Using linked data to classify web documents

Dominic Fripp

University of Bristol, Library Services & Cataloguing
dom.fripp@bristol.ac.uk

Graduated from the Department of Information Science and Digital Media, University of the West of England, Bristol, UK in 2010.

Abstract
Purpose – To find a relationship between traditional faceted classification schemes and semantic web document annotators, particularly in the linked data environment.
Design/methodology/approach – A consideration of the conceptual ideas behind faceted classification and linked data architecture is made. Analysis on selected web documents is performed using Calais' Semantic Proxy to support the considerations.
Findings – Technical language aside, the principles of both approaches are very similar. Modern classification techniques have the potential to automatically generate metadata to drive more precise information recall by including a semantic layer.
Originality – Linked Data has not been explicitly considered in this context before in the published literature.
Keywords Classification, Information retrieval, Semantic web, Linked data, Facet analysis, Metadata
Paper type Technical paper

Background
Classification lies at the heart of any attempt to organise. Whether it is books on the library shelf, food along supermarket aisles or genetic sequences in a database, the online world of today presents fresh challenges for classification schemes. The wealth of data and documents now available on the World Wide Web are there in mostly unstructured form. The notion of structure in this example relates to the presence of structure and/or descriptive metadata.

The traditional notion of classification in the library environment arose from concentration on printed materials and their arrangement within a physical space. However, in the world of digital documents, "there is no shelf" (Shirky, 2005). Classification relies on a more descriptive element (Broughton, 2006), i.e., the subject allocation (Langridge, 1992) that will aid information retrieval. The scale of documents available on the Web suggests the need for an automated approach to the task. As the web document corpus increases, the need for relevancy and precision in recall grows proportionately.

Recent ideas for document classification use semantic rather than syntactic approaches to enable subject analysis. The techniques have emerged from the burgeoning semantic web project: the drive to make all web data machine readable. Perhaps unexpectedly, at the heart of these projects are the principles upon which Ranganathan created his colon classification scheme. His focus on concepts and the

building up of relationships between them has clear parallels in the realm of ontology building and domain modelling.

To make this relationship more explicit, it is worth starting by considering three definitions (Schwarz, 2005):

1. Classifying as a verb is synonymous with domain modeling: the act of grouping together similar or related concepts and arranging the resulting groups in a logical way.
2. Classification as a noun is the resulting domain model.
3. A second meaning of classifying as a verb is used in relation to instances. Instances are classified according to an existing domain model in order to organize them, for example, classifying individual books in a library according to the Dewey Decimal Classification System.

Schwarz underlines the need to distinguish the act of classifying on the concept level from the act of classifying on the instance level. The World Cup Final is both a concept and an instance depending on the context. If the notion of time is added, the concept World Cup Final can be attributed with *every four years*, whereas World Cup Final attributed 1990 would be an instance of the concept.

A common example in the literature, (Broughton, 2006, and Rosenfeld and Morville, 2006, have examples) is that of classifying wine. Taking that as the domain, the concepts that best describe wine can be allocated and used as a way of grouping similar properties together. Four properties that a bottle is commonly "labeled" with are grape, region, price and year. As new instances of wine occur, the concepts can be extended to ensure that these wines also fall within the scope of the domain.

In Ranganathan's terminology, concepts can be read as facets. It is said that the inspiration for his colon classification scheme was the toy Meccano (Beghtol, 2008). On a visit to England during the 1920s, Ranganathan first saw the metal building kits and observed how many different structures could be built from a single box of pieces. The implication was that complex objects could be built up from a finite set of variables or facets.

Equivalently, for the purposes of classification, Ranganathan argued that any subject, no matter how complex, could be built from the same set of basic components. His universal system was PMEST, a five facet arrangement that focused on the concepts inherent in the subject matter of documents, rather than the allocation of a single place on a pre-existing branch of an enumerative structure such as Dewey. His schema was designed to cope with any conceivable document (Ranganathan et al., 1960).

