SKOS: a model for metadata representation and interoperability – Dutch Cultural Heritage Institution thesaurus conversion use case

Véronique Malaisé & the CHOICE team

Semantic interoperability between descriptions of digital documents from different collections can be achieved by different methods, including *a posteriori* converting or mapping between the vocabularies and models used for indexing these different collections, *a priori* complying to the same model or *a posteriori* converting or mapping the different data to a common generic (and preferably standard) model. Our research project, CHOICE[1], is focusing on applications and issues related to cataloging practices, in collaboration with the Netherlands Institute for Sound and Vision[2], in which indexing and retrieval is done based on a thesaurus called GTAA (a Dutch acronym for Common Thesaurus for Audiovisual Archives).

For interoperability purpose within other projects of the CATCH program[3] and for using this thesaurus in Semantic Web applications, we chose the third approach and converted it to the SKOS model. This paper describes the SKOS model in comparison with the ISO-standard way of representing thesaurus data, based on the Website of the W3C Semantic Web Deployment Working Group[4] and [1], and the advantages and drawbacks to comply with this model. We then detail the thesaurus that we have converted, the GTAA, describing its standard and specific features, and the conversion problems that we faced. A presentation of the conversion method that we followed and more details about this experiment can be found in [2]. We conclude on the advantages of such a conversion in terms of language integration possibility and of software application, mentioning the example of the SKOS Web Browser developed in our project.[5]

SKOS: the model, advantages and drawbacks

SKOS, Simple Knowledge Organisation System "*[…] provides a standard way to represent knowledge organisation systems using the Resource Description Framework (RDF). Encoding this information in RDF allows it to be passed between computer applications in an interoperable way*[6].". The SKOS Specifications are currently published as W3C Working Drafts, which means they are work in progress but on the way to become a W3C recommendation. In this respect, SKOS is interesting as a model for interoperability on the Web. The main advantage is that it proposes an RDF definition of a thesaurus's main construct, making them machine readable and usable in Semantic Web applications like vocabulary integration [3] or thesaurus browsers [4].

The main difference between a standard thesaurus, as described in the ISO norms, and the SKOS model is that the first is term-centered, whereas the latter is concept oriented, following RDF and ontologies' usual modeling features (Concepts or Classes and Properties or Relationships). A thesaurus distinguishes between preferred terms, meant to be used when indexing documents from a collection, and non preferred terms, which are considered as synonyms of the previous ones, but should not be used when indexing. These two entities become strings attached to a concept in a SKOS representation: a PrefLabel and AltLabel, for preferred label and alternative label. From the five core relationships in thesauri, namely broader term, narrower term (both building the thesaurus' hierarchical structure), related term (called associative relationship), use and use for (sometimes referred to as linguistic relationships), only

---

[1] http://www.nwo.nl/CATCH/CHOICE, CHOICE is one of the 10 projects of the CATCH program (see http://www.nwo.nl/CATCH), focusing on accessing, describing and integrating resources from Dutch Cultural Heritage Institutions.
[2] http://www.beeldengeluid.nl
[3] See footnote 1.
[4] http://www.w3.org/2004/02/skos/
[5] http://ems01.mpi.nl/CHOICE/, see the Demonstration page and [4].
[6] http://www.w3.org/2004/02/skos/

the first three apply to SKOS representation. In this case, they stand between two concepts, which have each preferred and non preferred labels attached to the conceptual root. This root is referred to by a unique identifier.

The picture below, extract from the UKAT thesaurus[7] and borrowed from the SKOS Specifications, features a thesaurus' classic representation and the corresponding SKOS graph:

