Signposting the Crossroads: Terminology Web Services and Classification-Based Interoperability

Gordon Dunsire* and Dennis Nicholson**

University of Strathclyde, CDLR, 26 Richmond Street, Glasgow, UK,
* g.dunsire@strath.ac.uk ** <d.m.nicholson@strath.ac.uk>

Gordon Dunsire is Head of the Centre for Digital Library Research at Strathclyde University in Glasgow, Scotland. He is a member of the Chartered Institute of Information Professionals (CILIP) and British Library Committee on *AACR* and the CILIP Committee on *DDC*, and is Chair of the Cataloguing and Indexing Group in Scotland. He is a member of Classification and Indexing Section of the International Federation of Library Assocations and Institutions (IFLA). He is the principal developer of the SCONE (Scottish Collections Network) collection descriptions service and other components of the Scottish Common Information Environment, and has been involved in several projects investigating the use of collection-level description and metadata aggregation in wide-area resource discovery.

Dennis Nicholson is a private consultant with expertise in the area of distributed digital libraries and library-related information technology. Between 1999 and 2009, he was Director of the Centre for Digital Library Research at the University of Strathclyde and Director of Research in Strathclyde University's Information Resources Directorate. He has been actively involved in research in the area of distributed digital libraries and information systems since 1991. He managed and led a range of funded research projects, including the High Level Thesaurus Project, and the Co-operative Academic Information Retrieval Network for Scotland project.

Dunsire, Gordon and Nicholson, Dennis. **Signposting the Crossroads: Terminology Web Services and Classification-Based Interoperability.** *Knowledge Organization, 37(4)*, 280-286. 20 references.

ABSTRACT: The focus of this paper is the provision of terminology- and classification-based terminologies interoperability data via web services, initially using interoperability data based on the use of a *Dewey Decimal Classification* (*DDC*) spine, but with an aim to explore other possibilities in time, including the use of other spines. The High-Level Thesaurus Project (HILT) Phase IV developed pilot web services based on SRW/U, SOAP, and SKOS to deliver machine-readable terminology and cross-terminology mappings data likely to be useful to information services wishing to enhance their subject search or browse services. It also developed an associated toolkit to help information services technical staff to embed HILT-related functionality within service interfaces. Several UK information services have created illustrative user interface enhancements using HILT functionality and these will demonstrate what is possible. HILT currently has the following subject schemes mounted and available: *DDC*, CAB, GCMD, HASSET, IPSV, *LCSH*, *MeSH*, NMR, SCAS, UNESCO, and *AAT*. It also has high level mappings between some of these schemes and *DDC* and some deeper pilot mappings available.

1.0 Introduction

It has become increasingly difficult for users to satisfy their information needs due to the rapid expansion of the Web and its sprawling nature; it is becoming progressively impractical for users to consult a wide range of sources to satisfy an information query. Consequently, it is of growing importance that users are able to search multiple distributed heterogeneous digital repositories simultaneously. With such a wide variety of resources available, however, the feasibility of achieving interoperability between them is gradually diminishing. Services employ different technical standards, indexing practices, search

facilities, and algorithms. There is wide variation in the language and terminology on which retrieval systems are founded. As a result, it is no longer sufficient for users to make decisions on whether to use keyword or phrase searching, employ Boolean operators, or try their luck with truncation; they must also now give consideration to the terminology they use. Problems relating to disparate terminology use have been an impediment to information retrieval for many years, but the growth of the Web, associated heterogeneous digital repositories, and the need for distributed cross-searching within information environments employing multiple terminologies has recently drawn the issue into sharp focus.

2.0 HILT project

The High Level Thesaurus (HILT) project comprised four phases of activity carried out by the Centre for Digital Library Research (http://cdlr.strath.ac.uk/) at the University of Strathclyde and funded by the UK's JISC (Joint Information Systems Committee, http://www.jisc.ac.uk/) with support from OCLC (http://www.oclc.org/). The project has been investigating mechanisms to assist the further and higher education community in the UK with problems associated with providing users the ability to find appropriate learning, research, and information resources by subject search-and-browse in an environment where most national and institutional service providers use different subject schemes to describe such resources. Those mechanisms, possibly applied through a JISC Shared Infrastructure Service, would help optimise the value obtained from expenditure on content and services by facilitating resource sharing. The environment is essentially monolingual with English as the predominant language in terms of resources and their users, although the project has carried out some investigation into non-English vocabularies.

