Terminology Registry Scoping Study (TRSS)

Final Report

Koraljka Golub UKOLN, University of Bath
Douglas Tudhope University of Glamorgan

Document details

Date:	22/10/08, revised 03/07/09
Version:	Revised final draft after comments
Notes:	Circulation to JISC Development Team for approval

Acknowledgement to funders

This work was funded as part of the Joint Information Systems Committee (JISC) Information Environment Programme.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

Acknowledgements

Thanks to all the national and international contacts who provided input to the study: Amanda Hill, Andrew Houghton, Ann Apps, Brian Matthews, Catherine Jones, Carole Goble, Dagobert Soergel, Dennis Nicholson, Debra Hiom, Diane Hillmann, Diane Vizine-Goetz, Emma Tonkin, Gail Hodge, James Reid, Jane Hunter, Joan Cobb, Jon Phipps, Lorna Campbell, Mahendra Mahey, Makx Dekkers, Marcia Zeng, Margherita Sini, Margie Hlava, Nick Poole, Paul Walk, Philip Carlisle, Philip Lord, Robert Stevens, Roy Lowry, Sean Bechhofer, Simon Coles, Sophia Ananiadou, Stella Dexter Clarke, and Traugott Koch. The views expressed, however, are the authors, informed by the various input. Special thanks to Emma Tonkin also for contribution on IEMSR.

Table of Contents

Executive Summary

As part of its Capital Funding Programme, the Joint Information Systems Committee (JISC) is supporting further work to realize a rich information environment within the learning and research communities. This scoping study analyses issues related to the potential delivery of a terminology registry as a shared infrastructure service within the UK's further and higher education's information environment (IE).

Purpose

The study's overall aims are:

- To inform the development of shared infrastructure for resource discovery;
- To describe the scope and potential use of a terminology registry;
- To analyse requirements for services based on a terminology registry; and,
- To help stakeholders understand the need for this component of a shared infrastructure.

The report is based on a review of related projects and literature, as well as data collected from a number of interviews and questionnaires. It proposes a terminology registry and describes its characteristics and components, underlying standards, architecture and governance.

Overview of report contents

The report defines terms and briefly reviews vocabularies of different types. Terminology Registries (TRs) are distinguished from other types of registry. The methodology of the study is described.

The main options for TRs are outlined and the functionality of TRs and terminology services generally is located within an information lifecycle framework for terminology services. The rationale for a general JISC TR is discussed. The use cases gathered, as part of the project, are presented.

Existing TRs and related registries are reviewed and categorised according to the main functionality options. Metadata for existing TRs are reviewed in depth, along with some suggested new metadata elements, and a core/optional set of TR metadata is recommended.

Relevant standards for representation, identifiers and protocols are briefly outlined. Governance issues and their implications are discussed. The report concludes with a set of options and recommendations for the JISC.

Key points

A TR allows discovery of suitable vocabularies for information or, potentially, use, by exposing rich metadata about them for navigation and retrieval. A TR might hold vocabulary level information only, or additionally comprise the member terms, concepts and relationships, and also provide or list services based on terminology.

The wide range of potential functionality and use cases for a TR demonstrate that a TR offers a distinctive set of potential benefits in its own right. There is significant interest in TRs both nationally and internationally.

There are two general architectural issues for TRs. The first is whether a TR is intended to support human access or m2m. Within the context of the JISC IE and eFramework, we assume that any JISC TR will support both human and m2m access. The second key architectural issue is whether the TR provides access solely to metadata on vocabularies or whether access is also provided to the vocabulary content (concepts, terms, relationships). Third party terminology services might also be

available, adding value to vocabulary content. We thus distinguish three broad elements of TR architecture functionality:

1. Registry provides metadata for each vocabulary and links to vocabulary owner/provider
2. Registry provides metadata on (and links to) any available terminology services
3. Registry provides access to vocabulary content (either by downloading the complete vocabulary, or providing access to a vocabulary's concepts, terms and relationships)

These three elements can be combined independently (in practice, most combinations would include element 1). The three elements correspond to options that JISC might choose from.

In the short term, based on the various governance arguments, the majority of respondents to the study tended to favour some version of Option 1 for any general JISC TR, with the registry maintaining rich metadata and possibly linking to terminology services.

Currently, there appear significant resource and cost/benefit implications in holding content of large, general vocabularies inside any JISC registry, along with possible IPR issues. Since major JISC projects tend to involve large general vocabularies, where some content is licensed and due to the management and governance issues, in our view it is not cost-effective in the immediate future to build and manage a registry that holds and distributes content for the large, general vocabularies. This may change and the incremental steps outlined could allow reconsideration of Option 3 at a later date.

Recommendations

Recommendation 1: JISC should consider a TR for UK HE purposes.

Recommendation 2: JISC should consider providing an Option 1 TR (provides metadata for each vocabulary and links to vocabulary owner/provider), as part of an extended IESR. The registry would be made available both for human inspection and m2m access.

A focused design (small) project should be set up for IESR and relevant stakeholders to consider the implications and, assuming it is considered practical, make a proposal of the design and tender for the work packages.

Option 1 would allow the situation to be reconsidered at a later date and decisions on further steps towards holding vocabulary content could be taken if warranted. Although detailed costs would be proposed by the focused design project, we anticipate that the costs for Option 1 (and also 2) would be fairly modest.

Recommendation 3: In the medium term, a pilot Option 2 (for both human and m2m access) should be considered after a collaborative study on an initial set of appropriate metadata elements for terminology services.

Recommendation 4: JISC should investigate the possibility of a licensing arrangement with OCLC to access vocabulary content and terminology services via an OCLC TR, augmented for JISC purposes.

Recommendation 5: JISC should track major international projects, involving a TR, including NSDL and Europeana. Major national projects include BODC and Lexaurus Bank/Editor, which should also be tracked.

Recommendation 6: JISC should consider the possibility of establishing some form of TR support and advisory effort that would act as a hub for management, inquiries, training, promotion and dissemination of any JISC TR. We anticipate that this would be a relatively modest cost, not exceeding 0.5 FTE effort. It would also investigate via small projects key future issues and potential future development of the TR.

Recommendation 6b: As an alternative to proceeding immediately with Option 1, JISC could consider an interim step where the TR Support Project was assigned an additional set of tasks that attempted to gauge the level of interest and support for a general TR within the JISC community.

Recommendation 7: TRs are advised to consider (as appropriate for their circumstances and functionality options) the vocabulary metadata element set tentatively recommended in Section 7.2.

Recommendation 8: A TR (Option 3) holding vocabularies internally should adopt SKOS as one of the representation formats for import and export.

Recommendation 9: Concept identifiers should be based on URIs (Option 3).

Recommendation 10: A TR (Option 2) should follow a service-oriented architecture and offer web service access, if possible via a variety of common standards.

1 Introduction

The need for controlled vocabularies or knowledge organization systems, such as thesauri and classification schemes, for resource discovery and terminology related services has been well recognized and established (see, for example, Lancaster 2003, Svenonius 2000).

Due to the large number of available vocabularies, the variety of potential applications and new possibilities offered by standards in digital representation and protocols, the issue of a terminology registry has become highly relevant. Even before the World Wide Web, comprehensive lists of vocabularies were collected. Today a number of related domain, national and international initiatives exist.

In 2007 JISC initiated a scoping study which is to analyse issues related to the potential delivery of a terminology registry as a shared infrastructure service within the UK's further and higher education's information environment (IE). Although certain existing terminology registries could be of some use to IE, they are not comprehensive but usually domain-specific, and authority and maintenance issues exist.

The study's overall aims are:

- To inform the development of shared infrastructure for resource discovery;
- To describe the scope and potential use of a terminology registry;
- To analyse requirements for services based on a terminology registry; and,
- To help stakeholders understand the need for this component of a shared infrastructure.

The report is based on a review of related projects and literature, as well as data collected from a number of interviews and questionnaires. It proposes a terminology registry and describes its characteristics and components, underlying standards, architecture and governance.

The report is structured as follows:

- **Section 2 Background** provides definitions of relevant concepts, types of vocabularies, with an indicative vocabulary list;
- **Section 3 Methodology** briefly describes the methodology for the study and its scope;
- **Section 4 Architecture and functionality of terminology registries** addresses the scope of TRs functionality and a rationale for a TR in the JISC IE context;
- **Section 5 Use cases for vocabularies and terminology services** presents collected use cases and related requirements;
- **Section 6 Review of terminology registries** examines existing terminology and related registries, their characteristics and components;
- **Section 7 Metadata** compares metadata in existing terminology registries and suggests metadata for use in the proposed terminology registry;
- **Section 8 Underlying standards** deals with relevant standards for representation, identification of concepts, terms and vocabularies, as well as protocols, profiles and APIs;
- **Section 9 Governance** discusses governance issues;
- **Section 10 Recommendations and options for JISC** is an overview of recommendations and options in the JISC context.
- Appendices follow and include:
    - **Appendix 1 Survey letter with questions**;

- **Appendix 2 Interview invitation letter**;
- **Appendix 3 List of people who provided input to the study**; and,
- **Appendix 4 Metadata with examples**.

2 Background

2.1 Definitions and examples

Controlled vocabulary (in further text: **vocabulary**). Vocabulary control aims to reduce the ambiguity of natural language when describing and retrieving items for purposes of information searching. Controlled vocabularies consist of terms, words from natural language selected as useful for retrieval purposes by the vocabulary designers. A term can be one or more words. A term is used to represent a concept.

Two features (synonyms and ambiguity) in natural language pose potential problems.

a) Different terms (synonyms) can represent the same concept.

b) The same term (homographs) can represent different concepts.

A controlled vocabulary can attempt to reduce ambiguity between terms by:

- defining the scope of terms - how they are to be used within a particular vocabulary.

- providing a set of synonyms or effective synonyms for each concept

- restricting scope so that terms only have one meaning (and relate to only one concept).

Not all vocabularies provide all three features above. Some are just simple lists of authorized terms (term lists). Section 2.2 briefly discusses different types of vocabularies.

Knowledge organization systems (KOS) are controlled vocabularies, which are organized and structured via different types of semantic relationships.

Terminology is often used in connection with registries and services. Dictionary definitions include “the technical or special terms used in a business, art, science, or special subject” (Merriam Webster Online), and “[t]he system of terms belonging to any science or subject; technical terms collectively; nomenclature”.

The scoping study focuses on vocabularies as defined above and mostly uses the term ***vocabulary***. The term ***KOS*** is used when that is the term applied in a particular registry or standard discussed. The term ***terminology*** is used when registries are discussed and with regard to services based on vocabularies.

Terminology services (TS) are (web) services that present and apply vocabularies, both controlled and uncontrolled, including their member terms, concepts and relationships (Tudhope, Koch, Heery 2006). Their major purpose is improving document and information discovery. They can be machine-to-machine (in further text: m2m) 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, browsing, query expansion, mapping, subject indexing and classification, semantic reasoning, etc.

See, for example, FAO’s AGROVOC web services, including services that enable one to: retrieve all terms that contain the search term; retrieve termcode, labels, synonyms, broader, narrower and related terms of the search term; retrieve the definitions, history or scope notes of a term; and, retrieve an extended search query with all synonyms of the search term.

HILT (HILT 2008) provides subject interoperability services between different controlled vocabularies used in different collections, mapping to the Dewey Decimal Classification (DDC) as the central spine. It started in 2000 with Phase I, and is currently in Phase IV with various pilots and demonstrators. Apart from aiming at allowing interoperable subject cross-searches, HILT also provides m2m information (via SOAP and SRW) on terminology sets which can be used to enhance the precision of subject searches. The HILT Phase 4 Demonstrators (2008) show various web services including vocabulary look up, browsing, searching. Terminology sets are structured using SKOS. HILT is

hoping to collaborate with IESR (see below) to select collections appropriate to a user's subject request.

Another example is GeoCrossWalk (GeoCrossWalk 2006). It is a JISC middleware which implements a digital gazetteer service and server designed to make geographic searching transparent by 'crosswalking' different geographical reference or search terms, e.g., turning place names into coordinates; postal areas into place names; coordinates into places etc. across different resources. For example, GeoCrossWalk would translate the user's postcode into whatever type of geographical reference is needed to search resources in a portal, which occurs invisibly to the user, who will enter a postcode and receive a list of relevant results, unaware that the geographical cross-walking has taken place.

2.1.1 Registries

Few definitions of a registry in our context exist; one example is the following: "Registries provide an index or description of the information held or maintained by an organization or community of interest…" (Kotok 2003). Special types of registries exist and include terminology registries, metadata registries, and collection registries.

Heery (2005) discusses the relationship between a metadata registry and a terminology registry, saying that there are obvious differences between 'metadata element sets' and 'subject vocabularies' as to different relationship between terms, different use cases and communities, different standards and different conventions. However, the two are also complementary since they contribute to same 'business processes', e.g., enterprise portal, records management, resource discovery, and contribute to same workflows and choreographed services. Metadata elements can be seen as existing within the 'attribute space', whereas the vocabulary elements that may comprise the metadata element content exist within the 'value space' (Baker *et al.* 2002).

A **terminology registry (TR)** lists, describes, identifies and points to sets of vocabularies available for use in information systems and services. It can cover free and publicly available, fee-based and restricted, or organisation-internal vocabularies. Different vocabulary types could range from ontologies, thesauri, classification schemes, authority files and synonym rings to lexical databases, encyclopaedias and others. The registry allows discovery of suitable schemes for information or, potentially, use, by exposing rich metadata about them for navigation and retrieval. The terminology registry can hold vocabulary level information only, or additionally comprise the member terms, concepts and relationships, and also provide or list services based on terminology such as the following:

- Searching;
- Disambiguation;
- Query expansion and reformulation;
- Browsing;
- Automated classification;
- Indexing and social tagging support;
- Mapping between vocabularies;
- Harvesting;
- Semantic reasoning;
- Text mining; and,
- Information extraction.

It could also include services supporting creation and maintenance of vocabularies, including suggestions from text mining and social tags. It should, if used as a digital infrastructure service, make their content available for both comfortable human

inspection and for m2m access.

Various registries exist within the JISC IE, including the IE Service Registry (IESR 2008). IESR is a machine readable registry of JISC collections of resources which contains information about these electronic resources, and details of how to access them and aims to make it easier for other applications such as portals and virtual learning environments to discover and use materials which will help their users' learning, teaching and research. It acts as a middleware and is primarily intended for m2m access. Collections are described using metadata based on RSLP and Dublin Core Collection Description schemas, and include elements such as title, description and controlled subject terms from different controlled vocabularies but with at least one DDC term to ensure interoperable searching. Services are described using a bespoke IESR scheme, and include a location address, technical method of accessing a collection or providing a service, and further description of technical access details. IESR metadata are supplied via several services: Z39.50, OAI-PMH, SRU/SRW, and OpenURL Resolution. There is also a Web search interface. Together with the US OCKHAM registry of the National Science Digital Library and Australian ORCA registry of repository collections, IESR has begun an initiative to enable sharing of collection descriptions and service details across registries (Apps 2008).

The IE Metadata Schema Registry (IEMSR) defines yet another type of a registry 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).

> MSRs are often seen as existing, invisibly, in the background, as a resource to be called on for any of a variety of purposes: for example, documentation of schemas, APs and elements; as an authoritative description of APs/schemas/elements at various stages in their respective lifecycles; design; development; cross-walk development; piecewise assemblage of novel APs from existing vocabularies and schemas; as a m2m lookup service for description of unknown elements. A metadata schema registry such as IEMSR can be seen as a 'compile-time' service - that is, one may see it as having a role only in the initial stages of building or applying a metadata schema or application profile. However, it can also be seen as having a more general role as a 'run-time' service. MSRs are often seen as having only an initial discovery, development or advisory role for the earliest stages in schema/AP design. A valid extension of that functionality might include various services supporting the use of schemas and APs throughout every stage of their lifecycles.
>
> *Emma Tonkin (IEMSR)*

The National Science Digital Library (NSDL) Registry (2008) provides access to both vocabularies and metadata schemas and defines itself as aiming to "identify, declare and publish" them (see below).

The ISO/IEC 11179 Metadata Registries family of standards (ISO 11179 2007) 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 XMDR project seeks to further build on this (see Section 6).

Other registry examples include the following: SchemaWeb (SchemaWeb 2005), a

directory of RDF schemas expressed in the RDFS, OWL and DAML+OIL schema languages; Dublin Core Metadata Initiative (DCMI) Registry (DCMI Registry 2008) which provides an up-to-date source of authoritative information about DCMI metadata terms and related vocabularies; METeOR (METeOR 2008), Australia's metadata registry for national data standards for the health, community services and housing assistance sectors, based on the ISO/IEC 11179 standard (see 5.3).

2.2 Types of vocabularies

There are various different types of vocabularies, each serving a different purpose. The major types include term lists, taxonomies, subject headings, thesauri, classification schemes, lexical databases and ontologies.

Descriptions and comparisons of the different types of vocabularies are often confusing because the terminology is not controlled and there is also a fair degree of overlap. We follow the analysis of vocabularies given in the JISC Terminology Services and Technologies Review (Tudhope, Koch, Heery 2006, p. 22-47), which also covers named entity authorities and folksonomies. In that report, vocabularies are considered by structural complexity and types of relationship and also discussed according to their main purposes or application areas, including retrieval, linguistic purpose, artificial intelligence, e-learning, and e-science. The NKOS network has discussed elements of a possible future classification of vocabularies according to several facets (summarised in Tudhope 2006).

