AUTOMATIC DOCUMENT-LEVEL SEMANTIC METADATA ANNOTATION USING FOLKSONOMIES AND DOMAIN ONTOLOGIES

By

Hend S. Al-Khalifa

A thesis submitted for the degree of Doctor of Philosophy

In the

Faculty of Engineering and Applied Science

School of Electronics and Computer Science,

University of Southampton,

United Kingdom.

June, 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

AUTOMATIC DOCUMENT-LEVEL SEMANTIC METADATA ANNOTATION USING FOLKSONOMIES AND DOMAIN ONTOLOGIES

by Hend S. Al-Khalifa

The last few years have witnessed a fast growth of the concept of Social Software. Be it video sharing such as YouTube, photo sharing such as Flickr, community building such as MySpace, or social bookmarking such as del.icio.us. These websites contain valuable user-generated metadata called *folksonomies*. *Folksonomies* are ad hoc, light-weight knowledge representation artefacts to describe web resources using people's own vocabulary. The cheap metadata contained in such websites presents potential opportunities for us (researchers) to benefit from.

This thesis presents a novel tool that uses folksonomies to automatically generate metadata with educational semantics in an attempt to provide semantic annotations to bookmarked web resources, and to help in making the vision of the Semantic Web a reality. The tool comprises two components: the tags normalisation process and the semantic annotation process. The tool uses the del.icio.us social bookmarking service as a source for folksonomy tags.

The tool was applied to a case study consisting of a framework for evaluating the usefulness of the generated semantic metadata within the context of a particular eLearning application. This implementation of the tool was evaluated over three dimensions: the quality, the searchability and the representativeness of the generated semantic metadata. The results show that folksonomy tags were acceptable for creating semantic metadata. Moreover, folksonomy tags showed the power of aggregating people's intelligence.

The novel contribution of this work is the design of a tool that utilises folksonomy tags to automatically generate metadata with fine gained and extra educational semantics.

Contents

List of Figures

List of Tables

Acknowledgements

The accomplishment of this thesis would never have come to light without the support of many people. In particular, I would like to thank:

My parents: for always being there to love me, for always reminding me that I am valuable, and for providing immeasurable support throughout my lifetime;

My Brother: for his support and making the PhD dream comes true;

My Advisor: for being my supporter over the last couple of years, for directing me and giving me the ability to investigate the topic of my interest;

My colleagues in LSL: Will, Asma, Onjira, Joe, Lester, Dave, Lauren and all other staff and colleagues who made may stay at the University of Southampton a rich experience.

King Saud University: for their scholarship and the Saudi cultural attaché for their services and support.

Teresa Jones and Tina Sullivan: for proofreading my thesis in its early stages.

And *finally*, to all who have crossed my path during my lifetime, given me spiritual or intellectual guidance, or have otherwise challenged me to think in new ways, I say thank you.

Definitions and Abbreviations Used

ARIADNE	Alliance of Remote Instructional Authoring and Distribution Networks for Europe
AICC	Aviation Industry CBT Committee
ADL	Advanced Distributed Learning Initiative
API	Application Programming Interface – a software interface that allows web applications to exchange data
CSS	Cascading Style Sheets
DC	Dublin Core
FOAF	Friend Of A Friend
MERLOT	Multimedia Educational Resources for Learning and Online Training
LOM	Learning Object Metadata
IMS	Instructional Management Systems
IEEE	Institute of Electrical and Electronic Engineering
RDF	Resource Description Framework
RSS	Real Simple Syndication
OWL	Ontology Web Language
Folksonomy	Is a blend of the words Folks + Taxonomy, which is a neologism for a practice of collaborative categorisation using freely chosen keywords.
Social Software	Let people connect or collaborate by use of a computer network.
Web 2.0	A term often applied to a perceived ongoing transition of the World Wide Web from a collection of websites to a full-fledged computing platform serving web applications to end users.

Metadata Elements, Fields and Descriptors are terms used interchangeably throughout this thesis to mean the same thing.

Chapter 1
Introduction

1.1 Research Overview

Metadata standards are used in many areas such as: library science, database systems and file systems. They can be defined as formal specifications used to semantically annotate electronic materials of any kind. They have been developed to support both machine interoperability (information exchange) and resource discovery by human users (Stratakis et al., 2003).

The importance of metadata has also evolved to include the domain of the Semantic Web. At the heart of the Semantic Web is the idea of adding formal metadata that describes the content, context and/or structure of a web resource (Berners-Lee et al., 2001).

Metadata are also used in the educational domain to describe learning materials (see chapter 2). There are two widely accepted metadata standards in education (Stratakis et al., 2003), namely:

1. DC (Dublin Core) educational version, and
2. IEEE-LOM (Institute of Electrical and Electronic Engineers/Learning Object Metadata).

Most eLearning developers do not adhere strictly to these standards, but prefer to use "application profiles" which more accurately reflect their application's metadata needs.

Duval et al. (2006) have defined application profiles as *"... mixing and matching metadata elements, in order to meet specific requirements for a particular context*". Examples of application profiles include CanCore[1], UK LOM[2] and ARIADNE[3].

To utilize application profiles, their elements need to be populated with appropriate descriptors. This brings us back to the main dilemmas of creating standard metadata, which are: the number of fields to be filled and the amount of time required to fill them.

A possible solution is "*Electronic Forms Must Die*" (Duval, 2004), Duval's famous slogan to evangelize the automation of metadata creation. Erik Duval, a well-known member in IEEE-LOM standardisation board, has realized the need for more automated process to create metadata so that the burden of creation can be alleviated by machines.

Despite Duval's vision of metadata automation, it is not possible within the existing standards to represent sufficiently fine grained semantic information about learning resources, which would allow the selection of appropriate learning materials from a number of resources within some domain. This drives the researcher to the use of semantic metadata techniques that employ ontologies to generate specific domain semantics.

Therefore, to remove the burden of metadata generation and to generate semantic metadata that handles particular domain semantics, the researcher proposes the use of folksonomies.

Folksonomies, as one of Web 2.0 signatures, are considered a free source of unstructured metadata. They can reveal a lot about a web resource subject, its type and possible applications. Social bookmarking services such as del.icio.us[4] are by definition good sources of folksonomies.

---

[1] http://www.cancore.org [last accessed 21/2/2007]

[2] http://www.cetis.ac.uk/profiles/uklomcore [last accessed 21/2/2007]

[3] http://www.ariadne-eu.org/ [last accessed 21/2/2007]

[4] http://del.icio.us [last accessed 21/2/2007]

The problem of metadata granularity and the need for automating the process of metadata generation are two important issues that led to the idea of using folksonomies in the process of creating semantic metadata. This realization can be exploited using the power of semantic metadata representations.

This thesis shows that folksonomies contain “good enough” indexing words that can create semantic metadata with added value. As Peterson (2006) said "*The overall usefulness of folksonomies is not called into question; just how they can be refined without losing the openness that makes them so popular*". In this work, rather than attempting to refine the tagging process, the researcher has taken the open vocabulary tags and mapped them against domain ontologies in order to derive structured semantic metadata from the folksonomy tags.

1.2 Significance of the Research

The significance of this research revolves around the following motives:

1- *Proof-of-concept*; to show that self-tagging (hereafter folksonomies) can be considered a good source of metadata to semantically annotate web resources; folksonomies can describe what a resource is about, and of which type it is (e.g. reference, slides) so that it can be used in specific fields.
2- *To benefit from the social aspect of the Web*, in other words, to harness the wisdom of the crowds. This can be achieved by customizing large social bookmark services to serve different domain requirements. In this thesis it will be the case of the educational domain.
3- Folksonomies are a new trend on the Web and their popularity is growing overtime, however, little has been written about them academically. This thesis will try and explore one aspect of folksonomies, using them to create semantic metadata, and report the results of the approach to the community.

1.3 Research Hypotheses

The hypotheses of this thesis can be stated as follows:

1. Folksonomies can be used in the process of semantic annotation of web resources; this implies the following sub-hypothesis:

   a. Folksonomies, as index keywords, hold more semantic value than keywords automatically extracted by machines.
   b. Searching by folksonomies mapped to ontologies retrieve more web resources than searching by folksonomies alone.
   c. Folksonomy annotations cover more contextual dimensions than a human subject-expert does.
2. Fine-grained metadata elements' values come from The Long Tail[5].

1.4 Research Scope

Figure 1.1 gives a snapshot of the various technologies utilised in this thesis.

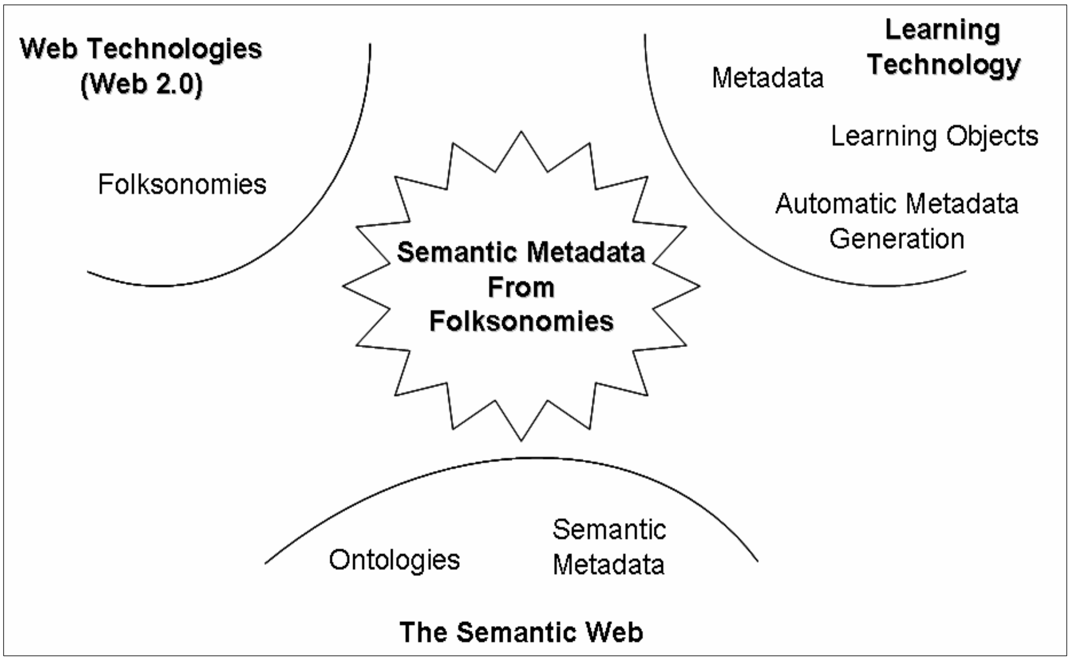


Figure 1.1: The research scope

From the Web 2.0 domain, the thesis exploits folksonomies, the light weight knowledge representation artefacts used in most contemporary web applications.

From the Semantic Web domain, the thesis employs the power of ontologies to generate semantic metadata using folksonomies.

---

[5] A theory that states "in statistical distribution the accumulated minority can be more important than the simple majority" (Grimes and Torres, 2006).

From the learning technologies domain, the thesis tries to fill the gap of automatic metadata generation for learning resources by utilising the domain of Web 2.0 and the Semantic Web to generate semantic metadata that automatically annotates learning resources.

1.5 Contributions

This thesis provides an empirical work for converting the unstructured folksonomy tags into structured semantic metadata. With this, the researcher believes that her key contributions can be highlighted as follows:

I. **Proposal for an Arabic LOM**: in the early days of researching the domain of metadata, the researcher discovered a lack of an Arabic metadata application profile that fulfils the functional requirements of the Arab region. This gap initiated the AraCore application profile initiative (Al-Khalifa and Davis, 2005).

II. **A comprehensive survey of the state-of-the-art folksonomy research**: in chapter 3, the researcher has compiled a comprehensive listing of the state-of-the-art folksonomy research covering different research themes.

III. **A model for identifying the semantics of Cascading Style Sheets (CSS)**: the researcher has created an ontology that captures the main semantics of the CSS domain.

IV. **A tool to convert folksonomy tags into semantic metadata**: this tool is comprised of two main modules: the first module is the normalisation pipeline which also introduces two new techniques, one to disambiguate tags senses and the other is to compute the ranking of the web resource based on people who have bookmarked the resource. Both techniques provide a novel mechanism to utilise folksonomy tags for different purposes. The second module is the semantic annotation pipeline which generates semantic metadata from the normalised tags.

V. **Accepted publications**: the researcher has successfully published ***nine*** *conference papers* (Al-Khalifa and Davis, 2006a; Al-Khalifa and Davis, 2006b; Al-Khalifa and Davis, 2006c; Al-Khalifa and Davis, 2006g; Al-Khalifa and Davis, 2006h; Al-Khalifa and Davis, 2006i; Al-Khalifa et al., 2007b; Al-Khalifa and Davis, 2007c) and ***two*** *journal articles* (Al-Khalifa and Davis, 2006j; Al-Khalifa and Davis, 2007a) and *withdrawn* one conference paper (Al-

Khalifa and Davis, 2006f) and two conference posters (Al-Khalifa and Davis, 2006d; Al-Khalifa and Davis, 2006e).

1.6 Outline of the Research chapters

Metadata, Learning Objects, Folksonomies, the Semantic Web, Ontologies and Social bookmarking system will be reviewed in this thesis. Each topic will lay a foundation for this research. In addition, references to existing and related work are made throughout this thesis.