An example can show how these different approaches relate to one another. The FAST (Faceted Application of Subject Terminology) schema (O'Neill *et al.*, 2001) leverages Library of Congress subject headings data from bibliographic records and adds them as searchable metadata. Taking this example, with Ranganathan's and the wine domain, the following correlation can be asserted:

FAST facets	**Ranganathan facets**	**Wine concepts**
Topic	Personality	Grape

Geographical	Space	Region
Period	Time	Year
Form	Matter	Price

Although the comparison is not perfect, it shows a conceptual similarity to the three approaches and an attempt to fulfill the criteria of Schwarz's point 1 regarding domain modeling.

At the heart of his writings, Ranganathan (Langridge, 1992: p.65) outlines a three step approach to document classification that can be added to Schwarz's definitions.

1. **Idea plane**. Subject analysis in one's own words, including form of knowledge, topic, and any lesser forms that apply.
2. **Verbal plane**. Examination of the schedules to find the necessary concepts.
3. **Notational plane**. Construction of notation for the subject according to the scheme's rules.

Step 1 is the analytical part to Step 3's synthetic. The analytico-synthetic is the decon/recon-structive process of facet analysis. Step 2 is the medium over which this process can take place. The PMEST formula occupies this space in that it provides the vocabulary by which the classification can be expressed. For wine, Zinfandel, Merlot and French would be examples of the vocabulary that would be used.

The steps can be reduced to the following fundamental elements:

1. Document analysis for topics (subjects).
2. Relating topics (subjects) to domain concepts.
3. Expressing the concepts using domain notation.

In this form, these rules of facet analysis can now be taken from library cataloguing theory in the midst of the twentieth century to the cutting edge of the semantic web: specifically the project of linked data.

Linked data

There are numerous definitions of ontology in the literature. Noy and McGuiness (2001) keep it simple:

> An ontology defines a common vocabulary for researchers who need to share information in a domain.

Their definition stipulates an additional computational requirement:

> [The ontology] includes machine-readable definitions of basic concepts in the domain and relations among them.

This readability is a key requirement for the semantic web project. Tim Berners-Lee has argued that most information on the Web is designed for human consumption and that "the Semantic Web approach… develops languages for expressing information in a machine processable form" (Berners-Lee, 1998).

There are many documents on the Web that can be read by humans and the semantic layer in which they operate resides in the interaction between the language of the document and the reader. Whatever the definition of understanding may be, it is sufficient to say that by adding metadata to documents, computers may also "understand" them. Floridi's (2009) objection to this implied artificial intelligence (and the problematic consequences thereof) can be avoided if the limit of machine comprehension is taken to mean the navigation of a document using clearly defined concepts in clearly defined machine languages.

In the world of linked data, these concepts are powered by RDF triples and URIs. RDF (Resource Description Framework) triples are tripartite expressions of relationships between different concepts based on a subject-predicate-object construction. URIs (Uniform Resource Identifiers) provide a fixed description of the subjects and objects. Within the whole ontology of the linked data cloud, the identifiers should be unique to ensure consistency over the whole domain.