2.3 Indicative vocabulary list

The following (partial) list of vocabularies potentially relevant to general JISC purposes was compiles from our survey and the vocabularies used by HILT, IESR, Intute and Jorum (for other vocabularies lists see Koch 2005, HILT Vocabulary resources 2008, Middleton 2008, University of British Columbia 2004. JISC Pedagogical Vocabularies Project Report 2005):

- ACM Computing Classification System http://www.acm.org/class/
- AOD (The Alcohol and Other Drug Thesaurus) http://etoh.niaaa.nih.gov/aodvol1/aodthome.htm
- APA (American Psychological Association) classification categories and codes http://www.apa.org/databases/training/classcodes.html
- Art & Architecture Thesaurus http://www.getty.edu/research/tools/vocabulary/aat/
- BIOSIS Controlled Vocabulary http://thomsonscientific.com/products/bsg/
- CAB Thesaurus http://www.cabi.org/cabthesaurus/
- CABICODES http://www.cabi.org/DatabaseSearchTools.asp
- CAS Registry Numbers http://www.cas.org/EO/regsys.html
- Dewey Decimal Classification http://www.oclc.org/dewey/
- EDINAUPDATE Wordlist http://edina.ac.uk/update/ (only available in the interface - see 'Wordlist Search')
- Pedagogic Terms taxonomy http://www.intute.ac.uk/publications/rdn-ltsn/pedagogic-terms/
- Educational Level classifications http://www.ukoln.ac.uk/metadata/education/ukel/
- Enzyme Commission Numbers http://www.chem.qmul.ac.uk/iubmb/enzyme/
- FAST http://www.oclc.org/research/projects/fast/

- Gemet Thesaurus http://www.eionet.europa.eu/gemet
- Geonames http://www.geonames.org/
- Getty Thesaurus of Geographic Names http://www.getty.edu/research/conducting_research/vocabularies/tgn/
- Global Change Master Directory (GCMD) (Science Keywords), NASA
- HASSET Humanities And Social Sciences Electronic Thesaurus (version 3.0) http://www.data-archive.ac.uk/search/hassetSearch.asp
- IBSS International Bibliography of the Social Sciences http://www.lse.ac.uk/collections/IBSS/
- INSPEC Thesaurus http://www.iee.org/publish/support/inspec/document/thes/
- INSPECCN Classification http://www.iee.org/publish/support/inspec/document/class/
- Integrated Public Sector Vocabulary (IPSV), e-Government Unit (UK)
- JACS Joint Academic Coding System of the Higher Education Statistics Agency http://www.hesa.ac.uk/jacs/completeclassification.htm
- JITA Classification Schema, E-Prints in Library and Information Science (E-LIS)
- Joint Academic Coding System (JACS), Universities and Colleges Admission Service (UK)
- LCC Library of Congress Classification http://www.loc.gov/catdir/cpso/lcco/lcco.html
- LCSH Library of Congress Subject Headings http://www.loc.gov/cds/lcsh.html
- Learning Directory Classification System http://www.advice-resources.co.uk/learningproviders/standards/ldcs_v3_nov.pdf/
- MDA-TMT Thesaurus of Monument Types http://www.fish-forum.info/i_tmt.htm
- MESH Medical Subject Headings http://www.nlm.nih.gov/mesh/meshhome.html
- MSC2000 Mathematics Subject Classification http://www.ams.org/msc/
- NLM (National Library of Medicine) Classification http://wwwcf.nlm.nih.gov/class/
- RCHME-APL RCHME Archaeological Periods List http://www.mda.org.uk/fish/i_apl.htm
- RCN thesaurus of nursing terms
- Thesaurus for Graphic Materials http://www.loc.gov/rr/print/tgm2/
- UDC Universal Decimal Classification http://www.udcc.org/
- UKAT UK Archival Thesaurus http://www.ukat.org.uk/
- UNESCO Thesaurus http://www.ulcc.ac.uk/unesco/
- Union List of Artist Names http://www.getty.edu/research/conducting_research/vocabularies/ulan/

UK cultural heritage community uses the following terminologies published online by:

1) MDA (SPECTRUM Terminology Bank 2008):

- MDA Archaeological Objects Thesaurus
- MDA Waterways Object Name Thesaurus

- MDA Railways Object Name Thesaurus
- MDA Codes
- British Museum Object Names Thesaurus
- British Museum Materials Thesaurus
- ICOM Costume Committee's Vocabulary of Basic Terms
- Royal Air Force Museum's Aircraft Types Thesaurus
- Pitt Rivers Museum, University of Oxford wordlists: Class; Continent; Country; Group; Keyword; Material; Process

2) National Monuments Record Thesauri (NMR), English Heritage

Various including:

- Monument Types
- Building Materials
- Defence of Britain
- Components
- Maritime Place Names
- Evidence Thesaurus
- Archaeological Sciences
- Historic Aircraft Type

2.4 Historical background

National, regional, local and domain organisations often created and maintained lists of vocabulary in use by their own organisation.

A larger and more recent list of this kind is the Thesaurus Guide (1993), published by the EU Commission, containing about 700 vocabularies available in at least one of the EU languages. It was also available as a database between 1993 and 1998. More than 2000 classification schemes, subject heading lists, and thesauri in the English language are physically collected at the University of Toronto Library and catalogued in its online catalogue (Subject Analysis Systems 2008). WorldCat (OCLC) (2008) also contains many catalogue records of vocabularies.

Since 1996 several lists of online available vocabularies in digital formats have been created, but most of them are not consistently enlarged or maintained (e.g., Koch 2007, HILT Vocabulary resources 2008, Middleton 2008, University of British Columbia 2004).

For some time (roughly 2003 -- 2004) the Dublin Core Metadata Initiative (DCMI) developed and tested a registry of "vocabulary encoding schemes", alongside its metadata registry, featuring a simple metadata schema to describe and label/name available vocabularies to be used in metadata records. The Usage Board developed a plan whereby people could fill in a Web form to apply for a DCMI-maintained URI identifying a controlled vocabulary for use as a Vocabulary Encoding Scheme. After developing a prototype service, they decided for various reasons -- mostly related to governance and long-term sustainability -- not to move into production, and stopped work on the prototype.

What DCMI has now is the DCMI registry, which stores terms defined by DCMI. These terms can be found by browsing "Summary of All Terms" on the DCMI registry at http://dcmi.kc.tsukuba.ac.jp/dcregistry/.

The NKOS network (2008) started an effort to design a terminology registry in 1998, emanating from discussions at the second NKOS workshop at the ACM Digital Library

Conference.

A small task force led by Linda Hill subsequently developed a very detailed metadata schema for the purpose, containing most of the information one would need to make an informed decision about the selection of an appropriate vocabulary. Version 2 was published on the NKOS website in 1998 (NKOS Registry 1998). Prior to the NKOS workshop 2001, Diane Vizine-Goetz from OCLC Research developed a more formal document as a draft, converting most of the descriptive data selected in the prior versions into a Reference document for data elements, based on Dublin Core elements described according to the ISO 11179 standard (NKOS Registry 2001). For a review of metadata elements in TRs, see Section 7.

Terminology registries were one of the main topics at the NKOS Special Session at DC 2005, bringing together the (DC) metadata and NKOS communities, featuring a main presentation by Rachel Heery (Heery 2005). The 2006 and 2007 European NKOS workshops (NKOS 2006 and NKOS 2007), again discussed the need for a registry. The NKOS workshop at ECDL 2008 discussed TR vocabulary metadata and the workshop at DC 2008 discussed the different types of registries (see NKOS website).

3 Methodology

3.1 Overall approach

The overall approach for the study involves identifying relevant information available from prior efforts and project documentation, supplemented by information obtained through consultation with key services, projects and executives across digital library, research and learning domains. Twenty-eight responses were collected.

See **Appendix 1** for the invitation letter and questionnaire, and **Appendix 2** for the interview letter. The interview was semi-structured and based around the questions asked in the emailed questionnaire, with a focus on any relevant specialist expertise. See **Appendix 3** for the list of people who provided input to the study, via a questionnaire (e-mail) or interview.

3.2 Scope of TRSS

The TRSS study focuses on requirements and potential delivery of a registry that would disclose information on vocabularies in common use within higher education in the UK, supporting use of terminologies in the context of “Discovery to Delivery”. Thus, while touching on all types of vocabularies, in this report we primarily focus on thesauri, classification schemes and subject headings, for the purposes of retrieval in the context of JISC Higher Education. Such vocabularies offer a cost effective approach to knowledge organisation, particularly for browsing, search and discovery purposes, and can be considered a middle ground between formal logic ontologies and folksonomies.

4 Architecture and functionality of terminology registries

There are two general architectural issues for TRs. The first is whether a TR is intended to support human access or m2m. If a TR supports m2m this is currently usually via some form of web service provision. Within the context of the JISC IE and eFramework, we assume that any JISC TR will support both human and m2m access. The second key architectural issue is whether the TR provides access solely to metadata on vocabularies or whether access is also provided to the vocabulary content (concepts, terms, relationships). Within the JISC IE, we assume that any access to content would be human and m2m. Third party terminology services might also be available, adding value to vocabulary content (see outline of terminology service functionality below). We thus distinguish three broad elements of TR architecture functionality:

1. Registry provides metadata for each vocabulary and links to vocabulary owner/provider (see Section 7 for review of metadata)
2. Registry provides metadata on (and links to) any available terminology services
3. Registry provides access to vocabulary content (either by downloading the complete vocabulary, or providing access to a vocabulary's concepts, terms and relationships)

These three elements can be taken as orthogonal facets; they can be combined independently (in practice, most combinations would include element 1).

The three elements correspond to options that JISC might choose from. One recommendation of this report is that JISC consider the possibility of funding a TR providing Option 1, as part of an extended IESR. In the medium term, a pilot Option 2 (for both human and m2m access) could be considered after a focused study on an initial set of appropriate metadata elements for terminology services.

JISC should also consider the possibility of establishing some form of TR support and advisory effort. As an alternative to proceeding immediately with Option 1, JISC could consider an interim step where the TR Support Project was assigned an additional set of tasks that attempted to gauge the level of interest and support for a general TR within the JISC community.

The various reasons are discussed in the following sections but the general rationale is in terms of cost benefit considerations. The recommendations are seen as a modest, incremental step, which could be taken further towards Option 3 if future circumstances warranted.

We now consider the functionality of TRs in general (across all three options). For technical architecture issues relating to registries, refer to the IESR general architecture (eg Apps 2007).

4.1 Scope of functionality of TRs

The major components of possible functionality include the following, arranged loosely according to a revised version of the information lifecycle framework for terminology services, in Tudhope, Koch, & Heery (2006, Section 4.2). This outline of functionality covers all three options and encompasses terminology services generally. Thus it should be considered a broad superset of possibilities, rather than any prescription.

Creation, modification and maintenance of vocabularies (Option 3)

Functionality to support the creation and editing/maintenance of vocabularies.

At the minimum, this includes an Import facility, which is capable of uploading a complete vocabulary in a variety of common 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 could be applied to terms, concepts, notes and possibly the relationships themselves. This might involve support for versioning. 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 community, oriented to a Web 2.0 perspective. See Section 9 on Governance of collaboration.

Maintenance can include support for versioning, which may be applied at different levels. Support could be provided to keep track of versions of complete vocabularies, or versioning could be applied at concept, or even term level. Again, this should be controlled by an appropriate governance model, which should be clearly specified.

Acquisition and publication of vocabularies (Options 1, 3)

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 most cases, this would require quality control and a review and selection process, according to the governance model of the TR. While the suggestion of vocabularies might conceivably be based on machine harvesting of vocabularies, in the foreseeable future, this will require human resources.

Publication is taken here to include support for licensing, where the vocabulary provider is a commercial entity or asserts some rights over the vocabulary. This might involve support for commercial charging for certain functionalities provided. Option 3 would normally provide support for export of a vocabulary in a variety of standard formats.

It is also appropriate to mention support for training and help in using the TR, and vocabularies generally. This could take the form of some hypertext help and tutorials but to be effective in a large TR it is likely to require human support.

For option 3, provision must be made to store the vocabularies. This may be in an internal TR representation format but it is assumed that import and export functions would be provided in standard formats (see Section 8).

Cataloguing (Options 2, 3)

Functionality to support indexing/classification/annotation activities. This could be via

- a cataloguing application
- direct provision of (web) services by the TR itself
- content open to third-party web services via a programmatic interface

For Option 2, where the content is not held in the TR, enough information must be given for a third party service to locate and access the vocabulary content, for example from the provider.

Integration (Options 2, 3)

This includes semantic interoperability support for mapping and possibly merging of vocabularies. This could be via:

- direct provision of mapping (web) services by the TR, or some form of crosswalk
- content open to third party web services via a programmatic interface

Access, search and discovery (Options 1, 2, 3)

This can be applied at the complete vocabulary level (Option 1). Support may be provided to search or browse at the vocabulary level (via metadata), 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.

Support may be provided to identify vocabularies that can be accessed via particular services, or that are used to index particular collections.

It can also be applied to discovery of individual concepts or terms (Option 3). At the

content level, support should be provided to match a user string with terms (or 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. Disambiguation is usually performed manually but there is scope for automatic disambiguation or selection of defaults. This could be used in a variety of applications, such as mapping, search, information extraction, automatic classification, etc.

Once a concept (or term) has been selected it may be used in search and retrieval applications to support querying. The query engine may operate over controlled vocabulary indexed collections, or a 'search thesaurus' may be used to assist query expansion for free-text search. A vocabulary may be used to support query expansion, either with synonyms or with semantically close concepts. The query engine may or may not be aware that a query term is from a controlled vocabulary and may or may not take advantage of the concept structure and hierarchy in the matching function. Search may be across single or multiple collections and may involve single or multiple vocabularies.

Option 3 also includes support for browsing and visualization of vocabulary content. Various user interface options are possible, including faceted browsing, polyhierarchy support, personalisation, etc.

Use (Option 2, 3)

This covers situations where, via the TR, an appropriate terminology service provides support for a wider application, such as information extraction, text mining, automatic classification, Semantic Web, e-Learning or e-Science applications. The application of vocabularies in social tagging and Web 2.0 applications is a growing area of interest.

Archiving and preservation of vocabularies (Option 3)

Currently this would form part of digital preservation generally. Long-term preservation of vocabularies has not been considered as part of the project scope but is an important issue. Standard XML-based representation formats for vocabularies (such as SKOS) would be a first step.

4.2 Rationale for a TR of vocabularies in common use within UK HE, in the context of the JISC IE

The main rationale for the immediate recommendation of this report (Option 1) is in providing a service to assist discovery of existing vocabularies, or the most recent version of a given vocabulary.

Several TRSS respondents and many use cases from the literature describe variants of a scenario, involving a user from a particular subject domain looking to see if a vocabulary with certain properties already exists. This may be for purposes of supporting access to a new repository or collection (via search and browse services). It may be to assist the design of a new vocabulary by first looking to see if anything similar already exists that can be used as is, or serve to inspire the construction of a specialist vocabulary.

Two simple use cases from the next section illustrate the general principle. (1) *An archivist on a tight budget wants to introduce a subject classification scheme and is unsure what schemes are available. She uses the Terminology Registry to discover schemes that are freely available.* (2) *A brain institute librarian is looking for useful vocabularies in the field of brain research. She uses the Terminology Registry to discover schemes that are freely available.*

The features of a vocabulary that afford discovery vary (widely) according to the user's search criteria. The user may have a rough idea of a particular vocabularies 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. The depth or breadth of topic coverage may be an issue. To assist discovery a rich set of metadata

should be available for the vocabulary. Section 7 reviews metadata for TRs.

This metadata should be open to both human and m2m access. 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 uer search is also desirable.

Various JISC shared services, projects and reports support some form of TR, including the JISC Pedagogical Vocabularies Project Report (2005), the Terminology Services and Technology Review (2006), HILT, IEMSR and IESR.

In general, a TR would contribute greater support for subject based access to collections and repositories in the JISC IE. Option 1 alone should provide cost savings in the resources required to discover relevant vocabularies and also in the construction of new vocabularies.

5 Use cases for vocabularies and terminology services

5.1 Overview

Vocabularies and services based on them enable users to undertake educational and research inquiries more effectively. When searching free text with uncontrolled terms, significant differences can stem from trivial variations in search statements and from differing conceptualisations of an information need. Different people use different words for the same concept or employ slightly different concepts. It can be difficult for non-specialists to employ technical vocabulary and variation in person or place names can frustrate consistent access. This may not be a problem if the purpose is just to obtain a few relevant items as examples of a topic. However, when the purpose is, for example, an in-depth educational review or systematic research on a specialized topic then it is undesirable to miss potentially relevant items. These problems can be helped by vocabularies and terminology services; applying them has been a common practice in libraries and indexing and abstracting databases, some for more than a century now.

At the simplest level, a controlled list of terms ensures consistency in searching and indexing, helping to reduce problems arising from synonym and homograph mismatches.

The following is an example from British Oceanographic Data Centre vocabulary server (BODC 2008), which provides the context of experimental data sets:

> "An example of how computers may benefit from the use of controlled vocabularies is in the summing of values taken from different data sets: one data set may have a column labelled 'Temperature of the water column' and another might have 'water temperature' or even 'temperature'. To the human eye, the similarity is obvious but a computer would not be able to interpret these as the same thing unless all the possible options were hard coded into its software. If data are marked up with the same terms, this problem is resolved.
>
> In the real world, it is not always possible or agreeable for data providers to use the same terms. In such cases, controlled vocabularies can be used as a medium to which data centres can map their equivalent terms."

Similarly, for digital curation of experimental data sets it is important to provide semantic alignment of terms used, e.g., for instrument calibrations, data units etc.

At a more complex level, the presentation of concepts in hierarchies and other semantic structures helps the indexer and searcher choose the most appropriate concept for their purposes. Browsing-based user interfaces become possible. A vocabulary can assist both precision (by allowing specific searching) and recall (by retrieving items described by related concepts or equivalent terms). It also provides potential pathways (for human and machine) that connect a searcher and indexer's choice of terminology. Also, a hierarchical thesaurus, for example, might represent the countries of the United Kingdom in a hierarchy, such that anything catalogued as being in or related to England would also automatically be in or related to the United Kingdom (Miller 2000).

Failure to adopt vocabularies and related services make it potentially impossible to effectively integrate different resources from within a single institution, or across multiple institutions and repositories or whenever there is a need for cross-disciplinary data sharing, re-use and integration such as e-research, e-government. Global warming is one example which requires the integration of data from numerous heterogeneous independent datasets.

Finally, no single thesaurus or other terminological tool will ever meet all the needs of all users. It is therefore important to hold reliable and consistent information about a range of vocabularies in a registry (Miller 2000). A terminology registry could provide information about existing terminology services, accessible to both humans and

machines. Machines could need to look up ways to access a certain terminology service and how compliant it is with the specified needs.

