Journal of the American Society for Information Science and Technology

Terminology Registries for Knowledge Organization Systems – Functionality, Use, and Attributes

Journal:	*Journal of the American Society for Information Science and Technology*
Manuscript ID:	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Golub, Koraljka; University of Bath, UKOLN Tudhope, Douglas; University of Glamorgan, Faculty of Advanced Technology Zeng, Marcia; Kent State University, Zumer, Maja; University of Ljubljana,
Keywords:	organization of information < information operations < (activities and operations), subject indexing < indexing < organization of information < information operations < (activities and operations), online searching < computer applications < computer operations < (activities and operations)

Terminology Registries for Knowledge Organization Systems – Functionality, Use, and Attributes

Koraljka Golub

UKOLN, University of Bath, Bath BA2 7AY, United Kingdom

Phone: +44 (0) 1225 383619, Fax: +44 (0) 1225 386838, E-mail: k.golub@ukoln.ac.uk

Douglas Tudhope*

Faculty of Advanced Technology, University of Glamorgan, Pontypridd, CF37 1DL, Wales, United Kingdom

Phone: +44 (0) 1443 483609, Fax: +44 (0) 1443 482715, E-mail: dstudhope@glam.ac.uk

Marcia Lei Zeng

School of Library and Information Science, Kent State University, P.O. Box 5190, 314 University Library, Kent, OH 44242-0001, United States of America

Phone: +1 (0) 330-672-0009, Fax: +1 (0) 330-672-7965, E-mail: mzeng@kent.edu

Maja Žumer

Department of Library and Information Science and Book Studies, Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI-1000 Ljubljana, Slovenia

Phone: +386 (0) 1 2411374, Fax: +386 (0) 2411371, E-mail: Maja.Zumer@ff.uni-lj.si

*corresponding author

Table of contents

Abstract

Terminology registries (TRs) are a crucial element of the infrastructure required for resource discovery services, digital libraries, Linked Data and semantic interoperability generally. They can make the content of knowledge organization systems (KOS) available both for human and machine access. The paper describes the attributes and functionality for a TR, based on a review of published literature, existing TRs and a survey of experts. A domain model based on user tasks is constructed and a set of core metadata elements for use in TRs is proposed. Ideally the TR should allow searching as well as browsing for a KOS matching a user's search while also providing information about existing terminology services, accessible to both humans and machines. The issues surrounding metadata for KOS are also discussed, together with the rationale for different aspects and the importance of a core set of KOS metadata for future machine-based access; a possible core set of metadata elements is proposed. This is dealt with in terms of practical experience and in relation to the Dublin Core Application Profile.

Introduction

The large number of knowledge organization systems (KOS) provided on the Web, together with the variety of potential applications facilitated via protocols and standards for digital representation, has made the notion of a terminology registry (TR) increasingly relevant. TRs have emerged since 2000 and have covered KOS of all types and complexities. A TR systematically registers KOS with standardized structures for both human inspection and machine to machine (in further text: m2m) access. It identifies, describes, and points to sets of controlled vocabularies available for use in information systems and services. Less frequently, it may optionally include the concepts, terms, and semantic relationships of a KOS vocabulary and may possibly provide terminology services that permit programmatic access by applications. TRs are a crucial element of an infrastructure for resource discovery. When adopting the Semantic Web standards, such as RDF (Resource Description Framework), SKOS (Simple Knowledge Organisation System) and OWL (Web Ontology Language), they promote the wider adoption, standardization, and overall interoperability of metadata by facilitating their discovery, reuse, harmonization, and synergy across diverse disciplines and communities of practice (Zeng & Chan, 2010, p. 4655).

The paper draws on a review of existing TRs, related projects, and published literature, and is supported by data collected from an email survey and a number of semi-structured interviews (28 responses in total from experts in related international projects and subject areas). The work originated in the Terminology Registry Scoping Study (TRSS) (JISC, 2009; Golub & Tudhope, 2009), being subsequently revised, extended and updated for this paper. A major, further extension draws on the work for the Dublin Core Application Profile, which resulted in recommended metadata elements for describing KOS resources (in further text: KOS-AP). While the original TRSS study took a bottom-up approach, i.e., from the analysis of existing terminology repositories to metadata elements, the KOS-AP approach employed a concept model based on the FRBR (Functional Requirements for Bibliographic Records) model. Through an analysis of the functional requirements in relation to user tasks in various scenarios, the KOS-AP reviewed the core and additional elements defined by the TRSS study and generated a list of metadata elements that would be meaningful and useful for use by TRs in the Linked Data environment.

The paper is structured as follows. In the second section the background of TRs and comparison with other kinds of registries and some definitions are given (Background). Then, the methodology is described (Methodology). Research findings are discussed around four sub-topics: users and use cases of TRs, existing TRs, attributes of TRs, and functionalities of TRs (Research Findings and Discussions). Concluding remarks summarise major findings (Conclusion).

Background

In general, this paper follows the definitions given in the JISC Terminology Services and Technologies Review (Tudhope, Koch, & Heery 2006, p. 22-47) and builds on work and discussions by the community of Networked Knowledge Organization Systems/Services (NKOS) in over twenty NKOS workshops since 1997.

The Scope of TRs

The scope of TRs has evolved with the increasing range of vocabularies and semantic tools used for organizing information and promoting knowledge management during the last two decades.

The term *knowledge organization system* (KOS) encompasses all types of vocabularies for vocabulary control, for organizing information and for promoting knowledge management (Hodge, 2000). The term came from the NKOS group when two workshops on terminology and classification tools were conducted in the 1997 and 1998 ACM Digital Libraries Conferences. The following workshop (in 1999) formally used the title of NKOS Workshop and workshops have been conducted each year since then (Networked Knowledge Organization Systems/Services/Structures, 2013). Different families of KOS, including lists, name authority files, subject heading systems, classification schemes, taxonomies, and thesauri, are applied in both modern and traditional information systems. They are also referred to as *controlled vocabularies* (ANSI/NISO Z39.19, 2005), *structured vocabularies* (BS 8723, 2005; ISO 25964-1, 2011; ISO 25964-2, 2013), *value vocabularies* (W3C Library Linked Data Incubator Group, 2013), *concept schemes* (Miles & Bechhofer, 2009), *semantic assets* (broader coverage) (Asset Description Metadata Schema, 2013), and *classification* (ISO/IEC 11179, 2013) by various standards. Some communities tend to use one type of KOS to cover all, such as the term *taxonomy* often used by government agencies (Lambe, 2007). The term *ontology* is often used rather loosely and ontologies are sometimes included under KOS. In this paper, registries of formal ontologies are considered a related but separate area than registries of vocabularies primarily designed for retrieval purposes (section Terminology Registry Review discusses prominent ontology registries). The W3C Recommendation *SKOS: Simple Knowledge Organization System Reference* considers knowledge organization systems to include thesauri, taxonomies, classification schemes, and subject heading systems (Miles & Bechhofer, 2009).

Nevertheless, taking an even broader view, there can be many more types of KOS. In addition to these most common types, inverted indexes of information retrieval systems, surrogate files, systematic nomenclatures, encyclopaedias, conceptual schemata of databases, and knowledge representation in knowledge bases, might also be seen as systems for knowledge organization (Hierppe, 1990; Souza, Tudhope, & Almeida, 2012). Nowadays, KOS products are regarded as semantic assets, together with others such as document DTDs, data models, code lists, XML schemas, and RDF models, as defined by the Asset Description Metadata Schema (ADMS) developed for the European Union's Interoperability Solutions for European Public Administrations (ISA) Program (Asset Description Metadata Schema Working Group, 2013). As Wright (2008) points out, communities of practice are an important organizing principle; different communities define KOS differently, according to their practical purposes. The TRs listed in section Terminology Registry Review reflect the different and overlapping coverage of their contents.

TR and Other Types of Registries

Efforts to provide comprehensive lists of controlled vocabularies by a variety of national, regional, and domain organizations predate the Web. A book, *Thesauri used in online databases: An analytical guide*, co-authored by Chan and Polland (1988), listed and annotated dozens of thesauri that conformed to the then international standard for constructing thesauri (ISO 2788, 1986) first published in 1974 and revised in 1986. The *Thesaurus guide: Analytical directory of selected vocabularies for information retrieval* published by the European Commission (1993) contained approximately 700 vocabularies available in at least one of the European Union languages at the time. The University of Toronto Faculty of Information maintains a North American Clearinghouse for English language thesauri and controlled vocabularies (also including multilingual thesauri with English language sections) published in print with over 2500 titles (University of Toronto, 2003; Dextre Clarke, 2005). Similar to this but functioning only as the catalogue of a virtual collection, WorldCat (OCLC, 2013) contains many catalogue records for vocabularies. More recently, the Open Knowledge Foundation's Data Hub hosts over 300 datasets that can be considered KOS, including self-registered types such as structured vocabulary, domain-specific ontology, list, and dictionary in addition to thesaurus, classification scheme, and name authority (Open Knowledge Foundation, 2013a; Zeng, 2012).

From the mid-1990s lists of online vocabularies have been provided on the Web, but typically are not consistently enlarged or maintained (e.g., Koch, 2007; Middleton, 2008). In general, such lists have focused on major vocabularies and have lacked the metadata that would facilitate their discovery and services that would allow access to individual terms and concepts.