The first phase (Nicholson et al. 2001) established that the preferred approach of the various services in the JISC domain to resolving the issue is one based on mapping the various subject schemes together through a central shared service that provides users with the correct alternative terms to use in the various different schemes. This architecture is referred to as a "spine", "switching language", or "hub-and-spoke." Phase II (Nicholson et al. 2004) built a pilot to illustrate the functions required of a terminologies service capable of taking a user-input subject term, identifying JISC collections relevant to the subject of the query, and providing the user with the correct subject term to use for the subject scheme employed by any given identified collection. The project then conducted a feasibility study for developing this into a machine-to-machine (M2M) pilot service to supply terminologies and mapping data for the use of other services, and scoped out an outline design for it. The third phase built the M2M pilot and scoped out a design for an initial entry-level service meeting the needs of a shared infrastructure.

HILT Phase IV (Nicholson, McCulloch, and Joseph 2009) developed pilot solutions for some of the problems encountered when cross-searching multi-scheme subject-based information environments, as well as providing a variety of other terminological searching aids. This phase delivered a range of simple M2M terminology services based on SRU (Search/Retrieve via URL, http://www.loc.gov/standards/sru/), SOAP (Simple Object Access Protocol, http://www.w3.org/TR/soap12/), and SKOS (Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/), using a database of terminologies and mappings of terms to the *DDC* (*Dewey Decimal Classification*, http://www.oclc.org/dewey/), along with an embryonic toolkit (CDLR 2009) to help developers of information services embed M2M interactions in user interfaces to improve subject retrieval, browse, and deposit functions. The project also developed a generic distributed subject interoperability and terminology services architecture and demonstrated its feasibility at a very basic level. A short extension project embedded interaction with HILT M2M services in the user interfaces of various information services serving the JISC community.

3.0 HILT architecture

A diagram of the architecture of the pilot HILT system is given in Figure 1.

Client applications are services such as information retrieval interfaces aimed ultimately at end-users. They access the content of the terminologies database using programmes containing functions from the HILT Application Programming Interface (API). These functions include:

a) **get_collections**: Takes a specified *DDC* notation and returns metadata about collections classified under the notation or its stems. The metadata include the subject scheme used by the collection's catalogue or other finding-aid.
b) **get_*DDC*_records**: Takes a subject term and returns *DDC* notations and captions mapped to

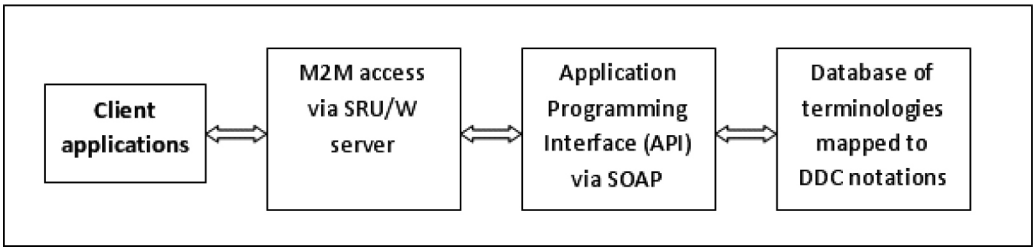

Figure 1. Architecture of the HILT pilot

terminologies containing the term. Terminologies include the *DDC* captions and relative index.

c) **get_non_*DDC*_records**: Takes a *DDC* notation and returns terms from terminologies mapped to the notation. Broader terms mapped to the stem chain of the notation are included in the output. Terminologies include *DDC* captions.
d) **get_all_records**: Takes a subject term and returns the combined outputs of get_*DDC*_records and get_non_*DDC*_records.
e) **get_filtered_set**: Takes a subject term and one or more specified terminologies and returns matching terms from the terminology, optionally together with their related terms, including broader, narrower, see also, and non-preferred terms.
f) **get_sp_suggestions**: Takes a subject term and return terms with similar spellings.
g) **get_wordnet_suggestions**: Takes a subject term and returns definitions and descriptions of the term.