The HILT M2M Feasibility Study (2005) described five use cases and possible technical solutions, including: resolving a user search term to controlled terminology, spelling correction, browsing, disambiguation, query expansion, mapping to Dewey Decimal Classification (and associated mappings), and querying. These are illustrated in Nicholson *et al.* (2006) which also shows how the HILT Phase 2 Pilot uses a Dewey Decimal Classification spine for subject interoperability over JISC collections. See also the pilot server examples at http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/.

The HILT general model involves a distributed view of a 'subject interoperability service', which involves some form of TR with metadata on multiple vocabularies (including crosswalks). It also envisages a set of registries (for different domains) that provide information on terminology services available for given vocabularies and which distinguishes between different types of terminology service. Client information services may have a preferred (local) service registry but may also need to discover (via appropriate metadata) other available terminology services.

According to Hodge et al. (2007), the purpose of a terminology registry would be the following:

- make traditional resources more visible,
- provide key characteristics of resources,
- encourage human assessment of these resources for applicability to semantic projects,
- provide characteristics needed to make these resources more computable,
- integrate resources with other data and resource descriptions, and
- promote information exchange and knowledge sharing.

In a specific domain, on the example of environmental science, she claims the following purposes: make environmental terminologies more interoperable and generally useful; standards regarding terminology will provide the integration of meaning and definitions across heterogeneous data and information systems allowing users of the data to understand the similarities and differences among terms and data; and, element set of resources used to identify environmental terminology resources to be integrated with semantic technologies.

A terminology registry can reduce 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; implementing a vocabulary that in the end proved not good enough involves huge costs. A terminology registry would list and describe all the different vocabulary in a domain and provide contact with existing users.

5.2 Use cases and their general requirements

The use cases (use scenarios) reported below were collected from related projects and literature as well as experts from related fields who responded to the survey (see Appendix 3). These present a wide set of possible requirements for terminology services generally. Use cases corresponding to all three options are reviewed, including both basic and more speculative, future-oriented scenarios.

In our study the most frequently collected use case for a terminology registry is discovering and examining vocabularies for a certain subject domain at the time of planning a repository or any digital collection. Other cases are given below.

5.2.1 Creation, modification and maintenance of vocabularies (Option 3)

In order to tackle reported issues such as vocabularies that are hard to keep up-to-date with new terminology unstructured content, maintenance difficulties, and duplication of effort (Lee 2004), providing a vocabulary development environment would be important. In addition, there are a lot of small vocabularies out there, that may even be local vocabularies, but they do not have the resources to build the system, to maintain the vocabulary, they may just be, for example, currently using Word or Excel to keep track of vocabulary, and having a service to maintain that information, would be useful.

Use cases

a) Managing local terminologies

An institution uses a set of local genre terms for assignment to resources. The genre terms are managed in a shared Excel spreadsheet. The institution elects to migrate from the Excel spreadsheet to the use of a terminologies service to manage genre terms and consolidate future terminologies. (Proffitt et al. 2007)

b) Establishing a project-specific subset of terms

In a library, a professional selects a list of descriptive terms from published and local terminologies for use by paraprofessionals and interns to describe resources for a specific project. (Proffitt et al. 2007)

c) Joint editing and annotation of local terminologies by experts

A digital library has established a list of local place names that a museum would like to use for a project cataloguing local artefacts. In the course of the project, the museum discovers that it needs local place names not represented on the existing list. In addition, some of the local place name metadata is missing or not sufficiently detailed. The library and museum agree to collaboratively update and annotate the list of local place names. (Proffitt et al. 2007)

d) Contributing to a published terminology

An institution has created a local set of terms that are extensions to a published terminology. The institution submits the terms for review and authorization so the terms can be incorporated into the published terminology. (Proffitt et al. 2007)

e) Capturing locally contributed end-user terminology

A researcher uses a finding aid at an archive to locate a collection of materials. While reading through the documents in the collection the researcher realizes that the finding aid does not list persons and places associated with the documents. The researcher updates and annotates the finding aid to include the missing information. (Proffitt et al. 2007)

f) Sharing local terminologies

A historical society creates a list of local place names not found in published sources and shares them with other institutions. (Proffitt et al. 2007)

g) To give provenance to the vocabulary one is using, something like “my repository is using this term from this taxonomy and this version of it”.

Requirements

Provide a vocabulary development environment with services that support vocabularies management and sharing. The functionalities should include the following:

- vocabulary registration and upload
- submission of metadata for submitted terminology
- validation of submitted terminology
- validation of metadata for submitted terminology

- provide URIs for each vocabulary
- editing
- revision and extension
- tracking and versioning
- submission of new versions
- collaborative support services (e.g., discussion board, wiki)
- tracking of who do users are (useful if you're looking for funding for a vocabulary or trying to build a community around maintenance)
- allow vocabulary users' registration and signing up for regular, configurable notification of changes in the vocabularies they use (e.g., in the form of files that can be used directly in update routines, human readable change listings)
- provide best practices for vocabulary development and management, for example, a road map of how to reuse existing vocabularies and their member terms/classes in future constructing of a new vocabulary

5.2.2 Acquisition and publication of vocabularies (Options 1, 3)

Use cases

a) Services for topical crawlers

A topical crawler uses controlled vocabularies as the basic mechanism. It uses a terminology services to get controlled terms from a specific domain.

b) Interworking between databases

Data warehouses – storing definitions of data elements and data types for the purpose of interworking between databases (Heery 2005).

Requirements

Provide vocabulary environment which provides services that support vocabularies publication and usage. The functionalities include the following:

- purchase licenses for making full vocabularies available to registered users
- allow viewing of vocabularies to registered users
- ensure easy user registration process (e.g., via IP without password)
- allow viewing of vocabularies to registered users before buying
- provide export/download of whole vocabularies
- provide export/download of parts of vocabularies
- provide export/download in a variety of standard formats
- provide web services for accessing individual terms and concepts of use to, for example, topical crawlers

5.2.3 Cataloguing (Options 2,3)

Use cases

a) Metadata validation

A repository manager has several items to deposit into the institutional repository, she knows that items are all medically related items and some partial metadata has been added to the individual items. She notices some inconsistencies with some of the

items, e.g. spelling mistakes, use of synonyms, she is concerned that appropriate keywords have not been applied to the items and submits the metadata to the terminology services for checking. The metadata is returned for the 10 items with items labelled consistently and correctly.

“we're making sure that where you have a particular vocabulary, covered a particular field in a metadata document, that that document, that that field, has legal terms in it, you know, that are spelled correctly, that they are in the list” (R. Lowry interview about BODC)

b) Browsing, searching and retrieving terms

A cataloguer is looking for a genre term associated with a resource. He types a term into the search box of a cataloguing tool and retrieves a list of terms from one or more terminologies. He selects the appropriate term and moves on to another task. (Proffitt et al. 2007)

c) Automated controlled terms suggestion

SWORD – for depositing at different repositories at the same time. At the time an author deposits her paper to different repositories via SWORD, she uploads the paper to a terminology service which returns automatically generated controlled term suggestions for each repository. She selects the terms she approves, modifies them or adds new ones, and moves on to complete the deposit.

Benjamin visits his institutional repository in order to deposit a set of images. The institutional repository presents him with a form with the appropriate application profile elements, as retrieved via an m2m query to the IEMSR. He has not previously used this particular form, and the AJAX layer connecting the IEMSR in real-time to the form recognises his hesitation and subtly presents some suggestions as to the sort of information that others generally use within each element, including formatting data. When a field is recognised as 'typically' containing elements from a given controlled vocabulary, suggesting that within the particular context of Benjamin's institutional repository, this is a convention that is widely used, the system offers a widget that allows him to select terms from a set of recommendations - allowing him of course the ability to override it, unless the repository designer or the AP definition states that the use of this controlled vocabulary is mandatory within that context. i.e. the system is able to infer - or apply - links between the metadata element and relevant controlled vocabularies, just as it is able to infer - or apply - links between the metadata element and acceptable formats or expression grammars.

Requirements

Provide services that support metadata creation. The functionalities include the following:

- provide services for browsing, searching and retrieving controlled terms to professional cataloguers
- provide services for browsing, searching and retrieving controlled terms to social taggers
- provide services for validating names and controlled terms in metadata, including a spell-check service
- provide services that support metadata creation
- provide links to related standards
- provide automatically generated controlled terms

5.2.4 Integration (Options 2, 3)

Use cases

a) Searching different collections through 'one' vocabulary

A user searches for an interdisciplinary topic inside AgriFor using CAB subject headings and does not get any hits. The system then suggests her to try get hits from search the Natural Selection collection. If she accepts the suggestion, hits from that collection are retrieved. In the background a terminology service was called that automatically mapped the CAB terms into DDC headings used by the Natural Selection. (adapted from Tudhope, Koch, Heery 2006)

You wish to search your institutional ePrints repository for articles on a particular subject. Since the coverage is wide, a general vocabulary is available for browsing access, in this case the top 2-3 levels of the Library of Congress Subject Areas, with associated postings. However, it is not clear from the main menu where your subject interest would fall – the terms you usually employ to describe your subject are not mentioned and you don't feel like browsing multiple sub-menus in the quite extensive browsing classification. In the browser, you try to Find on this page without success. There is no way of searching the vocabulary to find where your interest might fall. You can, of course, search the full text but this relies on a subject keyword appearing in the text. A TS that augmented the general classification with an entry vocabulary of synonyms and allowed search of this extended vocabulary would extend the utility of the retrieval functionality. This would provide additional entry points for browsing. This scenario assumes that subject search of a University publication repository is a sensible option. Given the probable patchy distribution of coverage in any one University, some form of known item search or author-based search may be more likely. However, subject-based access would be applicable to various types of aggregated repositories in the future. (Tudhope, Koch, Heery 2006)

b) Combining local, shared or published terminologies

A museum wants to use a set of local personal names in conjunction with names drawn from a published terminology, such as ULAN. The institution creates a combined terminology composed of the locally generated terms and the published terminology. (Proffitt et al. 2007)

c) Retrieving library holdings for courses

A portal wants to provide a list of books in the library catalogue for each course. It uses a terminology service with mapped course codes to JACS, with JACS mapped to LCSH and DDC, and retrieves the list of books.

Requirements

Provide services that support vocabulary mapping and merging. These include the following:

- provide services for (semi)-automated merging and mappings between vocabularies (e.g., data mining techniques, co-occurrence-based techniques)
- provide services with intellectual merging and mappings between vocabularies
- provide tools for producing merging and mappings between vocabularies
- provide above for merging and mappings between end-user vocabularies and published vocabularies

5.2.5 Access, search and discovery (Options 1,2,3)

Use cases

a) Discovering suitable vocabularies for a collection

An archivist on a tight budget wants to introduce a subject classification scheme and is unsure what schemes are available. She uses the Terminology Registry to discover schemes that are freely available.

A brain institute librarian is looking for useful vocabularies in the field of brain research.

She uses the Terminology Registry to discover schemes that are freely available.

An ontology developer wants to view what has changed between two versions of an ontology. (Mungal 2008)

Requirements

- allow searching and browsing of vocabularies metadata

5.2.6 Use (Option 2, 3)

Various terminology services could be integrated as an option in the search process, as sources for query terms. Google Toolbar already offers a dictionary service and similar forms of terminology services can be envisaged. A range of terminology services to improve query performance (both recall and precision) are possible. This includes various query expansion possibilities, where result ranking can be based on degree of semantic match. For example, you may wish to search with very specific terminology; you would be very interested in matches on those concepts and, failing that, would also be interested in matches on closely related concepts. Employing query expansion can combine several search 'moves' in the one query.

Use cases

a) Discovery of terminology services

b) Leveraging terminology for search optimization

A user formulates a query for the place name "Augsburg". The query is expanded to include equivalent terms, e.g., "Augusta Vindelicorum" (the original name of the Roman settlement). (Proffitt et al. 2007)

A user formulates a query for the place name "Bavaria" (state). The query is expanded to include terms in the same hierarchy, e.g., "Franken" (a district within Bavaria). (Proffitt et al. 2007)

A User is concerned about a specific type of cancer. She wants to discover any documents on the web (reliable and unreliable sources) about the disease, causes, treatment, victims, and researchers, and wants to find information that is related through generalization and specialization and other relationships. These are provided through terminology services. (XMDR Working Group 2005)

Paul is looking for resources about tuberculosis transmission. He uses his institutional portal which provides a search box for Zetoc (http://zetoc.mimas.ac.uk). He enters 'tuberculosis' into a Zetoc 'title' search and retrieves the expected Zetoc list of brief search results. The portal displays to Paul, in another part of its display, a list of other 'more-like-this' resources that may be of interest to him. It does this by using IESR, via its Z39.50 interface, to discover medical resources that potentially include tuberculosis and provide a Z39.50 service. In order to discover resources that may possibly cover tuberculosis the portal will need to use a thesaurus or terminology service (e.g. HILT) to find a suitable broader term for the search. The portal may also wish to translate this search term into Dewey for better searching of IESR. IESR will return a set of XML descriptions as a result of this search, which the portal will parse to elicit the significant details including the Z39.50 connection details for each resource. The portal then provides to Paul the 'more-like-this' list that is the result of a Z39.50 cross-search over these collections (excluding Zetoc itself), using 'tuberculosis' in the 'title' as a search term. Because the portal is using IESR dynamically, Paul will potentially find suitable, possibly new resources of which either he or the portal developer were unaware. (from JISC Circular, Call for projects to embed and to develop shared infrastructure services for the Information Environment, March 2008)

Your teacher has given an assignment to find information from the Intute on how vog is relevant to tomorrow's classes. Unfortunately your attention wandered momentarily at the point when this new word was explained. You do not know if it is something to do

with the morning class on Japanese culture and street style or the afternoon's class on volcanos and global warming. You do a search with Intute on vog and find no hits. Using a TS that searches a general subject vocabulary, you look up vog and find it is related to volcanic gases. You search Intute with these terms and find relevant resources. (Tudhope, Koch, Heery 2006)

Requirements

- allow searching and browsing of terminology services metadata
- provide service access details for terminology services (how to connect)
- for each KOS, provide KOS identifier and version identifier and cross-walk data
- allow searching and browsing of vocabularies for both cataloguers/indexers and end-users
- allow services that identify synonyms to automatically expand search for users
- allow services that provide hierarchies and other semantic relationships to automatically expand search for users
- allow services that offer the ability to search across multi-lingual repositories
- allow services to support disambiguation and cross-disciplinary searching
- allow services to support cross-disciplinary searching (via mappings)

5.2.7 Archiving and preservation of vocabularies (Option 3)

If Option 3, then appropriate preservation and archiving of contained vocabularies should be in place.

6 Review of terminology registries

We review some prominent existing examples of TRs in Section 6.2. Section 6.1 first considers some broadly related work, somewhat less central to the topic, and also projects in progress.

Perhaps the US National Library of Medicine Unified Medical Language System® (UMLS http://www.nlm.nih.gov/research/umls/umlsmain.html) is the oldest example of something like a TR. UMLS is a metathesaurus, bridging over 50 biomedical vocabularies including different language versions of the medical thesaurus, MeSH. Rather than a registry of vocabularies per se, UMLS can be considered more an integrated system (and set of tools) that offers access to health related concepts and their relationships, while maintaining information on a given concept's source vocabulary.

6.1 Broadly related work

Within the UK museum and heritage sector, the Collections Trust (http://collectionstrust.orangeleaf.org/ formerly MDA) have plans to build a Cultural Terminology Server, as part of the effort associated with the European Digital Library. (The MLA has phased out its Digital Futures division and given responsibility to the Collections Trust.) This effort will build on the existing Spectrum Terminology Bank (SPECTRUM Terminology Bank 2008), which lists online (html) vocabularies relevant to the sector. The aim is to assist the creation of resource metadata according to the UK museums standard SPECTRUM and the historic environment standard MIDAS, as developed by the MDA and the Forum Information Standards in Heritage (FISH) (Lee 2004). Work on the Cultural terminology Server is still at an early stage but the plans include an emphasis on collaborative development of structured vocabularies, in a variety of formats including SKOS. Vocabularies will be freely available, involving procurement of and provision of access to commercially published terminologies where necessary.

Within various domains in eScience, the value of linking data from experiments and studies to controlled vocabularies is becoming recognised. One prominent example is the Gene Ontology and its associated databases (see the BioPortal below). A registry facilitates reuse and awareness of work conducted by other research teams via search and discovery. Since such work may involve modelling a domain in terms of well defined objects (often data items) of scientific discourse and their properties, with possibilities for logical inferencing, it has been one of the areas where formal ontologies have been applied.

There are various formal ontology registries, typically holding their content in OWL or OBO format. Examples of related current projects include

- PRIDE (http://www.ebi.ac.uk/pride/) and see the related ontology lookup service http://www.ebi.ac.uk/ontology-lookup
- SeCO (http://www.seco.tkk.fi/)
- STITCH (http://www.cs.vu.nl/STITCH/)
- CO-ODE (http://www.co-ode.org/ontologies/) and also see http://owl.cs.manchester.ac.uk/repository/

Another example, the National Centre for Text Mining (NaCTeM), has a list (currently 5) of bio-medical ontologies (http://www.nactem.ac.uk/resources.php?view=5). The GRIMOIRES project was concerned with steps towards discovery and automated integration of services (http://www.grimoires.org). The Open Ontology Repository (OOR) Initiative (2008) aims to promote the global (re)use and sharing of ontologies (focusing mainly on formal ontologies): an "ontology repository is a facility where ontologies and related information artefacts can be stored, retrieved and managed".

Some online discussion seminars have been held but the initiative is still in its early stages. A recent communique reports on the initiative's annual summit (http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2008_Communique). The Ontology Metadata Vocabulary project has proposed some metadata elements for formal ontologies (http://omv.ontoware.org/). Ahmad and Colomb (2007) provide a review of existing ontology servers.

While formal ontology registries do not fall within the main scope of the TRSS, it is nonetheless important to maintain awareness and future compatibility in JISC oriented TRs. See recommendations on SKOS format, linked data and URI identifiers particularly, regarding compatibility with ontology registry work.

The Natural History Museum's Species Dictionary Project (http://nbn.nhm.ac.uk/nhm/), in collaboration with the National Biodiversity Network, is developing an exhaustive, standard reference for names of UK organisms from a wide range of datasets. It is possible to search by common or scientific names. The Integrated Taxonomic Information System (ITIS http://www.itis.gov/) aims to provide taxonomic information on world wide flora and fauna. Tools are available to check for duplicates or construct a hierarchy. As part of ITIS governance, stewards and specialist collaborators review and assign a Taxonomic Serial Number if appropriate.

