Enhancing information services using machine to machine terminology services

Gordon Dunsire

Presented to the IFLA 2009 Satellite Conference “Looking at the past and preparing for the future”
20-21 Aug 2009, Florence, Italy

Abstract

This paper describes the basic concepts of terminology services and their role in information retrieval interfaces. Terminology services are consumed by other software applications using machine-to-machine protocols, rather than directly by end-users. An example of a terminology service is the pilot developed by the High Level Thesaurus (HILT) project which has successfully demonstrated its potential for enhancing subject retrieval in operational services. Examples of enhancements in three such services are given. The paper discusses the future development of terminology services in relation to the Semantic Web.

Terminology services

A terminology server is defined in Wikipedia as “... software providing a range of terminology-related software services through an Applications Programming Interface to its client applications.”[i] The services are not intended for end-users. Instead, they are to be used by computer programmers to improve client applications; that is, specific end-user services such as subject-based information retrieval interfaces. A client application will typically submit data to the terminology server along with a request for them to be processed in a specified way and the results returned to the application. The application may then further process the results before displaying them to, or otherwise interacting with, an end-user.

The application software is run on the client computer, which is not the same machine as the terminology server computer. The interaction between the two sets of hardware and software is known as machine-to-machine (m2m) processing.

Terminology services have been defined as “Web services involving various types of knowledge organization resources [vocabularies], including authority files, subject heading systems, thesauri, Web taxonomies, and classification schemes … Web services are modular, Web-based, machine-to-machine applications that can be combined in various ways.”[1] An example service is given as mapping from a

---

[i] From the Wikipedia definition of 21 November 2010, available at: http://en.wikipedia.org/wiki/Terminology_Server

term in one vocabulary to one or more terms in another vocabulary. The OCLC Terminology Services project[2] has developed a set of simple services involving various English subject heading systems including Library of Congress Subject Headings (LCSH)[3], Medical Subject Headings (MeSH)[4], and Thesaurus for Graphic Materials (TGM)[5], although it does not currently include any mappings between these vocabularies. The services accept a client term (or its identifier) and return data about matching terms in a vocabulary specified by the client. Related terms from the vocabulary are included in the process. The client also specifies the format of the returned data, chosen to suit the needs of the client software; one of the available formats is Simple Knowledge Organization System (SKOS)[6], a component of the Semantic Web.

The previously cited Wikipedia article suggests several categories of m2m terminology service which can be expressed in terms of application functions as:

- Matching user-defined text with lexical resources, including dictionaries, authority files, and thesauri.
- Translations from one language to another, using controlled vocabularies and semantic mappings.
- Semantic relationships within specific vocabularies used in Knowledge organization systems (KOS).
- Semantic relationships between specific vocabularies using ontology mappings.

These functions can be used in client applications to improve subject information retrieval interfaces for end-users. Work with the Scottish Collections Network (SCONE)[7] and CAIRNS[8] has suggested examples of enhancements that would benefit users. One is spell-checking user input, to trap typing errors or match spelling variants. This might be done transparently, or with feedback to the user as a "Did you mean ... ?" message. Another example is clarifying a user's search term when it is ambiguous relative to the one or more KOS involved: does the user intend "tree" to refer to forest or family? This process has been referred to as "disambiguation"; it perhaps comes as no surprise to see that Wikipedia has to disambiguate it with the entry "Disambiguation (disambiguation)"[ii] although the default definition of "Word sense disambiguation" is the basis of its usage in relation to KOS. A further example is switching an uncontrolled user term to a controlled vocabulary term; as before, this may be achieved automatically, without reference to the user, or with intervention by means of a "Use: ... " message display.

An important enhancement for union catalogues such as CAIRNS is the ability to match a user-supplied subject term to the equivalent term in each of the different vocabularies used for subject access in the different library catalogues in scope. This helps control the precision of the subsequent "one-stop" search across multiple heterogeneous subject headings.

[ii] From the Wikipedia definition of 21 November 2010, available at: http://en.wikipedia.org/wiki/Disambiguation_%28disambiguation%29

HILT: High-Level Thesaurus project

The High-Level Thesaurus(HILT) project[9] started in 2000; its fourth phase was completed in May 2009. The project was funded by the UK's Joint Information Systems Committee and supported by OCLC. Its overall scope was to provide subject interoperability in a multi-scheme environment via inter-scheme mapping, with an additional goal of identifying a generic approach that could be developed through distributed collaborative action.

The main objectives of the fourth phase were to research and develop pilot solutions for problems in cross-searching multi-subject schemes. A terminologies server was implemented using the Dewey Decimal Classification (DDC)[10] as a "switching language" between different Anglophone subject schemes and other vocabularies, including the DDC captions and relative index, Art and Architecture Thesaurus[11], UNESCO thesaurus[12], LCSH, MeSH, and several others. Most of the mappings are partial, created for test purposes. Some non-English terms are also mapped for similar purposes. Several m2m protocols are used by the server; in particular, its output is made available in SKOS format.