4.0 HILT terminologies database

The terminologies database contains mappings to *DDC* notations from all or part of the following vocabularies:

Art & Architecture Thesaurus (*AAT*) (Getty Research Institute);
Commonwealth Agricultural Bureaux (CAB) thesaurus (CABI 2009);
Global Change Master Directory (GCMD) science keywords (NASA 2009);
Humanities and social science electronic thesaurus (HASSET) (UK Data Archive 2009);
Integrated public sector vocabulary (IPSV) (Great Britain. E-Government Unit. 2006);
Joint academic coding system (JACS) (UCAS 2007);
JITA classification schema of library and information science (JITA) (Barrueco Cruz et al. n.d.);
Library of Congress Subject Headings (*LCSH*) (Library of Congress, 2009);
Medical Subject Headings (*MeSH*) (US NLM 2009);
National monuments record thesauri (NMR) (English Heritage 1999);
Standard classification of academic subjects (SCAS), a precursor of JACS;
UNESCO thesaurus (UNESCO. 2003).

Most of the vocabularies are only partially mapped, to provide a pilot testbed. The database also contains the intrinsic mapping of the *DDC* captions to their *DDC* notations.

The *DDC* notations thus form the hub, spine, or switching language between the vocabularies. Any two vocabularies have an implicit cross-walk mapping via the *DDC* notation. This cross-walk is instantiated in an application programme using the **get_*DDC*_records, get_non_*DDC*_records,** and **get_all_records** functions. An advantage of this approach is that only one primary mapping between a vocabulary (a spoke) and the hub is required, and it can be maintained independently of other vocabularies. A disadvantage is the increased possibility of semantic misalignment in the two-stage correspondence between terms from different spoke vocabularies.

There are a number of benefits in using a classification schema as the hub, rather than a subject vocabulary based on natural language. Classification notations are independent of natural language and so avoid many of the problems associated with terminology such as case, plurals, antonyms, and synonyms. The notation is usually shorter than the corresponding caption and is unique to the concept, so it can readily form the basis of a term or concept identification system; homonyms (and translations) make this difficult to achieve with natural language. Classification systems can also provide methods for synthesising notations for new concepts from those that already exist.

The database also contains sample collection-level description metadata to support the **get_collections** function, and WordNet data to support **get_wordnet_suggestions**. The **get_filtered_set** function is supported by the inclusion of term relationships within

relevant vocabularies in the database. The **get_sp_suggestions** function uses an index of all terms recorded in the database.

5.0 Embedding HILT in end-user services

An extension project carried out between January and May 2009 had the aim of demonstrating enhanced functionality of a number of information services by embedding HILT M2M terminology and interoperability facilities within user interfaces. The services are:

1) The Depot (http://www.depot.edina.ac.uk/): An e-prints repository service for researchers who do not have access to an institutional repository. Metadata are user-generated by the e-print depositor, and includes subject headings taken from JACS. The service offers hierarchical browsing by subject heading.
2) Intute (http://www.intute.ac.uk/): An online finding-aid for web resources, selected by academics, to support study and research. The service offers browsing by a subject heading scheme of 19 categories, followed by browsing and keyword searching of the various subject heading scheme used by different component services.
3) Scottish Collections Network (SCONE, http://scone.strath.ac.uk/Service/Index.cfm): A collection-level descriptions service for identifying and accessing library, archive, and museum collections located in Scotland. Subject-based collections are classified by *DDC* and are assigned *LCSH* topics. The service offers browsing by *DDC* notation and *LCSH* topic, hierarchical browsing by *DDC* summary (top three levels or approximately 1000 classes) caption, and keyword searching by *LCSH* topic.

The Depot's embedding experiment (http://lucas.ucs.ed.ac.uk/cgi-bin/hilt-depot) displays headings from JACS which match a keyword search term entered by the user. If no headings are found directly, it displays JACS headings which are mapped to *DDC* captions (via the *DDC* notation) containing the search term. The user can then select one or more headings as the subject metadata for their e-print. The demonstrator developed by Intute (http://www.intute.ac.uk/search_hilt.html) displays up to 10 related terms for a keyword search term input by the user, along with metadata for resources matching the input term. It displays up to five terms with alternate spellings if the search term is not found. The displayed terms can be used to reiterate a search. The demonstrator also displays *DDC* captions and notations based on the search term; these are inactive, but have potential use as a source of keywords for searching component catalogues.