Other international scientific taxonomic online database projects include Species 2000 (http://www.sp2000.org/) and the Catalog of Life (http://annual.sp2000.org/search.php), which is a collaborative project between Species 2000, ITIS and the Global Biodiversity Information Facility (GBIF). The Species 2000 & Catalog of Life have implemented a validated index to known species in order to monitor biodiversity worldwide. The Catalog of Life provides online browse and search access to (currently) over 1,008,965 species and links to the original data supplier's database are provided where possible. The Catalog of Life provides the taxonomic underpinning for the recent Encyclopaedia of Life (EOL http://www.eol.org/) project. EOL is an ambitious project to collaboratively gather (from scientists and public) information on life on earth, worldwide in a large biodiversity database. See Tudhope, Heery, Koch (2006) for more information on taxonomic search engines.

The wide-ranging eXtended MetaData Registry (XMDR) project (http://www.xmdr.org/) is still in development. US Government agencies (DoD, EPA, USGS, National Cancer Institute, Lawrence Berkeley National Lab etc.), including some European partners such as EEA, are engaged in a five year project (started 2004/5). It aims to build upon and contribute to the further development of the ISO 11179 Metadata Registries family of standards. It is developing a prototype extended metadata registry, incorporating various terminologies and ontologies. The XMDR prototype is based on a REST API allowing SPARQL search, text search, and XMDR search. This effort has close links to the language engineering community and most related ISO subcommittees (SC 32, TC 37/SC 4). The focus seems to be on a (more ambitious) registry of individual terms than on vocabulary schemes and collections (Bargmeyer 2005).

The Australian DART Project at University of Queensland has implemented a prototype metadata schema registry and are seeking further funding to develop an associated terminology registry. The current purpose is cross-disciplinary data sharing, re-use and integration in the context of e-Research and e-Government. For example, the problem of global warming requires the integration of data from numerous heterogeneous datasets. Terminologies are seen as clearly relevant for validating metadata values associated with particular fields within a metadata schema. Planned functionality includes the ability for authorized users to upload terminologies (XML schemas, RDF schemas, ontologies, thesauri, controlled vocabularies) and attach metadata, to validate terminologies on upload, the ability to search, browse and retrieve terminologies based on terms and metadata, the ability to edit and upload new versions of terminologies.

http://www.itee.uq.edu.au/~eresearch/projects/dart/outcomes/metadataschemareg.php

English Heritage have recently developed a prototype EHKOS registry (http://www.ehkos.org) for their National Monuments Record vocabularies. It is an in-house development at national Monuments Record, still at a pilot stage, and further plans include web services and richer metadata. Currently, import format is CSV and access to complete vocabularies is via browsing an alphabetical list. Metadata is confined to Name and description of vocabularies. Interactive browsing and search is provided to vocabulary content, with the ability to search on definition, label and scope note. The intention is to support collaborative development (users wishing to submit or edit a vocabulary must register and be vetted by the registry team). There is a vocabulary mapping tool for the Admin role. As Admin it is also possible to create new relationships.

The AHRC funded STAR project operates an internal TR for project purposes, comprising seven English Heritage thesauri, represented in SKOS format (http://hypermedia.research.glam.ac.uk/kos/terminology_services/). Access is via a set of SOAP terminology web services, which currently provide term look up across the thesauri held in the system, along with browsing and semantic concept (and synonym) expansion within a chosen thesaurus. The service, based on a subset of the SWAD Europe SKOS API with extensions for concept expansion, currently consists of 7 function calls, which can be integrated into a textual or metadata based search system. A client demonstrator is available for download. Services based on URL calls have recently been developed, as part of a related Tagging Suggestion project (PERTAINS).

The HILT project (HILT 2008) operates an internal TR for project purposes (HILT Vocabulary resources 2008) and has implemented a number of terminology services, including term lookup, browsing and searching. It provides mapping services between different controlled vocabularies used in different collections, using Dewey Decimal Classification (DDC) as the central spine. Apart from aiming at allowing interoperable subject cross-searches, HILT also provides SOAP and SRW-based m2m information about terminology sets which is used to enhance the precision of subject searches. Terminology sets are structured using SKOS. For more information on HILT, see Section 2.1. Pilot web service examples can be seen at

http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/
http://hiltm2m.cdlr.strath.ac.uk/hilt4/hiltsoapclient.php

'Embedding project' demos of the services within three JISC services have recently been made available - http://hilt.cdlr.strath.ac.uk/.

6.2 Review of existing Terminology Registries

We now consider some prominent existing TRs, locating them as far as feasible within the framework of the major functionality options. Section 7 reviews the major efforts at vocabulary level metadata.

Lexaurus Bank (originated as the BECTA Vocabulary Bank)

Vocabulary Management Group http://www.vocman.com/

offers Lexaurus Bank and Lexaurus Editor

developed by Knowledge Integration http://www.k-int.com/

Options 1, 2, 3, interactive and m2m access

Becta Vocabularies Studio,

The product originated in Becta Vocabularies Studio, which supports creating, editing and maintenance of vocabularies via the Becta Vocabulary Bank. The TR is aimed at supporting (school National Curriculum) educational vocabularies for the various key-stage levels and authorities. There is a tagging tool and the service also allows vocabulary providers to assert mappings between elements of the vocabularies. This

employs a single (equivalence) relationship to a central spine.

The metadata is limited to authority, version, description. It is possible to interactively browse by vocabulary names sorted alphabetically or by the authority that owns them, search for words that appear in the vocabulary names or search for their identifier. A term (and ID) search facility is also supported, along with browsing an individual vocabulary. Vocabularies are represented in the Zthes XML DTD and an XML download of a complete vocabulary is supported. An SRU web services interface supports m2m access via the Zthes profile, with some additional indexes.

The VMS Studio service has been used operationally by around 10 authorities with 4 of these publishing their vocabularies to the Bank. We do not have information on the extent to which vocabularies are currently used for searching and indexing resources by e-Learning end-users. While the Bank is operational, we understand that BECTA does not have current plans to fund further development, although the situation is somewhat in flux.

Lexaurus Bank and Lexaurus Editor (recent development)

Part of the original Becta code was proprietary to SchemaLogic's SchemaServer engine. Recently, Knowledge Integration have developed a self contained version, available from VocMan, a joint venture between the Knowledge Integration and Schemeta companies. This comprises two cross-platform products, the vocabulary server, Lexaurus Bank and the Lexaurus Editor standalone application, with support for import, creation and editing of vocabularies. Lexaurus Editor now supports the SKOS format, in addition to ZThes and VDEX. Lexaurus Bank and Editor were both employed by National Strategies to underpin its gateway of teaching resources for primary and secondary schools. Lexaurus Bank is currently being used in the multilingual European Schoolnet initiative. Web service programmatic access is available.

BioPortal and OBO Foundry

http://www.bioontology.org/tools/portal/bioportal.html

http://www.bioontology.org/ncbo/faces/pages/ontology_list.xhtml

http://www.obofoundry.org/ http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html

Options 1, 2, 3, interactive and m2m access

The US OBO is an umbrella organisation for life-science ontologies. It can be accessed via the NIH Roadmap National Centre for Biomedical Ontology's BioPortal. OBO contains over 60 ontologies, mostly in OWL-DL and OBO formats. Metadata, including title, domain, description, relevant organism, source, status and various statistics, is available as XML or RDF.

BioPortal offers search and browsing access to its ontologies. It also offers access to experimental data via ontologies and annotations. Future plans include a set of developer tools, URI for content and SOAP Web services (some currently available). An alpha version of BioPortal 2 offers REST Web services, mapping and annotation tools.

The associated OBO Foundry is an influential ontology registry in the biomedical domain, which has a collaborative governance model and set of principles for collaborative development of ontologies.

Cendi Terminology Locator

http://www.cendi.gov/projects/termlocator.html

Option 1, interactive access

The US CENDI project is still in development. It aims to point terminology system developers, librarians, researchers, and others who are interested in scientific terms to

the terminology resources of the CENDI federal science research agencies, spanning agriculture to medicine to the environment. Current vocabularies include 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.

FAO KOS Registry

http://aims.fao.org/aims/en/home (multilingual)

Options 1, 2, 3, interactive access (and m2m access to Agrovoc)

The United Nation's Food and Agriculture Organization (FAO) has implemented a TR for its own vocabularies, including the influential Agrovoc. It holds over 90 KOS, in areas related to agriculture and administration. Vocabulary metadata is provided and so is either content when held internally or links to the vocabulary provider. Vocabularies are represented (and FAO owned ones available) in a variety of formats (including SKOS, OWL and native formats). They are organized by type and by subject domain. The different types cover authority files, classification schemes, dictionaries, glossaries, ontologies, taxonomies, terminologies, thesauri, and topic trees. Interactive access allows browsing by type (authority files, classification schemes, dictionaries, glossaries, ontologies, taxonomies, terminologies, thesauri, topic trees) and subject domain (11 subject areas). The project continues and the registry is operational.

The Agrovoc concept server allows wider functionality for accessing multilingual Agrovoc content. This includes interactive browse and search over terms. The current version is available in SKOS, MySQL, Postgres, MS Access, TagText and ISO2709 formats. Agrovoc is also available m2m via SOAP web services (http://www.fao.org/aims/ag_webservices.jsp). This includes multilingual term search and a term expansion call, which returns synonyms. Documentation and simple test clients are available for download. An Agrovoc concept server workbench is currently under development. There are also plans to create a registry of mappings.

NERC (BODC) Data Grid's Vocabulary Server

http://vocab.ndg.nerc.ac.uk/

http://www.bodc.ac.uk/products/web_services/vocab/

BODC Web Services http://www.bodc.ac.uk/products/web_services/

Options 1, 2, 3, m2m access

The British Oceanographic Data Centre (BODC) has implemented a TR, which underpins the NERC DataGrid vocabulary server. This operational system supports the management and interoperability of scientific datasets in collaborating international data centres (43 in Europe), including UK centres such as the British Antarctic Survey, the National Oceanographic Centre Southampton and the Plymouth Marine Laboratory. This is a fully operational system, with more than 100 vocabularies (and over 100,000 terms) and an umbrella vocabulary for semantic interoperability. It is adapted and being further developed for the SeaDataNet project.

Vocabulary level metadata is confined to short/long name, version/modification, description and url. However web service access is provided to content. The focus 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. The vocabulary server website gives a rationale:

> The NDG vocabulary server provides access to lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider

community.

Using standardised sets of terms (otherwise known as "controlled vocabularies") in metadata and to label data solves the problem of ambiguities associated with data markup and also enables records to be interpreted by computers. This opens up data sets to a whole world of possibilities for computer aided manipulation, distribution and long term reuse.

An example of how computers may benefit from the use of controlled vocabularies is in the summing of values taken from different data sets. For instance, one data set may have a column labelled "Temperature of the water column" and another might have "water temperature" or even "temperature". To the human eye, the similarity is obvious but a computer would not be able to interpret these as the same thing unless all the possible options were hard coded into its software. If data are marked up with the same terms, this problem is resolved.

In the real world, it is not always possible or agreeable for data providers to use the same terms. In such cases, controlled vocabularies can be used as a medium to which data centres can map their equivalent terms.

Interactive access is not currently provided. Web service access (via SOAP or HTTP-POX) to the vocabularies is provided via an API. Vocabularies are represented in SKOS and some support for versioning is in place. A mapping service is based on the SKOS mapping relationships. Current work involves investigating query expansion possibilities in the search interface. Considerable effort has been spent on training documentation and activities. A content governance model is in place, with layers of collaborative involvement (http://www.bodc.ac.uk/data/codes_and_formats/seavox/)

NSDL registry

NSDL registry home page http://metadataregistry.org/

Step-by-Step Instructions for Using the NSDL Registry http://wiki.metadataregistry.org/Step-By-Step_Instruction

Options 1 and 3, interactive access

The US National Science Digital Library (NSDL) TR registry contains vocabulary content, represented in SKOS. Content is mainly educational, along with RDA (Resource Description and Access) vocabularies. Development of the registry continues – it is intended as both a vocabulary and a metadata schema registry. A sandbox is available, as well as a development blog. The software is open source and available for download.

Administrator users can create and maintain their own vocabularies via forms for entering a concept and additional (SKOS) properties. An import function is planned. Support for namespaces and assigning URIs is also provided. Functionality includes support for storing vocabularies and their content and exporting as SKOS. It is possible to interactively retrieve vocabulary content as a list but conventional hierarchical browsing does not appear to be supported.

Ambitious support for versioning has been a focus of development effort. Every change is tracked and time-stamped which allows a snapshot to be taken of a vocabulary at a particular point in time.

Current access is mostly interactive. Ultimately, the registry is intended both for human inspection and m2m applications and future plans include import capability and web service access. Other future plans include evolving the registry to encompass an integrated metadata schema registry (now partly implemented) along with the terminology registry. We understand that NSF funding has terminated and the project is seeking funding to support the continuing further development activity.

OCLC's Terminology Services Pilot

OCLC's Terminology Services Pilot home page http://tspilot.oclc.org/resources/

Project overview http://www.oclc.org/research/projects/termservices/default.htm

Options 1, 2, 3, interactive and m2m access

The OCLC Terminology Services research project has implemented a registry, with both interactive and m2m access (SRU and CQL) available. The registry is currently being piloted and has been used in an experimental query expansion OPAC service by the University of Indiana.

Current vocabularies held include FAST, GSAFD, LC AC SH, LCSH, MeSH, TGM, with plans to include the Getty vocabularies via a license arrangement. Vocabularies are served as HTML, MARC XML, Zthes, and SKOS. Concept/headings records are in MARC 21 Authority data format. The vocabulary metadata are in MARC 21 Bibliographic data format and the intention is to further extend the metadata, drawing on the NKOS metadata element set developed by OCLC.

Web service functionality includes searching for complete vocabularies and also concepts via SRU CQL, based on the vocabulary metadata set and also the following fields:

- concept/heading identifier
- non-preferred terms
- words from a heading
- facets of a concept/heading
- identifier for a mapped concept/heading
- words in a mapped concept/heading
- MARC tags in a concept/heading record
- words in a note
- main heading
- words from preferred and non-preferred terms
- a KOS identification code.

The underlying SRU database has fully expanded hierarchies, with the aim of enabling web service calls to return composite vocabulary elements for building browsing interfaces without repeated calls to the server.

There is currently no provision for upload but the intention is to expand the vocabularies held in the registry and to investigate upload functionality. The source code has not been distributed to date but may become available as open source in the future. The service is being considered as a possibility for a full OCLC service.

Taxonomy Warehouse

http://www.taxonomywarehouse.com/

Option 1, interactive access

Taxonomy Warehouse is operated by Dow Jones Factiva's Taxonomy group, who also market the Synaptica Semantic Management Tool. Taxonomy Warehouse is probably the oldest (2001) dedicated TR and functionality is limited, although there is some coverage of vocabulary metadata elements (see Section 7). It claims to host more than 670 taxonomies (73 subject domains) from 288 publishers in 39 languages. Interactive access is provided to the vocabulary metadata, via browsing or search of vocabulary title and categories. Users can browse vocabularies, publishers, categories and an A-Z index.

Non-linked URLs allow access to the vocabulary publisher (and download site if freely available) and a (human) service is offered to facilitate obtaining a license to proprietary vocabularies and also for format conversion. A form is available for submitting a vocabulary to the registry. Some scholarly references are provided alphabetically as resources (the latest appears to be 2004).

7 Metadata

7.1 Metadata in terminology registries

We grouped the metadata into several categories, mostly based on the NKOS Registry (1998) metadata. *Names* of metadata and metadata categories are primarily taken from the same document, with additional ones from other registries. With each category group there is a table showing which of the metadata elements are used by individual registries or relevant documents. In both NKOS documents it is specified which elements are obligatory and which optional, but this is not clear from looking at other registries. A full list of metadata elements with examples for each registry is given in Appendix 4.

7.1.1 Product information

Product Information	A	B	C	D	E	F	G	H	I	J	K	L
Product Name/Title	+	+	+	+	+	+	+	+	+	+	+	+
Variant Product Name/Title /Acronym	+	+	+	+		+						
Type of Product	+	+	+		+	+		+		+	+	
Product Description	+	+	+*			+	+		+	+	+	+
Auxiliary Lists	+											
Author/Editor	+	+			+	+			+	+		
Current Version/Edition	+						+					+
Date of Current Version	+	+	+							+	+	
Product Update Frequency	+		+								+	
Available Format(s) and Size	+	+	+			+					+	
Online Availability	+		+		+	+	+**	+	+	+	+	
Notes	+											
URL for Examples	+											

*only have description but containing the marked elements

**also have URI base domain and token

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Product name or title is the name of the vocabulary. Terms used in other registries or

documents for this element are: name, name (includes acronym), title, and KOS title.

Variant product name or title refers to any other names or titles by which the product is known, including acronyms. Terms used in other registries or documents for this element are: alternative title.

Type of product refers to the type of terminology. In NKOS 1998, a choice needs to be made between the following: authority file, classification system, concept space, dictionary, hierarchical thesaurus, subject heading list, synonym list, topic grouping hierarchy (like the Global Change Master Directory), and other (with an explanation of the new type). Terms used in other registries or documents for this element are: type, resource type, KOS type.

Product description is additional information that does not appear in other metadata. The same term has been used in all registries or documents where present.

Auxiliary lists refers to names of any lists of terms or auxiliary tables that accompany the basic vocabulary product. Only used in NKOS Registry 1998.

Author/editor of a vocabulary. Terms used in other registries or documents for this or closely related elements are: owner/creator, creator, owner, organization name, contributor, author(s), authority.

Current version/Edition is an explanation of status if not published yet, or which edition it is. Terms used in other registries or documents for this element are: status, version.

Date of current version. Terms used in other registries or documents for this element are: date, creation date.

Product update frequency. Terms used in other registries or documents for this element are: revision cycle.

Available format(s) and size refers to the format in which available and size e.g. in bytes if available. Terms used in other registries or documents for format are: format, formats, model.