The project also developed pilot embedding of some the terminology services in the user interfaces of several operational information services. These were SCONE, Intute[13] and The Depot[14].

HILT case 1: SCONE

SCONE is a service which uses metadata for collections located in Scotland, from all heritage domains such as archives, libraries, and museums. The interface allows users to identify and locate Scottish collections, and access finding-aids such as catalogues which describe the items held within them.

Collections with a specific subject focus are classified with DDC and assigned LCSH entries to allow subject retrieval; multiple DDC notations as well as LCSH entries are used if necessary. The collection-level descriptions include metadata about the subject scheme used by a collection's finding-aids. An early experiment in the use of HILT showed that it was possible to direct different, but semantically-equivalent, terms from MeSH and LCSH to corresponding CAIRNS catalogues. The result was an improvement in recall, rather than precision, because in all cases both vocabularies were consolidated into a single search index within target catalogues. That is, the local subject index combined MeSH and LCSH terms in a single list. The experiment was not investigated further because these were the only vocabularies used by CAIRNS catalogues and available from the HILT server.

One enhancement developed for SCONE as a pilot during the HILT project accepts a subject term input by the user, and then displays the DDC caption hierarchies which match the term. The match is primary if the term is present in the caption, or secondary if the term is found in another vocabulary mapped to DDC. Figure 1 is a partial screen-shot where the user has entered the term "teeth", and is presented with a set of captions giving the different hierarchical contexts of the term in the DDC. Note that the last two captions displayed do not contain the user's term, so they are secondary matches. Note also that two distinct high-level contexts are given for the term, so this example of disambiguation is significant.

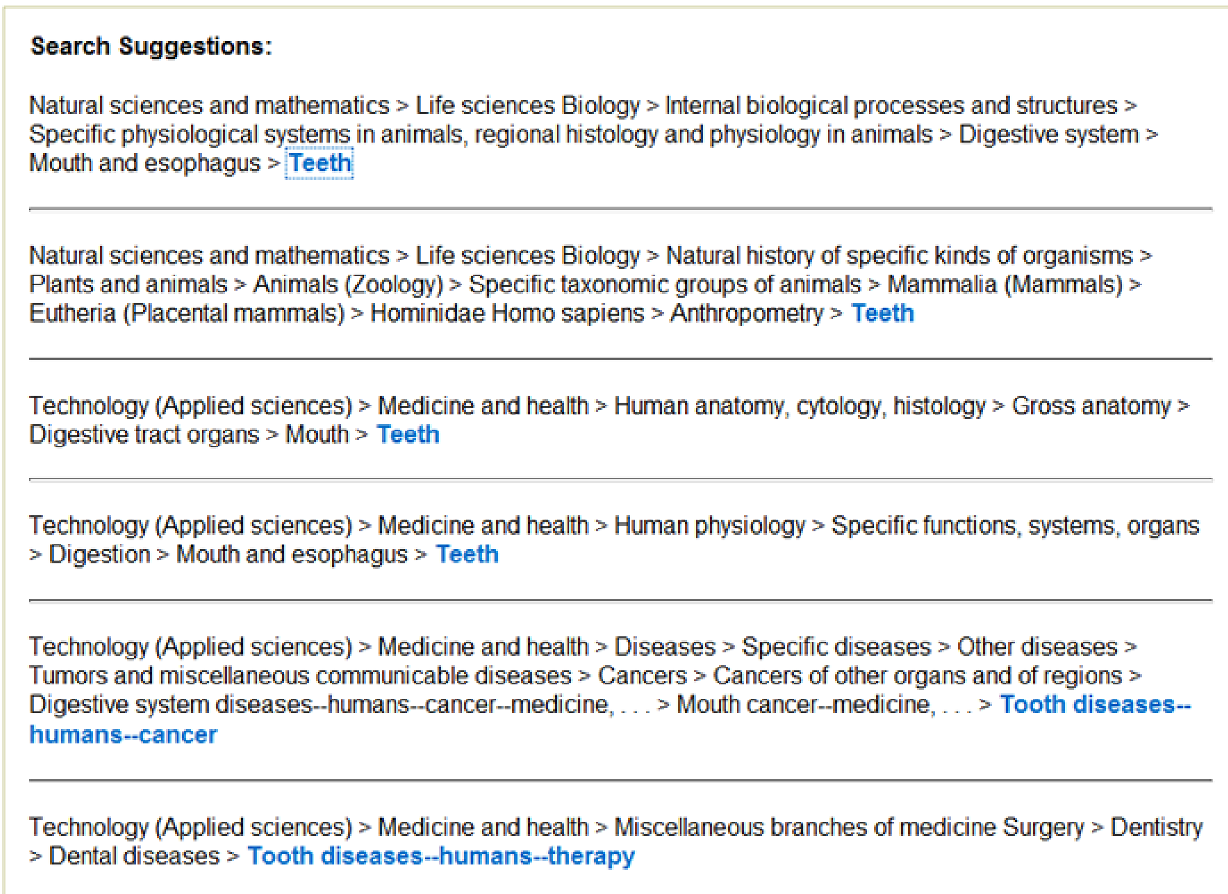

Figure 1: Partial screen-shot of search term disambiguation in the SCONE HILT pilot.