Term: Economic cooperation

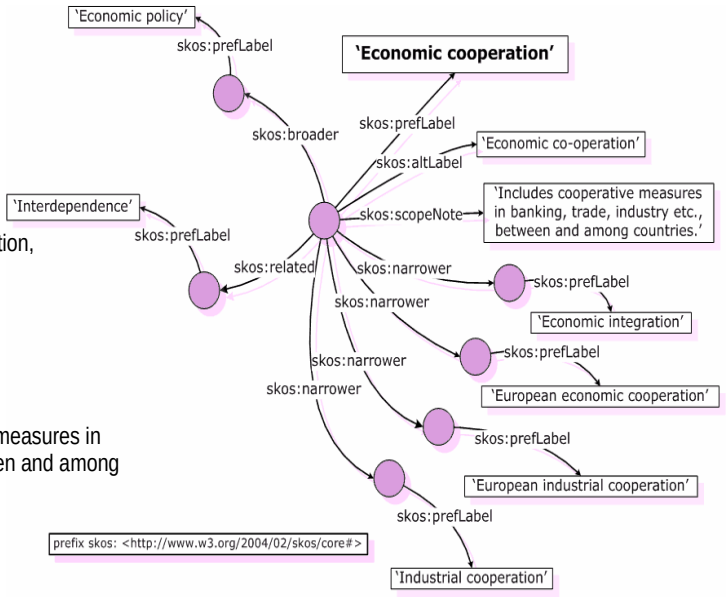
Used For: Economic co-operation

Broader terms: Economic policy

Narrower terms: Economic integration, European economic cooperation, European industrial cooperation, Industrial cooperation

Related terms: Interdependence

Scope Note: Includes cooperative measures in banking, trade, industry etc., between and among countries.



Grouping all the related information under one encapsulating concept makes the thesaurus updates easier: the indexing is done with the concept identifier, and the labels attached to it can easily be changed over the time, may it be to solve polysemy problems or because of terminological evolution. This concept-oriented modeling also helps getting the different versions of the annotations "backwards compatible". But the choice to restrict one thesaurus term to one string (or concept label) makes it impossible to attach additional information to this label, either preferred or not preferred. One such kind of information possibly present in the thesaurus is history notes, which can only be linked to the concept as a whole in SKOS, and not to the specific expressions as in the original thesaurus.

SKOS has been created to answer basic or more sophisticated thesaurus and controlled vocabularies modeling needs, but in some practical cases, it is not sufficient. For example, no standard mechanism is yet defined to express compound concepts or qualifiers. For interoperability purposes, it is suggested to extend the model for specific needs, instead of building local competing models. We will present the thesaurus used at Sound and Vision, the modeling issues that we faced when trying to convert it to SKOS, and in which respect we had to extend it.

The GTAA thesaurus

The GTAA is a faceted thesaurus: its terms are divided into 6 non overlapping groups. These groups are:

- Subject: terms from this facet are used to describe the main topic the TV program is about, or which is mentioned in the program;
- Person: to describe the main people the program is about, or people appearing on the screen;
- Location: to describe the main location the program is about, or the place(s) where it was shot;
- Name: to describe the name of companies, groups, bands etc that the program is about or who appear on screen;
- Genre: to qualify the genre of the program;

[7] http://www.ukat.org.uk/

- Maker: terms indicating the Maker(s) ' name(s).

Subjects and Genres are organised in broader term/narrower term hierarchies, Subjects, Genres and People have a use/use for relationship, Subjects have related terms (associative relationships) and the 6 facets can have scope notes. Besides these standard features, that have straightforward counterparts in SKOS, GTAA has also a number of more specific features and a set of *ad hoc* ones. We list the features of these two categories in the next section, with their SKOS counterpart.

SKOS conversion

The GTAA contains two standard features for which there is no conversion proposal in SKOS Core: facets and qualifiers. The facets can be described at the level of the Metamodel of the thesaurus, according to SKOS[8], but there is no specific construct or property that enables to link a particular instance of a thesaurus concept to one specific facet. Therefore we chose to extend the model and created 6 concepts as sub-classes of the generic skos:concept. Any instance is an instance of a sub-concept of skos:concept, being compatible with the model and keeping the semantics of the original thesaurus. The qualifier's problem is more complex, different modeling possibilities are still under discussion, but they mostly imply the fact that the qualifiers themselves (additional information attached to a term to disambiguate between different possibilities, like in the case of Amsterdam-Netherlands and Amsterdam-US) are entities of the thesaurus. This is not the case for the GTAA, and some qualifiers, like the role of a person in a TV program, are even added at indexing time. We did not choose a definitive modeling solution yet, and are waiting for concrete applications of the vocabulary (in semi-automatic indexing for example) to select the most appropriate one.