Online availability is the URL. Terms used in other registries or documents for this element are: URL, identifier. "**Identifier**" refers also to a unique identifier such as URL, ISBN, DOI.

Notes is for any explanations about the formats available, online availability, versions, and other such information.

URL for examples would be a file with examples of actual contents to give a better idea of the nature of the product, if the whole product is not online.

7.1.2 Scope and usage

Scope and Usage	A	B	C	D	E	F	G	H	I	J	K	L
Major Subject Coverage	+	+		+	+	+			+		+	
Minor Subject Coverage	+	+		+		+					+	
Used by (user community and applications)	+	+					+					
KOS relation		+										

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Major and minor subject coverage are recommended to be standardized from a controlled vocabulary as well. Terms used in other registries or documents for this element are: domain, environmental and non-environmental topics (Ecoterm), subject controlled, keywords, KOS subject and minor subject, subject, subjects covered, categories.

Used by (user communities and applications) lists actual publications or databases for which the vocabulary was designed or the general types of publications that use the vocabulary. Terms used in other registries or documents for this element are: application, community.

KOS relation is a reference in the form of an identifier to a related KOS.

7.1.3 Vocabulary characteristics

Vocabulary Characteristics	A	B	C	D	E	F	G	H	I	J	K	L
Language(s)	+	+			+	+	+	+	+		+	
Multilingual											+	
Type of Terms (e.g. concept terms, geographic names)	+	+								+		
Entity value		+										
Description of Overall Structure	+											
Source of New Terminology	+											
Number of Preferred Terms or Nodes	+										+	
Number of Non-preferred Terms	+										+	
Types of Relationships	+	+	+								+	
Arrangement of Displays (e.g., alphabetical, hierarchical)	+	+										
Depth of Hierarchy (maximum number of levels)	+										+	
Information given (e.g. Usage notes, References)		+										
Total terms											+	
Top terms											+	
Relationships											+	
Notation scheme											+	
Notation description											+	
Notes fields											+	
Additional information											+	

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Language(s) is an element for language(s) used by the vocabulary. Terms used in other registries or documents for this element are: language identifier.

Multilingual refers to stating whether more than one language is used by the vocabulary.

Type of terms refers to which terms are used by the vocabulary, e.g., concepts, geographic names, corporate names, etc. Terms used in other registries or documents for this element are: entity type, unit of information.

Entity value is listed but not defined yet.

Description of overall structure is an overview of the organizational structure of the vocabulary and any particular design particulars that potential users might need to know.

Source of new terminology serves for describing how new terms are added, e.g., by conversion from another source.

Number of preferred terms or nodes is the number or number range of preferred (valid) terms, the number of individual classification nodes, or the total number of "entry terms" if the vocabulary treats all terms the same. Terms used in other registries or documents for this element are: preferred terms.

Number of non-preferred terms is the number or number range of non-preferred terms. Terms used in other registries or documents for this element are: non-preferred terms.

Types of relationships is for labels or names of relationship types. Terms used in other registries or documents for this element are: relationships.

Arrangement of displays is for the presentation formats, for example alphabetical, hierarchical, tagged format, classification tree, rotated (permutated), facetted, graphical.

Depth of hierarchy is the maximum number of levels. Terms used in other registries or documents for this element are: levels.

Information given includes, for example, usage notes, conceptual relationships, references, date of entry, spelling variants, etc.

Total terms is the number of all terms.

Top terms is the number of top hierarchical terms.

Relationships is the number of terms with relationships to other terms.

Notation scheme is a yes/no for whether there is a notation system.

Notation description is description of the notation system.

Notes fields is what types of notes field there are (e.g., Other, Scope).

7.1.4 Vendor and contact

Vendor	A	B	C	D	E	F	G	H	I	J	K	L
Vendor Name	+	+	+		+	+	+	+		+	+	+
Vendor Street/Post Office Box	+			+							+	
Vendor City	+											
Vendor State/Province	+											
Vendor Country	+		+									
Vendor Postal Code/ZIP Code	+											
Vendor Voice Phone	+										+	
Vendor TDD/TTY Phone	+											
Vendor Fax	+										+	
Vendor Email	+					+		+				
Vendor Logo URL	+											
Vendor Web Site URL	+										+	
Vendor Hours of Service and Timezone	+											
Vendor Service Description	+											
Contact	A	B	C	D	E	F	G	H	I	J	K	L
Contact Name	+			+								
Contact Voice Phone	+											
Contact Fax	+											
Contact Email	+			+	+							
Contact Web Site URL	+											
More Contact Information	+											
Comments to Registry Editor	+											

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Vendor name is the name of the vendor who should be contacted about access to and use of the product. Terms used in other registries or documents for this element are: owner/creator, publisher, publisher(s), organization name, authority.

Vendor street/post office box, City, State/Province, Country, Postal code/ZIP code. Terms used in other registries or documents for these elements are: address.

Vendor voice phone. Terms used in other registries or documents for these elements are: phone.

Vendor TDD/TTY phone for if there is a special phone for the hearing impaired.

Vendor service description. The overall services of the producer or vendor so that potential users of the product(s) will have an understanding of the business environment of the organization.

Contact name for if a potential user needs to know the name or position title of a particular person to contact about the product. Terms used in other registries or documents for these elements are: technical contact, content contact.

Contact web site URL for if the contact has a different homepage connected to the vocabulary.

More contact information for any additional information that potential users of the product should know about how to contact.

7.1.5 Submission

Submission	A	B	C	D	E	F	G	H	I	J	K	L
Submission - organization name								+				+
Submission - organization mail address								+				
Submission - contact								+				
Date when added to registry									+			

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Submission – organization name. Terms used in other registries or documents for this element are: authority.

7.1.6 Terms and conditions

Terms and Conditions	A	B	C	D	E	F	G	H	I	J	K	L
Purchase Price by Format (or cost-free statement)	+	+				+			+			
Subscription Price by Format	+											
Licensing Availability	+											
Restrictions (or no-restrictions statement)	+											

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

Purchase price by format (or cost-free statement) is to provide purchase price information by product format or a statement that the product is freely available. Terms used in other registries or documents for this element are: rights.

Subscription price by format refers to licensing information by format.

Licensing availability is the actual licensing fees or an indication of the approximate fees or general availability for each product format or media that is available for licensing.

Restrictions refers to any restrictions on the use of the product(s) or a general statement about how the arrangements for use can be made.

7.1.7 Administration record

Administration record	A	B	C	D	E	F	G	H	I	J	K	L
Administration Record - creation date								+				
Administration Record - last change date								+				
Administration Record - effective date								+				
Administration Record - until date								+				
Administration Record - change description								+				
Administration Record - administrative note								+				
Administration Record - explanatory comment								+				
Administration Record - unresolved issue								+				
Administration Record - origin								+				

A: NKOS Registry 1998

B: NKOS Registry 2001

C: CENDI

D: Ecoterm (Environmental Terminology and KOS)

E: Food and Agriculture Organization (FAO) of UN

F: Hodge et al. 2007 (10th OFMR)

G: National Science Digital Library Registry

H: ISO 11179 (Information Technology - Metadata registries (MDR))

I: OCLC Terminology Services

J: SPECTRUM Terminology Bank

K: Taxonomy Warehouse

L: Vocman (Becta Vocabulary Bank)

7.1.8 Ontology metadata

In the ontology community, the Ontology Metadata Vocabulary (OMV) has been proposed for metadata about formal ontologies (http://omv.ontoware.org/). A report from 2008 (Palma et al.) defines OMV as a formalized OWL ontology itself. Metadata elements are divided into three categories: required, optional, and extensional. The latter refer to specialised metadata that are not part of the core metadata scheme. Core elements consist of the main class Ontology and various aspects related to creation, management and usage of an ontology such as OntologyTask (purpose), LicenseModel, OntologyLanguage etc. Each element has a number of attributes. For example, class Ontology has the following attributes: URI, version, resourceLocator, name, acronym, description, documentation, keywords, creationDate, modificationDate, naturalLanguage, numberOfClasses, numberOfProperties, numberOfIndividuals, and numberOfAxioms.

In OMV metadata elements are also divided based on the type and purpose of the information they contain: general (general information about the ontology), availability (information about the location of the ontology such as URI or URL), applicability (intended usage or scope of the ontology), format (representation languages in which the ontology is formalized), provenance (organizations contributing to the creation of the ontology), relationship (relationships to other resources such as versioning, extensions etc.), statistics (e.g., number of classes), and other (not covered in earlier categories).

7.2 Recommended metadata

This section tentatively proposes a terminology registry metadata schema based on metadata listed above as well as requirements collected based on use cases and through contacting experts for this scoping study. The majority of metadata listed in the above tables are recommended here (as core or optional). The ones that are left out are those for which no rationale was recognized in the JISC context. Also, some new metadata elements are added, as recognized through use cases or experts.

This is intended as an initial proposal. More work and empirical evidence is needed to test the suitability of these metadata in a real-life JISC TR. Another reason why these metadata need to be tested is that such rich metadata are time-consuming to create and maintain, which can be a particular issue when a third party, for example a vocabulary provider, would be responsible for creating and maintaining the metadata.

The metadata are grouped into the following suggested categories: general information, scope and usage, detailed characteristics, terms and conditions, and provider. While rich metadata are desirable, because of the time required to create them, some elements are tagged as optional.

We have published the recommended metadata as an outcome on the project website and have invited further feedback beyond the timespan of the TRSS project.

1 General information

Vocabulary name

Vocabulary alternative name or acronym (Optional)

Vocabulary type*

Author or editor

Current version/edition

Date of current version/edition

Update frequency (Optional)

Available format(s)

Available terminology services (Optional)

Vocabulary identifier

Vocabulary sample URL (Optional)

Vocabulary description

*See Section 2.2. A recommendation for future work is to further develop the classification of different vocabulary types.

Elements in this group are intended for creating metadata descriptions that will facilitate the discovery of vocabularies. This group of elements corresponds to the “Product information” group of elements from Section 7.1. The recommended elements are the same as the ones listed in the previous section apart from “Auxiliary lists” and “Notes”. The “Auxiliary lists” element was suggested only in NKOS Registry 1998 and refers to names of any lists of terms or auxiliary tables that accompany the basic vocabulary. It is excluded from the recommended elements because no need for it was recognized in this study. The “Notes” element was suggested only in NKOS Registry 1998 and refers to “explanations about the formats available, online availability, versions, and other such information”. This information is part of other elements recommended here. Also, a new element “Available terminology services” is added based on the recognized requirements, especially in Options 2 and 3.

2 Scope and usage

Language(s)

Major subjects covered

Minor subjects covered (Optional)

Purpose as given by author/publisher

Used by (Optional)

Description of collections where used (Optional)

Usage case study (Optional)

Use in application profiles (Optional)

Rating. Perhaps an automatically generated rating based on publisher, conformance to standards, spread of usage etc. (Optional)

URL to vocabulary users’ discussion board (Optional)

Change notification details (Optional)

Related vocabularies (Optional)

Overlap with related vocabularies (Optional)

Mappings to other vocabularies: which vocabularies, intellectual or automated (Optional)

URL to tutorial for applying vocabulary (Optional)

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. These two groups of elements correspond to the “Scope

and usage" and the "Vocabulary characteristics" groups of elements from Section 7.1. For changes from the previous section, see below under "Vocabulary characteristics".

3 Vocabulary characteristics

Type of display (Optional)

Description of overall structure (Optional)

Type of terms (Optional)

Types of relationships (Optional)

Total number of terms* (Optional)

Total number of classes* (Optional)

Number of preferred terms* (Optional)

Number of non-preferred terms* (Optional)

Depth of hierarchy (Optional)

Notes fields (Optional)

Information given (Optional)

*these could be updated automatically as vocabulary changes

In comparison with Section 7.1, the two groups of elements "Scope and usage" and "Vocabulary characteristics" contain most of the elements, apart from the following which were excluded since no rationale was recognized for them in the JISC context: Source of new terminology (only in NKOS 1998), Multilingual (used only by Taxonomy Warehouse), Entity value (listed but not defined yet in NKOS 2001), Top terms (used only by Taxonomy Warehouse and often left empty), Notation scheme and Notation description (also used only by Taxonomy Warehouse).

Based on requirements and in consultation with experts, the following were also recognized as important metadata and were added to the list under the Scope and usage group:

Description of collections where used

Usage case study

Use in application profiles

Rating. Perhaps an automatically generated rating based on publisher, conformance to standards, spread of usage etc.

URL to vocabulary users' discussion board

Change notification details

Overlap with related vocabularies

Mappings to other vocabularies: which vocabularies, intellectual or automated

URL to tutorial for applying vocabulary

4 Terms and conditions

Availability: free for all, free for registered users, costs

Import/download instructions (Optional)

Purchase/subscription price

Licensing options (Optional)

5 Provider

Vocabulary provider name

Vocabulary provider URL

Vocabulary provider contact details

Vendor provider and contact details were reduced here to these three elements as no rationale was seen for all the types of contact given mostly by NKOS 1998.

Submission metadata were left out of this section because they are related to Option 3 which needs more exploration when it comes to metadata.

Administration record is provided only by ISO 11179 and refers to the metadata about the created metadata. No rationale for it has been recognized in our context so far.

7.2.1 Defining data elements

As recommended in NKOS 2001, and following Dublin Core, each element could be defined using a set of ten attributes from the ISO11179 standard for the description of data elements. These include:

- Name - The label assigned to the data element
- Identifier - The unique identifier assigned to the data element
- Version - The version of the data element
- Registration Authority - The entity authorized to register the data element
- Language - The language in which the data element is specified
- Definition - A statement that clearly represents the concept and essential nature of the data element
- Obligation - Indicates if the data element is required to always or sometimes be present (contain a value)
- Datatype - Indicates the type of data that can be represented in the value of the data element
- Maximum Occurrence - Indicates any limit to the repeatability of the data element
- Comment - A remark concerning the application of the data element.

Five of the above ten attributes are common to all the elements. These are:

- Version
- Registration Authority
- Language
- Datatype
- Maximum Occurrence.

Further, each data element should be defined with a formal definition.

7.2.2 Terminology services metadata

Metadata on terminology services falls under Option 2 and would be developed by the collaborative project set up for that purpose (see Recommendation 3 in Section 10). This would form part of future development of IESR Service Metadata. The IESR Metadata Review February 2007 notes that the current list of Service Types needs revision and expansion. It refers to service function lists proposed for the eFramework and for ISO 2146 and mentions Terminology as a future addition to current IESR list.

Existing eFramework services, such as Find and Map, could be generalised and new additions defined. Possible terminology services include both high level and lower level services. For example, there could be provision for finding vocabularies as a whole, finding concepts/terms within vocabularies, returning subsets of vocabularies in order to dynamically create interface elements, describing mapping methods and provenance. The JISC Terminology Services and Technology Review (Tudhope, Koch and Heery 2006, Section 4.3) discusses a layered set of terminology services at different levels of granularity and gives some possible examples using the SKOS API as (one) low level protocol. As with IESR generally, service metadata should support various bindings to specific APIs or low level protocols.

The distinction between discovering (identifying) a suitable vocabulary and retrieving metadata about it, versus retrieving member concepts and terms of a vocabulary tends to be overlooked. However it is critical for enabling use of terminology services within other applications. As a partial list of example areas, terminology services might include provision for:

- browsing, searching for complete vocabularies;
- services related to member terms/concepts/relationships (and extracting subsets of a vocabulary), including known item search (eg via URI) and search matching a user string;
- various browsing services;
- various mapping services;
- various services supporting query expansion;
- services for validating names and controlled terms in metadata, including a spell-check service;
- services supporting disambiguation;
- (in the future) services for automatically generated (vocabulary based) metadata via automatic classification and information extraction.

8 Underlying standards

A TR should follow common standards for (externally) representing and accessing vocabularies, although it may hold data internally in a bespoke format. While it may not be feasible for any one vocabulary representation schema or access protocol to be universally adopted, TRs should orient to existing standards. Where possible it will be desirable to offer content in a variety of common formats. Some relevant standards are briefly reviewed below (for more information, see Tudhope, Koch, Heery 2006).

8.1 Representations

At a syntactical level, representations for import/export and vocabularies interchange formats should be based on XML as an underlying format. The MARC 21 Format for Authority Data in XML (MARCXML 2008) may be appropriate in some cases but is likely to prove cumbersome for many applications. The ADL Thesaurus Protocol (ADL Thesaurus Protocol 2003) has a light weight XML schema, employed by the Alexandria Digital Library project. The Zthes XML Schema (Zthes XML Schema 2006) is used by the Zthes profile (Zthes 2006). Work on Part 5 (interoperability issues) of the new BSI Thesaurus Standard is still ongoing. However a draft data model and associated BSI Thesaurus Standard XML Schema have been produced by the BS8723-5 working group (BSI 2007) and it is recommended that this be considered if a purely XML Schema is adopted.

In some cases, an RDF based representation may be appropriate. Simple Knowledge Organization System (SKOS) Core is a W3C Working Draft RDF/XML representation for KOS, based on a formal data model (SKOS 2008). It was originally conceived for thesauri but has potential for vocabularies generally, since it allows specialization (extension) for other types of vocabulary. SKOS vocabularies might thus include taxonomies and classifications, and less structured vocabularies for social tagging.

Some registry work, particularly in e-Science domains, has made use of formal ontology concept representations, modeling a knowledge domain with precise definitions and relationships. They are designed to be used by first order logic inferencing systems, such as Description Logic. In these situations, OWL (OWL Web Ontology Language Overview 2004) tends to be a standard representation format (though OBO (OBO 2006) is also used by the OBO Foundry).

Europeana, the EC Digital Library, specifies SKOS in its technical interoperability document (Dekkers 2007). It is emerging as the standard Semantic Web representation for vocabularies designed for information retrieval and browsing purposes, as opposed to logic-based ontologies. It has also received attention in Web 2.0 applications. The W3C Semantic Web Deployment Working Group is currently considering the appropriate mechanisms for combining SKOS and OWL representations.

For compatibility with future directions in Digital Libraries and the Semantic Web, it is recommended that SKOS be considered as one of the representation formats in any JISC TR.