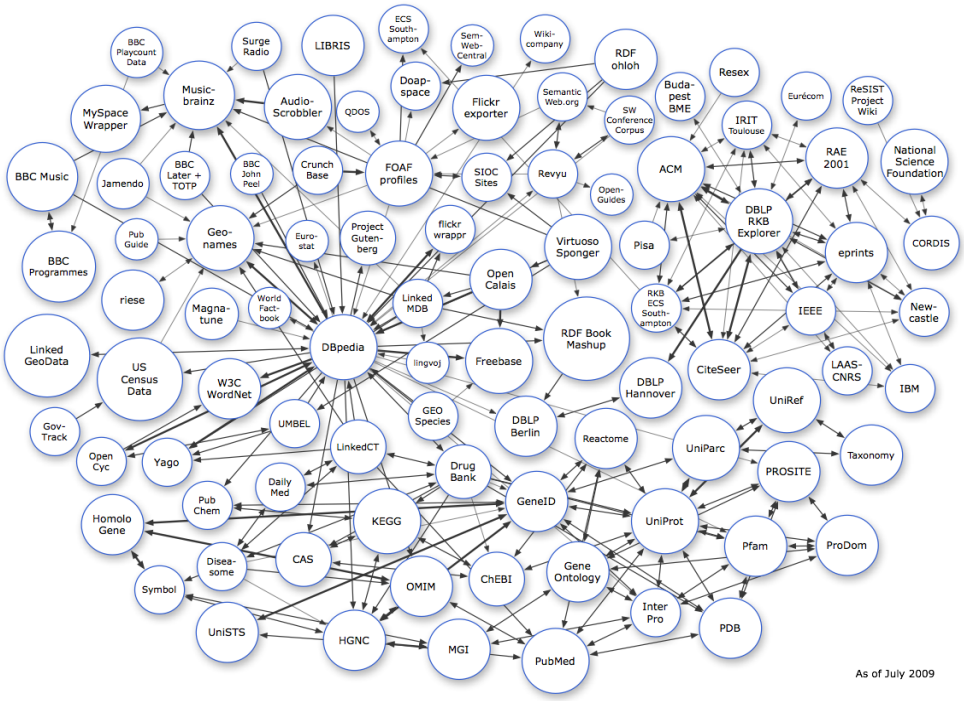
The power of linked data is the exposure of the data (anchored to URIs) to other resources that, in turn, can link to other data. This intermeshing of different data sets creates hyperdata (Idehen, 2009): a direct data correlation to hypertext.

Linked data is also able to deal with the information retrieval problems of polysemy and homonymy. Polysemy is not an issue providing all equivalent concepts point to the same URI. Thus synonym rings are infinitely extendable so long as the information target is the predefined entity indicator. For example, if *film* is the URI then the following are conceptually equivalent if anchored to the same URI. Film = cinema = movies = flicks = featurepresentation = fiml = cine = and any other possible equivalence that can be defined.

Homonymy is handled in much the same way. *Bank* can be defined by its target URI. The relational properties by which bank can be defined should distinguish between the side of a river and a financial institution. In the example shown later, Semantic Proxy differentiates between Tosca the opera and Tosca the character by examining the context of use based on suitable triples and URIs.

Richard Cyianek has created a visual representation of the databases that comprise the linked data cloud (Cyianek and Jentzsch, 2009).

Figure 1. The LOD dataset cloud



Source: Cyganiak, R. and Jentzsch, A. (2009)

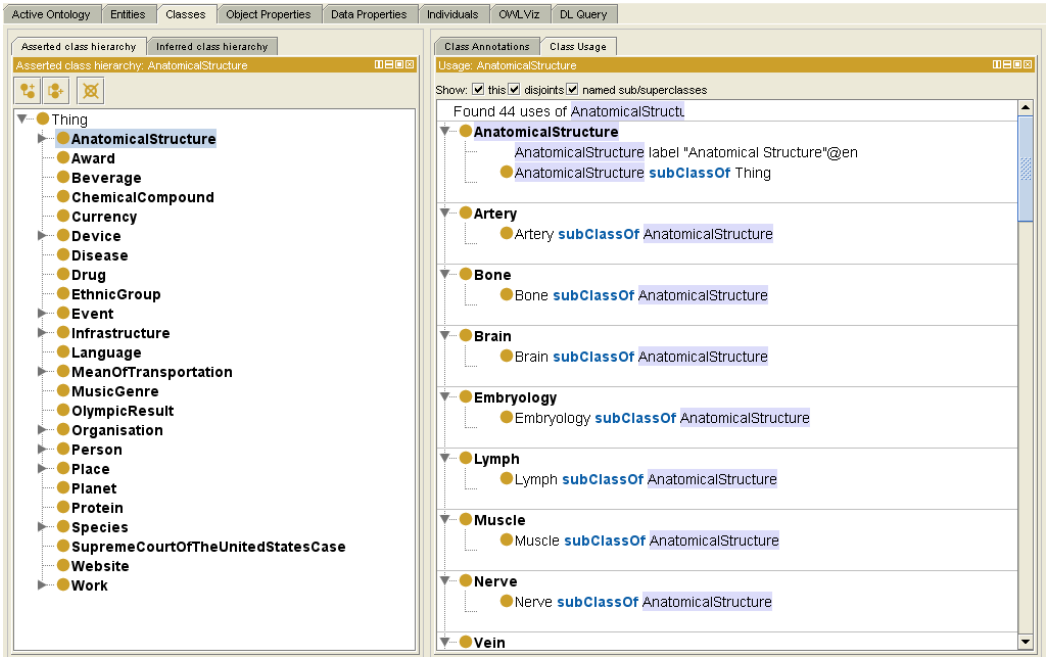
Cyganiek's circles (Figure 1) represent the different connected datasets. The size of the circle indicates the amount of RDF triples available in the set. The largest contain over one billion triples. The interconnecting arrows denote how URI's are used between datasets. Double headed arrows indicate that each set contains URIs that are used by triples in the other. The thicker the arrow, the more URIs are used. This is a considerable simplification of the linked data cloud, and further explanation isn't necessary here. The importance of datasets in the task of classification is the existence of billions of triples, each one with the potential to connect objects to the structural and descriptive metadata of the identifiers. From these building blocks, databases can build a large ontology that will be consistent as the data is interrelated. As an example**,** DBpedia is one of the largest datasets in the linked open data cloud (LOD cloud) and currently contains about 1,173,000 instances (information available at http://dbpedia.org/About). Table I lists the number of instances for several classes within its ontology.