Chapter 2 introduces the readers to metadata definition, types, principles, how are they generated and their purpose. Also, a brief discussion about application profiles, the relationship between metadata and the semantic web and the use of educational metadata will be outlined. Learning objects is another topic in this chapter, where the definition, the level of granularity and their relationship to the Semantic Web are discussed.

Chapter 3 discusses the concept of collaborative tagging (aka folksonomies), their types, reasons behind people tagging, and pros and cons. This is followed by a comprehensive review of state-of-the-art research, workshops, theses and case studies tackling folksonomies. The chapter concludes by discussing the gaps envisioned in the previewed folksonomy research and the potential direction this thesis can make to add a new contribution to the field.

Chapter 4 discusses social bookmarking systems, their definition, their architecture and characteristics; in particular, this chapter analyses the del.icio.us bookmarking service. The chapter also tackles the dissection of tagging patterns in del.icio.us users for a particular domain of interest. The chapter ends by making a comparison between social bookmarking services versus search engines.

Chapter 5 sheds some light on the Semantic Web, ontology languages, their design principles and the use of ontologies in the educational domain (Al-Khalifa and Davis, 2006c).

Chapter 6 introduces the readers in some detail to the different platforms, frameworks and tools used for semantic annotations. It also highlighted some important guidelines and requirements that need to be considered when designing an annotation tool for an eLearning domain.

Chapter 7 describes the experiment that has been carried out to compare the semantic value of folksonomies against automatic keyword extraction (Al-Khalifa and Davis, 2006b; Al-Khalifa and Davis, 2006g; Al-Khalifa and Davis, 2006j; Al-Khalifa and Davis, 2007a). This chapter starts by discussing similar works that have compared folksonomy tags to other indexing mechanisms. The setup of the experiment and data set selection are explained. Finally, the chapter concludes by reporting and discussing the results of the four experiment phases.

Chapter 8 describes the process of designing and developing the FolksAnnotation prototype tool (Al-Khalifa and Davis, 2006a; Al-Khalifa and Davis, 2006h). The processes involved in the tool are primarily described according to the following implementation layers: tags normalisation process and semantic annotation process. The tags normalisation process is responsible of cleaning the noise in tags assigned to a web resource. The semantic annotation process operates by mapping normalised folksonomy tags to ontology instances to create the semantic metadata. The chapter also discusses the implementation of the Tags Sense Disambiguation algorithm, which solves the problem of multiple meanings for a given tag. Finally, the chapter concluds with an example of generated semantic metadata and the general heuristic used to distinguish related CSS web resources from non-related ones.

Chapter 9 describes the process of designing and modelling the thesis domain ontologies. The chapter also presents the semantic metadata descriptors used to describe CSS resources and elaborates on their functionalities.

Chapter 10 presents a comprehensive framework to evaluate the usefulness, the quality, the searchability and the representativeness of the generated semantic metadata. For each phase of the evaluation a detailed analysis of the outcomes are accompanied with a thorough discussion. The evaluation starts with highlighting some descriptive statistics. This is followed by evaluating the semantic metadata searchability (Al-Khalifa and Davis, 2007b). Two broad searchability techniques are embraced in this phase: Browsing and Semantic Search. Then the semantic metadata assignment evaluation was carried out. This phase involvs two key measurements which are: evaluating the semantic metadata representativeness and evaluating the semantic metadata quality and validity (Al-Khalifa and Davis, 2007c). The chapter ends with analysis of unused folksonomy tags and inspects tags falling in the Long Tail region.

Chapter 11 presents an overview of related work in the area of automatic metadata generation (Al-Khalifa and Davis, 2006i). The chapter focuses on surveying systems that produce automatic metadata for educational purposes. The systems are categorised into three groups based on the type of the resultant metadata, which includes: standard metadata, semantic metadata and folksonomic metadata. The chapter discusses each system and compares and contrasts its functionalities with the thesis tool. Finally, the chapter concludes by summarising the main characteristics for each related system.

Chapter 12 concludes by reviewing the thesis key points and linking them to the achieved findings. The chapter also discusses the shortcomings of the thesis tool and suggests various enhancements. The chapter ends with some future research directions and open research questions.

1.7 Declaration

This thesis is based upon the work undertaken by the author within a collaborative research environment. It is all the original work of the author, except where explicitly stated otherwise.

Chapter 2
Metadata and Learning Objects

2.1 Introduction

In the past, metadata was often neglected and treated as a second-class citizen. However, once the computer era emerged and people started using computers to store their data, the need for techniques to retrieve these data from computers was established. Since then the metadata concept has evolved in the computer science paradigm, starting from the simple file systems (file names and types) in the early 60s, then database management systems (to describe database fields) in the early 70s, until the 21st century with the advent of the concept of metadata warehouses (Arun, 2004).

Metadata can take many forms and formats, they can be applied electronically to documents, applications and web services, or they can be presented physically such as the margins in a textbook. Metadata can also be expressed in a wide range of languages (formal or natural) by using a wide range of vocabularies (Corcho, 2006).

Metadata is a record that consists of structured information about a resource; it can be also defined as information about information or data about data; and it is structured in a manner that facilitates the management, discovery and retrieval of resources. Another useful definition for metadata is given by (Haase, 2004) as "*any data which conveys knowledge about an item without requiring examination of the item itself.*"

A metadata record typically consists of a set of elements (fields) which describe in detail the content of the resource, such as its intellectual property rights, and its 'instantiation' (e.g. date created) (LTSO, 2004).

In this chapter, metadata types, principles, applications and purposes will be discussed. Also a glance into metadata in education and The Semantic Web will be given. Finally, a short discussion about learning objects and their types will be presented.

2.2 Metadata Types

Metadata can be as simple as a set of keywords or as complex as a structured record. In principle, there are three types of metadata: descriptive, structural and administrative metadata (NISO, 2004).

Descriptive metadata describes what a resource is about to foster discovery and identification (e.g. title, author and keywords). Structural metadata describes how resources are related (e.g. how chapters are structured in a book). Administrative metadata describes how a resource can be managed (e.g. creation date, file type and who is allowed to access the resource).

Similarly, looking into the literature of metadata and its evolution (Al-Khalifa and Davis, 2006c), metadata can be classified based on recent research into:

1. **Standard metadata**: those are formal specifications used to semantically annotate materials of any kind. They have been developed to support both machine interoperability (information exchange) and resource discovery by human users (Stratakis et al., 2003). Examples include Dublin Core (DC) and IEEE-LOM.
2. **Semantic metadata**: "…*the process of attaching semantic descriptions to Web resources by linking them to a number of classes and properties defined in Ontologies*" (Scerri et al., 2005). More on semantic metadata in chapter 6.
3. **Attention metadata**: "… *concerns collecting detailed information about the relation between users and the content they access*." (Najjar et al., 2006).

Attention metadata uses the AttentionXML[6] open standard to track user interaction with web applications such as Blogs, Wikis, news, etc. The collected data from log files includes information about the user's preferences, context, goals and interests (Najjar et al., 2006). Najjar et al. is working on extending AttentionXML in order to collect rich data from eLearning applications. Their new attention schema is called CAM (Contextualized Attention Metadata) and it is used to collect and merge attention metadata of users from different educational tools.

2.3 Metadata Principles

Another important aspect of metadata is its underlying principles. Duval et al. (2002) have defined the principles in metadata context as: "*concepts that are judge to be common to all domains of metadata and which might inform the design of any metadata schema or application*". Applying these principles will provide the guidelines for developing practical solutions for semantic and machine interoperability for any domain using any metadata standard.

The first principle is *modularity*, which is a key organizing principle for managing multiple sources of content in metadata. It allows metadata schema designers to assemble data elements from different schemas rather than reinventing anew elements. They also benefit from vocabularies as well as other building blocks by combining them in a syntactic and semantic way to leverage interoperability.

The second principle is *extensibility*; this means that metadata schemas must be flexible enough to accept the addition of new elements to accommodate application needs. This also implies the notion of a base schema that has the basic elements which can be exchanged by different applications and the notion of local schema that has additional elements that tailor a given application to local or domain specific needs.

The third principle is *refinement*, which means the appropriate level of detail a metadata might have for a given application. This applies two notions which are:

[6] http://developers.technorati.com/wiki/attentionxml [last accessed 24/2/2007]

first, the addition of qualifiers that makes the meaning of an element more specific such as using the word author, composer or illustrator to mean the more general term creator. Such a refinement may be useful for a specific metadata application. The second is the use of controlled vocabulary to specify the value range of a given element. As an example, the use of an encoding standard to encode dates and times will remove ambiguity from the expression of a metadata value. The string 03/06/02 is interpreted as March 6, 2002 in North America and June 3, 2002 in Europe and Australia. So by using a standard such as W3C date and time format a date can be encoded in unambiguous manner (2002-03-06).

The fourth principle is *multilingualism*, which means adopting metadata architectures that respect linguistic and cultural diversity. This issue can be handled by a process called internationalization and localization. The former process relates to the creation of neutral standards while the latter refers to the adaptation of the neutral standard in a local context. There are many techniques proposed to apply multilingualism in metadata architecture one of these techniques is to translate relevant specification and standards documents into a variety of languages. Another proposed technique by CWA 14643, a CEN workshop agreement, is the use of universal canonical identifier as the encoding format for the data elements names in metadata (CEN, 2003). This approach will allow the exchange of metadata records made by cataloguing systems in different languages.

2.4 Metadata Purposes and Applications

Nowadays, metadata is important for indexing and describing what a resource is about. The applications of metadata extend the borders of a simple description of a resource to a vast variety of flavours. Cataloguing, content rating and electronic commerce are names of applications in which metadata plays an important role in their functionality (Lassila, 1998). Facilitating interoperability, digital identification and organizing e-resources are also other ways to use metadata (NISO, 2004).

2.5 How Is Metadata Generated/Created?

Sometimes the process of generating metadata and the economics of doing so is considered problematic. From a processing side, metadata can be generated either manually, semi-automatically or automatically; each of which requires a considerable amount of time and effort. From an economical viewpoint, generating metadata is an expensive process especially if done by an expert.

In the era of the World Wide Web metadata can be created in three ways (Mathes, 2004):

1- By professional librarians based on a specific scheme like the Dewey Decimal Classification System (DDCS). This approach has its drawbacks, such as the cost, the scalability and the complication of the tools used to create metadata.
2- By document author(s) using the Dublin Core schema. This approach has the problem of inadequate or inaccurate description of the resource.
3- By document users (chaotic), also known as user-created metadata, where the users of the material create metadata for their own use and for the community to share with (Chapter 3 and 4 will discuss this issue in depth).

2.6 Metadata and the Semantic Web

The Semantic Web is the vision of Tim-Berners Lee for making the Web 'machine understandable' (the Semantic Web will be discussed in chapter 5). Each layer of the Semantic Web has a connection to metadata or support for it (Greenberg et al., 2003).

For instance, URI, a major component of the base layer of the Semantic Web, is considered a metadata; as an analogy, it functions as an ISBN (International Standard Book Number) to a book. Furthermore, ontologies (another layer in the Semantic Web) are metadata systems, where domain concepts are defined and linked together using relations. This implies that metadata is a key player in the Semantic Web architecture, and to drive the Semantic Web to its potential, the challenge of creating semantic metadata needs to be addressed (hence the theme of this thesis).

2.7 Educational Metadata

A number of organizations are involved in producing metadata standards specifically for learning technologies. A list of the major ones includes: ADL[7], AICC[8], ARIADNE[9], IEEE LTSC[10] and IMS[11] (Robson, 2000; Redeker, 2003).

An educational metadata record extends the scope of regular metadata. It adds further fields to the metadata so that it describes information that has particular educational relevance (Recker and Wiley, 2001). Within education there are several groups focused on defining metadata structures specially designed to describe learning objects. For instance, ARIADNE education metadata is comprised of six categories, which are: general, indexation, annotation technical, semantic and pedagogical (Najjar et al., 2003b).

Similarly, Dublin Core Metadata Initiative (DCMI) Education[12] consists of 23 elements that resulted from adding the 15 base DC elements to the extended 8 educational specific elements.

Other important work has been conducted by IEEE standards committee; the Learning Technologies Standard Committee (LTSC). Their draft standard is called Learning Object Metadata (hereafter, IEEE-LOM) and it defines 80 fields within 9 categories as follows: 1-General, 2-Lyfecycle, 3-Meta-Metadata, 4-Technical, 5-Educational, 6-Rights, 7-Relation, 8-Annotation and 9-Classification (Recker and Wiley, 2001).

Sample fields of IEEE-LOM include:

Title	the name given to the resource.
Language	the language of the intended user of the resource.
Description	a textual description of the content of the resource.

[7] http://www.adlnet.org [last accessed 24/2/2007]

[8] http://www.aicc.org [last accessed 24/2/2007]

[9] http://www.ariadne-eu.org/ [last accessed 24/2/2007]

[10] http://ieeeltsc.org/ [last accessed 24/2/2007]

[11] http://www.imsproject.org [last accessed 24/2/2007]

[12] http://dublincore.org/groups/education/ [last accessed 24/2/2007]

Brasher and McAndrew (2004) state that metadata like IEEE-LOM is a collection of specified terms of descriptors related to particular aspects of the resource being described. The descriptors can be one of two distinct categories of sources: intrinsic sources, which are contained within the resource itself and compose a necessary part of it (e.g. title of a resource); extrinsic sources, which are not contained within the source itself (e.g. personal or organizational view about the expected use of the resource such as the difficulty field in IEEE-LOM).

2.8 Application Profiles

As more and more applications are implemented using educational metadata, it becomes obvious that it would be difficult for a single metadata model to accommodate the functional requirements of all applications (Chatzinotas and Sampson, 2004). This has created the need for what are known as application profiles.

