationships can be expressed with the Dublin Core element dct:relation. The following table demonstrates the sub-types of dct:relation, primarily "part-of" and "based-on", and can be used as object property in describing relationships as needed (refer to "KOS Relation-Type Vocabulary", DCMI-NKOS Task Group 2013).

Table 3. KOS Relation-Type Vocabulary -- Sub-types of dct:relation and examples
Note: Specializations of relationships may only be applicable to specific entity types.

Relation Type	Definition	Element	Example	
part-of:			**A**	**B**
is part of	**A is part of B.**	**dct:isPartOf**	Class H - Social Sciences of *Library of Congress Classification (LCC)*	*LCC*
	B has part A.	**dct:hasPart**		
. outline of	A is outline of B.	nkos:isOutlineOf	*DDC Summaries*	*DDC*
	B has outline A	nkos:hasOutline		
. excerpt of	A is excerpt of B.	nkos:isExerptOf	Table G (Geographic Notation) of the *National Library of Medicine (NLM) Classification*	*NLM Classification*
	B has excerpt A	nkos:hasExcerpt		
. fragment of	A is fragment of B.	nkos:isFragmentOf	entries from a scheme	a scheme
	B has fragment A	nkos:hasFragment		
. sample	A is sample of B.	nkos:isSampleOf	a sample entry or a page from a scheme	a scheme
	B has sample A.	adms:sample		
based-on:			**A**	**B**
is based on	**A is based on B.**	**nkos:isBasedOn**	*Canadian Subject Headings (CSH)*	*Library of Congress Classification(LCSH)*
is basis for	**B is basis for A.**	**nkos:isBasisFor**		
.translation of	A is translation of B.	nkos:isTranslationOf	*Dewey-Dezimalklassifikation 22*	*DDC 22*
	B has translation A.	adms:translation		
.abridgment of	A is abridgment of B.	nkos:isAbridgmentOf	*DDC Abridged Edition 15*	*DDC 23*
	B has abridgment A.	nkos:hasAbridgment		
.extension of	A is extension of B.	nkos:isExtensionOf	A localized version of *NLM Classification*	*NLM Classification*
	B has extension A.	nkos:hasExtention		
.version of	A is version of B.	dct:isVersionOf	*DDC 23*	*DDC*
	B has version A.	dct:hasVersion		

2.3 Core elements

The entities *Work*, *Expression*, and *Manifestation* all have their attributes. Built on the attributes defined by the previous NKOS group efforts, we defined a set of core metadata elements to be used in the KOS-AP (refer to "NKOS AP Core Attributes in the context of user tasks", DCMI-NKOS Task Group 2012b).

Table 4. Core Elements of KOS-AP within the Context of User Tasks

CORE ELEMENTS	NEEDED FOR:			TO SUPPORT USERS TO:				
	Work	*Expression*	*Manifestation*	Find	Identify	Select	Obtain	Explore
title	x	x	x	x	x			
identifier	x	x	x	x				
contact		x	x				x	
description	x	x	x		x	x		
type (of KOS)	x			x	x	x		
creator	x	x	x	x	x			
language		x		x	x	x		
publisher			x		x	x		
format			x	x	x	x		
size (of vocabulary)		x			x	x		
rights	x	x	x	x		x	x	
date (created)	x	x	x	x		x		
date (updated)		x		x		x		
subject	x			x	x	x		
relation (to other)	x	x	x					x
sample (a relation)					x	x		
Additional elements (Could be included in 'description')								
services offered			x			x		
used by (a relation)		x	x			x		
frequency of update		x				x		
audience	x	x		x	x	x		
supplementary doc (a relation)	x	x	x			x		

Definitions and best practice comments are provided separately for each property of W, E, and M, even though the element names are the same. For example, "title" would have three entries in the AP, specifically defined for a W, E, or M.

3. Discussion of Some Issues

During the testing of the model with real KOS systems that involve multiple editions, languages, delivery formats, and derivations, (such as the *Dewey Decimal Classification* 22nd edition and of the *ASIS Thesaurus* 3rd edition), we have identified some specific issues related to KOS description, for example the designation of a '*work'* in theory and practice, and the shift of management from the whole KOS expression level to concept and label level.

One of the issues goes back to the designation of a *work*. For example, what is the basic *work* when we try to describe the various products of DDC22? Do we consider the classification created by Dewey and continued by other editors the *work* ("Classification as *work*" in Figure 2) or the DDC Edition 22 as the *work* ("Edition as *work*" in Figure 2)?