The GTAA also contains specific features not described in the ISO norms: "Categories" and a relationship between terms called "linked term". Our first concern was how to interpret the Category relationship: either it is meant to disambiguate different aspects of a term (as a qualifier would do, for example in "Church-institution" *vs* "Church-building"), or it is a way of grouping terms sharing a specific aspect (as with "Milk by animal" and "Cow-milk", "Buffalo-milk", etc.). In the second case, "Milk by animal" is called a node label: it is a way of grouping terms, but the concept itself should not be used for indexing. These node labels are usually part of the term hierarchy. The experts indicated that this option was the intended usage of Categories: to provide a grouping of terms under a label that is not used in the indexing process. Nevertheless, they are meant to provide an alternative grouping of the GTAA terms, and thus are not part of the broader term/narrower term hierarchy. Although we mapped the Categories to the existing SKOS construct for these node labels, namely the skos:Collection, this modeling remains a non standard feature that cannot be processed by SKOS generic softwares. The Categories have explicit identifiers, from which we could infer their hierarchy (01 stands for Philosophy, and 01.01 is one of its subdivisions, for instance).
The linked term relationship connects related terms from different facets, like the name of a Queen with the Subject *Queens* and the country that she rules in the Location facet. These relationships were not instantiated in the original thesaurus, we added the links automatically using Natural Language Processing, and modeled them as sub-properties of the generic skos:related. Thus, we keep the compatibility with SKOS.

Summary: list of the GTAA features with their SKOS counterparts

GTAA unique items:

- Categories in the Subject facet: correspond to node labels and are modeled as skos:Collections, the terms belonging to them being skos:members;
- Linked term relationship: a sub-property of skos:related.

Standard thesaurus features present in GTAA but non addressed in SKOS-Core

- Facets: 6 sub-concepts of skos:concept;

[8] http://www.w3.org/TR/2005/WD-swbp-thesaurus-pubguide-20050517/#secExpressingMetadata

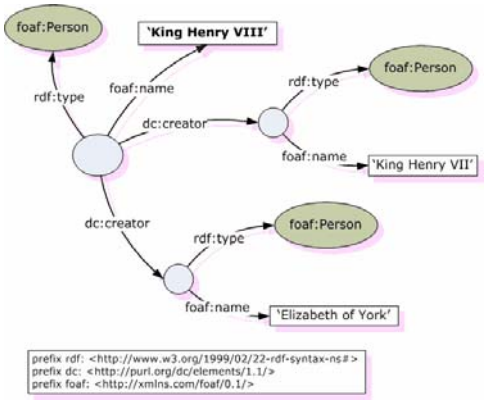
- Qualifiers: we will choose the modeling option that will suit the best our needs in terms of thesaurus usage, still work in progress.

GTAA standard items addressed in SKOS:

- Broader Term, Narrower Term, Related Term: skos:boader, skos:narrower, skos:related;
- Use, Use for: skos:prefLabel, skos:altLabel

Benefit of the conversion

The SKOS model being defined in RDF, turning a thesaurus into a SKOS compliant representation enables one to integrate RDF vocabularies (or other vocabularies also defined in RDF) in the document's description. For example, in the graph below, the thesaurus concept is associated with elements of information defined in Dublin Core[9] and in FOAF[10]:



Complying to SKOS also enables to process the data with generic tools like WebBrowsers (see the project's Demonstration section, for an example of such a Browser displaying the GTAA thesaurus). The page http://esw.w3.org/topic/SkosDev/ToolShed references tools based on or related to SKOS, as a wiki page where authors can freely add their work.

Acknowledgement

This work was partly supported by NWO's CHOICE projects. The authors wish to thank Mark van Assem for his support in the conversion of GTAA and the sharing of his knowledge on SKOS, and also our colleagues at the Netherlands Institute of Sound and Vision for their support and for providing a kind a stimulating working environment.

Bibliography

[1] *SKOS Core Guide*, 2nd W3C Public Working Draft 2 November 2005. Alistair Miles and Dan Brickley eds.
[2] Mark van Assem, Veronique Malaisé, Alistair Miles and Guus Schreiber.(2006). *A method to convert thesauri to SKOS*. In Proc. Third European Semantic Web Conference (ESWC'06), Budvar, Montenegro, June 2006.
[3] Marjolein van Gendt, Antoine Isaac, Lourens van der Meij and Stefan Schlobach. *Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study*. In: Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Julio Gonzalo, Constantino Thanos, M. Felisa Verdejo and Rafael C. Carrasco (eds.), Springer Verlag, LNCS vol. 4172, pp. 426-437, Alicante, Spain, September 17-22, 2006.
[4] Hennie Brugman, Véronique Malaisé and Luit Gazendam. *A Web Based General Thesaurus Browser to Support Indexing of Television and Radio Programs*. In: Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, May 24-26, 2006.

[9] http://dublincore.org/
[10] http://www.foaf-project.org/