Table I. DBPedia ontology content (January 2010)

Class	Instances
Place	339,000
Person	282,000
Work	234,000
Species	130,000
Organisation	119,000
Building	30,000

These instances are leveraged from Wikipedia articles and mapped onto 205 classes with 1,210 properties.

Figure 2. DBpedia data set examined with Protégé



Source :– Protégé available at: http://protege.stanford.edu/, DBPedia dataset available at: http://dbpedia.org/About

Figure 2 is a snapshot from the DBpedia ontology which shows the top class hierarchy on the left and some relational statements of instances to classes on the right. These are in the key triple format, for example, Artery subClassOf AnatomicalStructure.

Calais and Semantic Proxy

Calais is a project by Thomson Reuters designed to help realise Berners-Lee's vision of a machine readable web and is part of the linked data community. Its Semantic Proxy project aims to translate the content of any URL on the Web to its semantic representation in RDF, HTML or Microformats (Calais, 2009). Primarily designed to be used by machines, Semantic Proxy does provide information in a way that humans can understand too. In addition to document entity identification and extraction via linked data, Semantic Proxy aims to emulate human metadata creation by adding social tags to any document it processes.

The Calais service is, by its own admission, tailored towards the online media and enterprise markets. This explains some of the choice of facets, or entities that it extracts from a document. However, the list of facets should be as scalable as the open data sets permit and it can be argued that appropriate facets to the library and information environments (such as Dewey or Library of Congress subject headings) could be included when the concomitant datasets are made available for inquiry. The process by which it works has direct parallels with the three step document classification program outlined earlier.

Firstly, Semantic Proxy analyses the document for entities that belong to particular classes as defined by the supporting ontology. Secondly, based on the frequency of the appearance of these entities and the words around them, the software makes a calculation of relevancy of the instance to the concept and the document overall. Thirdly, the instances are expressed in terms of their related classes available through the ontology. Semantic Proxy can identify all of the following:

> Person, Company, Medical condition, Position, Natural feature, Social tag, Province or state, Organization, City, Continent, Facility, Country

By sorting the entities into facets such as these, Semantic Proxy creates a type of semantic index, a section of which is examined below.

Example: Facet (Concept) analysis of a Wikipedia document

Following the example in Steve Pepper's introduction to Topic Maps (Pepper, 2004), the URL for the Wikipedia article on Tosca was retrieved (http://en.wikipedia.org/wiki/Tosca). Note that this is a primary (and not disambiguated) document named Tosca.

Methodology
The URL was analysed using the Semantic Proxy. The results produced by the analysis were then fed into Thinkpedia to produce a visual representation of the data.