Sampson (2004) defines application profiles as "*an assemblage of metadata elements selected from one or more metadata schemas.*" Thus, an application profile will serve as an adaptor of a particular metadata schema or multiple schemas and it will be tailored to the functional requirements of a particular application taking into account interoperability with the original base schemas (Sampson, 2004). Likewise, Duval et al. (2006) defined application profiles as " ... *mixing and matching metadata elements in order to meet specific requirements for a particular context*."

Among the well-known application profiles is the UK LOM Core[13], an optimized version of IEEE-LOM standard designed for the use within the context of UK education. Also CanCore[14] the application profile used in Canada.

Table 2.1 shows some examples of the major application profiles along with their base scheme, number of elements and an enumeration of the educational elements field.

---

[13] http://standards-catalogue.ukoln.ac.uk/index/UK_LOM_Core [last accessed 24/2/2007]

[14] http://www.cancore.ca/en/ [last accessed 24/2/2007]

Standard	Base Scheme	Number of elements	Educational elements
Education Network Australia[15] (EdNa)	DC	23	Type, curriculum, document, event, audience, spatial
Gateway to Educational Materials[16] (GEM)	DC	23	Audience, format, grade, language, pedagogy, object type, subject
CanCore	IEEE LOM	30	Interactivity type, learning object type, semantic density, intended end-user role, context
UK LOM Core	IEEE LOM	46	Interactivity type, learning object type, interactivity level, semantic density, intended end-user role, context, difficulty, relation kind, purpose

Table 2.1: Major educational metadata application profiles [from (Qin and Hernández., 2006)]

From the previous table, it can be seen that there is an apparent lack of consensus among the vocabulary used in the education element for the different application profiles. As well the wording of the elements was not agreed upon as if GEM 'subject' element and CanCore 'context' element means the same thing.

The inconsistency of vocabulary use in the previous application profiles, although they have inherited much of the structure and semantic of the base schema, requires extra processing and mapping to convert from one application profile to another as experienced by (Najjar et al., 2003a). However, one solution the researcher has found invaluable is to use Resource Description Framework (RDF) (as will be

[15] http://www.edna.edu.au/edna/go [last accessed 24/2/2007]

[16] http://www.thegateway.org/ [last accessed 24/2/2007]

discussed in chapter 5) to represent metadata records in a more flexible and scalable manner.

The use of RDF as a preferable format for representing metadata can be justified by reading the seminal paper entitled "*Semantic Web Metadata for e-Learning - Some Architectural Guidelines*" by (Nilsson et al., 2002). Nilsson et al. highlighted some major differences between XML schema, which most standard application profiles use, and RDF schema. One important difference is that XML schema describes the syntactic structure of XML documents, while RDF schema describes the semantics of a vocabulary that can be reused in any setting. Moreover, when creating application profiles using XML, for each new application requirement the developer needs to create a new application profile, while in the case of RDF, for each new application requirement the developer needs just to add an extra RDF statement without the problem of reconstructing the RDF schema. These were just two samples of the benefit of RDF over XML, and for more about this topic the reader is referred to (Nilsson et al., 2002).

2.9 Issues Associated with Educational Metadata

By skimming through research that utilises standard metadata in the eLearning domain, the researcher has found that most researchers were unsatisfied with the capabilities provided by educational standard metadata. Among these recent complaints:

- "*The problem with metadata information like IEEE-LOM or IMS is mainly number of fields to fill (more than 50 fields) and the amount of time a user has to invest to describe a resource*" (Yin et al., 2003).
- "*LOM has a deficiency in semantic-awareness capability*" (Lee et al., 2006).
- "*… educational attributes of LOM are very difficult to produce*" (Motelet and Baloian, 2006).
- "*… LOM and SCORM, have emerged to annotate and package learning content. But they mainly deal with technical aspects and do not express much information about pedagogy*" (Dehors and Faron-Zucker, 2006).

Given these reasons and the variability of the educational elements used in most application profiles (as mentioned in the previous section), the researcher was inclined to create a new application profile using RDF format and is based on some of the IEEE-LOM elements. In addition, the thesis application profile has extends IEEE-LOM to include fine grained semantics that serves the functional requirements (Duval et al., 2006) of the thesis case study.

2.10 Learning Objects

There is currently a lack of common definition for learning objects, as many people and groups try to come up with a definition that suits their own needs. This caused a proliferation of terms, meanings, and definitions related to learning objects, and for a full discourse of learning objects definition the reader is referred to (Wiley, 2002). However, two of the well-known definitions are presented.

The IEEE defines a learning object as "*any entity, digital or non-digital, which can be used, reused or referenced during technology supported learning*" (LTSC, 2002). Similarly, Wiley (2002) has refined the IEEE definition to become "*Any digital resource that can be reused to support learning*".

From the previous definitions the researcher can understand that learning objects are small pieces of instruction with an educational objective that can be reused in various instructional contexts. In other words, a learning object is any resource with an explicit educational application. It can be digital, for example, a simple Microsoft Word, PDF, or text file, an e-book, or a Flash animation. Or it can be physical like a textbook or a CD-ROM. But the concern will be on digital representation of learning materials, due to the fact that they can be easily distributed and shared using a network, while physical learning materials do not have this capability (Stratakis et al., 2003).

However, in this thesis, the researcher will use the term *learning resource* instead of *learning object* to refer to any resource that can be helpful in the educational process. The rational for opting to use the term learning resource instead of learning object can be highlighted in the following motives. First, the term *resource* which refers to "*... anything that, for whatever reason, someone has found necessary or useful to* *describe*" (Downes, 2004) is more universal than the term *object*, which presupposes a specific type of software entity. Learning object also promotes the idea of self-explanatory and self-contained learning material. Secondly, most learning objects repositories such as MERLOT[17] have their learning objects as Web pages; this indicates a liberal and non-strict nature of Web pages compared to what is understood about the nature of learning objects. Since the concept of learning objects is still vague, and to avoid unfruitful discourse on the meaning of this concept, the researcher has chosen to use the term 'learning resource' throughout this thesis.

2.11 Taxonomy of Learning Objects Types

Wiley (2002) distinguishes between five types of learning objects:

- *Fundamental* – a single digital resource that stands by itself, e.g. image.
- *Combined-closed* – a small number of digital resources that interact with each other in whole and can not be modified e.g. a video with accompanying audio.
- *Combined-open* – a large number of digital resources that interact with each other in whole and can be modified, e.g. a web page dynamically combining the fundamental type and the combined-closed type.
- *Generative-presentation* – a digital resource that creates presentations for use in instruction, e.g. "*a JAVA applet capable of graphically generating a set of staff, clef, and notes, and then positioning them appropriately to present a chord identification problem to a student*".
- *Generative-instructional* - a digital resource that creates presentations and instructs and provides practice for any type of procedure e.g. giving a student a process to perform in a series of steps.

In this thesis, the learning resources that will be semantically annotated can be of any type of the previously mentioned classifications.

[17] http://www.merlot.org/merlot/index.htm [last accessed 24/2/2007]

2.11.1 Level of Granularity for Learning Objects

Learning objects can vary in size from a single slide in a PowerPoint presentation to a whole certificate program as has been discussed in the previous section. Thus, to deploy, reuse or author a learning object its level of granularity needs to be defined.

There are different levels of granularity for learning objects and many papers such as (Duval and Hodgins, 2003), (Redeker, 2003) and (Stratakis et al., 2003), have tried to define the boundaries between these levels. However, the issue remains fuzzy and it is hard to achieve consensuses due to the different perspectives of learning object authors and pedagogical specialists.

2.12 Learning Objects and the Semantic Web

The current set of elements in the IEEE-LOM standard is not sufficnet for intelligent discovery and assembly of learning objects. To verify this each learning object needs to specify how it is related to concepts in a particular domain and also clarify the types of learning outcomes possible in that domain (i.e. the need for an ontology). With this kind of knowledge Web agents can search and retrieve learning objects more intelligently (Mohan and Brooks, 2003). Further discussion about Semantic Web and ontologies in education will be addressed in Chapter 5.

2.13 Chapter Summary

This chapter has overviewed both metadata and learning objects as key players in learning technologies discipline. Metadata is used to describe learning objects for easy retrieval and discovery. Also, this chapter discussed the importance of metadata in the Semantic Web, which will be a major theme in this thesis.

Finally, the research in the area of metadata standards and application profiles has resulted in a proposal for an initiative to create the first Arabic metadata application profile called AraCore (Al-Khalifa and Davis, 2005).

Chapter 3

Collaborative Tagging

With the rapid explosion of information that is proliferating on the Web, it is not possible to create a professional set of metadata for resources without spending an enormous amount of time, money and effort. However, with the advent of contemporary services on the web that use an intuitive mechanism for describing web resources, people started participating in annotating their own resources in a process called *social tagging*. This process is easy to indulge in since it does not require any professional background; all that is needed is to freely choose keywords from an individual's vocabulary to annotate a web resource. This process of annotation has converted people into metadata generators.

This chapter, starts by discussing what is meant by the process of tagging and illustrates the different names used for this process. Next, the different types of tagging are explained in a section called 'types of folksonomy'. A discussion about some of the pros and cons of using folksonomies is highlighted, and before ending this chapter an outline of the state of the art of research on folksonomies is reviewed; as a new research discipline.

3.1 What is tagging?

Tagging is a simple, grassroots, ad hoc classification scheme and manual indexing mechanism that does not require any further processing; as opposed to the process of arranging resources into categories (Sinha, 2005; Tosic and Milicevic, 2006; Voss, 2007). The tag terms are selected from a flat namespace without any hierarchy, reflecting what a user thinks is the appropriate term to describe a resource. Notice that the tags' namespaces are user created and are usually uncontrolled.

There are many successful contemporary services on the Web that foster the concept of tagging. These include del.icio.us[18], flickr[19] and furl[20], to name but a few. In addition, tagging services fall into more specialised categories, like social bookmarking (e.g. de.icio.us), photo-sharing (e.g. flickr), and community-based news websites (e.g. Digg[21]), etc. Tags also play a prominent role in the Windows Vista OS, as reported on the Microsoft website[22]. Also, Amazon[23] is asking its customers to use tags for annotating its commodities (e.g. books, toys, etc.) and Google[24] is using tagging in its GMail[25] service.

Tagging can have other names that can be used interchangeably to mean the act of people assigning descriptions to resources, among these are: mob indexing, folk categorisation, social tagging, federated tagging, lazy tagging, folksonomy, tagsonomy, tagonomy, free tagging, distributed classification, post coordinate indexing, collective indexing, user-generated tagging and ethnoclassification (Hammond et al., 2005). However, the widely accepted and popular word is folksonomy; therefore, this term will be used throughout this thesis.

3.1.1 Why do people tag?

Hammond et al. (2005) have identified the motivation for tagging in four regions as shown in Figure 3.1. The figure splits the tagging players into a horizontal axis which denotes the creator of the content (either one or more) and a vertical axis which refers to the users of the generated tags.

---

[18] http://del.icio.us [last accessed 18/2/2007]

[19] http://www. flickr.com [last accessed 18/2/2007]

[20] http://www.furl.net [last accessed 18/2/2007]

[21] http://www.digg.com/ [last accessed 18/2/2007]

[22]http://www.microsoft.com/windows/products/~/productivity.mspx [last accessed 18/2/2007]

[23] http://www.amazon.com/gp/tagging/cloud [last accessed 18/2/2007]

[24] http://www.google.com [last accessed 18/2/2007]

[25] http://www.gmail.com [last accessed 18/2/2007]

Region 1 (self, self) represents an individual who is tagging his/her own content for their own benefit (as content creators and consumers) without taking into consideration the use of others; an example of such a tagging habit is evident in the Flickr photo-sharing service.

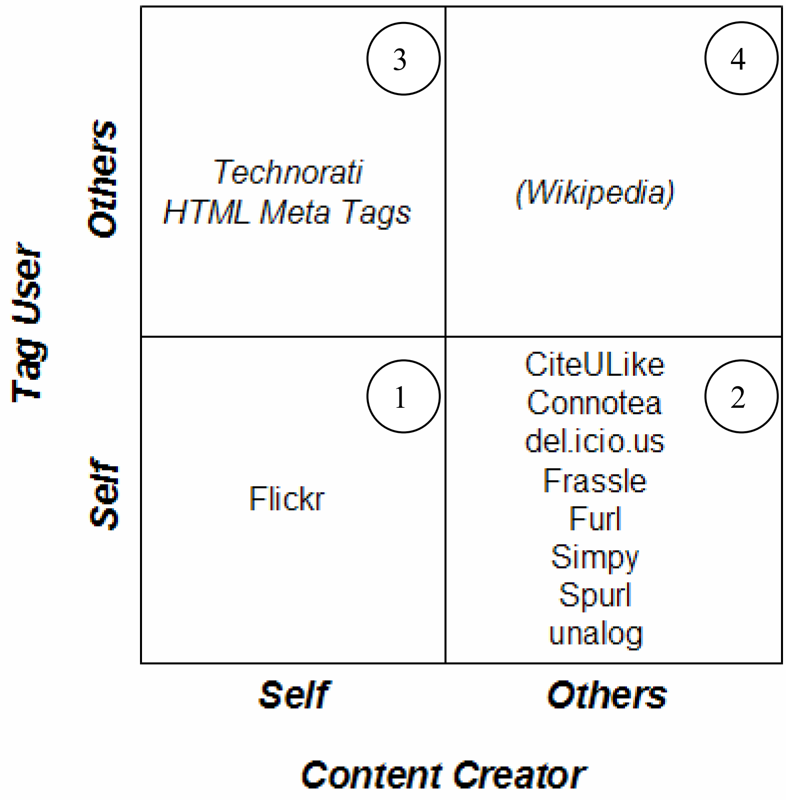


Figure 3.1: The four regions for the Motivation of Tagging alongside some examples of services that satisfy that motive [by (Hammond et al., 2005)]

Region 2 (others, self), represents an individual who is tagging others resources for his/her own use. An example of such a service is the social bookmarking system del.icio.us.