The SCONE subject retrieval pilot (http://scone.strath.ac.uk/Service/SCONEServiceHilt/*DDC*searchinput.cfm) displays the full hierarchy of *DDC* captions matching a keyword search term entered by the user. Captions are matched directly and via the other vocabularies mapped to *DDC* notations. The user can select a matched caption from one of the hierarchies as the input to a search for collections with *DDC* notations matching the notation, or its stem, for the chosen caption. This is equivalent to the **get_collections** function operating on non-HILT collection-level metadata. SCONE also developed a spellchecking pilot (http://scone.strath.ac.uk/Service/SCONEServiceHilt/IndexSpellCheck.cfm).

These demonstrators illustrate the utility of higher-level terminology functions such as:

a) Deriving terms from a target vocabulary;
b) Deriving alternate terms from a secondary vocabulary mapped to a target vocabulary;
c) Deriving related terms from a target vocabulary;
d) Deriving notations from a target classification scheme;
e) Deriving notations from a secondary vocabulary mapped to a target classification scheme;
f) Deriving terms with alternate spelling; and,
g) Disambiguating terms.

6.0 Developing a distributed approach

Hub-and-spoke mapping architectures are efficient to maintain and scale, but the sheer effort of mapping thousands of terms to the switching language prevents operational scaling of HILT, even within a monolingual environment, to cover all vocabularies likely to be significant for wide-area information retrieval. Direct mappings between vocabularies, such as in the MACS project (https://macs.hoppie.nl/pub/), are even more difficult to scale because of the combinatorial explosion: two vocabularies require 1 mapping, four vocabularies 6 mappings, six vocabularies 120 mappings, etc. Machine-generated mappings can considerably reduce the cost, but require a critical mass of instance data to be even remotely effective. The *LCSH* to *DDC* mapping in HILT is derived from WebDewey (http://www.oclc.org/dewey/versions/webdewey/) and is based on statistical asso-

ciations found in WorldCat (http://www.worldcat.org/) records. Reliability should increase with the amount of instance data processed, due to the effect of the law of large numbers. However, human review and amendment of such mappings is often required to make them successful; costs increase as a result. It seems, therefore, that all centralised approaches to improving subject interoperability at national and international scales are doomed to fail, leaving the retrieval environment littered with the corpses of incomplete and out-of-date mappings.

Fortunately, recent developments in the semantic web offer a way forward. SKOS, used by HILT when returning data called by a function, is a means of representing vocabularies and term relationships so they can be effectively processed by machines. It currently has the status of a W3C Proposed Recommendation. *LCSH* has recently become available in SKOS as a download or via a web service from Library of Congress Authorities and Vocabularies service (http://id.loc.gov/authorities/). The dewey.info service (http://dewey.info/) offers SKOS representations of the *DDC* Summaries in nine languages as an experimental linked data web service. There are proposals to treat the whole of *DDC* and its translations in a similar way (Panzer 2008). The European *DDC* Users Group (http://www.slainte.org.uk/edug/index.htm) is monitoring these developments with keen interest. Many other initiatives are underway world-wide to provide such services for other terminologies, including RAMEAU (http://www.cs.vu.nl/STITCH/rameau/) in French.

The process of representing vocabularies in SKOS assigns a unique identifier, the uniform resource identifier (URI), to each term as well as the vocabulary itself. Linked data, a semantic web concept exemplified by the LinkingOpenData project (SWEOIG 2009), uses URIs (with specified properties) to expose, share, and connect data on the world-wide web. Data is broken down into simple statements (called triples) such as "TermA has broader term TermB", represented in a machine-processable format (RDF/XML or Resource Description Framework in Extensible Markup Language). At the time of writing, the project had identified around 13 billion triples with around 150 million links, including RAMEAU linked to *LCSH* via the mappings created by the MACS project.

This distributed approach has the potential to replace the terminologies database in the HILT system architecture. SPARQL (W3C 2008), an RDF query language which has the status of a W3C Recommendation, can substitute for HILT's M2M access and API facilities, as shown in Figure 2.

Client applications still have to programme higher-level end-user functions using SPARQL, but would benefit from improved reusability and interoperability resulting from applying a common query language. A coordinating framework for such activity would be highly desirable.

