

SIMPLE KNOWLEDGE ORGANISATION AND THE SEMANTIC WEB
Notes on the Bliss Classification Association Annual Lecture 2005

Alistair Miles
CCLRC Rutherford Appleton Laboratory
a.j.miles@rl.ac.uk
+44 1235 445440

This article introduces the Simple Knowledge Organisation System (SKOS), a machine-readable representation format for controlled structured vocabularies and for subject indexes.

Scope, Design and Assumptions

The scope of SKOS (1, 2) includes those types of controlled structured vocabulary that are intended for use within retrieval applications. This includes thesauri broadly conforming to the ISO 2788:1986 guidelines, and both synthetic and enumerative classifications schemes, although some further work is required to establish the full requirements deriving from the application of synthetic classification schemes such as the BC2.

The model underlying the design of SKOS assumes that the basic purpose of a controlled structured vocabulary is to establish a set of distinct *meanings*, and to provide a way of referring to those meanings that is unambiguous at least within the scope of the vocabulary. For example, in a classification scheme such as the BC2 (6, 7), each class establishes a distinct meaning, and the notation for that class provides a way of referring to its meaning unambiguously within the scope of the BC2 schedule. In a thesaurus broadly conforming to ISO 2788:1986, each descriptor establishes a distinct meaning, and the lexical value of the descriptor itself is used to refer to that meaning unambiguously within the scope of the thesaurus.

Both a classification scheme such as the BC2 and a thesaurus broadly conforming to ISO 2788:1986 can be understood as consisting of a set of *conceptual units*, that is, a set of units each establishing a distinct meaning, and each providing a means of reference that is unambiguous at least within the scope of that classification scheme or thesaurus.

A SKOS representation of a controlled structured vocabulary begins with a description of its underlying conceptual units. Each conceptual unit may be allocated a Uniform Resource Identifier (URI), and it is recommended that this URI be used as the primary means of reference within computer systems. Note that, although URIs are commonly used to refer to “web sites”, it is a basic principal of the Semantic Web (3, 4, 5) that URIs may also be used to refer to anything from a make of washing machine to a country to an abstract concept such as multi-faith society. Using URIs as the primary means of reference within computer systems enables unambiguous reference in an open-ended system, which is vital where data from multiple sources is being combined and merged.

Each conceptual unit is then associated with *labels* and *documentation* that serve to explain the intended meaning, and provide a way of generating visual representations on screen or in print, and/or representations for other modalities (e.g. aural). The basic

type of label is a *lexical label*, i.e. a string of Unicode characters. For example, the character string “Animal physiology” is the preferred lexical label for the BC2 class with the notation “GBB”. Each lexical label may be associated with a particular natural language (such as British English), which provides a basis for *multilingual labelling* of conceptual units. Each lexical label is either *preferred*, *alternative* (i.e. non-preferred), or *hidden*. A conceptual unit must of course only have one preferred lexical label per language. Hidden lexical labels are usually not rendered when generating a visual or aural representation, but are made available to computer search applications – typically this feature is used for commonly mis-spelled or mis-typed words.

SKOS also currently provides support for *symbolic labels* (i.e. a label that is an image, rather than a sequence of characters), and in future may provide support for other types of label such as *speech labels* (i.e. labels that can be read by a speech synthesis engine) and *math labels* (i.e. labels that consist of complex mathematical notation).

Different types of *documentation* (also called *annotation*) may be associated with a conceptual unit. Currently supported are *scope notes*, *definitions*, *history notes*, *editorial notes*, *change notes* and *examples*. Additional types of documentation may be included via the extensibility mechanism that is built in to the design of SKOS. Documentation may of course be provided in multiple languages.

The basic structure of a controlled structured vocabulary may be represented via *semantic relationships* between conceptual units. SKOS offers built in support for three types of semantic relationship – *broader*, *narrower* and *related* – and custom relationship types may be defined via the extensibility mechanism. The meaning of the built-in semantic relationship types follows the guidelines given by BS 8723 part 2. Some further work is needed to explore whether the hierarchical relationships of some classification schemes fall within the broader/narrower paradigm.

SKOS also enables a *subject index* over a collection of items to be represented in a machine-readable format. A subject index essentially consists of a set of links between items in a collection and conceptual units in a controlled vocabulary. Two types of indexing link are built in to SKOS. The first is the basic *subject* link – this link establishes the given conceptual unit as a subject of a work, where the work may be about several subjects. The second is the *primary subject* link – this link establishes a given conceptual unit as the primary or principal subject of a work. These two types of link allow both the typical subject index constructed using a thesaurus (where the subjects of a work are described) and the typical subject classification constructed using a classification scheme (where works are classified by their primary subject) to be represented and differentiated. Note that it is quite reasonable to allow these two types of link to coexist within computer systems, where there is no need to locate items within a single physical region such as a shelf in a library.