8.2 Identification of concepts, terms and vocabularies

The unique identification of TR's resources is vital. Concepts, terms, vocabularies and the relationships between these various types of entities need to be identified so that they can be automatically referenced and processed. Identifiers should be persistent and unique, following standard conventions for dereferencing. It is recommended that a TR employ 'http' URIs (followed by the vast majority of TRs reviewed here). Attention should be paid to any future developments on standard formats for structuring URI strings, in order to facilitate automated processing.

Building on the possibilities of persistent URIs, the linked data initiative (Linking Open Data 2008) is a move towards the Semantic Web vision of a 'web of data'. Content is

made available in RDF, addressed via virtual but persistent URIs that allow HTTP clients to 'negotiate' their preferred representation of the content. This facilitates data reuse by RDF-aware applications and services. Summers *et al*. (2008) have experimented with making Library of Congress Subject Headings as linked data via SKOS and OCLC are also experimenting with linked data approaches.

8.3 Protocols, profiles and APIs

Protocols for retrieving vocabulary data are closely linked to representation formats. It is necessary to distinguish programmatic access to the vocabulary (e.g., searching or resolving to concepts) from vocabulary support for query (e.g., as a source of query terms) or browsing.

Generally, a protocol should be defined independently of any particular binding, allowing APIs or access methods to be defined for various platforms. There is discussion currently, as to the relative merits of SOAP based web services with an XML wrapper (SOAP 2007), light weight REST based approaches using only the standard http 'verbs' (Tilkov 2007) and URL based Remote Procedure Calls, where queries and parameters are communicated as part of the http address.

Zthes (Zthes 2006) was originally based on Z39.50 (Z39.50 2007) but is now also available as a profile for SRU and SRW (SRU 2008). Although based on the Z39.50 abstract model, SRW/U is less complex and is XML based. SRU is a URL REST-based alternative to the SOAP-based SRW. This can be combined with the CQL (Contextual Query Language, CQL 2007), Boolean query language, (not itself terminology aware). Access is normally via a single data record, which would necessitate repeated server calls, in order to dynamically construct a browsing interface. However, work at OCLC is ongoing on a fully exploded hierarchical index for the OCLC TR, allowing mini-trees to be requested in a single call.

The SKOS API (SKOS API 2004) defines a core set of methods for programmatically accessing and querying vocabularies, based on the SKOS-Core RDF schema. While intended as web service calls, the API itself remains independent of implementation details. One set of SKOS calls returns a concept(s) with its details via an ID, a preferred label, or matching a keyword or regular expression. Another call returns a list of supported semantic relations for the given vocabulary. Another set of calls returns concepts connected by a specified relation or all immediately connected concepts. It is possible also to get a set of concepts connected by a relation up to a given path length.

9 Governance

Governance and management is often the most problematic issue for any registry. This has been emphasised by several of the contacts interviewed for the study, in respect of TRs. While the issues cannot be separated from the particular functions and uses of any given TR, there are some general points. For example, we can distinguish technical and content governance issues.

Technical governance includes the usual computing best practice, maintenance, backup, mirroring, preservation of both the digital data and terminology services over the long term, as versions of software and operating systems change. It requires an infrastructure that can provide this, along with allocation of technical responsibility and authority.

Content governance, on the other hand, addresses similar issues from the specific point of view of vocabularies. The issues vary according to the characteristics of a particular TR (and which of the three options described in Section 4 it follows). They may include responsibility for the following:

- validation of correctness of content;
- versioning of representations according to standard: maintaining the vocabulary representations supported according to appropriate versions of any standards;
- versioning of the vocabulary intellectual content, which may include support for update of the whole vocabulary or of individual elements, together with evaluation of proposed additions (deletions), proposals for deprecated elements; and,
- evaluation (and selection) of new vocabulary offered to the registry, and judgment as to whether their suitability, quality, provenance justifies their inclusion in the registry;
- promotion of the TR and its services; and,
- 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.

Apparently, one of the reasons why the Dublin Core registry effort for vocabularies used with DC was abandoned was the maintenance and governance problem. One respondent mentioned a similar case in the eLearning domain:

> However, I have worked on small European projects which did attempt to build a registry or vocabularies for learning object metadata, and one of the issues that that project stumbled upon was where should the vocabularies reside, and who is going to be responsible for entering data into the registry about them. Because it can be very difficult to guarantee that the data that you're gathering is definitive in any way, shape or form. I mean, can anybody submit a vocabulary to the registry?

Another respondent said on this issue:

> … you would have a management problem, because you can't just open this up to the world and say "register your terminology here," because you'd get people putting rubbish in. So, you would need somebody to be, at least, moderating it.

These examples bring up the issue in open registries of the quality of the content and whether the registry is intended to be seen as a social collaborative effort (in which case an open access model might be appropriate), or as definitive and standard by some criteria.

Various contacts pointed out that the perceived longevity of shared services can be a problematic issue for uptake. It can be difficult for potential user organisations to have sufficient confidence that particular services or provisions will be maintained in the long-term, before basing important elements of their strategy and day to day work flow on the existence of such services. Of course, this issue is not confined to TRs.

Apart from stability of a registry, other issues affecting uptake included trust in the quality of information provided. In this context, see, for example, the Center for Research Libraries checklist of criteria for trustworthy repositories (Trustworthy Repositories 2007). Another factor which might positively influence uptake is whether a registry is part of a habitual workflow, for example an advice or information point. The quality of the user interface and tools, and any value-added services are clearly crucial.

Differing degrees of formality to the governance arrangements are found according to the domain. For example, the international Climate and Forecast group associated with BODC has a formal governance model, appropriate for a scientific international body.

The more complex the registry, the more resources required for governance and the more functionality to be maintained. It can be argued that dynamic editing of the vocabularies is not part of core registry functionality and that is better undertaken by the vocabulary provider or third party tool providers. Updated vocabularies could be uploaded in their entirety for Option 3. On the other hand, some registries (e.g., NSDL) see the ability to offer editing functionality, as a selling point of the registry. Some users may welcome the ability to evolve the vocabulary as part of the registry. The extent to which this is useful depends on the nature and internal resources of the vocabulary providers.

Thus there needs to be careful consideration of cost benefit issues. Some of the larger vocabularies have commercial business models where m2m use may raise issues of managing IPR and copyright. Several contacts highlighted the governance problems inherent in holding vocabulary content within the registry. In addition to maintaining current versions, the vetting, selection and quality control of vocabularies offered to the registry impose significant demands on resources.

10 Recommendations and options for JISC

10.1 General overview of a JISC TR

Sections 4 and 5 reveal the wide scope of application of TRs and the potential benefits. This ranges from support for an inquiry as to whether any suitable vocabulary exists for a particular domain or topic (for example, a recent post to the JISC Taxonomy list inquired about vocabularies describing human emotions) to support for information at the level of an individual term or concept. For example, a TR can be the source of validation of controlled elements in metadata profiles. The wide range of use cases demonstrates that a TR offers a distinctive set of potential benefits in its own right. These range from people searching for a vocabulary to adopt or modify for their project's purposes to developers wishing to make use of existing terminology service functionality for browsing or search applications.

The review in Section 5 indicates that there is significant interest in TRs both nationally (eg from JISC projects such as HILT, IEMSR, IESR and the UK cases in Section 6) and internationally. The JISC Pedagogical Vocabularies Project Report (2005), the Terminology Services and Technology Review (2006), HILT, IEMSR and IESR all support some form of TR. The major projects at FAO, NERC and OCLC, for example, demonstrate commitment of resources and, in the case of OCLC and Taxonomy Warehouse, possible commercial potential of a TR in their contexts. However, in the JISC context, any TR should be seen as a shared infrastructure service (it is unlikely to have immediate commercial potential).

For these reasons, a general recommendation of this report is that JISC should consider possibilities for moving towards a TR relevant to UK HE purposes in incremental steps, while taking account of national and international developments.

Recommendation 1: JISC should consider a TR for UK HE purposes.

The different possibilities are discussed below. The main options are

1. Registry provides metadata for each vocabulary and links to vocabulary owner/provider
2. Registry provides metadata on (and links to) any available terminology services
3. Registry provides access to vocabulary content (either by downloading the complete vocabulary, or providing access to a vocabulary's concepts, terms and relationships)

As discussed in Section 4, these three options should be seen as independent facets which can be combined. Option 1 is the logical starting point.

Based on the various governance arguments (see Section 9), the majority of respondents tended to favour some version of Option 1 for any general JISC TR, with the registry maintaining rich metadata and possibly linking to terminology services. For JISC HE purposes, holding vocabulary content of the common large vocabularies would create governance and licensing problems. One possible solution might be to make use of a future OCLC Terminology Registry Service, which is discussed in more detail below.

On the other hand, the problems of maintaining content appear more tractable within specific domains. It is no accident that most of the cases where registries maintain content are in specific domains, where either a parent organisation owns the vocabularies, or where a close community of users are motivated (or mandated) to develop their own vocabularies and offer them to the registry. Examples include the FAO's agricultural vocabularies, the BODC oceanographic registry, the educational (eLearning) audiences of the Becta Vocabulary Bank and NSDL. Where the governance and organisational issues can be overcome there are various use cases for a TR in supporting vocabulary development as well as indexing and searching

applications via m2m services.

Currently, there appear significant resource and cost/benefit implications in holding content of large, general vocabularies inside any JISC registry, along with possible IPR issues. Since major JISC projects tend to involve large general vocabularies, where some content is licensed and due to the management and governance issues discussed, in our view it is not cost-effective in the immediate future to build and manage a registry that holds and distributes content for such vocabularies. This may change and the incremental steps below could allow reconsideration of Option 3 at a later date.

Holding content may however be feasible in sub-domains where the vocabularies are owned or under active development by a particular community. These efforts should be tracked and interoperability encouraged by the TR Support Project group recommended below. For example, the operational experience of BODC could be shared more widely.

10.2 JISC TR as part of IESR

Recommendation 2: JISC should consider providing an Option 1 TR (provides metadata for each vocabulary and links to vocabulary owner/provider)**, as part of an extended IESR.** The registry would be made available both for human inspection and m2m access.

This development would be an extension of IESR, with a TR as a first class entity within IESR, with its own dedicated vocabulary level metadata. Thus it would support IESR services to data collections but the TR would also give information about vocabularies in their own right (independent of any particular data collection). It would thus be possible to have a TR view of the entire registry for some uses, without the rest of the current IESR collections. The details would be considered by a design project for that purpose.

A focused design (small) project should be set up for IESR and relevant stakeholders to consider the implications and, assuming it is considered practical, make a proposal of the design and tender for the work packages.

Option 1 would support the discovery of existing vocabularies, by applications or projects requiring a vocabulary. This would reduce the unnecessary duplication of effort. Sometimes the existing vocabulary might be adopted as it stands and sometimes it might be extended (specialised) or modified. If the resulting new versions were fed back into the TR then the collection of vocabularies in the TR would grow. Immediate users could include librarians, information architects and JISC developers in charge of vocabulary provision (see use cases outlined in Section 4.2). The cost savings in the resources required to discover relevant vocabularies and construction of new vocabularies would benefit end users generally.

Option 1 would allow the situation to be reconsidered at a later date and decisions on further steps towards holding vocabulary content could be taken if warranted. Although detailed costs would be proposed by the focused design project, we anticipate that the costs for Option 1 (and also 2) would be fairly modest. We anticipate the cost of the focused design (small) project to be approximately 0.1 FTE (say 0.1 effort for 4 months for a 3 person team). While the design effort would provide a proper scoping and costing of any IESR TR development, as a rough estimate at this initial stage, we would not expect the cost of any subsequent extension to IESR to exceed 0,5 FTE.

We consider this to be a cost-effective (incremental) step that builds on the expertise of the IESR shared service. IESR is a mature, operational registry whose functionality would be enhanced by a TR. The TR would benefit from the general registry architecture and functionality provided by IESR, while IESR could offer the use cases associated with Option 1 (outlined in Section 4.2). Such a move might also serve to attract additional users for the IESR collection registry.

Focus could initially be on UK vocabularies used by major projects (including the vocabularies listed in Section 2.3) and some key international vocabularies. This could potentially provide support for a wide variety of JISC projects. It would also form a useful basis for future international collaborations.

In the slightly longer term, the TR should also make provision for terminology services, with development of metadata for terminology services and links to those services to support m2m access (Option 2). With this comes the need for further research into the metadata for types of services and how they fit into IESR services and the e-Framework. Consideration should be given to how these services would be used by developers. This should be a collaborative effort by IESR, any TR Support Project (see below) and relevant UK projects reviewed in Section 6, such as HILT, and consultation internationally (e.g., e-Framework partners, OCLC, etc.). One initial step could be to arrange a focused workshop on this topic.

This should include further development of TR related use cases within IESR, particularly from the point of view of developers for future m2m access. While we consider that there is a strong case for support of terminology services in general, this is outside the scope of TRSS. The recommendation for Option 2 is confined to providing metadata for terminology services developed and provided outside the registry. We assume that information about vocabulary metadata can be provided by existing IESR services. (Option 3 would almost certainly involve development of internal terminology services but is not a recommendation, at this time).

Recommendation 3: In the medium term, a pilot Option 2 (for both human and m2m access) should be considered after a collaborative study on an initial set of appropriate metadata elements for terminology services.

The immediate beneficiaries of Option 2 would be JISC (and other) developers seeking to implement various forms of terminology services, such as browsing, mapping, query expansion (see the review of functionality for terminology services in Section 4). This might start with an initial collaborative step (see section 10.5) that further developed the metadata set, particularly for mappings, terminology services and a typology of vocabularies.

Rather than implementing lower level services from scratch (reinventing existing solutions) for each application, developers would be able to locate building blocks (program libraries, web services, APIs) that could form components in new m2m and end-user applications. Various projects have begun to develop web services for different aspects of terminology services (see Section 6 and also Section 4.3 of Tudhope et al. 2006). It is important to promote evolution of standards for interoperability of applications in this area. For example, the specification and validation of a layered set of terminology (web) services APIs could be a useful development for the eFramework (see Section 7.2.2).

10.3 OCLC services in JISC TR

Recommendation 4: JISC should investigate the possibility of a licensing arrangement with OCLC to access vocabulary content and terminology services via an OCLC TR, augmented for JISC purposes.

In our view, OCLC is currently one of the few organisations with a potential business case incentive to support the necessary management activities for technical and content governance of large vocabularies (via OCLC Services). The OCLC Terminology Services group has an established track record in this area. The TR Support Project (below) should track OCLC developments and investigate connections between any JISC IESR TR and OCLC services. An OCLC TR could potentially offer Option 1, 2 and 3. In other words, vocabulary metadata, content and web services could potentially be available. This could be an avenue for pursuing interoperability with international registry projects.

The advantage of this would be the ability to make use of an existing solution rather than bearing the cost of developing TR functionality. A full Option 3 might cost in the region of 2 FTE, possibly less depending on skill set of the development team. The cost of the OCLC route would depend on the licensing agreement reached with OCLC or other third party provider. We understand that the Getty vocabularies will shortly be available via this route, although they are not currently listed.

Possible disadvantages include some reliance on non-UK provision (if that is an issue), The OCLC TR is currently an OCLC Research service and not a full OCLC service. The mechanisms for adding more vocabularies would need to be investigated and the willingness or business case for OCLC to take on relatively small scale or UK-specific vocabularies would need to be explored. If it was decided that this route was to be explored seriously then it should be considered at a relatively senior level within JISC and OCLC. This should include any agreements on licensing, addition and maintenance of UK vocabularies, collaboration on terminology service development, whether content and code would be made available to JISC generally and in the event that OCLC's service was discontinued.

10.4 Track major international and national projects

Recommendation 5: JISC should track major international projects, involving a TR, including NSDL and Europeana. Major national projects include BODC and Lexaurus Bank/Editor, which should also be tracked. See Section 6 for details.

The SKOS-based NSDL is outlined in Section 6.2. We understand the developers are currently seeking funding and the status and take up of the project should be kept under review by any TR Support Project, along with opportunities for cooperation.

The Europeana cluster of projects is intended to develop a multilingual European Digital Library Portal, with associated API and opportunity for third party services and reuse of content (http://version1.europeana.eu/web/europeana-project/) Europeana is likely to involve some form of TR and web service API. A Thematic Network has recently been set up to provide wider support and liaison. Cooperation with the Collections Trust, who are involved in Europeana related efforts, could also be an option.

10.5 JISC TR Support Project

Recommendation 6: JISC should consider the possibility of establishing some form of TR support and advisory effort that would act as a hub for management, inquiries, training, promotion and dissemination of any JISC TR. A support project would thus benefit the JISC community generally. We anticipate that this would be a relatively modest cost, not exceeding 0.5 FTE effort. It would also investigate via small projects key future issues and potential future development of the TR, including the following:

- collaboration with IESR on running and populating any JISC TR
- collaboration with IESR, HILT, OCLC and other stakeholders (see relevant projects in Section 6) on the collaborative study of metadata and use cases for terminology services. This should include standards development work on the interoperability of terminology web services and how they might fit into the eFramework. Ideally, a protocol would be independent of any particular binding, allowing different solutions for different technical approaches (SOAP/REST etc) and platforms. Consideration should also be given to the international activities; Europeana is planning a web service based API including some form of terminology services and there has been discussion recently on the W3C SKOS list. Such an effort would begin to move towards addressing Recommendation 3 (see Section 10.2).
- coordination and cooperation with metadata schema registry projects, such as IEMSR, including web service programmatic interfaces for a TR to interact with metadata schema registry components. This could also be part of the standards development in the preceding bullet point.
- investigating collaboration with related national and international projects, e.g., with BODC, Collections Trust, DART, Europeana, FAO, HILT, Max Planck registry work, NDSL, domain specific TR projects, such as BECTA and National Strategies teaching resources for schools
- tracking and maintaining compatibility with e-Learning developments (e.g., JORUM developments)
- tracking and maintaining compatibility with Semantic Web developments (e.g., linked data)
- tracking and maintaining compatibility with Web 2.0 developments, looking to integrate social tagging and folksonomic elements with TRs
- tracking and maintaining compatibility with the wider area, including ontology registries, natural language oriented (terminology) registries such as XMDR, work on APIs to dictionaries and encyclopaedias

It is likely that we will see a landscape with a variety of flavours of TRs and one role for a TR Support Project is to facilitate coordination and reuse between different communities. Another role is to facilitate synergy between TR efforts and Web 2.0 social tagging and folksonomy developments. At one level, almost all vocabularies were originally products of social collaboration (with different scales/types of editorial teams) and many vocabularies continue to evolve via some form of user suggestions (or commentary). Folksonomies can serve to enrich more structured vocabularies with, for example, new or end-user terms, while controlled vocabularies can potentially be used to structure folksonomies.