With the advance of Web-based technologies, especially Web services, *terminology registries*, and other types of registries emerged in the 21st Century. The term *terminology* has been used when *registries* are discussed and with regard to services based on vocabularies. A *registry* is an authoritative, centrally controlled store of information (W3C Working Group, 2004). Dictionary definitions of *terminology* include "the technical or special terms used in a business, art, science, or special subject" (Merriam-Webster Online, 2013), and, similarly, "the body of specialized words relating to a particular subject; the study of terms" (Collins, 2013). Such a registry is a different sense of the term than when used in connection with resources for precise definitions of language use for translators or writers, or for computer-based linguistic tools. As explained in the introduction section, a terminology registry (TR) identifies, describes, and points to sets of KOS vocabularies available for use in information systems and services. Their content can be made available for human inspection and possibly m2m access. The scope of a TR can cover free and publicly available, or fee-based and restricted-access vocabularies. By exposing rich metadata, a TR facilitates the discovery of an appropriate vocabulary and potentially information about its use or terminology services based upon the vocabulary.

Heery (2005) discusses the relationship between a *metadata registry* and a TR, saying that there are obvious differences between "metadata element sets" and "subject vocabularies" as to different relationships between terms, different use cases and communities, different standards and different conventions. However, the two are also complementary since they contribute to the same "business processes" (e.g., enterprise portal, records management, resource discovery) and to similar workflows and choreographed services. Metadata elements can be seen as existing within an *attribute space*, whereas the vocabulary elements that may comprise the metadata element content exist within a *value space* (Baker et al., 2002). Metadata standards often specify vocabularies for use in value spaces, associated with certain metadata elements or fields. Consequently, metadata registries may also contain, or link to, terms and codes from these schemes (e.g., the DCMI Registry also includes the DCMI Type Vocabulary); thus the term "metadata registry" could also refer to an integrated structure housing both metadata and terminologies (Zeng & Chan, 2010). In 1999 the DCMI Registry Community (DCMI Registry Community, 2010) was established as a forum for service providers and developers of both metadata schema registries and controlled vocabulary registries to exchange information and experience. The Open Metadata Registry (2013) (formerly the NSDL Registry) is an example of both a TR and a metadata registry – it provides access to both vocabularies and metadata schemas.

The 9th International Forum on Metadata Registry (2006) held in Kobe City, Japan brought together researchers and implementers of metadata registries. Three standards, ISO/IEC 11179 Information Technology – Metadata Registries (2013), ISO 704 Terminology Work – Principles and Methods (2009), and ISO 12620 Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category

Registry for Language Resources (2009), with their associated technologies, were central to the forum. The characteristics of these overlapping registries were the focus of a special NKOS session on registries held at the International Conference on Dublin Core and Metadata Applications in 2008 (Zeng, Hillmann, & Sutton, 2008). Based on the discussions, registries related to KOS vocabularies can be categorized into four types:

- Metadata [Schema] Registries – registering metadata element sets, elements and refinements, application profiles, schemas in different bindings. UKOLN's CORES Registry (2003) is a good example.
- Terminology Registries / Repositories – may be considered at two levels: *basic* TRs contain only the metadata of KOS vocabularies, while *full* TRs contain also the members (e.g., concepts, terms, relationships) of the vocabularies. Examples are Terminologies Service by OCLC Research (2011) and the BioPortal ontology repository (National Center for Biomedical Ontology, 2013).
- Terminology services may be listed in a terminology registry or separately hosted in a service registry – known as Service (or Collection) Registries. They can be databases of descriptions of available services and, where appropriate, associated collections. An example is the JISC Information Environment Service Registry (IESR) (2011).
- Data [Standards] Registries – registries/repositories of all kinds of data standards (e.g., data dictionaries, data models, schemas, and code sets).

There are also integrated registries that could encompass all the registry contents listed above.

In addition to looking at *what* types of resources are covered by the registries (as stated above), the registries can also be characterized according to (1) *where*: community-based (e.g., museum, health, justice, environment), institution-based (e.g., US Environmental Protection Agency, US Cancer Institute, UN's Food and Agriculture Organization); (2) *who*: targeted audience (e.g., application developers, vocabulary developers, content providers, and end-users); (3) *when*: design-time or run-time, and (4) *how:* functions and services (e.g., persistent storage, management, m2m services, etc.). Additional variables are scale/size of a registry, data models a TR can handle (e.g., hidden semantics, relationship types), indexing and analysis requirements, extracting and downloading capabilities, and decentralization capabilities.

TRs may thus optionally include the vocabularies' concepts, terms, and semantic relationships and possibly provide *terminology services* that permit programmatic access by applications. *Terminology services* are web services that present and apply the content of vocabularies, either for m2m usage, or for human usage, and can be applied at various stages of the search process, e.g., for translating user terms to controlled terms, disambiguation of terms representing concepts, browsing, query expansion, mapping, subject indexing and classifying, etc. Their major purpose is improving document and information discovery.

Attributes of TRs

To register a KOS in a TR requires a set of common attributes that describe it, no matter what type it is. Discussions on the attributes of a TR required for the online environment started in the late 1990s. In 1996, Linda Hill (University of California at Santa Barbara) and Michael Raugh (Interconnect Technologies) drafted the attributes that would be needed in describing thesauri in a registry. The work was further developed by a working group formed in the NKOS community (Hodge, 1999). "Thesaurus-level metadata and thesaurus registries" was one of the four topics on the agenda of the 1998 NKOS Workshop (other topics were the data model, the function model and the business/intellectual property model) (Networked Knowledge Organization Systems and Services, 1998a). The questions of the registry are still relevant today and applicable to thesauri and beyond:

- What thesaurus-level metadata* are needed to represent the scope, structure, size, ownership, access constraints, etc. of a thesaurus so that potential users (for all applications) will know what is available and how to access and use it?
- What is the role of thesaurus-level metadata in enabling the interoperability of online-accessible thesauri?
- What role could thesaurus registries play in "advertising" the availability of thesauri and facilitating access and use?
- What tasks are involved in maintaining a registry?
- What kind of organizations would best fulfil the registry function?

*Note that "Metadata" is intended to mean not the actual attributes of individual terminology tools but the "collection-level metadata" that would describe the terminology tool as a whole (Networked Knowledge Organization Systems and Services, 1998a).

The first version of *NKOS Registry - Draft Set of Thesaurus Attributes*, developed in 1996 and last modified in 1998 (Networked Knowledge Organization Systems and Services, 1998b), lists 55 metadata elements grouped in 10 categories (product information, scope and usage, characteristics of descriptors, size of set of descriptors, labels for relationships, other product information, terms and conditions, vendor/provider information, contact information, additional information). In 2001, the registry reference document was extended to cover more types of KOS in the second version (Networked Knowledge Organization Systems and Services, 2001) and grouped metadata elements into five more general categories (product information, scope and usage, NKOS characteristics, terms and conditions, and vendors). Meanwhile, the Biological Resources Division (BRD) and the National Biological Information Infrastructure (NBII), in connection with the effort of building a CERES/BRD vocabulary for biodiversity and ecosystem science, developed an MS Access database according to the draft registry standard. The lessons learned from the testing indicated that these registry attributes worked better for thesauri than for other vocabularies. It was found that sometimes it was difficult for a cataloguer to complete the fields because more in-depth information was needed. There was also a belief that it would benefit from having owner/creators do the registration (Hodge, 1999). Soergel (2001) discussed characteristics for describing and evaluating KOS from various perspectives, including: purpose; coverage of concepts and terms, sources, quality of usage analysis; conceptual analysis and conceptual structure; terminological analysis; access and display, and format of presentation of the vocabulary.

The increasing use of the Dublin Core metadata elements in the beginning of the 21st century brought a parallel approach to the efforts of defining metadata for the TR. A proposal drafted by Vizine-Goetz (2001) of OCLC includes two groups of elements (each with mandatory and optional elements.). The first group matches the Dublin Core Metadata Element Set intended for creating metadata descriptions that will facilitate the discovery of KOS resources. The second group of elements is intended for recording of specific characteristics of a KOS resource that will facilitate evaluation of the resource for a particular application or use. Following Dublin Core's specification, each of the 20 data elements is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements. Similarly, Hodge, Salokhe, Zolly and Anderson (2007) proposed the following metadata elements for a TR: name (with acronyms), creator, description, subject controlled, keywords, resource identifier, language, resource type, rights, publisher, format, and contact email, as part of the Ecoterm environmental vocabulary and registry initiative. The NKOS community's efforts have been continued in the work of the DCMI/NKOS Task Group (see section Proposed Attributes for a Dublin Core Application Profile).

The ISO/IEC 11179 Metadata Registries family of standards has been in development since 1995 (ISO/IEC 11179, 2013). It aims to provide a theoretical model for metadata elements within registries, with a view to furthering reuse. There are six parts. Part 1 gives the general framework, while Part 2 provides a conceptual model for managing classification schemes (KOS) within a metadata registry. Part 3 defines a conceptual model for a metadata registry, expressing its data elements in terms of general attributes. Part 4 provides guidance on how to develop unambiguous data definitions, Part 5 on how to designate or identify a particular data item, and Part 6 on how a registration applicant may register a data item. The most relevant to TRs is Part 2, Classification. Here "classification schemes" include key words, thesauri, taxonomies, and ontologies. A comprehensive list of attributes is arranged by: Designation, Definition, Context, Classification Scheme, Classification Scheme Item, Classification Scheme Item Relationship, Administration Record, Reference Document, Submission, Stewardship, Registration Authority, and Registrar (ISO11179-2, 2005).