SKOS includes some support for representing meaningful groupings of conceptual units, known in BS 8723 as “arrays” with “node labels”, although there are some issues with this feature that remain to be resolved.

Currently there is no built in support for the synthesis of conceptual units to represent compound meanings, although proposals are under discussion at this time, and this is seen as an important feature.

Application, Context and Status

SKOS is intended to support the interoperation of three principal software components involved in the management of digital libraries: (a) a tool to manage the development of a controlled structured vocabulary; (b) a tool to manage the development of a subject index or classification for a particular collection of items; (c) a tool to enable retrieval of items from a collection using a vocabulary and an index. By representing both a controlled structured vocabulary and a subject index or classification in the SKOS format, and by publishing these *data* in the World Wide Web, the data may be shared and re-used across application boundaries.

Typically, several collections may use the same vocabulary to index or classify their items (for example several libraries use the BC2 schedule). Traditionally, retrieval services are provided locally per-collection, however there is a growing need to provide retrieval services across two or more collections, with a single virtual point of access. This allows a user to pose a single question, and have matching results returned from multiple sources, rather than having to individually query each of several collections. A more demanding situation is met were several collections use different vocabularies for indexing or classification, and a user needs to be able to search across those collections in a coherent way.

These are specific examples of a more general situation, where data are distributed across multiple locations, and where computer applications need to be able to *harvest* and *aggregate* those distributed data, in order to operate on the “bigger picture”. Another example, completely unrelated to the field of library science, is provided by a company where information about employees, project, products, and finance are managed in different databases, but where managers need to be able to pose questions such as, “which employees have worked on successful product development projects?” A day-to-day example is provided by the need to manage and to share personal data such as contacts, email, calendar, and bank details, across computer applications and also between friends, family, colleagues, or organisations, in a seamless and secure way.

The World Wide Web Consortium (W3C) has promoted the development of an underlying technological framework to enable solutions to this type of scenario, within the scope of its *Semantic Web Activity* (4, 5). To quote Tim Berners-Lee (3):

“The goal of the Semantic Web initiative is to create a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by people.”

To allow thesauri, classification schemes and subject indexes to be shared and re-used by many organisations and individuals, SKOS has been built using Semantic Web standards. SKOS is an application of a lower-level Semantic Web technology called the Resource Description Framework (RDF), which provides the logical “glue” needed to join data from different sources together. SKOS and RDF make extensive use of Uniform Resource Identifiers (URIs) to provide an unambiguous way of referring to abstract or physical things within computer systems. For example, if URIs were allocated to the classes of the BC2, any library system could use RDF to share its data with any other library, and their data could be meaningfully integrated and cross-references, without fear of ambiguity.

SKOS has been developed within the context of the World Wide Web Consortium's (W3C's) Semantic Web Activity (4). The initial development work was carried out by the Semantic Web Advanced Development for Europe project (SWAD-Europe), funded by the EU-IST 5th framework programme. Subsequently development continued within the W3C's Semantic Web Best Practices and Deployment Working Group (SWBPD-WG), which has published the SKOS Core Guide (1) and SKOS Core Vocabulary Specification as W3C Working Drafts.

At the time of writing, SKOS is a work in progress. The SKOS Core Guide and the SKOS Core Vocabulary Specification provide the normative documentation for the use of SKOS, and are published as W3C Working Drafts. The current intention is to develop SKOS towards a stable publication by the end of 2007, and therefore any comments and suggestions especially with respect to the requirements for the machine-readable representation of synthetic classification schemes would be warmly welcomed. All discussion relating to the development of SKOS is conducted via the public mailing list *public-esw-thes@w3.org* – comments and suggestions should be sent to this address, and all are welcome to subscribe and to participate in the discussions.

Acknowledgements

The members of the *public-esw-thes@w3.org* mailing list are gratefully acknowledged.

References

1. SKOS Core Guide.
<http://www.w3.org/TR/swbp-skos-core-guide>
2. SKOS Web Site.
<http://www.w3.org/2004/02/skos>
3. The Semantic Web – an Interview with Tim Berners-Lee.
<http://www.consortiuminfo.org/bulletins/semanticweb.php>
4. W3C Semantic Web Activity.
<http://www.w3.org/2001/sw/>
5. The Semantic Web, Scientific American.
<http://www.scientificamerican.com/linktous.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>
6. Bliss Classification Association.
<http://www.sid.cam.ac.uk/bca/bcahome.htm>
7. How to use the Bliss Bibliographic Classification system.
<http://www.sid.cam.ac.uk/indepth/lib/bc2guide.html>