The user can now select a highlighted term from one of the captions and use it to identify collections matching the term. The software achieves this by using the DDC notation for the selected term and matching it against those assigned to the collection descriptions. If no match is found, the DDC notation is shortened by one digit and the process is repeated. This is equivalent to broadening the semantic of the notation because it is decimal; a shorter notation usually implies a broader concept. The process is repeated until several collections have been found, or the top of the notation hierarchy is reached.

For example, the DDC notation for the fourth caption hierarchy in figure 1 is 612.311. There are no collections in SCONE classified with this notation, so the notation is truncated by its last digit to give

612.31, and the search repeated. This is done successively through 612.3, 612, and 610, at which point several collections are matched and displayed, as in Figure 2.

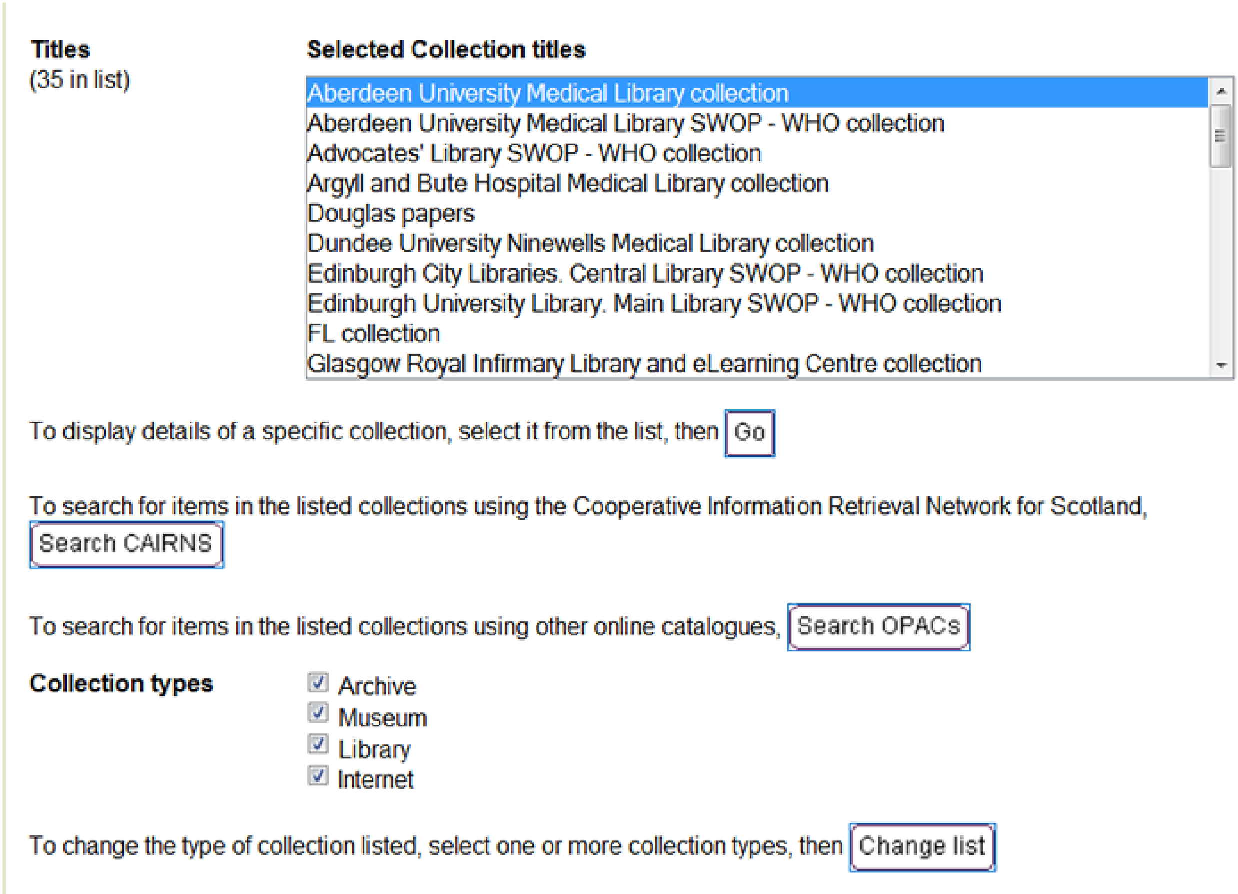

Figure 2: Partial screen-shot of SCONE collections matched by truncating the DDC notation 612.311.

HILT case 2: Intute

Intute is an online finding-aid for web resources supporting study and research. The resources are selected by academics. Metadata is created by subject-focused component catalogue services, each of which uses its own subject scheme. High-level subject retrieval is supported by a scheme of 19 categories.

The Intute pilot enhancement using HILT accepts a subject term input by the user and displays up to 10 related terms which can be used for another subject search. The service also displays any results from a search on the input term. The user can examine the results and select one of the related terms to redo the search if required. If no results are obtained, the pilot displays up to five terms with spellings related to the input term. Any of these can be selected by the user to carry out another search.