Methodology

TRs have been shaped by the advance of technologies, particularly the online environment, the Internet, and more recently, the Semantic Web. In order to develop an understanding of best practices, use cases, functionality, and attributes of TRs, the authors followed leading TRs and collected data from TRs, as well as the professionals and experts who conduct TR research and implement TRs. The research for this paper involved two major phases.

Phase I primarily focused on the status, architecture, and functionality of TRs. The objectives were set to: (1) gather a set of use cases that demonstrate how and why a TR as a shared infrastructure service is required; (2) gather requirements from various sources; (3) synthesize the outcomes of efforts to date; (4) include the international and commercial context; and, (5) analyze the potential costs, benefits and risks of TRs as shared infrastructure services.

The data collection process started from identifying and analyzing efforts of TRs and related reports. In addition, information was obtained through consultation with key services, projects, and executives across digital library, research, and learning domains. Data about attributes of TRs was collected from various sources, including TRs and associated reports or standards. Other cases studied include a wider set of research and operational TRs and repositories. Over 20 initiatives were included. A selected number of TRs are discussed in section Terminology Registry Review to demonstrate the architectural functions and the attributes used.

Additionally, key experts whose areas of interests were related to TRs were approached via an email invitation letter. The questions raised were open-ended and covered three major issues:

1) What should a terminology registry comprise and which functionalities should it offer?
2) What are the possible usage scenarios or use cases and is there a preference for an m2m access or for human inspection?
3) What are the major barriers and challenges to a terminology registry take-up and implementation?

Service providers were additionally asked to list KOS vocabularies used or planned to be used in the future. General comments were invited as well. The people were selected from various areas relevant to a terminology registry: related projects that included UK-based projects and international projects; various subject domains (cultural heritage, e-science, e-learning, e-framework); services with terminologies; terminology developers; and terminology experts. Out of 28 people contacted, there were 12 who were interviewed on site or over the telephone, while the other 16 responded via email; 20 people were from the UK, 5 from the United States, with 1 each from Australia, Germany, and Italy. The onsite or telephone interviewees were selected as the most knowledgeable in their specialty area and for the topic, therefore a more thorough discussion was held with them. The answers collected via email helped form a more complete picture of the state of the affairs. A full list of responders and original invitation letters and questions are available in the project report, together with details of the TRs reviewed for Phase I of this research (Golub & Tudhope, 2009).

Phase II of the research extended coverage to the TR efforts that have implemented Semantic Web technologies, especially the services established for Linked Data and that have adopted the W3C recommendation of

SKOS. These cases are presented together with the results of the Phase I study in section Users and Use Cases of TRs. Additionally, Phase II reports on the KOS metadata outcomes from the KOS-AP work as follows.

The Phase I comparison and recommendations of attributes of TRs took a bottom-up approach. A set of attributes of TRs were generated based on the structure of existing TRs and ISO 11179 guidelines (ISO/IEC 11179). The objects, i.e., the KOS vocabularies, were considered independently (no relationships between KOS vocabularies). In other words, the TRs register the vocabularies individually and provide the metadata describing each individual vocabulary. Taking a different path, in Phase II, the KOS-AP Task Group (the authors are key members) started from the top by defining a conceptual model according to the use cases and user-tasks. The team took representative KOS vocabularies to examine the dynamic and complex characteristics, selecting Dewey Decimal Classification (OCLC, 2013) and ASIS Thesaurus (Milstead, 1998; Redmond-Neal & Hlava, 2005), both of which have multiple editions, many translations of various editions, and are available as printed schemes, databases, SKOS-encoded datasets, and are distributed in various formats and media. Many KOS resources resemble these characteristics. A KOS scheme or system would lose its value and credibility if not constantly updated, hence they need to be continuously developed. In addition to micro-level updates, new versions with a significant amount of changes may be regularly released. More significantly, the KOS products are usually not developed or used as stand-alone resources. Reuse, mapping, re-alignments, and derivation are common use cases. It is important to know the relationships among the different KOS works to enable implementation and interoperability. Therefore, a multi-layered model is needed to present the complex relationships among KOS resources. The users of TRs were further defined and the study results from Phase I were expressed in the framework of models FRBR (Functional Requirements for Bibliographic Records) (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998) and FRSAD (Functional Requirements for Subject Authority Data) (IFLA Working Group on Functional Requirements for Subject Authority Records, 2011). The user tasks of TRs were summarized according to the functional requirements of different users at different use case scenarios. Recommended metadata elements for TRs which were verified based on Phase I research, were divided into core elements and additional elements. They were mapped into the FRBR- and FRSAD-based user tasks including find, identify, select, obtain, and explore (DCMI/NKOS Task Group, 2013b).

Research Findings and Discussions

Users and Use Cases of TRs

Based on the data collected from the research, three general types of users of TRs can be identified: KOS owners or creators, system developers, and end users (DCMI/NKOS Task Group, 2013b). The KOS creators may have two different roles.

The owner(s)/creator(s) of a KOS need a TR to publish and share their work while allowing their work to be reused and mapped by other users. They register and publish their KOS vocabularies and thus expose the KOS product(s) to interested parties. KOS creators may also have a role of a user of a TR. A use case at the time of developing a new vocabulary would be to see if any similar vocabulary could be adopted entirely or partially or be useful in the construction of a related vocabulary. A TR can assist discovery of existing vocabularies, or the most recent version of a given vocabulary. It can reduce the costs related to finding and implementing an appropriate vocabulary and learning by trial and error. Finding an appropriate vocabulary via, e.g., search engines, contacts, libraries etc. can be time consuming; developing a new vocabulary that proved unsuitable is costly.

For example, a domain-specific TR would list and describe various vocabularies in a domain and might ideally also provide contact with existing users. The research team at the Food and Agriculture Organization of the United Nations (FAO) reported that in developing their AGRIS Application Ontology, 40 terminological resources in agriculture and related domains were identified and studied. The team listed “Gather and Characterize Existing Terminological Resources in the Domain” as the first task of the project (Liang, Salokhe, Sini, and Keizer, 2007). Related use cases are the reuse, reference, and derivation of new works from other vocabularies in similar or related domains. Alternatively, the developer may be in the process of revising an existing vocabulary and would also have the above needs. From such a position, KOS creators and maintainers may be interested in an available KOS for reuse or as examples of good practice.

TR users also include information retrieval system developers who need to implement and evaluate a KOS, and/or to apply a KOS to a collection to support searching and/or navigation. In our study, the most frequently mentioned possible use case for a TR was reviewing and examining existing vocabularies and discovering whether any vocabulary currently existed that met, or approximated, a given set of requirements. This might be at the time of planning a digital collection or some other service that might be supported by a vocabulary. End users and researchers may be involved in terminology-related research and exploration within a subject domain. They may also want to evaluate, align, or compare KOS resources.

In any of the use cases, all three groups of users will need to find, identify, select, obtain, and explore KOS resources through the data provided by a TR. These user tasks can be considered in the contexts of the FRBR user tasks (find, identify, select, obtain) and the FRSAD extension (explore) of these tasks (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998; IFLA Working Group on Functional Requirements for Subject Authority Records, 2011):

- using the data provided by a TR 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 data retrieved from a TR to identify a KOS (e.g., to confirm that the KOS described in a record corresponds to the KOS sought by the user, or to distinguish between two KOS products or two editions that have the same title);
- using the data provided by a TR 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 data provided by a TR in order 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 computer);
- using the data provided by a TR to explore the different KOS that are available in a registry (e.g., get acquainted with the subject coverage of a KOS or discover available KOS in a specific domain) (DCMI/NKOS Task Group, 2013b).

Terminology Registry Review

This section reviews selected TRs that focus on the KOS vocabularies aimed at information retrieval. They are discussed with respect to functionality and metadata.

TR provides metadata for each vocabulary and links to owner/provider and/or services

The CENDI Agency Terminology Resources (CENDI, 2010) offers a basic TR. URLs are provided to online thesauri and indexing resources of the various federal scientific and technical agencies, to be of interest to those wishing to know about the scientific and technical terminology used in various fields, spanning agriculture to medicine to the environment. There are over 20 current vocabularies, including the Biocomplexity Thesaurus (USGS/NBII), the ERIC Thesaurus (NLE), MeSH (NLM) and the NAL Agricultural Thesaurus (USDA), amongst others. It is possible to interactively browse by subject. Use of SKOS is planned. Individual metadata are not provided. The descriptions of each vocabulary are detailed and include information such as name, URL, update, edition, number/type of terms, type of access, download format if available, publisher/editor, proposals for new terms email if available, type of product, formats, acronym, and online availability.

Taxonomy Warehouse (2013) provides what is probably the oldest (2001) dedicated basic TR. Interactive access to the vocabulary metadata are offered, via search of taxonomies metadata, subject categories, and publishers, although these three fields are offered in a search box together with other services offered (blogs, books etc.).The metadata include: title, publisher, type of vocabulary, description, informational URL at publisher's web site, online/download URL, number of total terms, revision cycle, formats in which available, notation scheme, additional information (such as conditions of use, characteristics of the vocabulary such as types of relationships between terms, details on hierarchical levels, etc.). Some metadata are interlinked as in an ontology: publisher (Published By), language (Has Language), categories (Is About), relationship to other controlled vocabularies (Use or UF).

TaxoBank is a more recent commercial TR, providing interactive access, established in 2009 (Access Innovations, 2013). It contains a range of vocabularies, described using metadata recommended in the TRSS report (Golub & Tudhope, 2009), excluding the optional ones. Vocabulary providers/owners are invited to register their vocabularies and provide the metadata; comments on how they have used the vocabulary, how it could be improved, etc., are also encouraged.