Region 3 (self, others) represents an individual who is tagging his/her own content for the benefit of other people. An example of this act is Technorati[26] service, an Internet search engine for searching blogs.

Region 4 (others, others) represents people who are tagging others resources for others to use. A well-known example is the Wikipedia[27] website.

[26] http://technorati.com/ [last accessed 18/2/2007]

[27] http://www.wikipedia.org/ [last accessed 18/2/2007]

In summary, the researcher believes that people practice the act of tagging to satisfy two motives: 1) find-ability, to be able to return to stuff that has been previously stored 2) knowledge organisation, to be able to organise content they find interesting.

3.1.2 Folksonomy: A Definition

The term folksonomy is a blend of the words folks and taxonomy; this term was first coined by the information architect Thomas Vander Wal in August of 2004. Folksonomy as (Wikipedia, 2007) defines is:

> *"... a user generated taxonomy used to categorize and retrieve Web pages, photographs, Web links and other web content using open ended labels called tags. Typically, folksonomies are Internet-based, but their use may occur in other contexts as well. The process of folksonomic tagging is intended to make a body of information increasingly easier to search, discover, and navigate over time. A well-developed folksonomy is ideally accessible as a shared vocabulary that is both originated by and familiar to its primary users."*

Also, Vander Wal (2007) has defined folksonomy as:

> "*… the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information. The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object. People are not so much categorizing, as providing a means to connect items (placing hooks) to provide their meaning in their own understanding.*"

One major feature of folksonomy tags is that they follow a power law[28] distribution (see Figure 10.13); power laws follow an asymptotic distribution[29] where the distribution curve approaches but never meets or crosses a given line or axis. Most used tags reflect consensus among users accessing a particular resource (Mathes,

[28] http://en.wikipedia.org/wiki/Power_law [last accessed 18/2/2007]

[29] http://en.wikipedia.org/wiki/Asymptotic_distribution [last accessed 18/2/2007]

2004), however, there will be tags that are used by only one or two users (Guy and Tonkin, 2006).

3.2 Folksonomy Types

Thomas Vander Wal has categorised folksonomies into two groups (Vander Wal, 2005): broad and narrow (Figure 3.2).

Broad folksonomy: is the result of many people tagging one resource (e.g. del.icio.us.) It is useful to know the tags that are agreed and preferred by most people to describe one item. This kind of agreement will help in finding out what is the common or emerging vocabulary of a community of people using such resources.

On the other hand, a *narrow folksonomy* is the result of one person tagging an item, or a smaller number of people tagging items (e.g. Flickr). While the goal and use of the narrow folksonomy differs from the broad folksonomy, the degree of visibility is higher in the broad folksonomy than the narrow one. In addition the narrow folksonomy loses the richness of the mass (Quintarelli, 2005).

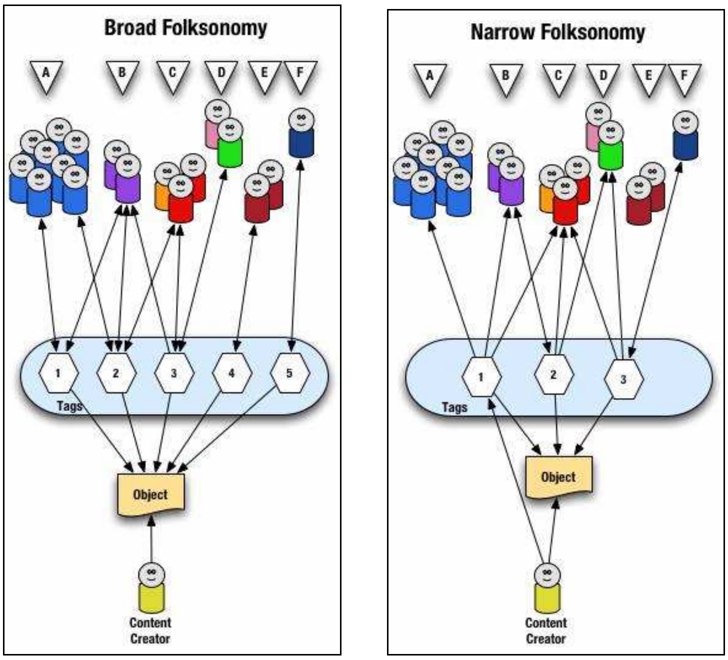


Figure 3.2: The illustration on the left depicts broad folksonomy while the one on the right depicts narrow folksonomy [by Vander Wal, 2005]

3.3 Folksonomies: Pros and Cons

There has been more debate concerning the value of folksonomies than there is space in this thesis to report. However, this section will try to shed some light on several of the arguments that have been raised about the strengths and weaknesses of folksonomies.

Kroski (2005), Quintarelli (2005) and Mathes (2004) discuss some of the folksonomy strengths as follows:

1. *Folksonomies are inclusive* - they can represent everyone's vocabulary, not like top-down taxonomies where it is restricted to a controlled vocabulary.
2. *Folksonomies are current* - people create tags as quickly as they create content, not like taxonomy classification where it takes varying time to classify content.
3. *Folksonomies are Non-Binary* - which means that when someone wants to categorise an item it can fit it in multiple categories, unlike traditional classification where an item can exist only in one place; this implies that folksonomies are multi-faceted.
4. *Folksonomies offer insight into user behaviour* - folksonomies help observe how people tag their own resources. Most of the time, people tags are considered subjective due to the personal nature of tagging. However, some people might find others' tags interesting; such an example will be to know what people have in their reading list.

In contrast, Guy & Tonkin (2006), Kroski (2005) and Mathes (2004) have summarised some of the weaknesses of folksonomies in the following points:

1. *Folksonomies are imprecise* - since the tags are added by the users of folksonomy sites (e.g. Flickr); the tags are usually ambiguous, personalised and inexact.
2. *Folksonomies are single-word metadata* -many folksonomy sites allow only single-word tags which result in many useless compound terms.
3. *Lack of synonym and homonym control* - synonym (different word, same meaning) and homonym (same word, different meaning), are rarely supported (in the case of synonyms) or never supported (in the case of homonyms) in

folksonomy sites. This result in a chaotic set of uncontrolled tags which do not support search.

4. *Folksonomies have a lack of recall* - due to the lack of synonym control, folksonomy search will not yield all results that have similar tags. For instance, the search for 'cat' will not return resources tagged with kitten, cats, etc.

Most of the previously mentioned advantages (4) and disadvantages (1, 3, and 4) will be evident in Chapter 10, when the researcher interprets the thesis experiment results.

3.4 Folksonomy and Taxonomy

From a categorisation perspective, folksonomy and taxonomy can be placed at the two opposite ends of categorisation spectrum. The major differences between folksonomies and taxonomies are discussed thoroughly in (Shirky, 2005) and (Quintarelli, 2005), although, a short discussion will be presented here.

Taxonomy is a top-down approach. It is a simple kind of ontology that provides hierarchical, domain specific vocabulary which describes the elements of a domain and their hierarchal relationships. Moreover, taxonomies are created by domain experts and librarians, and require an authoritative source.

In contrast, a folksonomy is a bottom-up approach. It does not hold a specific vocabulary nor does it have an explicit hierarchy. It is the result of people's own vocabulary, thus, it has no limits (i.e. open ended), and tags are not stable nor comprehensive. Moreover, folksonomies are generated by people who have spent their time exploring and interacting with the tagged resource (Wikipedia, 2006).

3.5 Folksonomy and the Semantic Web

One of the main debates is "*how to use folksonomies in the Semantic Web?*" in fact this question is what this thesis aims to answer (see Chapter 5).

Even though there is no definite answer on how a folksonomy can be used in or with the Semantic Web, the role of this thesis is to explore a potential area and try and to come up with a solution to this problem. The researcher's proposed answer is that folksonomies may be an enabler for developing semantic metadata; this will be demonstrated throughout this thesis.

Moreover, the knowledge implicitly held in folksonomy tags has lead to what has been known as *emergent semantics*, which "*result from the converging use of the same vocabulary*" (Hotho et al., 2006c). Paolillo and Penumarthy (2007) show how emergent semantics are present in folksonomy tags by saying "*Folksonomies are said to support emergent classification, where the semantic value of the tags and their relation to one another is worked out through a negotiated process of users applying their selected tags and seeing what others have tagged the same way*." Moreover, the researcher can demonstrate that by further exploring the emergent semantics in folksonomies, light-weight ontologies can be created; this will be illustrated in the next section.

3.6 State-of-the-Art Folksonomy Research

During the past couple of years, folksonomy research has gained a lot of attention from library and computer science researchers; also its themes have proliferated and spanned different disciplines. In this section, a comprehensive summary of the key research, workshops, theses and case studies are provided along with a short description of the academic research to date.

3.6.1 Research

In this section, the researcher tries to provide a short description of the academic research to date by classifying it based on the themes each has tackled.

Overview research:

Social bookmarking tools in general with special emphasis on folksonomies are described in (Hammond et al., 2005) and the strengths and weaknesses of folksonomies are discussed in (Mathes, 2004), (Quintarelli, 2005) and (Guy and Tonkin, 2006). Also, Marlow et al. (2006) provided a taxonomy of folksonomy systems, which was based on system design, attributes and user incentives. In system design, they talked about the dimensions of tagging systems' design that may have immediate and considerable effect on the content and usefulness of the tags generated by the system (e.g. tagging rights, tagging support, etc). When discussing user incentives, they claim that users' motivations, either personally or socially, play a significant role in affecting the tags that emerge from social tagging systems. They also present a preliminary analysis of tag usage within the photo-sharing and tagging system 'Flickr' to suggest potential future directions of research in tagging systems. Similarly, Wu et al. (2006) have proposed some enhancements that need to be considered when designing collaborative tagging systems. They also highlighted some key challenges encountered while building collaborative tagging systems and have developed a comprehensive evaluation methodology to be used in assessing the construction of collaborative tagging systems.

Ontology creation research:

A study by Mika (2005) has been carried out to construct a community-based ontology using del.icio.us as a data source. He created two lightweight ontologies out of folksonomies; one is the actor-concept (i.e. user-concept) ontology and the other is the concept-instance ontology. The goal of his experiment was to show that ontologies can be built using the context of the community in which they are created (the del.icio.us community). Despite the innovative approach that Mika follows, this thesis has not considered building ontologies from folksonomies. By the same token, Tom Gruber is working on a system called TagOntology to build ontologies out of folksonomies, and in his paper entitled "*Ontology of Folksonomy: A Mash-up of Apples and Oranges*" he casts light on some design considerations needed to be taken into account when constructing ontologies from tags (Gruber, 2005).

In addition, Ohmukai et al. (2005) proposed a social bookmark system, called 'socialware', using several representations of personal networks and metadata to construct a community-based ontology. The personal network was constructed using FOAF[30], RSS[31], and simple RDFS[32] formats, while folksonomies were used as the metadata.

---

[30] Friend Of A Friend

[31] Rich Site Summary

[32] To be discussed in chapter 5

Their system allows a user to browse friends' bookmarks on his/her personal network, and map their own tag onto more than one tag from multiple friends, so that they are linked by the user. This technique will allow for efficient recommendation for tags because it is derived from personal interest and trust. They also used their social bookmark system to design an RDF-based metadata framework to support open and distributed models.

Christiaens (2006) devised a mechanism to convert folksonomy tags into a taxonomy and then combine them with ontologies. The process of creating a taxonomy was not explicitly clear in his paper; however, the author claimed that trying this approach in a system called Guide proved valuable. His idea originated from the need to bridge the gap between restricted vocabulary (i.e. ontologies) and free vocabulary (i.e. folksonomies).

Folksonomy patterns, linguistics and analysis research:

Golder and Huberman (2006), from HP Labs, analysed the structure of collaborative tagging (aka folksonomies) to discover the regularities in user activity, tag frequencies, the kind of tags used and bursts of popularity in bookmarked URLs in the del.icio.us system. They also developed a dynamic model that predicts the stable patterns in collaborative tagging and relates them to shared knowledge. Their results show that a significant amount of tagging is done for personal use rather than public benefit. However, even if information is tagged for personal use, other users can benefit from it. They also state that del.icio.us, for most users, functions as a recommendation system even without explicitly providing recommendation. This argument supports the design decision that the researcher has followed when developing her annotation tool.

Sen et al. (2006) presented a user-centric model of vocabulary evolution in tagging communities based on community influence and personal tendency. They collapsed Golder's classes into three general classes and used the modified classification metric to evaluate the MovieLens recommender system. They also used four tag selection algorithms to recommend tags to users of the MovieLens recommender system and to evaluate the effect of the algorithms on vocabulary evolution, tag utility, tag adaptation and user satisfaction. The modified categorisation that Sen et al. proposed was used in this thesis to evaluate the unused folksonomy tags that were not utilised in the process of semantic annotation.