The interface also displays DDC notations and captions related to the subject term, to demonstrate the potential of using HILT to identify appropriate terms for searching the different subject schemes used in the Intute component catalogues. This has not been developed further, and the links displayed are inactive.

Figure 3 is a partial screen-shot where “tree” has been entered as a search term. The related terms displayed indicate at least two distinct subject contexts, genealogy and forestry. If the current results, displayed at the bottom of the screen, are not in the expected subject domain, the user can click on one of the related terms, for example “genealogy”, and carry out a search on that topic.

intute

Helping you find the best websites for study and research

Home | Web resources | Internet training | All services | Support for... | About us | Feedback

Home › Web resources › Search results

Search results

Search web resources

tree

Search Advanced search | Help

Search term(s)	tree [subject(s): All > All]
Intute catalogue results	530
Related terms [max 10]	genealogy broadleaved forests coniferous forests countryside conservation forest management forestry mixed forests orchards soft fruit sylviculture
DDC suggestions	christmas tree industry (338.1759775) trees--ornamental agriculture--law (343.0765977) trees (511.52) trees and petrified wood (561.16) trees (582.16) eucalyptus (gum trees) (583.766) haemodoraceae (bloodwort family), hypoxidaceae, velloziaceae (tree lily family) (584.354) pinophyta (gymnosperms) coniferales (conifers) (585) hylidae (tree frogs) (597.878)
Harvester results	View results (What's this?)

Filter by resource type: All Go

Now showing: 1 - 25 of 530 records
Order by Relevance | Alphabetical

Page: 1 2 3 4 5 ... 22 >> Next
Save record
More details The ultimate tree-ring web pages
Dr Henri Grissino-Mayer of the University of Tennessee has compiled a comprehensive online resource on all aspects of dendrochronology, aimed at a wide audience, from the general public to professional dendrochronologists. The website offers an introduction to the basic principles of the science, tips on the care and use of laboratory and field equipment, as well as a portal to related resources on the Internet, including tree-ring databases, software, suppliers of equipment, conferences and workshops, and bibliographies. An image gallery of trees and tree-rings refers mostly to North American trees. The site is indexed and can also be searched by keyword.
http://web.utk.edu/~grissino/

Click here to rate this resource or leave a comment.

Within this section

- Web resources
- Advanced search
- Harvester
- Subject A-Z

MyIntute

Email:

Password:

Login

Forgotten password?

Saved records
0

Email records

MyIntute accour

Register for an account to:

- Save records a searches
- Email alerts o new records f your subject a
- Export record your web pag

Figure 3: Partial screen-shot the Intute HILT pilot showing disambiguation and alternate term suggestions.

HILT case 3: The Depot

The Depot[iii] is an e-prints repository service aimed at researchers who do not have access to an institutional repository to deposit their papers. It relies on self-deposit, and the depositor is expected to generate metadata as part of the process. In particular, the user is required to assign one or more subject terms taken from the Joint Academic Coding System (JACS) scheme[15].

The pilot enhancement developed using HILT services helps the user to identify the relevant JACS captions. It accepts a term input by the user and displays all JACS captions containing the term. If the input term is not found in the JACS vocabulary, the pilot searches for it in the DDC captions and displays JACS captions mapped to the corresponding DDC notations. The user is then asked to select one or more of the displayed JACS captions as the subject metadata for the deposit.

Figure 5 shows the pilot enhancement inserted into the user-generated metadata workflow at the point where subject classifications are to be added. The depositor has input “teeth” as a search term to identify appropriate JACS captions and notations. Figure 6 shows the results. The term has been matched to the JACS headings by finding it in DDC captions mapped to JACS. The user can further identify context by exposing the code hierarchies, or simply check the box against as many of the captions as deemed relevant.

[iii] Now OpenDepot.org.