Collection registries and extended metadata registries covering KOS vocabularies

The JISC Information Environment Service Registry (IESR, 2011), is a registry of JISC collections of electronic resources, together with associated services and agents, and associated metadata. Collections are described with metadata which include controlled subject terms from different vocabularies but with at least one DDC (Dewey Decimal Classification) term to ensure interoperable searching. IESR acts as middleware and is primarily intended for m2m access. Services are described using a bespoke scheme which includes a location address, technical method of accessing a collection or providing a service, and further description of technical access details.

The JISC Information Environment Metadata Schema Registry (IEMSR) is an example of a metadata registry defined as "an application that provides services based on information about metadata vocabularies, the component terms that make up those vocabularies, and the relationships between terms. This information about metadata vocabularies and their components is provided in the form of schemas" (Johnston, 2004). Functions might include discovery of information about terms, usage in metadata application profiles, guidelines for use, bindings, provenance of terms, support for mapping or inferencing (Heery, 2005).

METeOR (2013) is an Australian metadata registry for national data standards for the health, community services and housing assistance sectors, based on the international standards for metadata registries ISO/IEC 11179 (2013), while the DART Project at University of Queensland (DART) implemented a prototype metadata schema registry in the context of data sharing in e-Research and e-Government.

Dublin Core Metadata Registry (2011) provides an up-to-date source of authoritative information about DCMI metadata terms and related vocabularies. The registry has metadata for nine KOS vocabularies (referred to as Value Encoding Schemes). For the DCMI Type Vocabulary, a controlled list, not only the metadata about this vocabulary are available but also the individual terms in the list are registered.

Linked Open Vocabulary (LOV) describes RDF vocabularies, including mostly the metadata element sets that have been published as RDF vocabularies and OWL ontologies used by the Linked Data datasets (Vatant & Vandenbussche, 2013). Those descriptions contain metadata either formally declared by the vocabulary publishers or added by the LOV curators. Each vocabulary is described by metadata. The information of how vocabularies rely on, extend, specify, annotate or otherwise link to each other and the update history are all visualized through its own platform. According to the developer, its "LOV Aggregator" feature aggregates all vocabularies in a single endpoint/dump file. The last version of each vocabulary is checked on a daily basis. This endpoint is used to extract data about vocabularies, generate statistics ("LOV Stats" feature), and support research ("LOV Search" feature).

Data Hub (2013a) is a community-run catalogue containing over six thousand datasets as of May 2013. It uses an open-source data cataloguing software called CKAN, written and maintained also by Open Knowledge Foundation (2013b). Users can browse the available datasets through dozens of groups (e.g., Linking Open Data Cloud, BioPortal, Economics Datasets, etc.), learn about each dataset through the metadata and descriptions, access to the services provided by dataset providers through the links, and download Linked Data datasets in various formats, specifications and documentations. Since BioPortal is included, many domain-specific ontologies are also registered here. In addition there are more than 100 KOS registered, including term lists, dictionaries, name authorities, classification schemes, subject headings, thesauri, ontologies, and vocabulary registries (Zeng, 2012).

TR provides access to vocabulary content

Perhaps the oldest example of what could broadly be referred to as a TR dating from the 1980s is the Unified Medical Language System® (UMLS®) (U.S. National Library of Medicine, 2013). While UMLS® is an integrated system of over 50 biomedical vocabularies rather than a TR as such, it offers a set of tools allowing access to biomedical concepts and their relationships and maintains information on a given concept's source vocabulary. UMLS® is used in a variety of applications including information retrieval, natural language processing, creation of patient and research data, and the development of enterprise-wide vocabulary services.

The eXtended MetaData Registry (XMDR) project seeks to build upon the ISO 11179 Metadata Registries family of standards (Lawrence Berkeley National Laboratory, 2009). US Government agencies (DoD, EPA, USGS, National Cancer Institute, Lawrence Berkeley National Lab etc.), as well as some European partners such as EEA, are involved. It is developing a prototype-extended metadata registry, incorporating various terminologies and ontologies.

This effort has close links to the language engineering community and related ISO subcommittees (ISO/IEC JTC 1/SC 32, 1997; ISO/TC 37/SC 4, 2011). The focus seems to be on a registry of individual terms than on vocabulary schemes and collections (Bargmeyer, 2005).

OCLC's Terminology Services (OCLC Research, 2012) provides both interactive and m2m access (including OAIPMH Web services) to a selection of prominent vocabularies (Library of Congress Subject Headings, Medical Subject Headings (MeSH®), Faceted Application of Subject Terminology (FAST subject headings)). An operational version is available to OCLC member institutions worldwide. The vocabulary metadata are stored in MARC 21 Bibliographic data format. The metadata elements on the project website include name, description, date, identifier and links to external URL about the vocabulary, MARC statistics on fields and subfields, SRU interface. The concepts and terms can be retrieved in multiple representations including HTML, MARC XML, SKOS, and Zthes.

Lexaurus Bank (Knowledge Integration, 2011) is a commercial terminology management system for publishing vocabularies. Example applications include the Lexaurus Bank public vocabulary service in the field of education and also the more recent Culture Grid vocabulary bank, in collaboration with the Collections Trust (Collections Trust & Knowledge Integration, 2013). There is an alerting function which provides details of changes to vocabularies in RSS and ATOM formats. The metadata varies with the vocabulary but typically include: identifier, name, description, authority, language, category, date, rights, version and term count. Both interactive and m2m interfaces are available.

The United Nation's Food and Agriculture Organization (FAO) has implemented a combined registry of vocabularies, metadata sets and tools related mostly to agriculture, including over 100 controlled vocabularies with access to vocabulary content for most (Food and Agriculture Organization of the United Nations, 2012a). The metadata somewhat vary with the vocabulary and include: name of the vocabulary, URL, acronym, description, organization (owner/creator), languages available, URL for more info, additional URL, contact email, the list of tools that support its use, subject coverage, and vocabulary type. It includes AGROVOC, an influential thesaurus containing over 40,000 concepts in up to 22 languages (Food and Agriculture Organization of the United Nations, 2012b). The editing of AGROVOC is distributed using an open source editing tool. It is expressed in SKOS and published as Linked Data. An automated indexing tool based on AGROVOC is freely available as part of the registry. AGROVOC can be searched or browsed for terms, new terms can be suggested, it can be downloaded or accessed via Web services, the latter including about 50 methods.

The Open Metadata Registry (formerly NSDL registry) (Open Metadata Registry, 2013) provides interactive access. As both a metadata and terminology registry, it contains vocabulary content together with metadata element sets. The metadata elements used to describe vocabularies include owner, name, URL, note, community, status, language, URI base domain, URI token, URI, users name and whether he or she is an administrator, maintainer, or registrar. While originally built to support the US National Science Digital Library (NSDL), the Registry is freely available and the software is open source. Administrator users can create and maintain their own vocabularies via interactive forms. There are around 100 authors, both organizations and individuals, and about 300 controlled vocabularies.

BioPortal (National Center for Biomedical Ontology, 2013) is a prominent example of ontology registries, typically holding their content in OWL (Web Ontology Language) or OBO (Open Biomedical Ontologies) formats. BioPortal provides both interactive and m2m access. Metadata include: ontology identification number, Bioportal's PURL, status, format, categories, groups, contact, URLs for home page, for publications page, and for documentation page, description, license information, reviews, versions, views created by users, views-specific metadata, projects using an ontology, metrics which includes number of classes, number of individuals, number of properties, maximum depth, maximum number of siblings, average number of siblings, classes with a single subclass, classes with more than 25 subclasses, and classes with no definition.

Another ontology registry is the PRoteomics IDEntifications database (PRIDE) (European Bioinformatics Institute, 2013a) which contains proteins, peptides and spectra. The related ontology lookup service currently lists over 80 ontologies (European Bioinformatics Institute, 2013b). The Ontology Metadata Vocabulary (OMV) project has proposed detailed metadata elements for formal ontologies (Palma, Hartmann, & Haase, 2009), formalized as an OWL ontology. The metadata are grouped into a number of categories: availability (location of the ontology), applicability (intended usage or scope), format, provenance (organizations contributing to the creation of the ontology), relationship (relationships to other resources, versioning, extensions, generalization/specialization and imports), statistics (e.g., number of classes), and, other (information not covered in previous categories). Ontology metadata are also grouped into required, optional (important but not strongly required), and extensional (specialized metadata entities, which are not considered to be part of the core metadata).

ONKI (Semantic Computing Research Group SeCo, 2013) is a Finnish registry of ontologies as well as some controlled vocabularies such as MeSH. It is a part of the Finnish effort to build a national Semantic Web infrastructure. The 80 ontologies and vocabularies listed can be searched by name and browsed by subject (upper, domain, business, cultural, health, nature, public administration), by structure (class ontology, instance ontology, advanced vocabulary, simple vocabulary), publishing status, and publisher. Many of them can be downloaded.

Services providing mainly access to vocabulary content

With the increasing number of vocabularies that are published in SKOS, services that focus on the access to the vocabulary contents emerged. They are not considered TRs here, as they do not provide much in the way of structured metadata describing the vocabularies, though they may provide title and explanation of a vocabulary, the available downloading formats and links, and modification date.