Recommendation 6b: As an alternative to proceeding immediately with Option 1, JISC could consider an interim step where the TR Support Project was assigned an additional set of tasks that attempted to gauge the level of interest and support for a general TR within the JISC community.

Given appropriate background knowledge in vocabularies, the Support Project could investigate possibilities for developing a community that might use a future JISC TR. This could include gauging the willingness of vocabulary owners to provide metadata for Option 1 and potential audiences for locating vocabularies via a TR. It would also be possible to test and refine core/optional judgments for metadata elements (see Section 7).

Some limited form of ‘Wizard of Oz’ prototyping provision of a TR might be possible via a new Vocabulary Mailing List, where inquiries to the Support Project via the mailing list acted in some aspects as a very simple surrogate for a TR. The Support Project could then email responses (partially corresponding to results from future searches of a TR for vocabularies meeting certain criteria). This process could also be used to refine the initial list of vocabularies appropriate for a JISC TR (in Section 2.3). As part of this

effort, an initial set of metadata for various vocabularies could be gathered and prototyped in some XML-based storage format.

One important issue for the Support Project would be to begin consideration of appropriate governance structures / options, in order to develop an understanding of the issues around governance, as part of this prototyping activity (see Section 9). For example, the project should develop a process for explicitly reflecting on issues surrounding:- the formation of a governance policy / review body, the principled selection of vocabularies, a process/criteria for evaluating and deciding whether to accept (uninvited) offered vocabularies, asking for metadata and reviewing metadata returned by vocabulary providers, the cost/benefits in how rich a metadata set to recommend (see section 7: a richer set might be more useful but deter vocabulary providers) promotion of the prototype registry and development of appropriate education/training. The various governance issues should be related to the relevant TR option(s) outlined above.

Another useful immediate task might be to compile a list of vocabulary development tools commonly available, along with their main characteristics (including cost and conformance to emerging standards such as SKOS). The list of tools at http://www.willpowerinfo.co.uk/thessoft.htm is a good starting place for thesauri. Note that some are part of large complex collection management systems (and relatively expensive if only limited functionality is required). Some software is also mentioned in Section 6, including the commercial Lexaurus Editor.

One issue for consideration is the extent to which tools supporting collaborative development are available. Consultants in the area could be asked to give their perspective. Another issue to consider is whether vocabulary creation/editing functionality is better separated from any TR functionality.

10.6 JISC TR metadata elements

These are general metadata recommendations to be considered beyond the immediate context of this report by TRs generally, as appropriate for their situation.

The field has reached a level of maturity where it is possible to make tentative recommendations on metadata elements for vocabularies in a TR (see Section 7.2). Where possible, a rich set of metadata should be maintained, allowing a diverse range of use cases for identifying a suitable vocabulary for different purposes. For example, the registry could offer a search by: domain, type of vocabulary, owner, languages, etc. Apart from contacts with the vocabulary provider, contact details of user groups and application implementers could possibly be listed, along with tutorials on how to apply the vocabulary in different ways.

Recommendation 7: TRs are advised to consider (as appropriate for their circumstances and functionality options) the vocabulary metadata element set tentatively recommended in Section 7.2.

10.7 JISC TR technical recommendations

These are general technical recommendations to be considered beyond the immediate context of this report by TRs generally, as appropriate for their situation.

Recommendation 8: A TR (Option 3) holding vocabularies internally should adopt SKOS as one of the representation formats for import and export. SKOS is an emerging standard and this will facilitate compatibility with the EDL (Europeana), e-Science specialised registries and Semantic Web developments.

Recommendation 9: Concept identifiers should be based on URIs (Option 3).

Recommendation 10: A TR (Option 2) should follow a service-oriented architecture and offer web service access, if possible via a variety of common standards (see Section 8).

11 References

Ahmad, Mohammad Nazir; Colomb, Robert M. (2007). Overview of Ontology Servers Research. Webology, Volume 4, Number 2, June, 2007. http://www.webology.ir/2007/v4n2/a43.html

Apps, Ann. (2007). Using an application profile based service registry. Proc International Conference on Dublin Core and Metadata Applications.

Apps, Ann. (2008). "Register locally – discover globally". Blog posting 03 April 2008. http://iesr.ac.uk/service-registries-blog/

Baker Thomas, Blanchi Christophe, Brickley Dan, Duval Erik, Heery Rachel, Johnston Pete, Kalinichenko Leonid, Neuroth Heike, Sugimoto Shigeo, (2002). Principles of Metadata Registries: A White Paper of the DELOS Working Group on Registries. http://delos-noe.iei.pi.cnr.it/activities/standardizationforum/Registries.pdf

Bargmeyer, Bruce (2005). eXtended Metadata Registries (XMDR). Presentation at the 7th NKOS Workshop at JCDL 2005. http://nkos.slis.kent.edu/2005workshop/Bargmeyer.ppt

BODC: British Oceanographic Data Centre's Vocabulary Server. (2008). http://vocab.ndg.nerc.ac.uk/

CENDI Agency Terminology Resources. (2008). http://www.cendi.gov/projects/proj_terminology.html

Chapman, A.; Russel, R. (2006, p. 15-20). JISC Shared Infrastructure Services Synthesis Study: A review of the shared infrastructure for the JISC Information Environment. http://www.jisc.ac.uk/Shared_Infrastructure_Services_Review_Sep_06

DCMI Registry. (2008). http://dcmi.kc.tsukuba.ac.jp/dcregistry/

Dekkers, Makx et al. (2007). EDLnet: Initial Semantic and Technical Interoperability Requirements.http://www.europeana.eu/public_documents/EDLnet_D2_2_Initial_Semantic_and_Technical_Interoperability_Requirements_final.pdf

Ecoterm. *See Hodge et al. 2007*

FAO Knowledge Organization Systems. (2008). http://www.fao.org/aims/kos_intro.htm

GeoCrossWalk. (2006). http://www.geoxwalk.ac.uk/

GRIMOIRES – Grid Registry with Metadata Oriented Interface: Robustness, Efficiency, Security. (2005). http://www.ecs.soton.ac.uk/research/projects/grimoires

Heery, Rachael. (2005). (Metadata and) Vocabulary Registries. http://www.ukoln.ac.uk/terminology/events/NKOSatDCMI05.html

Hillmann, Diane I.; Sutton, Stuart A.; Phipps, Jon; Laundry, Ryan. (2006). A Metadata Registry from Vocabularies Up: The NSDL Registry Project. International Conference on Dublin Core and Metadata Applications, 3 - 6 October 2006. Available at: http://arxiv.org/abs/cs.DL/0605111

HILT. (2008). http://hilt.cdlr.strath.ac.uk/index2.html

HILT M2M Feasibility Study. (2005) http://hilt.cdlr.strath.ac.uk/hiltm2mfs/

HILT Phase 4 demonstrators. (2008) http://hilt.cdlr.strath.ac.uk/hilt4/demonstrators.html

HILT Vocabulary resources. (2008). http://hilt.cdlr.strath.ac.uk/Sources/vocabulary.html

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. http://www.clir.org/pubs/reports/pub91/contents.html

Hodge, G.; Salokhe, G.; Zolly, L.; Anderson, N. (2007). Terminology Resource Registry: Descriptions for Humans and Computers. Presentation at Integrating Standards in Practice, 10th Open Forum on Metadata Registries, New York City, NY USA, July 9-11, 2007. http://www.metadataopenforum.org/index.php?id=21,74,0,0,1,0

IESR: Information Environment Service Registry. (2008). http://iesr.ac.uk/

IESR Metadata Review February 2007. (2007) Apps. A. http://iesr.ac.uk/metadata/reviews/review-200702.html

ISO 11179. (2007). http://metadata-standards.org/11179/

JISC Pedagogical Vocabularies Project Report (2005) http://www.jisc.ac.uk/elp_vocabularies.html

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

Koch, Traugott (2007). Controlled vocabularies, thesauri and classification systems available in the WWW. http://www.mpdl.mpg.de/staff/tkoch/publ/koslist.html

Kotok, A. (2003). Metadata rules - a report from the Open Forum on Metadata Registries. http://www.webservices.org/categories/technology/standards/metadata_rules_a_report_from_the_open_forum_on_metadata_registries/(go)/Articles

Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. London: Facet Publishing.

Lee, Ed (2004). Terminology resource discovery for the UK heritage sector – a case study. Presentation at the 3rd European NKOS Workshop. http://www2.db.dk/nkos-workshop/pp%20presentationer/NKOS-Ed.pdf

Merriam Webster Online, s.v. "vocabulary", http://www.merriam-webster.com/dictionary/vocabulary)

Merriam Webster Online, s.v. "terminology", http://www.merriam-webster.com/dictionary/terminology

METeOR. (2008). http://meteor.aihw.gov.au/content/index.phtml/itemId/181162

Middleton, M. (2008). Controlled vocabularies. http://www.imresources.fit.qut.edu.au/vocab/

Miller, Paul. (2000). I say what I mean, but do I mean what I say? Ariadne, 23 http://www.ariadne.ac.uk/issue23/metadata/

Mungal, Salvatore (2008). Terminology metadata. In *Metadata Open Forum, Sydney 2008.* http://www.metadataopenforum.org/index.php?id=34,168,0,0,1,0

National Monuments Record Thesauri (2008), English Heritage http://thesaurus.english-heritage.org.uk/

Nicholson, D. and McCulloch, E. (2006) Design of a Pilot SRW-compliant Terminologies Mapping Service (HILT). 5th European Networked Knowledge Organization Systems (NKOS) Workshop at the 10th ECDL Conference, Alicante, Spain. http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2006/presentations/d-nicholson.ppt

NKOS network. (2008). http://nkos.slis.kent.edu/

NKOS Registry, Version 2 with Documentation for Data Elements – Draft. (1998). http://nkos.slis.kent.edu/registry2.htm

NKOS Registry. Reference document for data elements. Version 3, Draft. (2001). http://staff.oclc.org/~vizine/NKOS/Thesaurus_Registry_version3_rev.htm

NKOS workshop at ECDL 2006. (2006). http://www.ukoln.ac.uk/nkos/nkos2006/

NKOS workshop at ECDL 2007. (2007). http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/

NSDL Registry. (2008). http://metadataregistry.org/

OCLC. (2008). Terminology Services. http://www.oclc.org/research/projects/termservices/

Open Ontology Repository (OOR) Initiative. (2008). http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository

Oxford English Dictionary, s.v. "vocabulary", http://dictionary.oed.com/cgi/entry/50278672?query_type=word&queryword=vocabulary

Oxford English Dictionary, s.v. "terminology", http://dictionary.oed.com/cgi/entry/50249385?single=1&query_type=word&queryword=terminology

Palma, Raul et al. (2008). O M V: Ontology Metadata Vocabulary for the SemanticWeb. OMV repor, v. 2.4, January 2008. http://ontoware.org/frs/download.php/418/OMV-Reportv2.4.pdf

Proffitt, Merrilee; Waibel, Günter; Vizine-Goetz, Diane; Houghton, Andrew. (2007). Terminologies Strawman. http://www.oclc.org/programs/events/2007-09-12a.pdf

SchemaWeb. (2005). http://www.schemaweb.info/

SPECTRUM Terminology Bank. (2008). http://www.mda.org.uk/spectrum-terminology/termbank.htm

Subject Analysis Systems (SAS) Collection. (2008). Faculty of Information Studies Inforum Library, University of Toronto. http://www.fis.utoronto.ca/index.php?option=com_content&task=view&id=386&Itemid=134 *The physical collection is catalogued and searchable in the University of Toronto Library Online Catalogue* at: http://webcat.library.utoronto.ca/

Summers Ed, Isaac, Antoine, Redding Clay, Krech Dan (2008). LCSH, SKOS and Linked Data, Proc. DCMI 2008. International Conference on Dublin Core and Metadata Applications

Svenonius, E. (2000). The Intellectual Foundation of Information Organization. Cambridge, Massachusetts and London, England: MIT Press.

Taxonomy Warehouse (2003-). Factiva. http://www.taxonomywarehouse.com/

Taylor, Mike. Becta VMS. (2008). http://www.slideshare.net/cetismdrsig/becta-vms

Thesaurus guide: analytical directory of selected vocabularies for information retrieval, 1992. 2nd ed. / prepared by EUROBrokerS for the Commission of the European Communities. Luxembourg : European Communities, 1993. 1033pp. (EUR/92/14006) ; (Rapports EUR 14006).

Thesaurus guide: analytical directory of selected vocabularies for information retrieval, 1992. (**1993**). 2nd ed. / prepared by EUROBrokerS for the Commission of the European Communities. Luxembourg : European Communities, 1993. 1033pp. (EUR/92/14006) ; (Rapports EUR 14006).

Tilkov, Stefan. (2007). A Brief Introduction to REST. http://www.infoq.com/articles/rest-introduction

Trustworthy Repositories Audit & Certification: Criteria and Checklist. (2007). http://www.crl.edu/PDF/trac.pdf

Tudhope, D. (2006). A tentative typology of KOS: towards a KOS of KOS? NKOS Workshop at ECDL 2006. http://www.ukoln.ac.uk/nkos/nkos2006/

Tudhope, D.; Koch, T.; Heery, R. (2006). Terminology Services and Technology: JISC state of the art review.

http://www.jisc.ac.uk/Terminology_Services_and_Technology_Review_Sep_06

University of British Columbia. (2004). Indexing Resources on the WWW. http://www.slais.ubc.ca/RESOURCES/indexing/database1.htm#online

Vocman. (2008). Becta Vocabulary Bank. http://bank.vocman.com/

WordNet. (2008). http://wordnet.princeton.edu/perl/webwn

WorldCat (Open Web Access). (2008). http://www.worldcat.org/

XMDR. (2007). http://www.xmdr.org/

XMDR Working Group. (2005). Direction of proposals for new edition (E3) of ISO/IEC 11179. http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N0851-N0900/WG2-N0883-Presentation-about-direction-of-11179-E3-proposals.ppt

Some Standards References

ADL Thesaurus Protocol. (2003). http://www.alexandria.ucsb.edu/thesaurus/protocol/

BSI. (2007). Website for BS8723-5 working group for resources such as XML Schemas. http://schemas.bs8723.org/2007-06-01/Documentation/Home.html

CQL. (2007). http://www.loc.gov/standards/sru/cql/index.html

Linking Open Data. (2008). W3C SWEO Linking Open Data community project Wiki. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

MARCXML: MARC 21 XML Schema. (2008). http://www.loc.gov/standards/marcxml/

OBO Flat File Format Specification, version 1.2. (2006). http://www.geneontology.org/GO.format.obo-1_2.shtml

OWL Web Ontology Language Overview. (2004). http://www.w3.org/TR/owl-features/

SKOS Simple Knowledge Organisation Systems. (2008). http://www.w3.org/2004/02/skos/

SKOS API. (2004). http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html

SOAP. (2007). http://www.w3.org/TR/soap/

SRU. (2008). http://www.loc.gov/standards/sru/

Zthes. (2006). http://zthes.z3950.org/

Zthes XML Schema. (2006). http://zthes.z3950.org/schema/index.html

Z39.50. (2007). http://www.loc.gov/z3950/agency/

12 Appendices

Appendix 1. Survey letter with questions

General

Dear _____,

We are contacting key experts in areas potentially related to the JISC's Terminology Registry Scoping Study (http://www.ukoln.ac.uk/projects/trss/). As an initial part of the study, we are looking for feedback on requirements, usage scenarios, and comments generally.

A terminology registry describes, identifies and points to sets of vocabularies available for use in information systems and services. The registry allows discovery of suitable schemes for information or, potentially, use, by exposing rich metadata about them for navigation and retrieval. Terminology registries can hold vocabulary scheme level information only, or also include a vocabulary's terms, concepts and relationships. They could also provide services based on terminology (such as crosswalks, browsing, query expansion, disambiguation, automatic classification/indexing, reasoning). They can make their content available for human inspection and machine-to-machine access.

Many basic lists of vocabularies have been made freely available on the Web. An example of a simple terminology registry from the commercial world is the Factiva's Taxonomy Warehouse (http://www.taxonomywarehouse.com/). Various international efforts with some form of direct access to underlying vocabulary elements are currently underway. More complex options include the possibility of providing machine-to-machine web services as part of a terminology registry.

Our study analyses issues related to the potential delivery of a terminology registry as a shared infrastructure service within the JISC Information Environment (IE). The role of a terminology registry will be considered in relation to other components of the information landscape and relevant experience in other domains. The study aims to describe usage scenarios and use cases, investigate requirements and sustainability, study costs and benefits. Architectural issues will be explored, in particular the potential for co-ordination of registry efforts within the JISC IE and across domains.

We would greatly appreciate it if you could take the time to answer the questions below, addressing requirements and usage scenarios, by Tuesday, 29 April 2008. (Please feel free to pass this email on to someone else in your organisation who might be better placed to respond.) For any further queries, please feel free to contact us via email (dstudhope@glam.ac.uk | k.golub@ukoln.ac.uk) or telephone (Doug 01443 483 609 | Kora 01225 383 619).

Thank you!

With kind regards,

Koraljka Golub & Doug Tudhope

_____

Questions:

1) What should in your view a terminology registry comprise, which functionalities should it offer? Please briefly describe what you would see as the main requirements of a terminology registry.

2) Might you see yourself as a potential user of a terminology registry? Which of the two potential scenarios would be more useful to your needs, machine-to-machine or for human inspection? Please outline possible usage scenarios (use cases) and explain how they could fit into your work practice.

3) What do you see as major barriers and challenges to a terminology registry take-up and implementation?

4) Any general comments are welcome.

Services with KOS

Dear _____,

We are contacting key experts in areas potentially related to the JISC's Terminology Registry Scoping Study (http://www.ukoln.ac.uk/projects/trss/). As an initial part of the study, we are looking for feedback on requirements, usage scenarios, and comments generally.