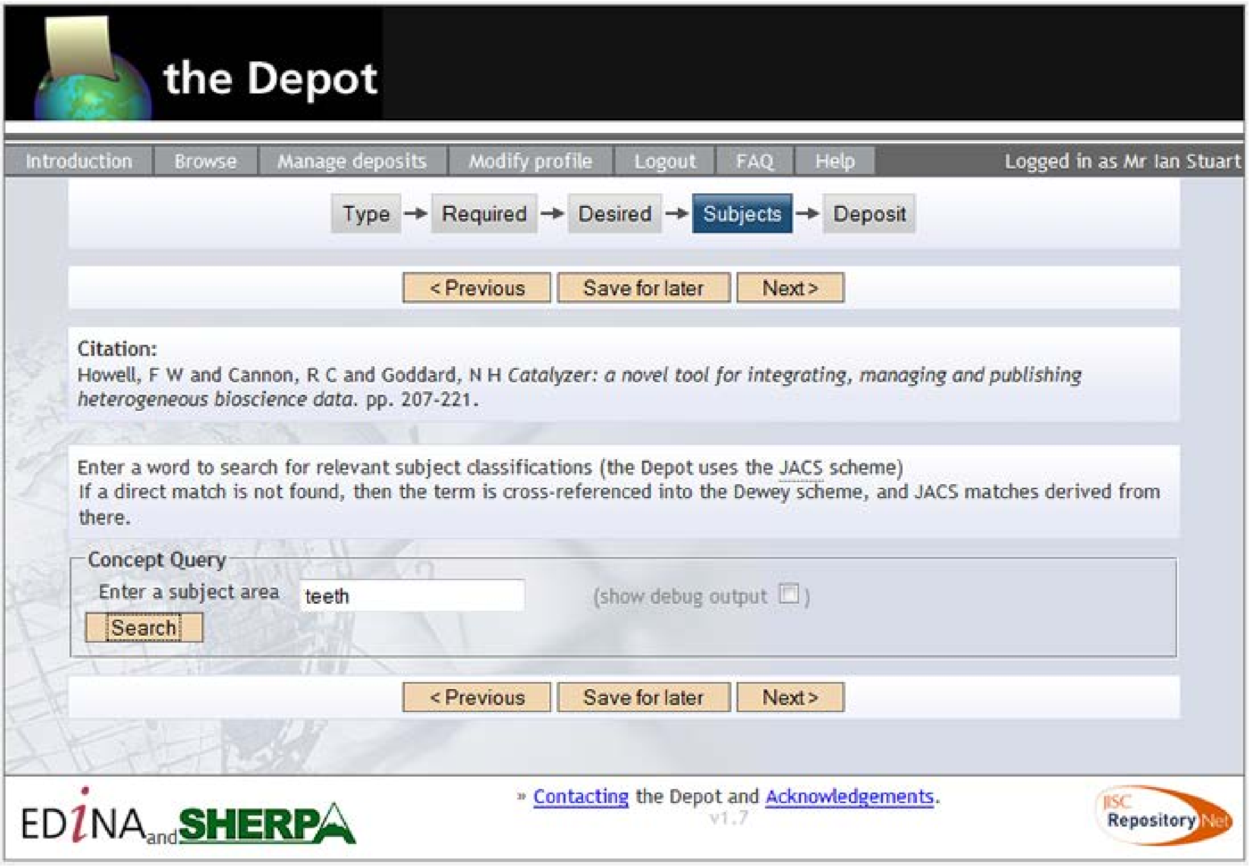

Figure 4: Partial screen-shot of the The Depot HILT pilot for helping user assignment of subject terms.