An experiment in MIT labs was carried out by (Liu et al., 2006) to generate a 'taste fabric' of social networks. Folksonomies were used in the experiment to weave the taste fabric. Their idea was based on philosophical and sociological theories of taste and identity to weave a semantic fabric of taste. They mined 100,000 social network profiles, segmented them into interest categories and then normalised the folksonomies in the segments and mapped them into a formal ontology of identity and interest descriptor (Liu et al., 2006). Their work supports the researcher's idea of using folksonomies in the process of semantic annotation.

Kipp and Campbell (2006) analysed the tagging patterns revealed by users of the del.icio.us bookmarking service. Their aim was to assess how collaborative tagging supports and enhances traditional indexing and classification of documents. They used Multi-Dimensional Scaling (MDS) for co-word clustering. This approach helped them visualise the relationship between tags for a given URL. Thus, their findings show that tagging practice to some extent mimics the ways used to classify documents in conventional indexing systems.

Veres (2006) presented a study in which the linguistic properties of folksonomies demonstrated that users engaged in resource tagging are performing classification according to principles similar to formal taxonomies. To prove his findings, Veres analysed the kinds of classification observed in user tags using the non-taxonomic categories proposed by the linguist Anna Wierzbicka. He then compared users' patterns to those observed for two well known sources of classification schemes on the Internet: the open directory project (DMOZ) and the Yahoo directory. His findings showed that there is a clear difference between folksonomy tags and the two classification schemes. Tags are drawn from most categories while DMOZ and YAHOO were biased only towards one category (namely functional category). Also Veres (2006b) used folksonomies to model concepts in a domain. He used a method, based on the linguistic properties of the tags, to extract structural properties of free-form user tags to construct an ontology. The resultant ontology is a simple conceptual domain model built from automatically mediated collaboration; this ontology has been used to facilitate interoperability between application-dependent tag sets.

Folksonomy statistical research:

Hotho et al. (2006a) presented a new search algorithm for folksonomies, called ***FolkRank***, which exploits the structure of the folksonomy tags. Their proposed algorithm is used to support the retrieval of resources in the del.icio.us social bookmarking services by ranking the popularity of tags. They demonstrated their findings on a large-scale data set (around 250k bookmarked resources) and showed that their algorithm yielded a set of related users and resources for a given tag. Therefore, 'FolkRank' can be used to generate recommendations within a folksonomy system. In the same vein, Dubinko et al. (2006) introduced the '*interestingness'* algorithm, which is based on the characterisation of the most interesting tags associated with a sliding interval of time. They experimented with a large number of tags in the Flickr online photo-sharing community to visualise the interesting tags over time. The Dubinko et al. *interestingness* algorithm was used by (Hotho et al., 2006b) to rank the interesting resources in the del.icio.us bookmarking service for an interval window size of one month. They compared the results of the *interestingness* algorithm to the results of the FolkRank algorithm and found that, the *interestingness* algorithm is more sensitive to temporary changes in folksonomy tags than FolkRank. In contrast, FolkRank algorithm was more useful for long-term observations.

Both the Hotho et al. and Dubinko et al. proposal for computing a recommendation (ranking) value from folksonomy tags seems practical and very robust, however, their underlying algorithms were very complicated and they require large data sets to come up with reasonable values. These two requirements have put off the researcher from trying to use either algorithm in computing the recommendation value proposed for the folksonomic semantic metadata, as will be seen in Chapter 8.

Similarly, Szekely and Torres (2005) from Harvard University have developed a system called "gourmetvillage.org" that uses folksonomies as a vehicle for sharing and classifying information in order to evaluate restaurants. The system is based on two algorithms: 'UserRank' and 'TagRang'. Szekely and Torres define **UserRank** as "*… an algorithm based on Google's PageRank that provides a ranking of users* *based on whose taggings are most often followed*" and they define **TagRank** as "*…provides a ranking of tags based on the ranking of users.*" Their technique and their implemented system are not adequate for the current thesis objective, because their system relies on user popularity which is not a factor in this thesis. Also, their system targets a certain community with a special interest.

Moreover, Zhang et al. (2006) have used a bottom-up approach to semantically annotate web resources using folksonomies. The semantics of the folksonomies were statistically inferred using the asymmetric Separable Mixture Model (SMM) statistical model for data co-occurrence to resolve the ambiguity of folksonomy tags. Their model has succeeded in resolving tags' ambiguity and in grouping synonym tags together. In the same way, Cattuto et al. (2006) have studied the semantic breadth of tags by investigating the statistical properties of tag co-occurrence. Their findings provided social bookmarking systems an improved tag suggestion during search or navigation.

Folksonomy visualization, search and recommendation research:
Choy and Lui (2006) provided a possible way to navigate through the tagging space by understanding the semantics of tagging. They proposed the use of Self-Organisation Map (SOM) technique to visualise multi-dimensional data onto a 2D map. SOM helps provide a graphical map that reveals important information in tagging space for the users of the collaborative tagging systems. Similarly, Russell (2006) has developed an online visualisation tool called (Cloudalicious[33]) that gives insight into how folksonomies are developed over time for a given URL. Also, Aurnhammer et al.(2006) used a tag visualisation technique to ease the exploration of image databases.

Han et al. (2006) experimented with an exploratory system that uses folksonomy tags to enhance searching. They integrated Google's search functionality and the URL check functionality provided by del.icio.us to provide adaptive guidance to users. As the user uses the exploratory system interface to input his/her keywords, the system sends the keywords to Google and also to del.icio.us to extract the corresponding tags. Then the keywords contained in the tags are displayed as clickable hints. Their

[33] http://cloudalicio.us/ [last accessed 18/2/2007]

preliminary evaluation showed that their technique has increased the accuracy of search.

Niwa et al. (2006) have constructed a new web recommender system that is not limited to particular websites. Their system was based on large amount of public bookmarked data on social bookmarking systems. The system utilises folksonomy tags, by clustering them, to classify web pages and to express users' preferences. The evaluation results showed that the precision rate of their system was about 40% to 60%.

Folksonomy visualisation, search and recommendation are not among the key themes in this research; however, the researcher has introduced them to show the diverse range of research that folksonomies are involved in.

Finally, two unpublished research works by (Shen and Wu, 2005) and (Lambiotte and Ausloos, 2005) were carried out to study the nature of collaborative tagging as a complex network. The topic of complex networks is too advanced and it is outside the scope of this thesis.

3.6.2 Workshops

In the World Wide Web 2006 conference, a workshop under the name "Collaborative Web Tagging"[34] explored the social and technical issues and challenges involved in Web tagging. The themes covered by the workshop ranged between tagging in the enterprise (Farrell and Lau, 2006; John and Seligmann, 2006), mining the TagSpace such as tags clustering (Begelman et al., 2006) to improve search and tags suggestions (Xu et al., 2006), tags applications such as tags in museums (Trant and Wyman, 2006) and tag visualisation (Dennis, 2006).

Similarly, the 17th SIG/CR Classification Research Workshop[35] (CRW) was dedicated to research about social classification. The topics covered in this workshop ranged from comparing social tagging with subject cataloguing as in (Tennis, 2006),

---

[34] http://www.rawsugar.com/www2006/taggingworkshopschedule.html [last accessed 18/2/2007]

[35] http://www.slais.ubc.ca/users/sigcr/events.html [last accessed 18/2/2007]

or with controlled vocabulary and title-based automatic indexing as in (Lin et al., 2006), or finding the relationship between the Semantic Web and social tagging as in (Campbell, 2006). Also the workshop had a panel about social classification of visual resources, where it discussed the use of social tagging in museums and photo-sharing sites.

In spite of the diverse themes and topics tackled in the both workshops, the WWW06 and CRW, none of these have proposed any possible usage of folksonomy tags in the domain of eLearning; for example, using folksonomy tags to create semantic metadata for annotating learning resources. However, these workshops gave the researcher an insight of the line of research that both computer science and library science researchers are embracing.

3.6.3 Case Studies

Elke Michlmayr has conducted a case study on the properties of metadata provided by folksonomy; her domain of research was in social networks (Michlmayr, 2005). In her paper Elke provided an in-depth study of the properties of tags produced by folksonomies. She investigated how metadata produced by folksonomies can serve as simulation data in peer-to-peer environments. To accomplish her goal she developed a method for selecting subsets of folksonomies tags, from the del.icoi.us bookmark service, that adhere to the principle of interest-based locality. Her result shows that folksonomies can be applied for simulating peers and their content in peer-to-peer environment.

Another case study was carried out by (Lawrence and Schraefel, 2006) on the amateur fiction community. The study analysed how folksonomies evolve inside these communities and considered how ontologies and folksonomies can be used together to add the easy usability of free tagging to ontology descriptions and the richness of conceptual ontologies to folksonomies.

3.6.4 Thesis

In a master degree thesis completed by (Bielenberg and Zacher, 2005), a prototype system called "GROOP.US[36]" was developed to explore the potential of social navigation based on tagging. The system uses the del.icio.us bookmarking service to visualise a person's social context based on the tags (s)he has used. It shows a person's areas of interest and relates people with similar interests to groups with shared tags and resources. The underlying concepts that Bielenberg and Zacher used in their thesis lend themselves to four areas, which are: 1) metadata, 2) social networks, 3) the three levels of context (i.e. individual context, social context and shared context) and 4) social navigation; further discussion regarding these areas can be found in (Bielenberg and Zacher, 2005).

Another master thesis dealing with social bookmarking and personal organisation on the web was completed by (Trevino, 2006). She has investigated social bookmarking in detail by focusing on the users of del.icio.us. She used in-depth interviews along with content analysis to discover more about how people understand the information in del.icio.us and the implications of the site's structure. Her findings suggest that people use del.icio.us as a sign of what web resources they value and as a memory aid. She also suggests that posts of others, when aggregated, provide an insight into what interests del.icio.us users have as a group.

3.6.5 Discussion

From the previous overview the reader can observe that most research on folksonomies is either user-centric e.g. (Mika, 2005) and (Ohmukai et al., 2005) or tag-centric e.g. (Hotho et al., 2006a) and (Choy and Lui, 2006). Little research has been addressed towards the URL-centric perspective, which this thesis tackles. By URL-centric the researcher means that knowledge about a specific URL is constructed from the tags associated with it. Similarly, most research concentrated on either understanding the underlying statistical model e.g. (Zhang et al., 2006) or mathematical model of folksonomies e.g. (Shen and Wu, 2005), or interpreting and visualising social tags e.g. (Russell, 2006), or trying to structure tags to create

[36] http://groop.us/ [last accessed 18/2/2007]

ontologies e.g. (Christiaens, 2006), or even knowing who is using social tagging e.g. (Trevino, 2006).

In addition, research on folksonomy tags came mostly from computer science and library science; there is an apparent lack of input from sociology, psychology and cognitive science (Voss, 2007).

Finally, the previous overview has highlighted an escalation in folksonomy research during 2006 where most thorough folksonomy research has taken place. This means that the area of folksonomy research will gain more attention in the forthcoming years.

3.7 Chapter Summary

Folksonomy is still a nascent technology. As more systems embrace this approach, by developing applications to use tagging, better solutions will occur over time.

Users usually tag a web resource to help them to find it later. They use social bookmarking systems to discover interesting web pages they have not seen before. Also, they usually tag a resource differently from the resource author/creator; this may imply that user tags can be useful (e.g. in eBay things get a reputation from the number of people tagging an item).

This chapter has also presented a comprehensive summary of the key research, workshops, theses and case studies carried out in the academic research to date.

In the next chapter, the del.icio.us bookmarking service, which has been referenced throughout this chapter, will be studied thoroughly.

Chapter 4
Social Bookmarking Services

Unlike simple browser-based bookmarking functionality, the new trend of bookmark management services is based on server-side web applications; where people can save their favourite links for later retrieval.

People usually use bookmarks to save web resources that they feel useful or interesting, or as Millen et al. (2005) said "… *people create bookmarks based on the quality and personal interest of the content, high frequency of current use and a sense of potential for future use*". Bookmarks started as a service integrated within the browser then as the web evolved new services have emerged which enabled anyone to easily save, access and share their bookmarks from anywhere, using any browser. These services are referred to as 'social bookmarks' to signify their social and shareable nature.

A plethora of bookmarking services already exists, be it general purpose services (e.g. del.icio.us[37], Furl[38], Simpy[39] and Ma.gnolia[40]), enterprise services (e.g. dogear (Millen et al., 2005)), or reference management services (e.g. BibSonomy[41] (Hotho et al., 2006c), Connotea[42] and CiteUlike[43]).

---

[37] http://del.icio.us [last accessed 15/2/2007]

[38] Furl (File Uniform Resource Locators) http://www.furl.net/ [last accessed 15/2/2007]

[39] http://www.simpy.com/ [last accessed 15/2/2007]

[40] http://ma.gnolia.com/ [last accessed 15/2/2007]

[41] http://www.bibsonomy.org/ [last accessed 15/2/2007]

[42] http://www.connotea.org/ [last accessed 15/2/2007]

This chapter, however, is dedicated to analyzing the del.icio.us social bookmarking service for two main reasons: 1) del.icio.us is the largest social bookmarking service on the web; since its introduction in December 2003, it has gained great popularity and there are more than 90,000 registered users using the service and over a million unique tagged bookmarks (Menchen, 2005; Sieck, 2005); and 2) del.icio.us shares the same characteristics and underlying concepts that other social bookmarking services use, such as: tagging, web-based storage and the social nature of these Web applications (Millen et al., 2005).

Therefore, this chapter will start with a comprehensive overview of the del.icio.us service; then an anatomy of the tags stored within the del.icio.us service will be carried out. Finally, the chapter will conclude with a brief comparison between bookmarking services and search engines.