A terminology registry describes, identifies and points to sets of vocabularies available for use in information systems and services. The registry allows discovery of suitable schemes for information or, potentially, use, by exposing rich metadata about them for navigation and retrieval. Terminology registries can hold vocabulary scheme level information only, or also include a vocabulary's terms, concepts and relationships. They could also provide services based on terminology (such as crosswalks, browsing, query expansion, disambiguation, automatic classification/indexing, reasoning). They can make their content available for human inspection and machine-to-machine access.

Many basic lists of vocabularies have been made freely available on the Web. An example of a simple terminology registry from the commercial world is the Factiva's Taxonomy Warehouse (http://www.taxonomywarehouse.com/). Various international efforts with some form of direct access to underlying vocabulary elements are currently underway. More complex options include the possibility of providing machine-to-machine web services as part of a terminology registry.

Our study analyses issues related to the potential delivery of a terminology registry as a shared infrastructure service within the JISC Information Environment (IE). The role of a terminology registry will be considered in relation to other components of the information landscape and relevant experience in other domains. The study aims to describe usage scenarios and use cases, investigate requirements and sustainability, study costs and benefits. Architectural issues will be explored, in particular the potential for co-ordination of registry efforts within the JISC IE and across domains.

We would greatly appreciate it if you could take the time to answer the questions below, addressing requirements and usage scenarios, by Tuesday, 29 April 2008. (Please feel free to pass this email on to someone else in your organisation who might be better placed to respond.) For any further queries, please feel free to contact us via email (dstudhope@glam.ac.uk | k.golub@ukoln.ac.uk) or telephone (Doug 01443 483 609 | Kora 01225 383 619).

Thank you!

With kind regards,

Koraljka Golub & Doug Tudhope

_____

Questions:

1) What should in your view a terminology registry comprise, which functionalities should it offer? Please briefly describe what you would see as the main requirements of a terminology registry.

2) Might you see yourself as a potential user of a terminology registry? Which of the two potential scenarios would be more useful to your needs, machine-to-machine or for human inspection? Please outline possible usage scenarios (use cases) and explain how they could fit into your work practice.

3) What do you see as major barriers and challenges to a terminology registry take-up and implementation?

4) In the service(s) you provide, please list any vocabularies used, or planned to be used in the future.

5) Any general comments are welcome.

Appendix 2. Interview invitation letter

Dear_____

We are contacting key experts in areas potentially related to the JISC's Terminology Registry Scoping Study (http://www.ukoln.ac.uk/projects/trss/). As an initial part of the study, we are looking for feedback on requirements for a terminology registry, usage scenarios, and comments generally.

In our view, a terminology registry describes, identifies and points to sets of vocabularies available for use in information systems and services. The registry allows discovery of suitable schemes for information or, potentially, use, by exposing rich metadata about them for navigation and retrieval. Terminology registries can hold vocabulary scheme level information only, or also include a vocabulary's terms, concepts and relationships. They could also provide services based on terminology (such as crosswalks, browsing, query expansion, disambiguation, automatic classification/indexing, reasoning). They can make their content available for human inspection and machine-to-machine access.

Our study analyses issues related to the potential delivery of a terminology registry as a shared infrastructure service within the JISC Information Environment (IE). The role of a terminology registry will be considered in relation to other components of the information landscape and relevant experience in other domains. The study aims to describe usage scenarios and use cases, investigate requirements and sustainability, study costs and benefits. Architectural issues will be explored, in particular the potential for co-ordination of registry efforts within the JISC IE and across domains.

We would greatly appreciate it if you could take about 30 min of your time for a conference call with us over the next few weeks? If you are willing to do this, as a first attempt would any time suit you on the following days: 24 April, 25 April and 01 May? If possible, please respond by the end of this week, Friday 18 if this would suit you.

We are very much looking forward to hearing from you.

Thank you!

With kind regards,

Koraljka Golub & Doug Tudhope

Appendix 3. List of people who provided input to the study

The following people provided input to the study, via a questionnaire (e-mail) or interview.

Related projects

	JISC		
1	IESR	Ann Apps	(interview)
2	HILT	Dennis Nicholson	(interview)
3	IEMSR	Emma Tonkin	(interview)
4	NAMES	Amanda Hill	(e-mail)
5	GeoXwalk + Edina	James Reid	(e-mail)
	International		
6	NSDL registry	Diane Hillman and Jon Phipps	(interview)
7	OCLC registry	Diane Vizine-Goetz and Andrew Houghton	(interview)
8	DART	Jane Hunter	(e-mail)
9	FAO registry	Margherita Sini	(e-mail)

Subject domains

	Cultural heritage		
10	Nick Poole		(interview)
11	Philip Carlisle		(e-mail)
	E-science		
12	Roy Lowry		(interview)
13	Sophia Ananiadou		(e-mail)
14	Simon Coles		(e-mail)
15	Brian Matthews		(e-mail)
16	Carole Goble		(e-mail)
17	Sean Bechhofer & Robert Stevens		(e-mail)
	E-learning		
18	Lorna Campbell		(interview)
	E-framework		
19	Paul Walk		(interview)

Services with terminologies

20	UK repositories	Mahendra Mahey	(interview)
21	EDL	Makx Dekkers	(e-mail)
22	Intute	Debra Hiom	(e-mail)
23	Carmen	Philip Lord	(e-mail)

Terminology developers

24	Joan Cobb (Getty)		(e-mail)

Terminology experts

25	Gail Hodge		(interview)
26	Traugott Koch		(interview)
27	Stella Dexter Clarke		(e-mail)
28	Marcia Zeng		(e-mail)

Appendix 4. Metadata with examples

1) CENDI

not separate metadata but descriptions that seem to include the following:

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
online availability

Example:
NAL Agricultural Thesaurus http://agclass.nal.usda.gov/agt/agt.shtml

The NAL Agricultural Thesaurus (NALT) is annually updated and the 2007 edition contains over 65,800 terms organized into 17 subject categories. NALT is searchable online and is available in several formats (PDF, ASCII text, XML, SKOS) for download from the web site. NALT has standard hierarchical, equivalence and associative relationships and provides scope notes and over 2,400 definitions of terms for clarity. Proposals for new terminology can be sent to thes@nal.usda.gov. Published by the National Agricultural Library, United States Department of Agriculture.

2) Ecoterm (Environmental Terminology and KOS)

Name (includes acronym)
Environmental Topics
Non-Environmental Topics
Technical Contact (Name, adress, e-mail)
Content Contact (Name, address, e-mail)

These are based on the NKOS and XMDR Content elements (Hodge et al. 2007)

3) Food and Agriculture Organization (FAO) of UN

FAO Knowledge Organization Systems	
Element name	**Explanation**
	name of the KOS
Domain	subject domain
Owner/ Creator	
Description	several sentences describing what the KOS covers
Language	list of languages covered
Type	type of KOS (e.g., taxonomy, thesaurus)
URL	
Model	URL to OWL represenation of FAO-produced KOS
Contact Email	

FAO Knowledge Organization Systems	
Vessel types and size	
Domain	Fisheries and Aquaculture
Owner/ Creator	Food and Agriculture Organization of the …
Description	This ontology organizes the information …
Type	Ontology
URL	http://www.fao.org/a....
Contact Email	FAO-AGRIS-....

4) Hodge et al. 2007 (10th OFMR)

Name (with acronyms)
Creator
Description
Subject Controlled
Keywords
Resource Identifier
Language
Resource Type
Rights
Publisher
Format
Contact Email

5) NKOS Registry 1998

Product Information *element required

- Product Name/Title *
- Variant Product Name/Title
- Type of Product *
- Product Description *
- Auxiliary Lists
- Author/Editor
- Current Version/Edition *
- Date of Current Version *
- Product Update Frequency *
- Available Format(s) and Size *
- Online Availability
- Notes
- URL for Examples

Scope and Usage

- Major Subject Coverage *
- Minor Subject Coverage
- Used by (user community and applications)

NKOS Characteristics

- Language(s) *
- Type of Terms (e.g. concept terms, geographic names, corporate names, etc.) *
- Description of Overall Structure *
- Source of New Terminology *
- Number of Preferred Terms or Nodes *
- Number of Non-preferred Terms
- Types of Relationships *
- Arrangement of Displays (e.g., alphabetical, hierarchical, graphical)
- Depth of Hierarchy (maximum number of levels)

Terms and Conditions

- Purchase Price by Format (or cost-free statement) *
- Subscription Price by Format
- Licensing Availability
- Restrictions (or no-restrictions statement) *

Vendor

- Vendor Name *
- Vendor Street/Post Office Box *
- Vendor City *
- Vendor State/Province *
- Vendor Country *
- Vendor Postal Code/ZIP Code *
- Vendor Voice Phone *
- Vendor TDD/TTY Phone
- Vendor Fax
- Vendor Email
- Vendor Logo URL
- Vendor Web Site URL
- Vendor Hours of Service and Timezone *
- Vendor Service Description *

Contact

- Contact Name
- Contact Voice Phone
- Contact Fax
- Contact Email
- Contact Web Site URL
- More Contact Information

Comments to Registry Editor

5) NKOS Registry 2001

KOS Title (R)
Alternative Title (O)
Creator (O)
KOS Subject (R)
Description (O)
Publisher (O)
Date (R)
KOS Type (R)
Format (R)
Identifier (O)
Language (R)
KOS Relation (R)
Rights (O)
Entity Type (R)
Entity Value (O)
Relationships (R)
Information Given (O)
Arrangement (R)
Application (O)
Minor Subject (O)

where R stands for Required, and O for Optional

Following Dublin Core, each element is defined using a set of ten attributes from the ISO/IEC 11179 (ISO 11179 2007) standard for the description of data elements:

Name
Identifier
Version
Registration Authority
Language
Definition
Obligation
Datatype
Maximum Occurrence
Comment

6) National Science Digital Library Registry

NSDL registry

Element name	Explanation
Owner	
Name	
URL	
Note	description of content
Community	who it is aimed at
Status	e.g., published
Language	
URI Base Domain	
URI Token	
URI	
Users Name, Administrator, Maintainer, Registrar	user's name and whether she is an administrator, maintainer, or registrar

NSDL registry example

Owner	GEM Exchange		
Name	21st Century Skills		
URL			
Note	Partnership of 21st Century Skills vocabulary of skills		
Community	Education		
Status	Published		
Language	English		
URI			
Base Domain	http://purl.org/ASN/scheme		
Token	P21		
URI	http://purl.org/ASN/scheme/P21		
Users			
Name	Administrator	Maintainer	Registrar
sas1	Tick	Tick	Tick

7) ISO 11179 (Information Technology – Metadata registries (MDR))

From ISO 11179-2: Classification

Attribute	**Occurrences**
Designation - name	One per Terminological Entry Language Section
Designation - preferred designation	Zero or one per Terminological Entry Language Section
Designation - language identifier	One per Language Section in each Terminological Entry
Definition - definition text	One per Terminological Entry Language Section
Definition - preferred definition	Zero or one per Terminological Entry Language Section
Definition - source reference	Zero or one per Terminological Entry Language Section
Definition - language identifier	One per Language Section in each Terminological Entry
Context - administration record	One per context
Context - description	One per context
Context - description language identifier	Zero or one per context
Classification Scheme - type name	One per classification scheme
Classification Scheme Item - value	One per classification scheme item
Classification Scheme Item - type name	Zero or one per classification scheme item
Classification Scheme Item Relationship - type description	One per classification scheme item relationship type description
Administration Record - item identifier	One per classification scheme
Administration Record - registration status	One per classification scheme
Administration Record - administrative status	One per classification scheme
Administration Record - creation date	One per classification
Administration Record - last change date	Zero or one per classification
Administration Record - effective date	Zero or one per classification
Administration Record - until date	Zero or one per classification
Administration Record - change description	Zero or one per classification
Administration Record - administrative note	Zero or one per classification
Administration Record - explanatory comment	Zero or one per classification
Administration Record - unresolved issue	Zero or one per classification
Administration Record - origin	Zero or one per classification
Reference Document - identifier	One per reference document
Reference Document - type description	Zero or one per reference document
Reference Document - language identifier	Zero or more per reference document
Reference Document - title	Zero or one per reference document
Reference Document - organization name	One or more per reference document
Reference Document - organization mail address	Zero or one per reference document
Submission - organization name	One per classification scheme
Submission - organization mail address	Zero or one per classification scheme
Submission - contact	One per classification scheme
Stewardship - organization name	One per classification scheme
Stewardship - organization mail address	Zero or one per classification scheme
Stewardship - contact	One per classification scheme
Registration Authority - organization name	One per classification scheme
Registration Authority - organization mail address	Zero or one per classification scheme
Registration Authority - registration authority identifier	One per classification scheme
Registration Authority - documentation language identifier	One or more per classification scheme
Registrar - identifier	One or more per classification scheme
Registrar - contact	One or more per classification scheme

8) OCLC Terminology Services

OCLC Terminology Services

Element name	Explanation
dc.contributor	An entity responsible for making contributions to a controlled vocabulary
dc.description	A description of a controlled vocabulary
dc.identifier	An unambiguous reference to a vocabulary metadata record
dc.language	A language of the controlled vocabulary
dc.rights	Information about rights held in and over a controlled vocabulary
dc.subject	A subject focus of a controlled vocabulary
dc.title	A name given to a controlled vocabulary
oclcts.marcTags	MARC tags in a controlled vocabulary metadata record
oclcts.vocabularyId	A code assigned to a controlled vocabulary
cql.resultSetId	An index defined by the CQL context, required by the SRU protocol
cql.serverChoice	The default index, defined by the CQL context
At the project website, KOS are described using the following elements:	
Name	same as dc.title??
Description	same as dc.description
Date	date when added to the registry
Identifier	same as oclcts.vocabularyId
Links	include About, SRU Interface, Examples, MARC Statistics

OCLC Terminology Services Example

Name	Form and genre headings for fiction and drama
Description	Form and genre terms from the Guidelines On Subject Access To Individual Works Of Fiction, Drama, Etc., 2nd ed.
Date	2008-03
Identifier	gsafd
Links	About, SRU Interface, Examples, MARC Statistics

MARC example:
http://tspilot.oclc.org/meta/?query=oclcts.vocabularyId+exact+%22gsafd%22&version=1.1&operation=searchRetrieve&recordSchema=info%3Asrw%2Fschema%2F1%2Fmarcxml-v1.1&maximumRecords=10&startRecord=1&resultSetTTL=300&recordPacking=xml&recordXPath=&sortKeys=

9) SPECTRUM Terminology Bank

SPECTRUM	
Element name	**Explanation**
Title:	
Resource Type:	e.g., thesaurus
Author(s):	
Publisher(s):	
Creation Date:	
Description:	
URL:	
SPECTRUM Unit of information:	

SPECTRUM example	
Title:	Pitt Rivers Museum - University of Oxford Group Thesaurus
Resource Type:	Simple Wordlist
Author(s):	Pitt Rivers Museum documentation staff
Publisher(s):	Pitt Rivers Museum
Creation Date:	2001
Description:	Keyword list for Group.
URL:	http://www.mda.org.uk/spectrum-terminology/pitt-rivers/group
SPECTRUM Unit of information:	People name

10) Taxonomy Warehouse

Taxonomy Warehouse

Element name	Explanation
Name	KOS name
Publisher	publisher name
Type	type of KOS (e.g., taxonomy, thesaurus)
Categories	subjects covered
Description	several sentences describing what the KOS covers
Total Terms	number of total terms
Top Terms	number of top hierarchical terms
Preferred Terms	number of preferred terms
Non-Preferred Terms	number of non-preferred terms
Relationships	number of terms with relationships to other terms
Levels	number of hierarhical levels
Notation Scheme	yes/no for whether there is a notation system
Notation Description	e.g. for Eurovoc: numeric, two-digit numbers identify 21 fields (subject areas), four-digit numbers indicate microthesaurus.
Relationship Types	types of relationships between terms (e.g., associative, hierarchical)
Notes Fields	types of notes field (e.g., Other, Scope)
Multilingual	yes/no for whether it is multilingual
Languages	list of languages covered
Additional Information	
Revision Cycle	how frequent the KOS is updated
Last Updated	last update date
Formats	formats in which it can be available
Informational URL	information at publisher's web site
Online/Download URL	
When ordering,further publisher info:	
Address	
Phone	
Fax	
URL	

Taxonomy Warehouse Example

Element name	Value
Name	KOS name
Publisher	publisher name
Type	type of KOS (e.g., taxonomy, thesaurus)
Categories	subjects covered
Description	several sentences describing what the KOS covers
Total Terms	number of total terms
Top Terms	number of top hierarchical terms
Preferred Terms	number of preferred terms
Non-Preferred Terms	number of non-preferred terms
Relationships	number of terms with relationships to other terms
Levels	number of hierarhical levels
Notation Scheme	yes/no for whether there is a notation system
Notation Description	e.g. for Eurovoc: numeric, two-digit numbers identify 21 fields (subject areas), four-digit numbers indicate microthesaurus.
Relationship Types	types of relationships between terms (e.g., associative, hierarchical)
Notes Fields	types of notes field (e.g., Other, Scope)
Multilingual	yes/no for whether it is multilingual
Languages	list of languages covered
Additional Information	
Top Terms	
Preferred Terms	28000
Non-Preferred Terms	10900
Relationships	
Levels	
Notation Scheme	No
Notation Description	
Relationship Types	Associative, Equivalency, Hierarchical
Notes Fields	Scope
Multilingual	Yes
Languages	Arabic, Chinese, Czech, English, French, Japanese […]
Additional Information	Available in SKOS, MySql, Postgres, MsAccess, TagText and ISO2709 formats
Revision Cycle	updated quarterly
Last Updated	200604
Formats	Public Website
Informational URL	http://www.fao.org/aims/ag_intro.htm
Online/Download URL	http://www.fao.org/aims/ag_download.htm
Publisher information	
Address	Viale delle Terme di Caracalla, 00100, Rome, Italy
Phone	+39 06 5705 1
Fax	+39 06 5705 3152
URL	http://www.fao.org/

11) Vocman (Becta Vocabulary Bank)

Becta Vocabulary Bank

Element name	Explanation
Authority	Ogranisation in charge of creating and maintaing the KOS
Version	
Description	

Becta Vocabulary Bank Example

ACLearn

Authority: SkillsWeb
Version: 2
Description: The Adult Community Learning vocabulary