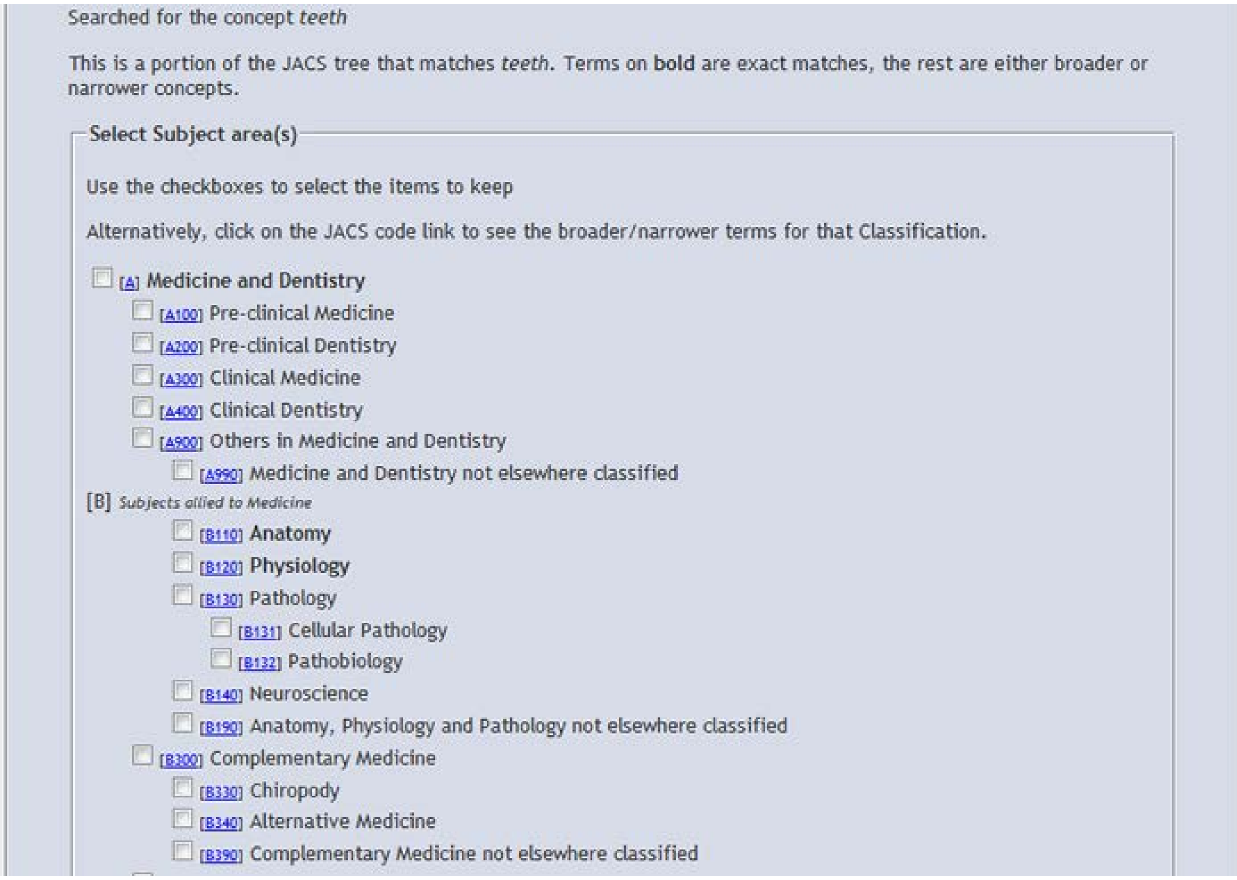

Figure 5: Partial screen-shot of JACS captions matched to user input via mappings to DDC.

Beyond HILT

The HILT approach is too expensive to scale across all subject schemes, despite the efficiency gains of using the hub-and-spoke architecture of a switching language. In such an architecture direct mappings between pairs of vocabularies are avoided by using the indirect mapping from spoke to spoke via the hub. This was recognized from the beginning of the project and is reflected in the “High level” part of its name.

Hub-and-spoke mapping architectures are themselves less expensive to scale than direct, one-to-one, mappings. As each new vocabulary is added, a new set of mappings is required for each existing vocabulary, rather than a single set of mappings to the hub or switching vocabulary. General, large-scale terminology services are therefore likely to employ hybrid architectures which complement a basic hub-and-spoke core by adding one-to-one mappings as cross-spoke links, where such one-to-one mappings are available. In some instances terms from two spokes may be linked directly by such a one-to-one mapping and also indirectly via the hub. In other instances, terms in one spoke may map only to the other spoke and not directly to the hub.

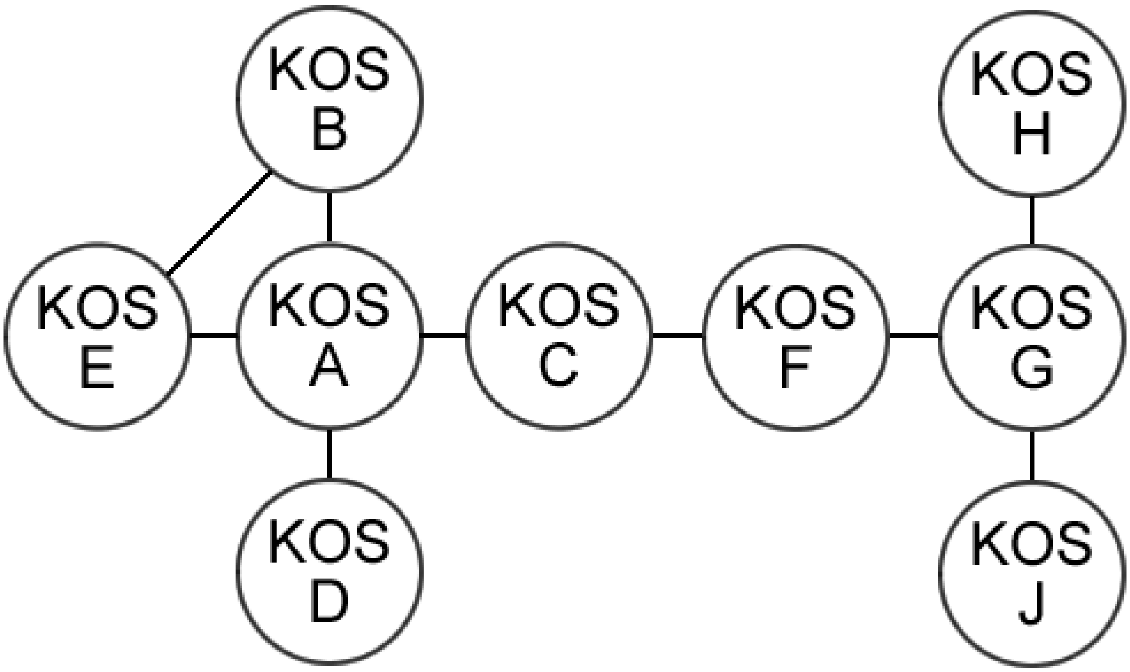

Figure 6: Hybrid architecture of hub-and-spoke and direct mappings between vocabularies.

In the example of a hybrid architecture given in Figure 7, KOS A is a hub for the vocabularies of the KOS B, C, D, and E spokes. There is also a direct mapping between KOS B and KOS E. KOS F is mapped directly to KOS C, and therefore indirectly to the KOS A hub. But KOS F is itself a spoke, along with KOS H and KOS J, to another hub KOS G.