4.1 The del.icio.us Social Bookmarking Service

Every day hundreds of URLs are bookmarked online using the del.icio.us bookmarking service. Each bookmarked URL is accompanied by a line of text describing it and a set of tags assigned by people who bookmarked the web resource (as shown in Figure 4.1).

Figure 4.1: Excerpt from the del.icio.us service showing the tags (Blogs, internet, ... ,cool) for the URL of the article by Jonathan J. Harris, the last bookmarker (pacoc, 3mins ago) and the number of people who bookmarked this URL (1494 other people).

Visitors and users of the del.icio.us service can browse the bookmarked URLs by user, by keywords (tags) or by a combination of both techniques. By browsing others' bookmarks, people can learn how other people tag their resources; thus, increasing their awareness of the different usage of the tags. In addition, any user can create an inbox for other users' bookmarks, by subscribing to the other user's

[43] http://www.citeulike.org/ [last accessed 15/2/2007]

del.icio.us pages. Also, users can subscribe to RSS feeds for a particular tag, group of tags or other del.icio.us users.

4.1.1 The del.icio.us Data Model

The del.icio.us data model is composed mainly of three interconnected components, as shown in Figure 4.2, which are: URLs, tags and users.

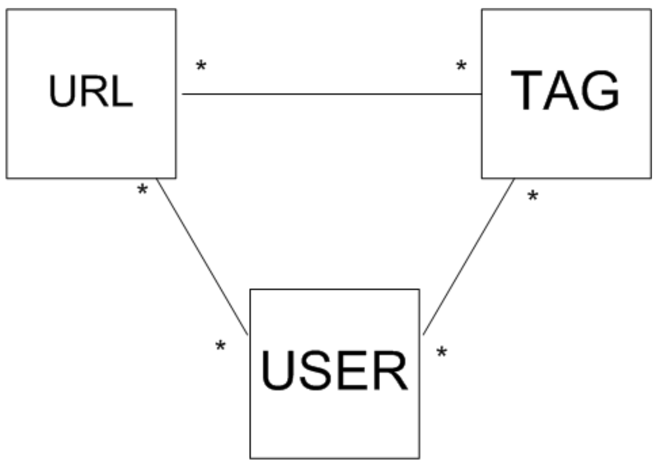


Figure 4.2: The relation between the three del.icio.us components

URLs are the main assets of the del.icio.us service. A bookmarked URL can have multiple tags and can be bookmarked by many users. It can also point to a website, a Word file, PDF document, a Video, Audio or Flash resource.

Bookmarked URLs cover a variety of topics such as web development, media, business and entertainment. Over time, tags are accumulated depending on the number of people who bookmarked the same URL. Each bookmarked URL is listed in a backward-chronological order.

In the del.icio.us service, what makes a URL so valuable is the fact that tags have been assigned to it. Tags can be treated as kind of metadata; they can tell what a resource is about without further investigation. So, as a URL gains more popularity overtime, the bookmarking service can be thought of as a collaborative information filtering (i.e. recommendation or voting) system for the best web resources on the web.

Tags are one-word descriptors that are assigned to a URL. A Tag is usually associated to one or multiple URLs. The process of annotation (aka tagging) is straightforward. A user is presented with one line text box where (s)he can type in tags. Each tag can be entered using a white space as a delimiter; which restricts the use of single keywords. Users can get around this problem by using punctuation such as hyphens or underscores, and some might prefer to combine more than one word by using camel case format (e.g. OpenSource).

When tags are accumulated for a given URL, they appear in a portion called 'common tags' (see Figure 4.3). This indicates the level of agreement on vocabulary and meaning within the underlying community who bookmarked the resource.

Figure 4.3: A Screenshot showing the 'common tags' portion

It is worth mentioning that during the annotation process the system provides the user with suggested tags where the most popular tags used for the currently bookmarked URL are displayed (see Figure 4.4).

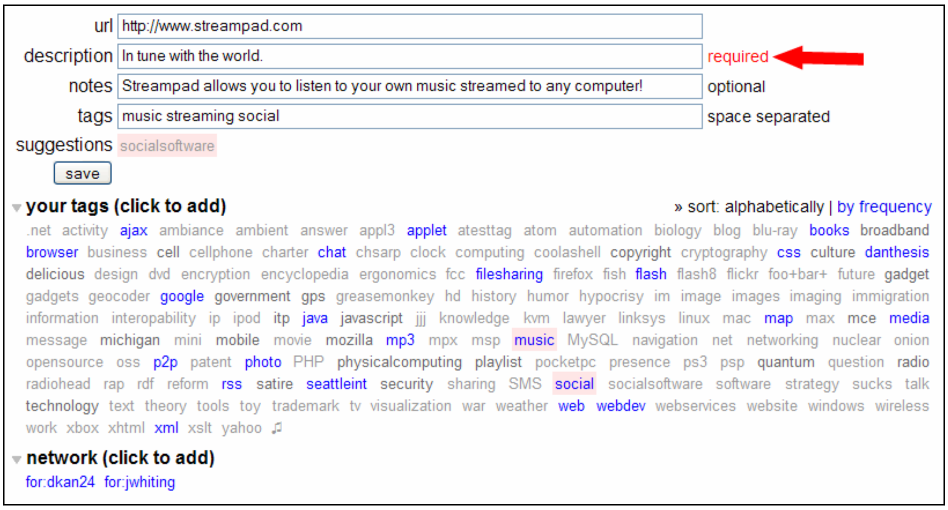


Figure 4.4: A Screenshot showing the ‘suggestions’ field and the ‘tags’ box

Tags listed under a URL can be clicked; this will take the user to a page which lists all the URLs given the same tag. The page will also display a list of ‘related tags’ that have been used with the given tag, but in a different context (see Figure 4.5).

Figure 4.5: A Screenshot showing the ‘related tags’ portion

Users are the engine of the del.icio.us service. With their social efforts del.icio.us has been widely used. del.icio.us provides each user with his/her own page that shows his/her bookmarked web resources displayed in a chronological order together with the associated tags. The web page also list all the tags used by the user.

In a pilot research by (Menchen, 2005) to identify the occupation of the del.icio.us users, she found that the predominant occupations for a sample of the del.icio.us users were in the information technology industry and education or research. Another indicator of the IT nature of URLs bookmarked in del.icio.us is an experiment

carried out by (Wu et al., 2006) to study the semantics of tags; they found that most web bookmarks collected in the del.icio.us bookmarking service were mainly from the field of IT.
Both findings are valid and can be easily verified by browsing the popular tags page, where most tags are in the domain of Web technologies.

4.1.2 Anatomy of the del.icio.us Tags

(Golder and Huberman, 2005; Guy and Tonkin, 2006) have both analysed the kind of tags used in the del.icio.us service. For instance, Golder and Huberman have identified seven functions tags can perform for the bookmarked URLs which are: identifying what it is about, identifying what it is, identifying who owns it, refining categories, identifying qualities or characteristics, self reference and task organisation. In addition, Wash and Rader (2006) analysed the del.icio.us service extensively, and compared the tags that were applied to a site with the actual contents of the site webpage. They found that the average site has only 26% of its tags appearing in the webpage at all. As a result they cite this as evidence that the tags provide useful metadata that is not directly available in the webpage.

Based on the previous overviews, the researcher has conducted her own analysis to identify the main aspects of social tagging. Thus, after analysing the tags used over time in the del.icio.us service, the researcher found valuable information regarding tagging as a process and how people develop folksonomies.

People usually perform four actions when tagging a URL. They tag a URL so they describe what it is about – its 'about-ness' (e.g. programming), or/and to show how to use the URL or what to do with it (e.g. to read), or/and to describe its usefulness (e.g. cool, good) and/or describe its type (e.g. tutorial), as shown in Figure 4.6. Moreover, people usually give different tags to mean the same concept (e.g. ontology vs. topic maps); this might be because they use the same concept in different context.

Figure 4.6: An excerpt from the del.icio.us service, showing that the resource "jQuery: New Wave Javascript" is about 'Programming' in 'Ajax', and the person who bookmarked it will going to read it 'toRead' and (s)he describe how useful the resource was ('Cool') and defines the type of the resource as being a 'library'

Many patterns have been observed after analysing people's vocabulary, which include:

- Specific Words with distinct meaning (e.g. CSS, Mac, OS).
- Acronyms and abbreviations (e.g. UI means User Interface, CompSci means Computer Science).
- Compound words or phrases (e.g. computerscience, computer_sceince, computer.science or ComputerScience).
- Misspelled tags.
- Singular and plural.
- Synonyms.
- Capitalisation (e.g. CSS or Css or css).
- Non-English tags and symbols (e.g @site).

In other words, people vocabulary can be categorised into:

- Domain specific tags, either broad or narrow (e.g. *broad*: Programming, *narrow*: Javascript).
- Type of a resource (e.g article, tutorial, reference)
- Subjective (opinion or expression) that provides judgment-related context (e.g. fun, funny, cool)
- Attitudes, functional tags (e.g. toread, 2read, tovisit, learn-later)
- Colloquial phrases and localisation (Motive, 2005).
- Others that only make sense to the tag creator. Hence, people usually tag for themselves (WeBreakStuff, 2005; Stock, 2006).

Furthermore, some users of del.icio.us have adopted a private convention to indicate the tag's hierarchy (i.e. structural relation between tags e.g. Dev/Perl). Also, another known approach is the use of tag bundles where users tag a set of tags to create hierarchical folksonomies (Hammond et al., 2005).

4.1.3 Users' Patterns in del.icio.us for a Domain of Interest

One requirement of this thesis is to develop domain specific ontologies (ontologies will be discussed in Chapter 5). Therefore, a thorough analysis of the specific domain of interest, in particular Web Design and Cascading Style Sheets (CSS), was carried out.

The researcher was inclined to investigate the domain of Web Design and CSS for the following reasons: 1) the popular[44] tags in the del.icio.us service came from the domain of Web Design; hence, most bookmarked web resources are from this domain. 2) Since the researcher comes from a Computer Science background she wants to acquire knowledge from her area of expertise.

After analysing bookmarked entries over time in the researcher's domain of interest, the researcher observed three main patterns of tagging: general purpose for Web design, domain-specific for the CSS domain, and the type of the web resource. Sometimes the quality of the web resource is explicitly expressed (e.g. cool). These patterns (general, domain concepts and resource types) will then be considered as the thesis ontologies (will be discussed in Chapter 9).

Figure 4.7 shows an example of a set of tags that illustrate the emerging three patterns; the (A) rectangle represents concepts from the general domain ontology, the (B) rectangle represents concepts from the resource type ontology, and the (C) rectangle represents concepts from the specific domain ontology. Notice that the expression 'Useful' was used to represent the quality of the resource.

[44] Note that Web Design and CSS were among the popular tags in del.icio.us, http://del.icio.us/tag/ [last accessed 15/4/2005]

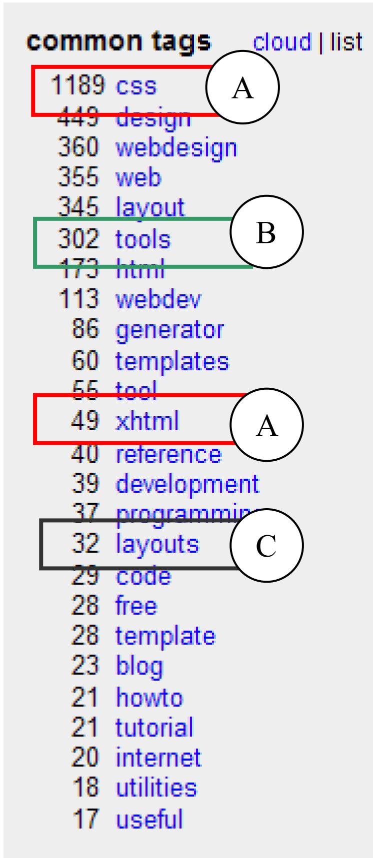


Figure 4.7: Set of Tags assigned to a website with the title "Layout-o-matic[45]"

The researcher has also taken into consideration the importance of the number associated with each tag; hence the number represents how many people have used that tag. Detailed discussion about this topic and the design decisions will be addressed in Chapter 9.

4.1.4 Social Bookmarking Services versus Search Engines

A comparison between search engines and social bookmarking services is not quite equitable since each system provides a different service. On one hand, search engines are purely machine-centric but on the other hand, social bookmarks are purely human-centric. However, one problem with search engines comes from the results they give. Search engine results usually include 'noise' in the form of unrelated results. Their results differ from tag search results, as search engines are not based on user-assigned keywords.

[45] http://www.inknoise.com/experimental/layoutomatic.php

Moreover, despite the success of some search engines, they still suffer from coverage bias (Vaughan and Thelwall, 2004). In a study by Vaughan and Thelwall to check the search results of some search engines (Google[46], AllTheWeb [47]and Altavista[48]), they found that big search engines have coverage biased toward the U.S.A. Another noticeable deficiency is that search engines do not 'know' the top issues for a given day nor what the priorities are for people searching the web (Brunton, 2006).

Tags used in social bookmarking services, on the other hand, help reduce noise experienced in search engine results. Besides, social bookmarks are good for discovering new web resources (serendipity). However, up to writing this thesis the researcher has not heard of any serious spam[49] problems in social bookmarking service despite their usage of spam defence armour[50], except for minor incidents such as the one reported by (Stutzman, 2006) and (Bosworth, 2006). But, as soon as social bookmarks' spam issue becomes of a greater concern, the underlying thesis work would need to be re-designed to handle such an issue.