User tag data from Delicious was obtained from http://bit.ly/6YsGDO

Findings
A selection of the Semantic Proxy analysis is shown below.

> Topic: Entertainment/Culture (63%)

Semantic Proxy states that, in its current incarnation, it works best with news (http://www.semanticproxy.com/about.html). Topic, as used here, is best understood as a news category into which the analyzed content could be placed.

> **Entertainment_Culture** (socialTag) importance: 2
> **Tosca** (socialTag) importance: 1
> **Operas** (socialTag) importance: 1

For comparison, the top three user generated tags on the Delicious bookmarking system (as of 10th January 2010 and excluding "wikipedia") for the same page are:

Tag	**Used**
Opera	7
Tosca	2
Puccini	2

The following instances give at least one example of where the instance was found in the text. The amount of times that an instance is found appears directly proportional to the relevancy score awarded. The name of the attendant class is supplied in brackets after the instance.

> **Rome** (City) relevance: 57% resolutions: Rome,Province of Rome,Italy
> Instance Info
> detection: The work premiered at the Teatro Costanzi in **Rome**
>
> **Italy** (Country) relevance: 29.8% resolutions:Italy
> Instance Info
> detection: served as Director. Queen Margherita of **Italy**
>
> **Roman prison** (Facility) relevance: 50.4%
> Instance Info
> detection: of his escape from Castel Sant'Angelo (papal **Roman prison**),
>
> **food** (IndustryTerm) relevance: 35.6%
> Instance Info
> detection: Mario!). Cavaradossi gives Angelotti some **food** and helps him return

pain (MedicalCondition) relevance: 12.8%
Instance Info
detection: he knows where Angelotti is hiding. In his **pain** and humiliation, Cavaradossi denounces Tosca

Alps (NaturalFeature) relevance: 27.3%
Instance Info
detection: The following spring, Napoleon crossed **the Alps** with an army

Mario falls (NaturalFeature) relevance: 7%
Instance Info
detection: with new hope.") The soldiers fire; **Mario falls**.

Austrian-Russian army (Organization) nationality: Austrian organizationtype: governmental military relevance: 28.3%

Floria Tosca (Person) nationality: N/A persontype: N/A relevance: 77.6%

Note the limitations of this approach in the NaturalFeature class, which has included the instance of "Mario falls". It is also worth noting the low relevancy score which does not correct the mistake in itself, but does discount it from appearing within the visual metadata record shown in figure 3. Semantic Proxy does have the capacity to identify generic relations between objects as demonstrated by:

GenericRelations
relationobject: the knife **relationsubject: Floria Tosca** verb: plunge
Instance Info
detection: him to kiss her. As he advances to embrace her, **she plunges the knife** into him.

It can also deal with more complex relationships, as shown here:

EmploymentRelation
person_employee: Spoletta person_employer: Scarpia position: deputy status: current
Instance Info
detection: , when Scarpia, chief of police, arrives with **his deputy Spoletta**.