The Library of Congress Linked Data Service: Authorities and Vocabularies (Library of Congress, 2013) makes publically available standards and vocabularies promulgated by the Library of Congress. The Linked Data approach is followed, in that each vocabulary possesses a resolvable URI, as does each datum value within it. For human inspection a search web interface for individual values is provided, and a visualisation interface of related concepts; in addition, a form to suggest terminology is available with each term. Individual metadata are not provided per se, but summary descriptions of each vocabulary appear to include information on: purpose, usage and function, types of terms included, relationships between terms, number of terms, update information, and standards definitions. M2m access is enabled as URI over HTTP requests. Support for download of both bulk vocabularies and individual concepts and headings is available. The vocabularies include Library of Congress Subject Headings and Thesaurus of Graphic Materials, as well as 10 more authorities such as ISO and MARC standards.

Helping Interdisciplinary Vocabulary Engineering (HIVE) (Greenberg et al., 2011) supports SKOS-based searching and browsing access to six controlled vocabularies. It also provides an automated indexing service whereby terms from the selected vocabularies are assigned to a document uploaded or found at a given URL. Metadata for each vocabulary include acronym, the number of concepts included, the number of relationships and date of last update.

NERC (National Environment Research Council) Vocabulary Server is an operational TR which supports the management and interoperability of scientific datasets in collaborating international data centres via m2m access (British Oceanographic Data Centre, 2013). The focus of the service is on providing support to data managers to assign and (automatically) validate scientific metadata by means of vocabularies, such as those describing instrumentation, geographic locations, temperature or measure units. Support is also provided to map from a term used in a local centre to an overarching term interoperable with other data centres. A significant subset of vocabularies is available for individual interactive search where the listed vocabularies are assigned the following metadata: key, long name, short name, version, and last modified.

Attributes of TRs

We first briefly report on the attributes resulting from Phase I of the research and then on the metadata proposed by Phase II.

Proposed attributes based on TR resource analysis in Phase I

From 1996 the NKOS community began an effort to design a TR, which resulted in several versions of a detailed metadata schema, and in numerous discussions at NKOS workshops in the years that followed. A Dublin Core based version became the third version of NKOS registry attributes (see Background). Apart from the two NKOS documents, Phase I reviewed the ISO/IEC 11179 standard, part 2: Classification (ISO/IEC 11179-2, 2005) which provides a conceptual model for managing classification schemes within a metadata registry and lists 44 elements, together with the elements proposed by Hodge, Salokhe, and Anderson (2007). The proposed attributes grouped into five categories are as follows (O indicating optional):

(1) General information. Elements in this group are intended for creating metadata descriptions that will facilitate the discovery of vocabularies and terminology services. They include: Vocabulary name; Vocabulary alternative name or acronym (O); Vocabulary type (whereby a recommendation for future work is to further develop the classification of different vocabulary types); Author or editor; Current version/edition; Date of current version/edition; Update frequency (O) (how often the vocabulary is updated); Available format(s); Available terminology services (O); Vocabulary identifier (e.g., URL, ISBN, DOI); Vocabulary sample URL (O) (a file with examples of actual contents to illustrate the nature of the product, in particular if the whole product is not freely available online); Vocabulary description (additional information that does not appear in other metadata).

(2) Scope and usage. Elements in this and the following group are intended for recording specific characteristics of vocabularies that will facilitate the evaluation of the vocabulary for a particular application or use. They include: Language(s) (in which the vocabulary is available or languages which it covers if multilingual); Major subjects covered; Minor subjects covered (O); Purpose as given by author/publisher; Used by (O) (a list of actual application contexts, e.g., document collections for which the vocabulary was designed or document collections in the vocabulary is used); Description of collections where used (O); Usage case study (O) (to further illustrate potential usage and (dis)advantages); Use in application profiles (O); Rating (O) (perhaps an automatically generated rating based on publisher, conformance to standards, spread of usage etc.); URL to vocabulary users' discussion board (O); Change notification details (O); Related vocabularies (O); Overlap with related vocabularies (O); Mappings to other vocabularies (O) (which vocabularies, whether mappings are intellectual or automated); URL to tutorial for applying vocabulary (O).

(3) Vocabulary characteristics. Type of display (O) (e.g., alphabetical, hierarchical, tagged format, classification tree, rotated (permutated), facetted, graphical); Description of overall structure (O) (overview of the organization structure, e.g., hierarchical, whether terms can belong to one or more hierarchies, whether plural forms are used, disambiguation device(s) used); Type of terms (O) (e.g., concept terms, geographic names, corporate names);

Types of relationships (O) (e.g., broader, narrower, related); Total number of terms (O); Total number of classes (O); Number of preferred terms (O); Number of non-preferred terms (O); Depth of hierarchy (O) (maximum number of levels); Notes fields (O) (types of notes fields available); Information given (O) (e.g., whether any of the following are provided: usage notes, conceptual relationships, references, date of entry, spelling variants etc.).

(4) Terms and conditions. Availability (free for all, free for registered users, costs); Import/download instructions (O); Purchase/subscription price; Licensing options (O).

(5) Provider. Vocabulary provider name; Vocabulary provider URL; Vocabulary provider contact details.

Proposed attributes for a Dublin Core Application Profile

The attributes proposed at Phase I (discussed in the above section) were generated based on the TR recourse analysis, with no conceptual model. In Phase II when developing the Dublin Core Application Profile for KOS resources (KOS-AP), the DCMI-NKOS Task Group followed the *Guidelines for Dublin Core Application Profiles* (Coyle & Baker, 2009) and the first task was to establish a domain model that characterizes the types of things described and their relationships in the context of the user tasks.

It is important to recognize that almost all KOS products are constantly updated and new versions are released, many have translations or extracts and are reused; and, last but not least, they are available as different deliverables and in different formats. A concept model needs to model such a network of entities and relationships. The Task Group adopted the FRBR model developed by a working group of the International Federation of Library Associations and Institutions (IFLA). According to this model the KOS product as a whole is a *work*, different versions in time and/or language are modelled as *expressions,* and the *manifestation* level covers different formats in which the KOS is published. Taking the ASIS Thesaurus example, the thesaurus as a whole is a *work*. Different versions (such as Version 1994 in English, Version 2005 in English, Version 2012 in French) are different *expressions* of this *work*. The printed edition of the 2010 English version and the SKOS Linked Data representation of the same version are examples of *manifestations* (DCMI/NKOS Task Group, 2012). The following figure presents such a model (DCMI/NKOS Task Group, 2013a). At the centre are the three entities and the relationships between the entities. The outlying entities have certain relationships with *work, expression,* and/or *manifestation*.

Insert Figure 1 here

Figure 1. KOS-AP concept model.

Two major types of relationships can be found based on this model. The first type contains the basic FRBR relationships between a *work* and its *expressions* and between an *expression* and its *manifestations*. The second type is between entities of the same type: *work*-to-*work, expression*-to-*expression, and manifestation*-to-*manifestation,* as listed in Table 1. All relationships listed below have inverse relationships. For example, is-part-of has an inverse relation has-part. Some of these relationships were considered in the TRSS report (Golub & Tudhope, 2009) as the attributes among those in Group 2 "Scope and usage".

Table 1. KOS-AP defined relations.

Insert Table 1 here

There are more detailed relationships between *expressions* which can be used as needed. For the "is part of" relation, the detailed ones are: "outline", "excerpt" and "fragment of". For "based on", the detailed relationships include "translation of", "abridgment of", "extension of" and "version of". For other relationships currently the only specific one is "aligned with". These relationships, in addition to basic FRBR relationships between *works, expressions* and *manifestations*, cover all usual scenarios of updating, development, transformation and reuse of KOS.

The attributes listed in Table 2 were chosen to represent the essential ones present in current KOS-AP, which also support the basic user tasks *find, identify, select, obtain* and *explore*. While some of them are assigned to only one entity type (e.g., "language" is an attribute of *expression*), several are appliccable to two or all three entity types. For example, *work, expression* and *manifestation* have separate "titles"; "rights" can be assigned to all three levels as well.

Table 2: Attributes of KOS-AP.

Insert Table 2 here

Although KOS-AP set the user task and conceptual model first, and then mapped the attributes available in TRSS report of Phase I, the mapping was very easy and straight-forward because the KOS-AP defined core attributes that are common. There are fewer elements in the KOS-AP in comparision with the TRSS attribute list. One consideration was the concern not to discourage vocabulary providers from contributing metadata by requiring information that may be hard for them to obtain. Additional elements, such as "frequency of update", intended "audience" and "used-by", are also defined because they were considered important in the TRSS study.

Functionality of Terminology Registries

We conclude by discussing the possible functionality of TRs in general; the major components are arranged loosely according to a TR information lifecycle framework. The framework is a version of a framework for terminology services proposed in Tudhope, Koch, and Heery (2006) which was based on an earlier review of semantic interoperability in digital libraries that had synthesized lifecycle models from knowledge representation and information science (Patel, Koch, Doerr, & Tsinaraki, 2005). The framework is here revised to accommodate TR purposes. This outline of functionality is a broad superset of possibilities; a particular TR might only include a selection from it. The relevant options for each lifecycle element are indicated and some indicative use cases are included, drawing on Proffitt, Waibel, Vizine-Goetz, and Houghton (2007).

Acquisition, creation and modification of vocabularies

This option encompasses the functionality to support the creation and editing or maintenance of vocabulary content. At the minimum, this includes an import facility supporting an upload of a complete vocabulary in a variety of formats. A more ambitious provision would support the ability to edit and modify the individual elements of vocabularies, with functions for addition, deletion, modification. These functions could be applied to terms, concepts, notes and, possibly, to the relationships themselves. Depending on the domain context, support may be needed for selection of vocabularies to be included in the registry. Usually individuals or groups will propose vocabularies to be supported by the TR. In some cases, this could require quality control as part of a review and selection process, which may entail resource overheads (see discussion on governance below).