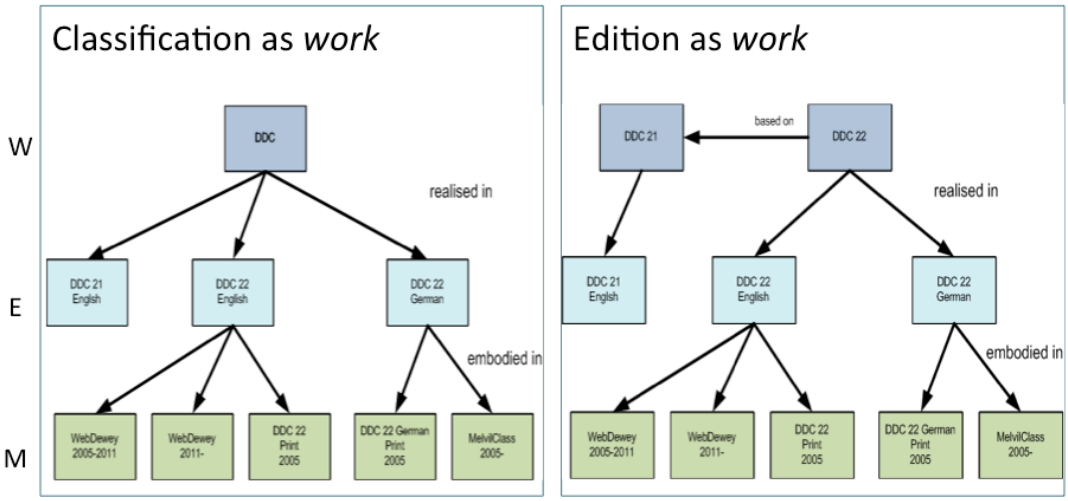

Figure 2. Classification as Work *vs.* Edition as Work—Using DDC22 as an example.
Source: Žumer, Zeng, and Mitchell, 2012

In the right side of the figure, "DDC22" is considered to be the *work*. DDC22 English and DDC22 German are *expressions* of DDC22. The embodiments of each *expression* in different published formats (print, web, linked data) are *manifestations.* An *item* (not modeled here) is the particular print copy of the classification in hand, the accessed web version as it appears on the screen, etc. The comparison of both figures shows no dramatic differences, though, and both approaches enable basically the same functionality.

Should the classification system itself be the *work*, and the "editions" presented as nested *expressions*? The argument for the "Edition as *work*" view is that "edition" has been the traditional identifier for DDC data, the situation has always been far more fluid, though. Editions are updated constantly; translations based on editions usually reflect an updated (and sometimes expanded or contracted) view of the base edition (Žumer, Zeng & Mitchell 2012).

This leads to the second issue: in the current information environment where KOS editions are disappearing, does it make sense to declare a particular edition a *work*? In the previous discussion on the characteristics of KOS resources, we indicated a feature "Towards micro-level management". In recent years more and more subject heading systems and thesauri update their contents online, releasing updated portions as a database, in XML, or in RDF frequently; or they are simply open for downloading at record/entry level, collectively or individually. For example, LCSH provides for each single record multiple downloading formats, together with the management data for each entry. If we consider LCSH as a whole concept scheme on the Web, the scheme available on the Website each day might be different. The notion of "edition" is no longer applicable. Thus the challenges of describing a concept scheme are obvious because the current way of describing the editions and manifestations are all at the whole scheme level. How could the emphasis shift to the micro-level management situation? This remains an open question.

4. Conclusion

This paper reports on the end-products of a Dublin Core Application Profile for KOS Resources developed by a DCMI Task Group, which builds on the work done by the NKOS community during the last decade. All related documents are available on the Task Group wiki, including user scenarios, a FRBR-based domain model, the attributes of KOS *work*, *expression*, and *manifestation*, and the associated metadata elements defined in the context of user tasks. In addition, a KOS-Types vocabulary (defined based on the existing national and international standards), examples of the KOS resource relationships explained using this model, and other related documents and tools are also available.

A KOS-AP for describing and accessing KOS is becoming more meaningful with the fast development of Linked Data, the success of which depends heavily on using, sharing, and interlinking of standardized value vocabularies. The primary usage of this application profile would be the registries for KOS vocabularies which cover many types of KOS resources, from thesauri to classification, from mono-lingual to multilingual (symmetrical and non-symmetrical), from independent single scheme to those with multiple editions, versions, variations, and derivations, and from direct expressions of an original work to aggregated products of multiple KOS works, which, needless to say, are usually delivered in multiple formats. New challenges exist as the KOS management tends to shift from the whole scheme to the individual concept and label, and releases of born-digital concept schemes in multiple manifestations have become common. We believe that the theoretical exploration, the conceptual model, and the core elements to be used in KOS registries that are introduced in KOS-AP should also be applicable in microdata of KOS Websites, although further testing and adjustments are probably needed. The KOS-AP could also be adopted for use by other types of specifications that share common characteristics of KOS, such as frequently updated, translated, and derived handbooks, technical manuals, and schemas.

References

ADMS Working Group 2012. *Asset Description Metadata Schema*. ISA programme, European Union. Available at <http://joinup.ec.europa.eu/asset/adms/home>.