To sum up, the researcher reasons that both social bookmarking services and search engines are two ends in the spectrum of "find-ability". In terms of website coverage, search engines such as Google are more dominant than social bookmarking services, given the size of Google database index compared to the size of del.icio.us bookmarked links database. In terms of content identification, search engines fail to identify hidden aspects of non-textual web resources; in contrast, social bookmarks present human understanding of the content of the bookmarked web resource. In terms of web resource importance, search engines use link popularity and sophisticated algorithms to compute the ranking of the search results; on the other hand, social bookmarking services rely on human attention and peer judgment, i.e. what web resources people pay attention to. In terms of currency, search engines are unlikely to be able to update their database index as fast as people bookmark time-

[46] http://www.google.com [last accessed 15/2/2007]

[47] http://www.alltheweb.com/ [last accessed 15/2/2007]

[48] http://www.altavista.com/ [last accessed 15/2/2007]

[49] http://en.wikipedia.org/wiki/Bookmark_spam [last accessed 15/2/2007]

[50] Such as the use of **CAPTCHA,** which stands for "Completely Automated Public Turing test to tell Computers and Humans Apart"

sensitive web resources. This can be witnessed from people's experience in using del.icio.us to pitch their websites e.g. (Martino, 2005).

4.2 Chapter Summary

Every day, hundreds of URLs are bookmarked using the del.icio.us service; these URLs represent what people think worth bookmarking for later use. Among the bookmarked URLs there exists some sign of web resources that can be nominated as being useful in an educational context.

To further investigate the usage and quality of folksonomies two experiments will be presented. The first experiment was carried out to explore the value of folksonomies compared to automatically extracted keywords (Chapter 7). The second experiment was carried out to use folksonomies in the process of semantic annotation (Chapter 8).

Chapter 5
The Semantic Web and Ontologies in Education

The problem of the current Web is that, it is only understandable by humans. Machines cannot interpret information on the Web as people do. To illustrate this problem, suppose a person wants to search the Web for the term ‘Apple’ by which (s)he means the fruit. The search engine will return results with no semantic relations. It may give web pages on ‘Apple’ as the computer company, or as a fruit, or even as an online shop named ‘Apple’. This ambiguity in the results needs to be solved by semantically annotating resources on the web so that intelligent results can be retrieved. This can be done by adding an extra layer of semantics to the current Web to enable machines understand what a web page is about.

The Semantic Web intends to improve the existing Web with a layer of machine-interpretable metadata so that a computer program can understand what a web page is about, therefore draw conclusions. The Semantic Web as defined by its creator Tim Berners-Lee (Berners-Lee et al., 2001) implies:

> *"... an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation."*

To add the layer of semantics to the existing web, three challenges need to be achieved (Harmelen, 2004):

- A syntax for representing metadata,
- Vocabularies for expressing the metadata, and
- Metadata for lots of Web pages.

The Semantic Web includes the following technologies (Antoniou and van Harmelen, 2004):

- *Explicit Metadata*: the Semantic Web does not rely on text-based manipulation, but on machine-processable metadata.
- *Ontologies*: an ontology can be defined as an explicit and formal specification of a conceptualization (this topic is discussed in depth in the next section).
- *Logic and Inference*: where automated reasoners can infer conclusions from the given knowledge.
- *Agents*: are computer programs that work autonomously on behalf of a person. They receive tasks to accomplish, make certain choices and give answers.

As ontologies represent a core component in the Semantic Web, this chapter will give a thorough definition of ontologies and their applications in learning technologies. Thus, ontology types, design principles, ontology languages and the different approaches to build ontologies are discussed. Finally, some applications of ontologies in education will be reviewed.

5.1 Ontologies: Definition and Design Principles

In a survey paper entitled "Exploring Ontologies", Kalfoglou gave a through definition of ontologies as follows (Kalfoglou, 2001):

> *"an explicit representation of a shared understanding of the important concepts in some domain of interest. The role of an ontology is to support knowledge sharing and reuse within and among groups of agents (people, software programs, or both). In their computational form, ontologies are often comprised by definitions of terms organised in a hierarchy lattice along with a set of relationships that hold among these definitions. These constructs collectively*

impose a structure on the domain being represented and constrain the possible interpretations of terms" (p.3).

Kalfoglou also described the criteria for the proposed principles in designing ontologies as*: clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment.*

Clarity means that the meaning of the ontology can be easily captured. This can be achieved by minimizing ambiguity and by giving examples to help the reader/user understand the ontology. Coherence means that the ontology should be internally and logically consistent. Extendibility means that the ontology terms can be extended without the revision of existing definitions. Encoding bias means that the representation is made purely for the convenience of notation or implementation. It should be minimized to enhance knowledge-sharing between agents. Finally, minimal ontological commitment means that the ontology should model the domain with fewer claims. It is worth mentioning that the previous criteria's are not potentially meet by ontology designers. Therefore Kalfoglou thinks that the main notion that needs to be tackled is the criteria of ontological commitment due to its important role in software systems (Kalfoglou, 2001).

This thesis adheres to the previous criteria as much as possible, to ensure that the generated ontologies can be reused within a wider audience.

5.2 Types of Ontologies

Ontologies can be categorized into four different types (van Heijst et al., 1997), namely: *application ontologies, domain ontologies, generic ontologies and representation ontologies*.

Application ontologies contain the knowledge needed to model a particular application. Domain ontologies express concepts that are specific for a particular domain. Generic ontologies define concepts that are generic across different disciplines. Finally, representation ontologies provide a representational framework with a neutral view of the world.

There is also another classification of ontologies based on their generality (i.e. scope) and expressiveness (i.e. level of details) (Bruijn and Fensel, 2005). In the level of generality there are three different types of ontologies: top-level ontologies (e.g. CYC[51], WordNet[52]) which are shared by many people in different domains, domain ontologies (e.g. UNSPSC[53], The United Nations Standard Products and Services Code for classifying products and services) which are shared between stakeholders in a particular domain and finally application ontologies (e.g. an ontology for a course) which are used for a particular application.

The other orthogonal classification of ontologies is based on their expressiveness. Ontologies can be distinguished by their different levels of expressiveness such as: thesaurus (e.g. WordNet), controlled vocabulary (e.g. Dublin Core[54]), informal/formal taxonomy (e.g. Yahoo directory[55]/UNSPSC), frames (e.g. RDFS), value restrictions (e.g. OWL data-type), limited logic constraints (e.g. OWL DL[56]) and general logic constraints (e.g. CyCL[57], OWL DL).

Finally, Bruijn and Fensel (2005) also mentioned that the level of expressiveness can be seen as two distinct categories: light-weight ontologies, which include the concepts and the relations between them and heavy-weight ontologies, which include axioms and constraints.

This thesis is going to focus on the use of application ontologies with a light-weight level of expressiveness.

---

[51] http://www.opencyc.org/ [last accessed 11/2/2007]

[52] http://wordnet.princeton.edu/ [last accessed 11/2/2007]

[53] http://www. unspsc.org [last accessed 11/2/2007]

[54] http://dublincore.org/ [last accessed 11/2/2007]

[55] http://dir.yahoo.com/ [last accessed 11/2/2007]

[56] http://www.w3.org/TR/owl-guide/ [last accessed 11/2/2007]

[57] http://www.cyc.com/cycdoc/ref/cycl-syntax.html [last accessed 11/2/2007]

5.3 Ontology Languages

Prior to the initiative of the Semantic Web by Tim Berners Lee, many systems existed that used different languages to represent ontologies like SCL[58], CyCL and LOOM[59] (Bruijn and Fensel, 2005). Although they offer a powerful expression and reasoning mechanism, they still lack intimate support of RDF (the key language in the Semantic Web).

To express the semantics of a resource on the Web so that humans, as well as machines, can understand it, a set of formal languages are used. These languages can be stacked on top of each other to form what Berners-Lee (2000) called "The Semantic Web Language Layer Cake".

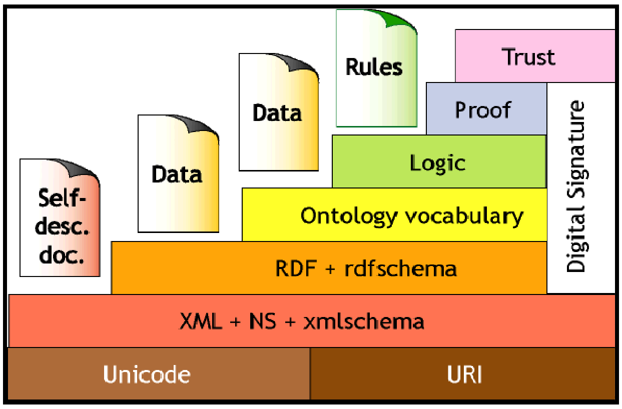


Figure 5.1: The Semantic Web Language Layer Cake [Berners-Lee, 2000][60].

Figure 5.1 depicts the layers of the Semantic Web starting with: the Unicode and URI layer which forms the base for the upcoming layers. The second layer is the XML and XML Schema, which forms the syntactical basis for the Semantic Web languages. The third layer is the RDF and RDF Schema which represents the expressive language for the Semantic Web. The next layer is OWL, which represents the ontology language for the Semantic Web. An overview of each of the three languages used in the Semantic Web is presented in the following sub-sections, with

---

58 Simple Common Logic (SCL) http://www.ihmc.us/users/phayes/SCL-december.html [last accessed 11/2/2007]

59 http://www.isi.edu/isd/LOOM/ [last accessed 11/2/2007]

60 http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html [last accessed 11/2/2007]

information derived from (Antoniou and van Harmelen, 2004), (W3C, 2001), (W3C, 2002) and (W3C, 2004).

5.3.1 XML/DTD/XML Schema

Although XML is not an ontology language; it is a core technology in the ‘Layer Cake’ and all the subsequent layers are built on top of it.

XML (eXtensible Markup Language) is an application of SGML (ISO 8879). It is a structured language and was developed due to the shortcomings of HTML.

XML is used to exchange data between web applications. To accomplish the exchange, applications need to agree on common vocabulary to support communication. From these vocabularies are MathML[61] (for mathematics) and NewsML[62] (for news).

An XML document consists of the following parts (Figure 5.2): Prolog, elements and attributes.

```
<?xml version="1.0" encoding="UTF-8"?>
<Book>
  <Authors>
     Grigoris Antoniou and Frank van Harmelen.
  </Authors>
  <title ISBN=”1234567”>
     A Semantic Web Primer
  </title>
</Book>
```

Figure 5.2: XML Snippet.

A prolog is a line that appears before the root element in an XML document. It contains XML declaration and reference to other documents (e.g. <?xml version=”1.0”?>). Elements represent the ‘things’ the XML document talks about (e.g. Authors). Finally, an attribute is a value inside the opening tag of an element (e.g. ISBN).

---

[61] http://www.w3.org/Math/ [last accessed 11/2/2007]

[62] http://www.newsml.org/pages/index.php [last accessed 11/2/2007]

An XML document is a well-formed document, which means it is restricted to a set of syntactic rules. XML structure can be defined using two methods:

- **Document Type Definition (DTD)** "old and hard": the document can be within the XML file (internal) or on a separate file (external).
- **XML Schema** "new and easy": its syntax is based on XML documents and it provides a richer language to define the structure of an XML document. The important feature in an XML schema is its ability to be reused and refined.

5.3.2 RDF/RDFS

RDF (Resource Description Framework) is a data model that consists of object-attribute-value triple called a statement. RDF triples can be expressed in different ways: by using XML syntax (Figure 5.3).

```
<rdf:Description rdf:ID=" http://www.example.com/Adam">
  <Name>Adam</Name>
</rdf:Description>
```

Figure 5.3: RDF serialization using XML.

Or using N3/Notation3[63] (Figure 5.4).

```
<http://www.example.com/Adam> <Name> "Adam" .
```

Figure 5.4: RDF in N3.

Or using binary predicate form, e.g. Property(object,value), Figure 5.5.

```
Name("http://www.example.com/Adam","Adam").
```

Figure 5.5: RDF as a binary predicate.

Or using a directed labelled graph (Figure 5.6).

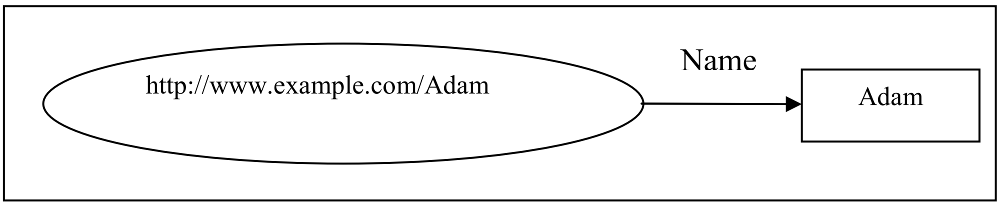


Figure 5.6: RDF as a graph.

[63] http://www.w3.org/DesignIssues/Notation3 [last accessed 11/2/2007]

RDF is a domain independent language, where no claim about a specific domain is made. This implies the need to define someone's own terminology using RDF Schema (RDFS). Also, RDF does not contain vocabulary to author metadata, thus an RDF Schema is needed to define a predefined vocabulary to be used with metadata generation.

RDFS is intended to model a domain in a hierarchical fashion which poses the problem of representing some ontological knowledge, thus RDFS is very limited. One example of this problem is the representation of the '*disjointness of classes*', sometimes we wish to say that classes are disjoint, however, RDFS is not capable of representing this relation, it can only show the subclass relationship. Due to this limitation a more expressive language is needed (hence OWL).