In some situations, providing a vocabulary development environment can be important. Local organisations wishing to provide small vocabularies may not have the resources to build and maintain a vocabulary (possibly relying on word processing or spreadsheet applications). Concrete use cases might include managing local terminologies; establishing a project-specific subset of terms; joint editing and annotation of local vocabularies by experts; contributing to a published vocabulary; capturing locally contributed end-user vocabulary; and, sharing local vocabularies. Corresponding functionalities could include: vocabulary registration and upload; submission of metadata for submitted vocabulary; validation of submitted vocabulary; validation of metadata for submitted vocabulary; provision of identifiers (URIs) for each vocabulary and for vocabulary elements; editing; revision and extension; tracking and versioning; submission of new versions.

Maintenance can include support for versioning, which may be applied at different levels: to keep track of versions of complete vocabularies, or be applied at concept or term level. Support for collaboration might be offered to allow a community to jointly maintain and evolve a vocabulary. The community might be a tight-knit group of domain experts or a wider Web 2.0 community.

Publication of vocabularies

Publication is here taken to include support for selecting an appropriate license, or possibly for charging, where the vocabulary provider is a commercial entity. For full TRs, provision must be made to store the vocabularies and make them available. This may be in an internal TR representation format but export or download of whole or parts of vocabularies should be provided in a variety of standard formats.

Access, search and discovery

Support may be provided to search or browse metadata about vocabularies when the use case requires an unknown vocabulary to be discovered. For example, a user may search to see whether any vocabulary's subject coverage matches a search string, or is in a particular language. This requires appropriate metadata. Support may also be provided to identify vocabularies that are used to index particular collections or to identify vocabularies that can be accessed via particular services. It can also be applied to discovery of individual concepts or terms, where support should be provided to match a user string with terms (and optionally scope notes). For example, a list of candidate concepts may be offered, taken from a selected vocabulary or from all vocabularies held in the TR. There may also be scope for automated disambiguation assistance. It should also be remembered that support can be provided for both human and machine agents by providing web services for accessing individual terms and concepts. The latter could be of use to topical crawlers, for example.

Use

Once a concept (or term) has been identified and selected it may be used in a variety of applications, such as mapping, search, information extraction, text mining, automatic classification, data interlinking, personalisation, etc. Various forms of "toolbar terminology services" can also be envisaged. This may require functionality for searching and browsing of terminology services metadata. Some key uses are outlined below.

(1) Search. Vocabulary concepts, if uniquely identified, may support semantic search. They may also support query expansion, either via synonyms or with semantically close concepts. Employing query expansion can combine several search "moves" in the one query.

(2) Cataloguing. This includes functionality to support indexing (classification, annotation, tagging) and metadata creation activities. It could be achieved via: a cataloguing application associated with the TR; direct provision of web services by the TR itself; or making available information on third party services, e.g., by the vocabulary provider. Example use cases might include: metadata validation (e.g., inconsistency of controlled index terms in a repository and name validation), spell-checking, browsing, searching and retrieving terms, with more advanced options of automated controlled terms suggestion (e.g., at the time of deposit to a repository), or perhaps automatically generated metadata. Different services might be offered to professional cataloguers versus social taggers.

(3) Integration. This includes semantic interoperability support for mapping and possibly merging of vocabularies, including interoperability and/or merging between end-user vocabularies and published vocabularies. Both automated and intellectual techniques may be involved. This could be achieved via direct provision of terminology services for mapping or crosswalk by the TR, or making information available via third party services, e.g., by the vocabulary provider. Advanced options might include query expansion via automated mapping between vocabularies, combining local, shared or published vocabularies, disambiguation and multi-lingual services.

Archiving and preservation of vocabularies

Currently this could be subsumed under digital preservation generally. Long-term preservation of vocabularies is an important issue but is outside this paper's scope.

More generally, governance and management is a critical issue for TRs, encompassing the usual best computing practice technical governance but also the governance of vocabulary content. While issues will vary with the particular situation, they can include assigning responsibility for issues such as validation of correctness of content, versioning and maintenance (both vocabulary content and representation formats), which may include support for update of the whole vocabulary or individual elements, proposals for deprecated elements, evaluation of new vocabulary offered to the registry and judgment as to their inclusion, promotion of the TR and its services, education and training in the resources and services.

Content governance requires a responsible body in charge of the registry, with sufficient resources, longevity, and authority recognised for its purposes. There must be sufficient reason to justify allocation of the resources necessary for this by the parent body or funders. In the Phase I survey, several contacts highlighted the governance problems inherent in holding vocabulary content within the registry as a critical factor. In addition to maintaining current versions, the vetting, selection and quality control of vocabularies offered to the registry can impose significant demands on resources. This issue continues to be highly relevant today to the publication of vocabulary Linked Data, where appropriate assignment of intellectual property rights and copyright and a long term strategy for versioning are highly important.

Conclusion

Terminology Registries are discussed in terms of practical application in this article. TRs can, if used as a digital infrastructure service, make their vocabulary content available for both intuitive human inspection and for m2m access.

The paper has summarized the characteristics of various types of TRs, and presented a generalized view of the functionality of a TR. Ideally, it should be possible to both search and browse for a vocabulary matching a user's search. The capability to sort, by various criteria, a result list of vocabularies in a registry matching a user search is also desirable. A TR could also provide information about existing terminology services, accessible to both humans and machines, including information related to use of the service. Governance is an important issue.

The features of a vocabulary that allow for discovery vary widely, depending on the user's criteria. The user may have a rough idea of a particular vocabulary's title; the user may require a vocabulary covering a particular subject domain (to greater or lesser degree of specificity); it may be critical that the vocabulary be free to use; it may be important that the vocabulary be available in a particular language; or the depth or breadth of topic coverage may be an issue. To assist discovery to satisfy all these needs, a rich set of metadata should be available for the vocabulary. This metadata should be open to both human and m2m access. The challenge in attempting to promote a standard set of metadata elements is to build on best practices while focusing on a core set that vocabulary providers are likely to provide in practice.

The work for this paper in defining the most useful and common metadata attributes led the researchers to first investigate best practices and documentation from TR owners and previous initiatives, analysing the attributes of the TRs. In the second phase of the research, a domain model based on user tasks was constructed and a set of core metadata elements for use in TRs was proposed. The research findings thus result from a combination of bottom-up and top-down approaches. Although aimed at TR implementations, the finding results may also be applicable for use by KOS descriptions outside of TRs, for example, as microdata to be embedded in a website of a KOS resource.

Whether embedded within broader registry frameworks, or existing as independent registries, TRs are a crucial element of the infrastructure required for resource discovery services, digital libraries, Linked Data and semantic interoperability generally. It is hoped that this paper may play some part in helping to encourage further work towards the integration of both traditional library vocabularies and emerging vocabularies in the wider networks made possible by current technology.

Acknowledgments

The TRSS project was funded by United Kingdom's Joint Information Systems Committee (JISC), Repositories and Preservation Programme, Shared Infrastructure Services strand. Thanks also to the various national and international contacts who provided input to the study. The KOS-AP project is an on-going project which involved the members of the DCMI/NKOS Task Group, especially Gail Hodge, Marjie Hlava, Joseph Busch, and Traugott Koch.

References

Access Innovations. (2013). TaxoBank: Access, deposit, save, share, and discuss taxonomy resources. Retrieved May 22, 2013, from http://www.taxobank.org/

ANSI/NISO Z39.19-2005: R2010. (2005). Guidelines for the construction, format, and management of monolingual controlled vocabularies. Retrieved May 22, 2013, from http://www.techstreet.com/standards/niso/z39_19_2005?product_id=1262086

Asset Description Metadata Schema Working Group. (2013). Asset Description Metadata Schema (ADMS). Retrieved May 22, 2013, from http://joinup.ec.europa.eu/asset/adms/home

Baker, T., Blanchi, C., Brickley, D., Duval, E., Heery, R., Johnston, P., … Shigeo, S. (2002). Principles of metadata registries: A white paper of the DELOS Working Group on Registries. Retrieved May 22, 2013, from http://web.archive.org/web/20110818100958/http://delos-noe.iei.pi.cnr.it/activities/standardizationforum/Registries.pdf

Bargmeyer, B. (2005, June). eXtended Metadata Registries (XMDR). Paper presented at the 7th NKOS Workshop of Joint Conference on Digital Libraries (JCDL), Denver, CO. Retrieved May 22, 2013, from http://nkos.slis.kent.edu/2005workshop/Bargmeyer.ppt

British Oceanographic Data Centre. (2013). NERC vocabulary server. Retrieved May 22, 2013, from http://www.bodc.ac.uk/products/web_services/vocab/

BS 8723: Structured vocabularies for information retrieval – Guide. (2005). London: British Standards Institution.

CENDI. (2010). CENDI Agency terminology resources. Retrieved May 22, 2013, from http://www.cendi.gov/projects/proj_terminology.html

Chan, L.M., & Polland, R. (1988). Thesauri used in online databases: An analytical guide. New York: Greenwood Press.