Coyle, Karen and Thomas Baker 2009. *Guidelines for Dublin Core Application Profiles*. Available at <http://dublincore.org/documents/profile-guidelines/>.

DCMI. 2009. DCMI/NKOS Task Group Home Page. Available at <http://dublincore.org/groups/nkos/>.

DCMI-NKOS Task Group 2012a. DCMI NKOS Task Group Wiki site. Available at <http://wiki.dublincore.org/index.php/DCMI_NKOS_Task_Group>.

DCMI-NKOS Task Group 2012b. KOS AP Worksheet. (Last updated March, 2013). Available at < http://wiki.dublincore.org/index.php/NKOS_AP_Worksheet>.

DCMI-NKOS Task Group 2012c. Core Elements. (Last updated March 2013). Available at <http://wiki.dublincore.org/index.php/Core_Elements>.

DCMI-NKOS Task Group 2012d. KOS example. Available at <http://wiki.dublincore.org/index.php/KOS_example>.

DCMI-NKOS Task Group 2013. NKOS Vocabularies. Available at <http://wiki.dublincore.org/index.php/NKOS_Vocabularies>.

Functional Requirements for Subject Authority Data, A Conceptual Model (FRSAD) 2011. IFLA Working Group on Functional Requirements for Subject Authority Records (FRSAR). Eds. Zeng, Marcia L, Maja Zumer, and Athena Salaba. Berlin/Munich: De Gruyter Saur.

Functional Requirements for Bibliographic Records - Final Report 1998. IFLA Study Group on the Functional Requirements for Bibliographic Records (FRBR). Munich: K.G. Saur.

Getty Vocabularies Program 2013. History of the AAT. (Revised 27 March 2013). Available at <http://www.getty.edu/research/tools/vocabularies/aat/about.html#history>.

Golub, Koraljka and Tudhope, Douglas 2008. *JISC Terminology Registry Scoping Study (TRSS) - Final Report*. (Revised 2009). Available at <http://www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf>.

Harpring, Patricia 2013. Getty Vocabularies and Linked Data. (Revised 17 June 2013). Available at <http://www.getty.edu/research/tools/vocabularies/Linked_Data_Getty_Vocabularies.pdf>.

NKOS 1998. *NKOS Registry – Draft Set of Thesaurus Attributes.* (Last modified July 30, 1998). Available at <http://nkos.slis.kent.edu/Thesaurus_Registry.html>.

Vizine-Goetz, D. 2001. *Networked Knowledge Organization Systems (NKOS) Registry: Reference document for data elements*. Available at <http://staff.oclc.org/~vizine/NKOS/Thesaurus_Registry_version3_rev.htm>. and <http://nkos.slis.kent.edu/registry3.htm>.

Zeng, Marcia Lei and Gail Hodge 2011. Developing a Dublin Core Application Profile for the Knowledge Organization Systems (KOS) Resources. *Bulletin of the American Society for Information Science and Technology*, 37(4):30-34. Available at <http://www.asis.org/Bulletin/Apr-11/AprMay11_Zeng_Hodge.html>.

Žumer, Maja, Marcia Lei Zeng, and Joan S. Mitchell 2012. FRBRizing KOS relationships: Applying the FRBR model to versions of the DDC. In: *Categories, Contexts and Relations in Knowledge Organization. Proceedings of the Twelfth International ISKO Conference, 6-9 August 2012, Mysore, India*. 191-194.

Žumer, Maja, Marcia Lei Zeng, and Marjorie MK Hlava 2012. A Domain model for describing and accessing KOS resources: Report of processes in developing a KOS description metadata application profile. In: *Metadata for Meeting Global Challenges. Proceedings of the 2012 International Conference on Dublin Core and Metadata Applications, Kuching, Sarawak, Malaysia, Sept. 3-7, 2012.* Available at <http://dcpapers.dublincore.org/pubs/article/view/3656>.

KOS References

Art & Architecture Thesaurus 1994. Second Edition. Edited by Toni Petersen. New York: Oxford University Press.

Art & Architecture Thesaurus® Online 1998-. Los Angeles, Calif. : J. Paul Getty Trust. Available at <http://www.getty.edu/research/tools/vocabularies/aat/index.html>.

ASIS&T Thesaurus of Information Science, Technology, and Librarianship 2005. Third Edition. Edited by Alice Redmond-Neal and Marjorie M. K. Hlava. Information Today, Inc.

Dewey Decimal Classification and Relative Index. Ed. 22. 2003. Melvil Dewey. Edited by Joan S Mitchell *et al*. Dublin, Ohio: OCLC.

Library of Congress Subject Headings. Linked Data version based on Edition 32. 2008?-. Washington, D.C.: Library of Congress. Available at <http://id.loc.gov>.