The true power of using the semantic web will be realised when bibliographic records contain subject metadata are also represented as linked data. Client applications will be able to retrieve records seamlessly, processing user input terms using terminology services and then linking directly to metadata for relevant resources for display. As yet, only a tiny fraction of the world's online catalogues and other types of resource description records in machine-readable formats has been exposed in this way, although many initiatives are underway or being planned. One of the prerequisites for recasting records as linked data and subsequently accessing them via specific metadata elements is to expose the metadata schemas in use to the semantic web. Again, several significant projects are working towards this, including the ISBD/XML Study Group (http://www.ifla.org/en/events/isbdxml-study-group) with respect to the International Standard Bibliographic Description (ISBD) and DCMI RDA Task Group (http://dublincore.org/dcmirdatskgroup/) with respect to the RDA: resource description and access element set. Discussions about the MARC21 (http://www.loc.gov/marc/) and UNIMARC (http://www.ifla.org/en/unimarc) formats are also in progress. Eventually, the schemas can be used

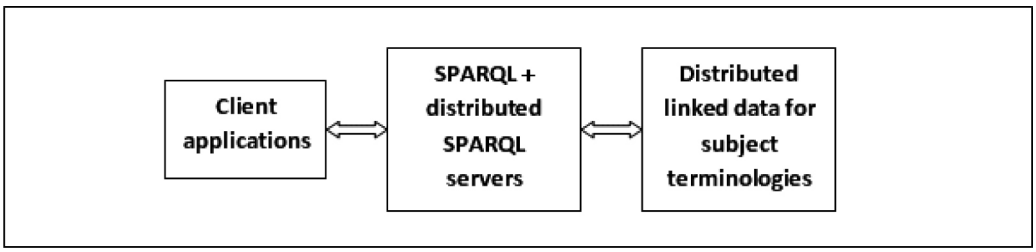

Figure 2. Distributed architecture base on the semantic web

to parse instance records into triples, and semantic equivalence between schema elements can be established as linked data, for example MARC21 tag 651 is the same as the Dublin Core dc.spatial element.

Names and "work" titles as subjects are not being neglected. It is likely, for example, that the Library of Congress Name Authority File (LCNAF) will be made available via the same service as *LCSH*. And when a large amount of instance records using, or otherwise linked to, two or more subject schemes, such as a classification and controlled subject heading, is made available, it provides critical mass for statistical mappings between schemes.

It should be noted that the linked data environment does not favour any specific mapping architecture. If Scheme A is directly mapped to Scheme B, and Scheme B is directly mapped to Scheme C, then Scheme A is indirectly mapped to Scheme C; Scheme B automatically becomes a hub with two spokes. If all of the existing equivalence mappings between terms in different schemes, along with all of the semantic mappings between terms in the same scheme, were published as linked data the result would be a net of mappings, a mixture of hubs and spokes linked to other spokes. And in many cases there would be more than one chain of links between any two terms. In this scenario, the reliability of each mapping (the authority of its creators) and the granularity of equivalence (exact, near, partial, etc.) will become important indicators of "best" pathways between terms.

7.0 Potential role of UDC

A prerequisite for a hub connecting subject schemes for a wide range of disciplines and domains is that it encompasses the range; it is necessary for the classification to be universal, covering all areas of human knowledge and endeavour. It is desirable that the classification has already been mapped to one or more subject schemes. Added value is available if the scheme is in widespread use; the hub can be used to retrieve catalogue records directly, instead of using an indirect link via a spoke. The Universal Decimal Classification (UDC) therefore qualifies as a potential hub. It is already in a machine-readable format from which a semantic web representation can be made. Like *DDC*, UDC has been translated fully or partially into between 30 and 40 languages. UDC is used by many special libraries, so the potential for machine-generated associative mappings with special subject vocabularies is high.

The Renardus project (Koch, Neuroth, and Day 2001) rejectied UDC as its hub, for reasons summed up as "When it comes to digital library applications … the UDC system and its development efforts are clearly insufficient and fall far behind the *DDC*." This assessment would change significantly if UDC could catch up with semantic web developments. In particular, exposing the UDC schedules as open linked data, or at second-best allowing UDC instance data to be used without licensing restrictions, will allow UDC to become a passive hub as direct and associative mappings are added or created as linked data.