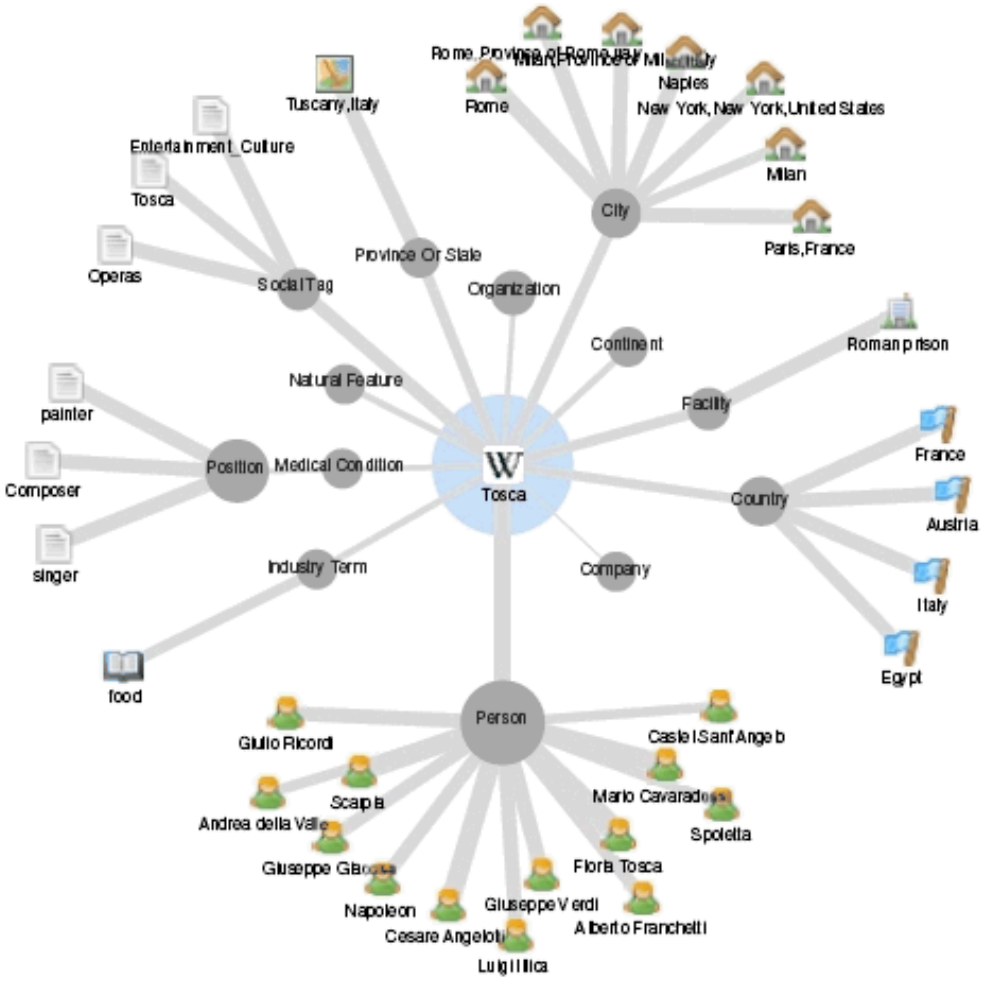
Visualising metadata using Thinkpedia

Thinkpedia was developed by Christian Hirsch in 2008 and uses the semantic information leveraged by Semantic Proxy and fed through Thinkmap software to create a visual graph of the document metadata. The thickness of the line between the main classes and the instances indicate the relevancy of each piece of metadata to the document. Each instance in each class is interactive, making Thinkpedia a navigational tool, not just between document metadata arrangements but between the instances and classes themselves.

The "Tosca" document map is shown in figure 3. With recourse to Ranganathan's classification prescription, the main facets of the document are the dots closest to the

central hub. Rather than the five of his PMEST formula, there are thirteen classes identified in the data output. The size of the dot is the statement of its relevancy to the document. It is easy to spot the more relevant facets, although lesser facets are also included. This keeps the dimensionality of the structure high but the metadata rich.

Figure 3. Thinkpedia Map of "Tosca" Wikipedia document



Source: created by Thinkpedia (http://thinkpedia.cs.auckland.ac.nz) | powered by www.wikipedia.org, www.semanticproxy.com and www.thinkmap.com

Conclusion

Classification still plays an important role in organising information. More than ever, the surfeit of information coupled with the growing size and interconnectivity of the Web provide conditions in which many unstructured documents can be marginalized

and become irretrievable by searches. By utilising facet analysis, modern information retrieval techniques promise to add a semantic layer to search querying. As Ranganathan opened up a new type of classification scheme by examining documents in terms of their subject components, linked data has the potential to index, annotate and connect web documents (and data) automatically by using semantic concepts as a means of organising metadata. Visual data representations of this technique expand Ranganathan's Meccano metaphor into a complex network of fully searchable and scalable classes and instances. This network arrangement has strong visual similarities with Broughton's (2007) molecular analogy of faceted knowledge organisation.