Mappings between vocabulary terms are usually created with human intervention to ensure that nuances of meaning within and between different languages are preserved. This is the major component of the cost of developing and maintaining mappings, which could be significantly reduced with the use of machine-processing. Statistical analysis of associations of terms from different vocabularies used to index the subjects of the same resource can be used to determine strong correlations between terms, as in the mapping between DDC and LCSH found in OCLC's WebDewey service[16]. Such analyses require a critical mass of test data, and become more accurate as the amount of data increases. OCLC's Classify[17] service shows that consensus about the correct DDC number for a resource emerges from analyzing sufficiently large numbers of separate records for the resource. It might therefore be expected that the number of machine-generated mappings between terms in different vocabularies will increase as more and more metadata records are brought together in union catalogues and digital library aggregations.

Another source of mappings may lie in user-generated metadata. Users are encouraged in many social networking websites such as the photograph-sharing service Flickr[18] to "tag" information resources with their own terms describing what the resource is "about". Again, statistical clustering techniques can be used to ignore terms used very few times and arrive at a group consensus. It is not difficult to imagine that the hybrid architecture of Figure 7 will involve hundreds (or more) of sets of mappings between controlled and uncontrolled vocabularies in an effective terminology service covering a general range of subjects.

Semantic Web

The Semantic Web is "a group of methods and technologies to allow machines to understand the meaning … of information on the World Wide Web"[iv]. It therefore relies on machine-processing of metadata, or data about data, as a source of "meaning" or aboutness. Machine-processing requires that the metadata is marked-up and identified for programmes supporting semantic-based services. The Semantic Web uses Resource description framework (RDF)[19] as a metadata model for the most basic possible type of metadata statement: something has-some-property (with a value of) something else. This three-part statement is known as a triple. Triples can be chained to together by using special types of identifier for each part, to create webs of so-called linked data.

The markup of metadata into simple triples is essentially conceptual. The property forming the central part of a triple can be given a human-readable label, definition and scope note to ensure that cataloguers and retrieval system developers apply it in the correct semantic context. These meta-properties of label, definition and scope note are also essential to KOS and are available as RDF properties in SKOS.

SKOS was primarily designed for RDF representation of terms in thesauri, classification schemes, subject heading lists and taxonomies. As its name suggests, it can model simple relationships between terms, such as equivalence and hierarchy, but it does not provide capabilities for advanced structures such as faceted classification and subject heading schemes. These, however, can be marked-up using other RDF applications such as RDF schema (RDFS)[20] and Web Ontology Language (OWL)[21].

The Semantic Web therefore offers a number of features of use to terminology services:

- An environment optimized for machine-processing.
- An underlying framework (RDF) that can be scaled from single to multiple controlled vocabularies.
- A model for representing simple structures within and between vocabularies (SKOS).
- A means of representing more complex structures within and between vocabularies (OWL).

Terminology maintenance services

The contents and structure of many controlled vocabularies and subject schemas change through time, as a result of the need to accommodate new subject topics, and expand or contract the definition and

---

[iv] From the Wikipedia definition of 21 November 2010, available at: http://en.wikipedia.org/wiki/Semantic_web

scope of existing topics. The Semantic Web discourages the deletion or removal of anything that has been identified and published, to prevent the breaking of established links. Instead, terms which are no longer current should be retained, but marked as deprecated for future use. An important consideration for terminology services is the currency of a vocabulary; how does the programmer of an application based on the terminology service find out about old and new versions of a term, and how can use of the current version be ensured?

Version control is therefore important, and terminology services should be prepared to identify the date of last update of their constituent vocabularies within the service, and store similar information about the amendment of individual terms within a vocabulary by its maintainers. In essence, the vocabulary itself and each term it contains should be assigned a series of time-stamp properties which is available to developers and end-users of applications based on the terminologies.

RDF offers a useful mechanism for maintaining translations of vocabulary terms into other languages. A machine-readable identifier for a single term can be associated with labels, definitions, and scope notes in multiple languages, using a simple auxiliary identifier for the language. This allows applications using a terminology service to switch languages by applying the auxiliary identifier without altering the underlying programmes requesting output from the service. An application displaying a list of terms in English from a particular vocabulary can easily switch to displaying the Italian translations of those terms, if such translations exist. This approach is not useful, however, for the mapping of terms from one controlled vocabulary to another controlled vocabulary in a different language, because the relationships between terms within each vocabulary must be preserved to maintain semantic integrity and cohesion. Instead, the service needs to maintain a mapping between two different terminologies as a separate component which may require amendment if a term in either of the vocabularies is changed. In other words, the relationship of a term to its equivalent in another language can be modelled intrinsically within a single vocabulary and set of identifiers, while the relationship between terms from different vocabularies, whether in the same or different languages, can be modelled extrinsically using SKOS or some other set of mapping properties.