A significant milestone for UDC in this respect was the launch of the UDC Summary service (http://www.udcc.org/udcsummary/php/index.php) in November 2009. This provides online access to a selection of around 2000 classes from the UDC scheme, comprising main numbers, common auxiliary numbers, and special auxiliary numbers. The captions are available in 16 languages, with plans to add a further 7. The data provided by the service is released under a Creative Commons license allowing copying and reuse on condition that attribution is assigned to the UDC Consortium and any redistribution uses the same licence. A further milestone will be reached with the planned release of export formats and mappings to other schemes early in 2010. Export formats will include RDF/XML, making the classes compatible with the semantic web and allowing them to be published as linked data. These developments will make UDC as viable as a potential hub as *DDC*. Indeed, if the existing mapping between the high-level classes of UDC and *DDC* is included in the UDC Summary service in an RDF representation, there is further potential for hybrid hubs using a mix of UDC and *DDC*, and super-hubs (hubs of hubs based on UDC or *DDC*).

Classification is indeed at the crossroads. It has the potential to become the spaghetti junction (Wikipedia 2009) of the information superhighway. But, like so many other professional library and information practices, services, and systems, classification is also at an administrative, business and technical crossroads, where decisions must be made as to which direction is best for the future.

References

Barrueco Cruz, José Manuel, et al. (n.d.). *JITA classification schema of library and information science*. Available http://eprints.rclis.org/jita/.

CABI. 2009. *CAB thesaurus*. Available www.cabi.org/cabthesaurus/.

CDLR. 2009. *[HILT toolkit]*. Centre for Digital Library Research. Available http://hilt4.cdlr.strath.ac.uk/toolkit.zip.

English Heritage. 1999. *National monuments record thesauri*. Available http://thesaurus.english-heritage.org.uk/.

Getty Research Institute. No date. *Art & architecture thesaurus online*. Available http://www.getty.edu/research/conducting_research/vocabularies/*AAT*/

Great Britain. E-Government Unit. 2006. *IPSV - Integrated public sector vocabulary*. Version 2.00. Available http://www.esd.org.uk/standards/ipsv/.

Koch, Traugott; Neuroth, Heike; and Day, Michael. 2001. *DDC mapping report: Renardus D7.4.* Available http://homes.ukoln.ac.uk/~tk213/Mapping report-d74.htm.

Library of Congress. 2009. *Library of Congress authorities*. Available http://authorities.loc.gov/.

NASA. 2009. GCMD's science keywords and associated directory keywords. Available http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html.

Nicholson, Dennis, et al. 2001. *HILT: High-Level Thesaurus project final report to RSLP & JISC*. Available http://hilt.cdlr.strath.ac.uk/Reports/Final Report.html

Nicholson, Dennis, et al. 2004. *HILT: High-Level Thesaurus project phase II : a terminologies server for the JISC Information Environment : final report to JISC. Main report*. Available http://cdlr.strath.ac.uk/pubs/nicholsond/HILT2FinalMain.pdf

Nicholson, Dennis; McCulloch, Emma; and Joseph, Anu. 2009. *HILT IV and embedding extension. JISC final report*. Available http://hilt.cdlr.strath.ac.uk/hilt4/documents/finalreport.pdf

Panzer, Michael . 2008. *Cool URIs for the DDC: towards web-scale accessibility of a large classification system*. In Greenberg, J. and Klas, W., eds. *Metadata for Semantic and Social Applications. Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, 22-26 September 2008*, pp. 183-190. Available http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf.

SWEOIG. 2009. *LinkingOpenData*. W3C Semantic Web Education and Outreach Interest Group. Available http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

UCAS. 2007. *JACS 2.0*. Available http://www.ucas.com/he_staff/datamanagement/jacs/jacs20.

UK Data Archive. 2009. *Humanities and social science electronic thesaurus*. Available http://www.data-archive.ac.uk/search/hassetSearch.asp.

UNESCO. 2003. *UNESCO thesaurus*. Available http://www2.ulcc.ac.uk/unesco/.

US NLM. 2009. *Medical subject headings*. United States National Library of Medicine. Available http://www.nlm.nih.gov/*MeSH*/.

W3C. 2008. *SPARQL query language for RDF*. Available http://www.w3.org/TR/rdf-sparql-query/.

Wikipedia. 2009. Spaghetti junction. Available http://en.wikipedia.org/wiki/Spaghetti_junction.