Acknowledgements

Many thanks to Tom Tague of the Calais Initiative and Kingsley Idehen of OpenLink Software, both of whom have provided invaluable guidance in the writing of this paper.

Software

Protégé – Ontology builder used to expose the main class fields and class relations in the DBpedia database. Available at: http://protege.stanford.edu/

Semantic Proxy – part of the Thomson Reuters Calais initiative. Currently at Beta testing stage. Available at: www.semanticproxy.com

Thinkpedia – Thinkpedia is developed by Christian Hirsch, PhD student at the University of Auckland, under the supervision of John Hosking and John Grundy. Requires Java. Available at: www.thinkpedia.cs.auckland.ac.nz

Thinkmap – available at: http://www.thinkmap.com/ - requires Java.

References

Beghtol, C. (2008), "From the Universe of Knowledge to the Universe of Concepts: the structural revolution in classification for information retrieval", *Axiomathes*, Vol. 18 No. 2, pp. 131-144.

Berners-Lee, T. (1998), "Semantic Web Road map", available at: http://www.w3.org/DesignIssues/Semantic.html (accessed 1 January 2010).

Broughton, V. (2006), "The need for a faceted classification as the basis of all methods of information retrieval", *Aslib Proceedings*, Vol. 58 No. 1-2, pp. 49-72.

Broughton, V. (2007) "Meccano, molecules, and the organization of knowledge - The continuing contribution of S.R. Ranganathan", available at: http://www.iskouk.org/presentations/VandaBroughtonNov2007.ppt (accessed 1 January 2010).

Calais (2009), "About Semantic Proxy", available at: http://semanticproxy.com/about.html (accessed 1 January 2010).

Cyganiak, R. and Jentzsch, A. (2009), "About the Linking Open Data dataset cloud", available at: http://richard.cyganiak.de/2007/10/lod/ (accessed 1 January 2010).

Floridi, L. (2009), "Web 2.0 vs. the Semantic Web: a philosophical assessment", *Episteme*, Vol. 6 No. 1, pp. 25-37.

Idehen, K. (2009), "5 very important things to note about HTTP based linked data", Kingsley Idehen's Blog Data Space, 2010, available at: http://www.openlinksw.com/dataspace/kidehen@openlinksw.com/weblog/kidehe

n@openlinksw.com%27s%20BLOG%20%5B127%5D/1591 (accessed 1 January 2010).
Langridge, D.W. (1992), *Classification: its Kinds, Elements, Systems and Applications (Topics in Library and Information Studies)*, Bowker-Saur, London.
Noy, F. and McGuinness, D. (2001), "Ontology Development I0I: A Guide to Creating Your First Ontology", available at: http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf (accessed 1 January 2010).
O'Neill, E.T., Childress, E., Dean, R., Kammerer, K. and Vizine-Goetz, D. (2001), "FAST: Faceted Application of Subject Terminology", available at: http://www.oclc.org/research/activities/fast/dc-fast.doc (accessed 1 January 2010).
Pepper, S. (2004), "The TAO of Topic Maps - Finding the Way in the Age of Infoglut", available at: http://www.ontopia.net/topicmaps/materials/tao.html (accessed 1 January 2010).
Ranganathan, S. R., Palmer, B. I. and Association of Assistant Librarians. (1960) *Elements of library classification : based on lectures delivered at the University of Bombay in December 1944, and in the schools of librarianship in Great Britain in December 1956,* 2nd ed.*,* London: Association of Assistant Librarians.
Rosenfeld, L. and Morville, P. (2006), *Information Architecture for the World Wide Web*, 3rd ed., O'Reilly, Farnham.
Schwarz, K. (2005), "Domain model enhanced search - a comparison of taxonomy, thesaurus and ontology", unpublished thesis, University of Utrecht, available at: http://homepages.cwi.nl/~media/publications/masterthesis_kat_domainmodel_2005.pdf (accessed 1 January 2010).
Shirky, C. (2005), "Ontology is overrated: categories, links, and tags", Writings, available at: http://www.shirky.com/writings/ontology_overrated.html (accessed 1 January 2010).