Another important consideration in the maintenance of terminology services is quality assurance. This is partially met by version control, but information about the source of a vocabulary is also an indicator of quality. Vocabularies maintained by large professional organizations such as the Library of Congress are likely to be of higher quality than those from small amateur organizations, and application developers may want to be able to identify and prefer or avoid some vocabularies in favour of others. This requirement is likely to increase if machine-generated and user-sourced vocabularies are part of the service. This does not imply that such vocabularies will always be of lower quality, but applications must have sufficient information to allow the appropriate vocabulary to be chosen to meet their functional requirements. The Semantic Web environment itself provides no level of quality control or indication. RDF is not designed to ascertain the truth of a triple: the simple statement "pigs" is a narrower term of "flying animals" can be expressed as a valid RDF triple. A semantic reasoning application would detect a conflict with the statement "Pigs cannot fly", but by itself would not be able to determine which is true and which is false. A human programmer would need to know that, say, the first triple came from a user-generated vocabulary about cartoons and the second statement from the Pig Breeders' Association before accepting or rejecting the metadata for the application.

Conclusion

Terminology and vocabulary services have an important role to play in computer-assisted information retrieval systems. They effectively bridge the semantic gap between humans and machines by encoding intellectual concepts and their organization into machine-processable representations that human programmers can use to build subject-based applications for end-users. In particular, the Semantic Web requires such services to develop utility from large numbers of basic metadata statements about terms and the relationships between them. Terminology services can provide complex building-block functions for interfaces matching user input to metadata about information resources, including disambiguation and monolingual and multilingual translation between specific vocabularies on a global scale.

General terminology services themselves require access to as many vocabularies as possible, including fully-controlled terminologies and mappings from professional organizations, semi-controlled or uncontrolled terminologies from amateur, end-user sources, and machine-generated mappings from critical masses of metadata. These vocabularies are best represented in RDF in order to exploit and contribute to the utility of the Semantic Web. Open-access publishing of vocabularies and schemas in Semantic Web formats is likely to encourage uptake and development of terminology services.

Several important vocabularies in wide-spread use in legacy metadata records are already available as part of the linked open data environment, as shown in the Linking open data cloud[22], including LCSH, Rameau subject headings in French, and the Schlagwortnormdatei (SWD) subject headings in German, together with mappings between them. The addition of linked data versions of other major subject heading and classification schemes in widespread use, and the development of terminology services, are essential to unlocking the world's subject catalogues and indexes for the benefit of the Semantic Web and its users.

References

[1] Vizine-Goetz, D. and others. 2004. Vocabulary mapping for terminology services. Journal of digital information, vol. 4, no 4 .(2004). Available at: http://journals.tdl.org/jodi/article/viewArticle/114/113

[2] OCLC. No date. Terminology services. Available at: http://www.oclc.org/research/projects/termservices/

[3] Library of Congress. 2009. Library of Congress authorities. Available at: http://authorities.loc.gov/

[4] United States National Library of Medicine. 2009. Medical subject headings. Available at: http://www.nlm.nih.gov/mesh/

[5] Library of Congress. 2007. Thesaurus for graphic materials. Available at: http://www.loc.gov/rr/print/tgm1/

[6] W3C. 2009. SKOS simple knowledge organization system - Home Page. Available at: http://www.w3.org/2004/02/skos/

[7] Scottish Collections Network (SCONE). 2009. [Homepage]. Available at: http://www.scotlandsinformation.com/scone/

[8] Co-operative Information Retrieval Network for Scotland (CAIRNS). 2009. [Homepage]. Available at: http://www.scotlandsinformation.com/cairns/

[9] High-Level Thesaurus Project (HILT). 2009. [Homepage]. Available at: http://hilt.cdlr.strath.ac.uk/index.html

[10] OCLC. 2009. Dewey services. Available at: http://www.oclc.org/dewey/

[11] Getty Research Institute. No date. Art & architecture thesaurus online. Available at: http://www.getty.edu/research/tools/vocabularies/aat/

[12] UNESCO. 2003. UNESCO thesaurus. Available at: http://www2.ulcc.ac.uk/unesco/

[13] Intute. 2009. [Homepage]. Available at: http://www.intute.ac.uk/

[14] OpenDepot.org. No date. [Homepage]. Available at: http://opendepot.org/

[15] UCAS. JACS 3. No date. Available at: http://www.ucas.com/he_staff/courses/jacs/jacs3

[16] OCLC. WebDewey. 2010. Available at: http://www.oclc.org/dewey/versions/webdewey/

[17] OCLC. Classify: an experimental classification web service. 2010. Available at: http://classify.oclc.org/classify2/

[18] Flickr. 2010. Available at: http://www.flickr.com/

[19] W3C. Resource description framework (RDF). 2004. Available at: http://www.w3.org/RDF/

[20] W3C. RDF vocabulary description language 1.0: RDF schema. 2004. Available at: http://www.w3.org/TR/rdf-schema/

[21] W3C. OWL 2 Web ontology language: document overview. 2009. Available at: http://www.w3.org/TR/owl2-overview/

[22] Cyganiak, R. and Anja Jentzsch. Linking open data cloud diagram. Available at: http://lod-cloud.net/