Collections Trust & Knowledge Integration. (2013). Culture grid vocabulary bank. Retrieved May 23, 2013, from http://culturegrid.lexaurus.net/culturegrid/home

Coyle, K., & Baker, T. (2009). Guidelines for Dublin Core Application Profiles. Retrieved May 23, 2013, from http://dublincore.org/documents/profile-guidelines/

DCMI Registry Community. (2010). Retrieved May 22, 2013, from http://dublincore.org/groups/registry/

DCMI/NKOS Task Group. (2012). KOS example. Retrieved May 22, 2013, from http://wiki.dublincore.org/index.php/KOS_example

DCMI/NKOS Task Group. (2013a). Core elements: NKOS-AP. Retrieved May 22, 2013, from http://wiki.dublincore.org/index.php/Core_Elements

DCMI/NKOS Task Group. (2013b). DCMI/NKOS Task Group Wiki. Retrieved May 22, 2013, from http://wiki.dublincore.org/index.php/DCMI_NKOS_Task_Group

Dextre Clarke, S.G. (2005). Organising access to information by subject. In A. Scammell, Handbook of information management (pp. 75-113). London: Routledge.

Dublin Core Metadata Registry. (2011). Version 3.3.8. Retrieved May 23, 2013, from http://dcmi.kc.tsukuba.ac.jp/dcregistry/

European Bioinformatics Institute. (2013a). PRoteomics IDEntifications database (PRIDE). Retrieved May 23, 2013, from http://www.ebi.ac.uk/pride/

European Bioinformatics Institute. (2013b). OLS – Ontology Lookup Service. Retrieved May 23, 2013, from http://www.ebi.ac.uk/ontology-lookup

European Commission. (1993). Thesaurus guide: Analytical directory of selected vocabularies for information retrieval. 2nd ed. Luxembourg: Euratom.

Food and Agriculture Organization of the United Nations. (2012a). VEST registry: Vocabularies. Retrieved May 23, 2013 from http://aims.fao.org/vest-registry.

Food and Agriculture Organization of the United Nations. (2012b). AGROVOC. Retrieved May 23, 2013 from http://aims.fao.org/standards/agrovoc/about.

Golub, K. & Tudhope, D. (2009). Terminology Registry Scoping Study (TRSS): Final report. Retrieved May 22, 2013, from http://www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf

Greenberg, J., Losee, R., Agüera, J. R. P., Scherle, R., White, H. & Willis, C. (2011). HIVE: Helping Interdisciplinary Vocabulary Engineering. Bulletin of the American Society for Information Science and Technology, 37, 23–26.

Heery, R. (2005, September). (Metadata and) vocabulary registries. Paper presented at the NKOS Special Session of Special Session of the International Conference on Dublin Core and Metadata Applications, Madrid, Spain. Retrieved May 22, 2013, from http://www.ukoln.ac.uk/terminology/events/NKOSatDC2005/heery-nkos-v2.ppt

Hierppe, R. (1990). A framework for characterizing systems for knowledge organization: A first basis for comparisons and evaluations. In Fugmann, R. (Ed.), Tools for knowledge organization and the human interface, 1st International ISKO-Conference (pp. 21–46). Frankfurt/Main, Germany: Indeks.

Hodge, G. (1999, August). NKOS thesaurus registry update. Paper presented at NKOS Workshop of ACM Digital Libraries Conference, Berkeley, CA.

Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. Washington, DC: Council on Library and Information Resources. Retrieved May 22, 2013, from http://www.clir.org/pubs/reports/pub91/contents.html

Hodge, G., Salokhe G., Zolly, L., & Anderson, N. (2007, July). Terminology resource registry: Descriptions for humans and computers. Paper presented at 10th Open Forum on Metadata Registries "Integrating Standards in Practice", New York City, NY.

IESR: JISC Information Environment Service Registry. (2011). Retrieved May 22, 2013, from http://iesr.ac.uk/

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). Functional Requirements for Bibliographic Records (FRBR): Final report. München: KG Saur. Retrieved 22 May, 2013, from http://www.ifla.org/files/cataloguing/frbr/frbr.pdf

IFLA Working Group on Functional Requirements for Subject Authority Records (FRSAR). (2011). Functional Requirements for Subject Authority Data (FRSAD): A conceptual model. M.L. Zeng, M. Zumer, & A. Salaba (Eds.). Berlin/München: De Gruyter Saur. Retrieved 22 May, 2013, from http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-model.pdf

ISO 12620:2009, Terminology and other language and content resources -- Specification of data categories and management of a data category registry for language resources. (2009). Retrieved May 22, 2013, from http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=37243

ISO 25964-1, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. (2011). Geneva: International Organization for Standards, August 8, 2011.

ISO 25964-2, Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies. (2013). Geneva: International Organization for Standards, March, 2011.

ISO 2788:1986,Documentation -- Guidelines for the establishment and development of monolingual thesauri (1986). Geneva: International Organization for Standards, 1986.

ISO 704:2009, Terminology work -- Principles and methods. (2009). Retrieved May 22, 2013, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38109

ISO/IEC 11179: Information Technology – Metadata Registries. (2013). Retrieved May 22, 2013, from http://metadata-standards.org/11179/

ISO/IEC 11179-2: Classification. (2005). Retrieved May 22, 2013, from http://standards.iso.org/ittf/PubliclyAvailableStandards/c035345_ISO_IEC_11179-2_2005(E).zip

ISO/IEC JTC 1/SC 32: Data management and interchange. (1997). Retrieved May 23, 2013, from http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/iso_technical_committee.htm?commid=45342

ISO/TC 37/SC 4: Language resource management. (2011). Retrieved May 23, 2013, from http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=297592

JISC. (2009). Terminology Registry Scoping Study (TRSS) Website. Retrieved May 22, 2013, from http://www.jisc.ac.uk/whatwedo/programmes/reppres/sharedservices/terminology.aspx

Johnston, P. (2004). JISC IE Metadata Schema Registry: Functions of the IE Metadata Schema Registry. Retrieved May 22, 2013, from http://www.ukoln.ac.uk/projects/iemsr/wp2/function/

Knowledge Integration. (2011). Lexaurus Bank. Version 5. Retrieved May 23, 2013, from http://www.vocman.com/?q=lexaurusbank

Koch, T. (2007). Controlled vocabularies, thesauri and classification systems available on the WWW. Retrieved May 22, 2013, from http://web.archive.org/web/20110512152252/http://www.mpdl.mpg.de/staff/tkoch/publ/koslist.html

Lambe, P. (2007). Organising knowledge: Taxonomies, knowledge and organizational effectiveness. Oxford: Chandos Publishing.

Lawrence Berkeley National Laboratory. (2009). eXtended MetaData Registry (XMDR) Project. Retrieved 3 January, 2012 from http://hpcrd.lbl.gov/SDM/XMDR/

Liang, A., Salokhe, G., Sini, M., & Keizer, J. (2007). Towards an infrastructure for semantic applications: Methodologies for semantic integration of heterogeneous resources. In J. Greenberg & E. Mendez Rodriguez (Eds.), Knitting the Semantic Web (pp. 161-189). New York: The Haworth Information Press.

Library of Congress. (2013). LC Linked Data Service: Authorities and Vocabularies. Retrieved May 22, 2013, from http://id.loc.gov/

METeOR: METadata Online Registry. (2013). Retrieved May 22, 2013, from http://meteor.aihw.gov.au/content/index.phtml/itemId/181162

Middleton, M. (2008). Controlled vocabularies. Retrieved May 22, 2013, from http://web.archive.org/web/20091025234954/http://www.imresources.fit.qut.edu.au/vocab/

Miles, A. & Bechhofer, S. (2009). SKOS: Simple Knowledge Organization System Reference, W3C Candidate Recommendation. Retrieved May 22, 2013, from http://www.w3.org/TR/skos-reference/

Milstead, J. (1998). ASIS thesaurus of information science and librarianship. 2nd ed. Medford, NJ: Information Today.

National Center for Biomedical Ontology. (2013). Bioportal. Retrieved May 22, 2013, from http://bioportal.bioontology.org/

Networked Knowledge Organization Systems and Services (NKOS). (1998a). Results of the NKOS '97 & '98 Workshops. Retrieved 22, May 2013, from: http://nkos.slis.kent.edu/workshop_results.html

Networked Knowledge Organization Systems and Services (NKOS). (1998b). NKOS Registry -- draft set of thesaurus attributes. Retrieved 22, May 2013, from http://nkos.slis.kent.edu/Thesaurus_Registry.html

Networked Knowledge Organization Systems and Services (NKOS). (2001). Networked Knowledge Organization Systems (NKOS) Registry: Help document. Retrieved 22, May 2013, from http://nkos.slis.kent.edu/registry2.htm

Networked Knowledge Organization Systems/Services/Structures (NKOS). (2013). Retrieved May 22, 2013, from http://nkos.slis.kent.edu/

OCLC (Online Computer Library Center). (2013). Dewey services. Retrieved May 22, 2013, from http://www.oclc.org/dewey.en.html

OCLC (Online Computer Library Center). (2013). WorldCat. Retrieved May 22, 2013, from http://www.worldcat.org/

OCLC Research. (2011). Terminology services: Experimental services for controlled vocabularies. Retrieved May 22, 2013, from http://tspilot.oclc.org/resources/

Open Knowledge Foundation. (2013a). Data Hub. Retrieved May 22, 2013, from http://datahub.io/

Open Knowledge Foundation. (2013b). CKAN. Retrieved May 23, 2013, from http://ckan.org/

Open Metadata Registry. (2013). Retrieved May 22, 2013, from http://metadataregistry.org/

Palma, R., Hartmann, J., Haase, P. (2009). OMV: Ontology Metadata Vocabulary for the Semantic Web. Version 2.4.1. Retrieved May 23, 2013, from http://sunet.dl.sourceforge.net/project/omv2/OMV%20Documentation/OMV-Reportv2.4.1.pdf

Patel, M., Koch, T., Doerr, M., & Tsinaraki, C. (2005). Report on semantic interoperability in digital library systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1. Retrieved May 23, 2013, from http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/SI-in-DLs.pdf

Proffitt, M., Waibel, G., Vizine-Goetz, D., Houghton, A. (2007, September). Terminologies strawman. Terminologies Services Meeting, New York City, NY. Retrieved May 23, 2013, from http://www.oclc.org/programs/events/2007-09-12a.pdf

Redmond-Neal, A. & Hlava, M.M.K. (2005). ASIS&T Thesaurus of Information Science, Technology, and Librarianship. Third Edition. Information Today, Inc.

Semantic Computing Research Group SeCo. (2013). ONKI: Ontology library service. Retrieved May 23, 2013, from http://onki.fi/en/

Soergel, D. (2001, June). Evaluation of Knowledge Organization Systems (KOS): Characteristics for describing and evaluating KOS. Paper presented at Classification crosswalks: Bringing communities together, the 4th NKOS Workshop of ACM-IEEE Joint Conference on Digital Libraries (JCDL), Roanoke, VA. Retrieved May 22, 2013, from http://nkos.slis.kent.edu/2001/SoergelCharacteristicsOfKOS.pdf

Souza, R.R., Tudhope, D., & Almeida, M.B. (2012). Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems. Knowledge Organization, 39, 3, 179-192.

Taxonomy Warehouse. (2013). Retrieved May 22, 2013, from http://www.taxonomywarehouse.com/

Terminology. (n. d.). In Collins English Dictionary. Retrieved May 22, 2013, from http://www.collinsdictionary.com/dictionary/english/terminology.

Terminology. (n. d.). In Merriam-Webster Online. Retrieved May 22, 2013, from http://www.merriam-webster.com/dictionary/terminology

The 9th International Forum on Metadata Registry: Harmonization of Terminology, Ontology and Metadata. (2006, March), Kobe City, Japan. Retrieved May 22, 2013, from http://taalunieversum.org/agenda/754/9th_international_forum_on_metadata_registry/

Tudhope, D., Koch, T., & Heery, R. (2006). Terminology services and technology: JISC state of the art review. Retrieved May 22, 2013, from http://www.jisc.ac.uk/Terminology_Services_and_Technology_Review_Sep_06

U.S. National Library of Medicine. (2013). Unified Medical Language System® (UMLS®). Retrieved May 23, 2013, from http://www.nlm.nih.gov/research/umls/

UKOLN. (2003). CORES Registry. Retrieved May 22, 2013, from http://web.archive.org/web/20120407065154/http://www.cores-eu.net/registry/

University of Toronto. (2003). Subject Analysis Systems (SAS) collection. Retrieved May 22, 2013, from http://plc.fis.utoronto.ca/resources/inforum/sas.htm

Vatant, B. & Vandenbussche, P. (2013). Linked Open Vocabulary (LOV). Retrieved 23, May 2013, from http://lov.okfn.org/dataset/lov

Vizine-Goetz, D. (2001). Networked Knowledge Organization Systems (NKOS) Registry: Reference document for data elements. Retrieved May 22, 2013, from http://staff.oclc.org/~vizine/NKOS/Thesaurus_Registry_version3_rev.htm

W3C Library Linked Data Incubator Group (LLD-XG). (2011). Datasets, value vocabularies, and metadata element sets. A. Isaac, W. Waites, J. Young, & M. Zeng (Eds.). Retrieved May 22, 2013, from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

W3C Working Group. (2004). Web services architecture. In Editors D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, & D. Orchard (Eds.). Retrieved May 22, 2013, from http://www.w3.org/TR/ws-arch/

Wright, S. E. (2008, September). Typology for knowledge representation resources. Paper presented at NKOS-CENDI Workshop, The World Bank, Washington, DC. Retrieved May 22, 2013, from http://nkos.slis.kent.edu/2008workshop/SueEllenWright.pdf

Zeng, M.L. & Chan, L.M. (2010). Semantic interoperability. In M. Bates (Ed.), Encyclopedia of Library and Information Sciences, 3rd ed., 1, 4645-4662.

Zeng, M.L. (2012, October). The state of KOS in the Linked Data movement - The publishing, management, and interoperating of KOS for the Semantic Web. Panel presentation at the American Society for Information Science and Technology (ASIS&T) 75th Annual Meeting, Baltimore, Maryland. Retrieved May 22, 2013, from http://www.slideshare.net/MarciaZeng/zeng-asist2012

Zeng, M.L., Hillmann, D.I., & Sutton, S. (2008, September). Metadata registries vs Terminology registries vs Service/Collection registries: Synergies and differences. Paper presented at the NKOS Special Session of the International Conference on Dublin Core and Metadata Applications, Berlin, Germany.

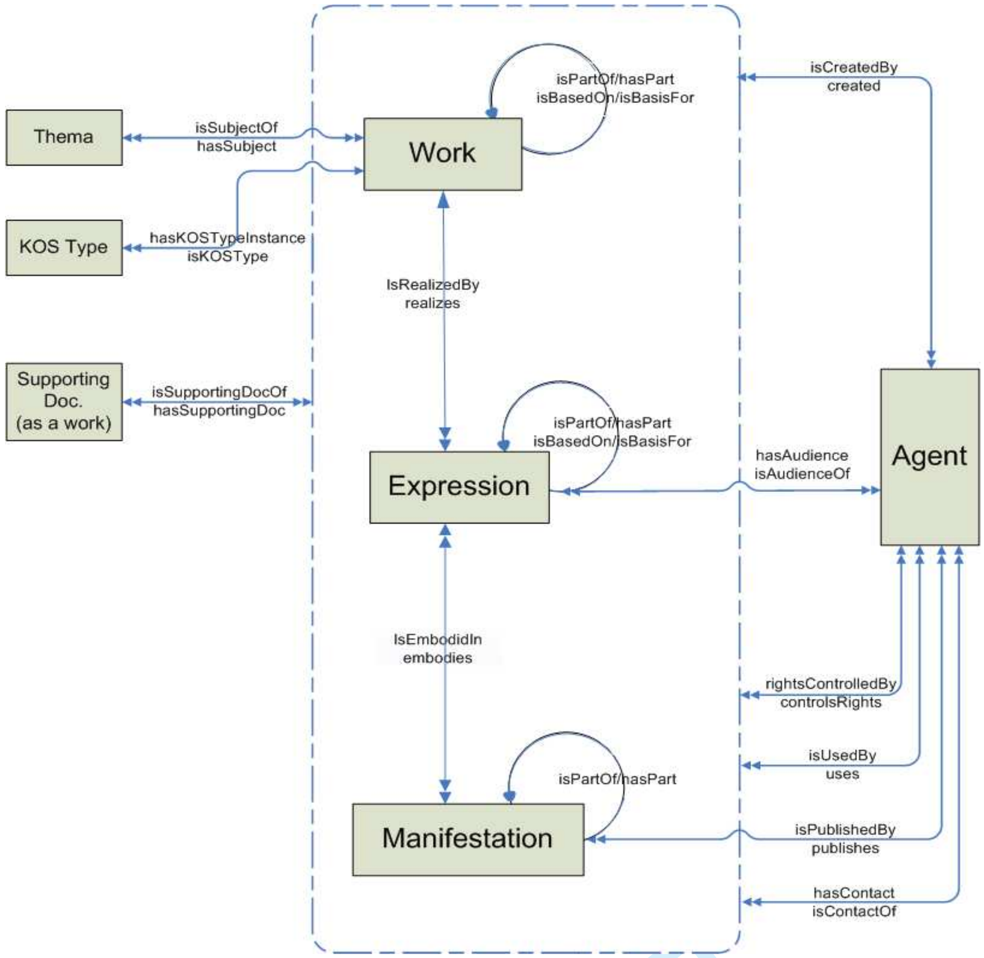

Figure 1. KOS-AP concept model.

Table 1. KOS-AP defined relations.

Between entities of different types	*work (W)-to-expression (E):*	(E) realizes (W)
	expression (E)-to-manifestation (M):	(M) embodies (E)
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)
	manifestation (M)-to-manifestation (M):	part of (M)

Table 2: Attributes of KOS-AP.

Core Attributes	*Associated with*			*Metadata Elements*
	Work	*Expression*	*Manifestation*	
Title	X	X	x	dct:title
Identifier	X	X	x	dctidentifier
Description	X	X	x	dct:description
Type (of KOS)	X			nkos:kosType
Language		X		dct:language
Creator	X	X	x	dct:creator
Contact		X	x	adms:contactPoint
Rights	X	X	x	dct:rights
Publisher			x	dct:publisher
Format			x	dct:format
Date (created or issued)	X	X	x	dct:created, dct:issued
Date (updated)		X		dct:modified
Subject	X			dct:subject
Relation (to other KOS)	X	X	x	dct:relation
Sample (a relation)		X	x	adms:sample
Supporting doc. (a relation)	X	X	x	wdrs:describedBy
Used by (a relation)		X		nkos:usedBy
Additional Attributes:				
Frequency of Update		X		nkos:updatePrequency
Audience	X	X		dct:audience
Size (of vocabulary)		X		nkos:size
Service offered			x	nkos:serviceOffered