5.3.3 OWL

Research groups both in the United States with their DAML ontology language and Europe with their OIL ontology language, identified the need for more powerful ontology language. Their joint efforts produced what is called as DAML+OIL ontology, which then was taken as a starting point for the W3C ontology working group in developing OWL (Web Ontology Language.) This language is aimed to be the standard ontology language for the Semantic Web.

OWL comes in three different flavours namely: *OWL Lite, OWL DL and OWL Full*. Each flavour differs in its level of expressiveness and reasoning.

- **OWL Lite**: is a limited version of OWL that provides a classification hierarchy and simple constraints. Examples of OWL Lite include thesauri and taxonomies.
- **OWL Description Logic (DL)**: includes all OWL language constructs, and provides the maximum expressiveness while maintaining a finite computation time.
- **OWL Full**: includes all OWL language constructs, and provides the maximum expressiveness with no computational guarantees.

5.4 Building Ontologies

Ontologies can be generated either manually or semi-automatically (Gómez-Pérez and Manzano-Macho, 2004). Manual ontology building is a tedious, time-consuming and error-prone task. Semi-automatic building of ontologies is more appropriate for speeding up the process of ontology generation.

The process of semi-automatic generation of ontologies is usually referred as an ontology learning process, which can be defined as

> "*the application of a set of methods and techniques used for building an ontology from scratch by enriching, or adapting, an existing ontology in a semi-automatic fashion using distributed and heterogeneous knowledge and information sources, allowing a reduction in the time and effort needed in the ontology development process*" (Gómez-Pérez and Manzano-Macho, 2004, p.187).

The process of ontology learning from text includes a number of methods that came from complementary disciplines (e.g. Natural Language Processing 'NLP' and machine learning) and is applied to different types of unstructured, semi-structured, and fully structured data. These methods can be summarized as follows (Gómez-Pérez and Manzano-Macho, 2004):

- *Approaches based on linguistic techniques*: These include NLP techniques such as pattern-based extraction, semantic relativeness, etc. An example of a system using this technique is SOAT (WU and HSU, 2002).
- *Approaches based on statistical techniques*: These methods rely on calculating several statistical measures (e.g. Term Frequency Inverse Document Frequency 'TFIDF') to help the ontologist detect new concepts and the relationships between them. As an example of a system based on this technique is WOLFIE (WOrd Learning From Interpreted Examples) (Thompson and Mooney, 1999).
- *Approaches based on machine learning algorithms*: These algorithms include all methods from the machine learning domain to assist the ontologist in detecting new concepts and their relations, and to help in placing them in the correct position in the taxonomy. As an example of a system that uses this technique is OntoLearn (Navigli et al., 2003).

Despite the wide area of ontology semi-automatic generation, this thesis used the manual technique to build its domain ontologies. This was due to two reasons: (1) the non-existence of pre-constructed ontologies in the thesis domain of interest and (2) to speed up the process of building the domain ontologies without the need to evaluate the validity of the resultant ontologies.

5.4.1 Existing Ontologies on the Web

Pre-constructed ontologies can be found either in ontology libraries, specialized search engines or portals. Many ontology libraries do exist on the web, for instance DARPA[64] (DAML Ontology Library) contains around 280 ontologies written in the DAML ontology language. Ontologies in this library range from medical research to business. In addition, Stanford University holds another library of ontologies created using Protégé editor; this library is called Protégé Ontology Library[65]. Schemeweb[66] is another source for pre-created ontologies. Also, OntoSelect[67] is an ontology repository that harvests ontologies from the web. The user can browse ontologies according to size (number of classes, properties), representation format (DAML, RDFS, OWL), connectedness (score over the number of included and referring ontologies) and human languages used for the class/property labels. The library also supports ontology search. When searching using OntoSelect the returned results are ranked based on ontologies relevance.

Specialized search engines such as Swoogle[68] can also help find ontologies on the web. Swoogle is capable of searching around 10,000 ontologies. Similarly, ONTOSEARCH2[69] is "a search and query engine for ontologies and ontological data on the Semantic Web. It allows ad-hoc queries across hundreds of OWL files using the SPARQL query language".

---

[64] http://www.daml.org/ontologies/ [last accessed 11/2/2007]

[65] http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary [last accessed 11/2/2007]

[66] http://www.schemaweb.info/schema/BrowseSchema.aspx [last accessed 11/2/2007]

[67] http://olp.dfki.de/OntoSelect/index.php?mode=select [11/2/2007]

[68] http://swoogle.umbc.edu/ [last accessed 11/2/2007]

[69] http://www.ontosearch.org/ [last accessed 11/2/2007]

Finally, ONTHOLOGY[70] ("anthology of ontologies") is a portal for ontologies contributed by users. It uses the proposed Ontology Metadata Vocabulary (OMV) that is used to provide metadata descriptors to identify ontologies. OMV is like DC for documents. The portal contains 127 ontologies which someone can browse or search.

5.5 Ontologies in Education

Ontologies in education can be classified into three categories (Stojanovi et al., 2001): content (domain) ontologies, context ontologies and structure ontologies.

Content (domain) ontologies; define the content of a learning document in the process of searching for the learning material as well as in the process of providing learning materials. Context ontologies, define the place where the learning material will be presented. Structure ontologies, define learning materials as small chunks and connect these chunks to each other in order to build up a complete course.

The approach that the researcher adopted in this thesis was based on the first type of ontologies (i.e. content ontologies).

5.6 Ontology Applications in Education

Ontologies have proven their success in many educational systems and in different applications; however it is impossible to produce a complete listing of all educational applications that uses ontologies. This section, however, will give pointers to some applications in the learning technologies discipline. To do this effectively, the applications have been clustered according to their area of research.

For instance, in the area of learning objects, (Gasevic et al., 2004) proposed an approach to enhance learning object content using ontologies and Semantic Web languages. They implemented a simple educational web application using content structure ontologies and domain ontologies to illustrate their approach. The application was based on the *Petri net* ontology. Furthermore, (Verbert et al., 2005)

[70] http://www.onthology.org/ [last accessed 11/2/2007]

used the previous approach to apply ontologies in repurposing learning object components. They did that by decomposing learning objects into their components and attaching metadata to each component so the ontology can be used to automatically assemble different components based on its educational purpose.

In the area of metadata, the Edutella P2P network (Brase and Painter, 2004) used learning object annotated with a subset of Dublin core and LOM metadata using RDF(S). The learning objects were classified using domain specific ontologies.

In the area of eLearning courses, (De Nicola et al., 2004) developed an ontological system integrated with an eLearning platform to support teachers in building courses and students in accessing courses content. The system is part of the Italian project '*Web Learning'*. The project is ongoing and further elaboration will be carried out to improve the retrieval of learning resources.

In the area of educational web portals, (Woukeu et al., 2003) developed '*Ontoportal',* an ontological hypertext framework for building educational web portals based on simple domain ontologies. The ontological web portal contains links to educational resources that are semantically interconnected.

5.7 Chapter Summary

Ontologies play a great role in the Semantic Web discipline. This chapter highlighted the importance of ontologies in different educational context. Besides, this chapter discussed the definition, languages, types and engineering of ontologies.

In the next chapter, another core process of the Semantic Web (hence semantic metadata annotation) will be discussed.

Chapter 6
Semantic Metadata Annotation

Annotation is a mechanism to associate metadata with web resources (Bechhofer et al., 2002). Annotating a web resource with semantic metadata provides meaning to its content.

This chapter starts by clarifying the meanings of 'semantics', 'annotation' and 'semantic metadata annotation', as these three terms formulate a cornerstone for understanding what is meant by semantic metadata annotation. Next, a comprehensive discussion about the different semantic annotation techniques and methods that have been used in most semantic annotation tools is laid out. Finally, the chapter ends with some concluding remarks concerning the development of the FolksAnnotation tool.

6.1 What is Semantics?

> **Semantics** [noun]: the study of meanings; the meaning or relationship of meanings of a sign or set of signs; especially: connotative meaning (*From Merriam-Webster online Dictionary*[71]).

Different areas of computer science have different interpretations of what 'semantics' mean (Sheth et al., 2005; Lytras and Naeve, 2006). For instance, in the

---

[71] http://www.m-w.com/dictionary/Semantics [last accessed 11/2/2007]

domain of databases, metadata is thought of as a conceptual schema that describes the structure of a database. In the domain of information retrieval, metadata might be consider as the set of keywords that describe the main theme of a document, or as a record that confirms to a specific schema (e.g. Dublin Core).

Sheth et al. have described the different depictions of metadata, organizing them into three types of semantics (Sheth et al., 2005): implicit, formal and powerful.

Implicit semantics appear in unstructured text that has been loosely defined and less formally structured (e.g. Information Retrieval). *Formal semantics* appear when the data representation takes a more rigid form (e.g. Knowledge Representation). Finally, *powerful semantics* imply the combination of simple syntactic structures to represent the meaning of complex ones.

6.2 What is Annotation?

> **Annotation** [*noun*]: 1) is a note added by way of comment or explanation, 2) the act of annotating (*From Merriam-Webster online Dictionary*[72]).

In the computer context, annotation has been defined as "…*a set of instantiations attached to an HTML document*" (Handschuh and Staab, 2003a). Euzenat (2002) has also defined 'annotation' as a function from document to formal representations. By the same token, Euzenat defines 'indexing' as a function from formal representations to documents.

Bechhofer et al. (2002) have classified annotation into three types: textual annotation, link annotation and semantic annotation.

Textual annotation is the process of adding comments or notes to a text. This type of annotation has been used for many years in communities such as biology especially in biology databases, where the protein's sequences information is described using annotation. *Link annotation* extends text annotation by adding links rather than text

[72] http://www.m-w.com/dictionary/Annotation [last accessed 11/2/2007]

to the content. Finally, the *semantic annotation* is where the content of the annotation contains more semantic information taken from ontologies.

6.3 What is Semantic Metadata Annotation?

The definition of semantic metadata annotation, based on what has been addressed in the Semantic Web literature, can be divided into three distinct terms: '*semantic annotation', 'semantic metadata'* and *'metadata annotation'*.

Semantic annotation means specifying some machine processable meanings about a web resource. This can be done by committing a web resource to some domain ontologies (Zhihong and Mingtian, 2003). Semantic annotation can be also defined as "*… a specific metadata generation and usage schema, aiming to enable new information access methods and to extend the existing ones*" (Kiryakov et al., 2003). In addition, Ding (2005) defined semantic annotation as "*…a process … to label web page content explicitly, formally, and unambiguously using ontologies*". In short, semantic annotation can be named ontology-based metadata (Handschuh and Staab, 2003b).

On the other hand, semantic metadata can be defined as "*…[linking] related terms to one another*" (Haase, 2004).

Finally, metadata annotation can be defined as "*…the process of attaching semantic descriptions to Web resources by linking them to a number of classes and properties defined in Ontologies.*" (Scerri et al., 2005).

It seems from the previous discussions that most of the definitions address the same concept; adding semantic descriptors to a document based on an ontology. Therefore, the more expressive term 'Semantic Metadata Annotation' will be used to describe the theme of this thesis.

6.3.1 Categories and Levels of Semantic Annotation

Semantic annotation methods can be categorized into two groups (Scerri et al., 2005): *Internal annotation* which involves the embedding of the semantic markup elements inside the HTML document, and *external annotation* that involves storing the metadata in a separate file.

It is worth mentioning that there are several levels of semantic annotation used with content-level annotation in natural language processing domain (aka Part-Of-Speech[73] annotation/tagging); this includes: word-level, sentence-level, paragraph-level and section-level annotation. However, this thesis will concentrate on document-level annotation.

6.4 Semantic Annotation Research

In a developing field such as the Semantic Web, it is impossible to complete a comprehensive survey of new tools and new versions of existing tools due to the rapid changes in this area. This section will attempt to summarize the main techniques in the semantic annotation field.

In a comprehensive survey by both (Uren et al., 2005) and (Reeve and Han, 2005) about the different tools used in semantic annotation, both surveys tried to categorize the types of tools used in semantic annotation from different perspectives. On one hand Uren et al. separate the semantic annotation techniques into semantic frameworks and semantic tools. On the other hand, Reeve and Han state that semantic annotation platforms can be classified based on the type of annotation method used, this includes: pattern-based, machine learning based and multi-strategy based, which uses a combination of pattern-based and machine learning methods. However, both surveys agree on the different approaches of semantic annotation, which include: manual, semi-automatic and automatic annotation.

Manual annotation requires a user to manually annotate a document content using a predefined ontology. An example of this type of annotation is the OntoMat-

[73] http://en.wikipedia.org/wiki/Part-of-speech_tagging [last accessed 11/2/2007]

Annotizer tool (Handschuh et al., 2001). The most significant drawback of manual annotation is that it is prone to errors due to many factors such as annotator unfamiliarity with the domain and/or his/her lack of motivation (Bayerl et al., 2003). Also manual annotation is an expensive process in terms of time and effort.

Semi-automatic annotations analyze a text to identify instances and then relate them to their corresponding ontological concept. These systems are not completely automatic; hence human intervention is required to clarify ambiguous terms. An example of this type of annotation is SemTag (Dill et al., 2003a; Dill et al., 2003b).

Reeve and Han have claimed that complete automatic annotation tools do not exist, based on the fact that in an early stage of the annotation process a human intervention is required to bootstrap the process (Reeve and Han, 2005). However, the researcher will show in section 6.4.3.2 an example of a complete automatic annotation tool called C-PANKOW.

6.4.1 Platform Classification

As mentioned previously, Reeve and Han classified annotation platforms based on the type of annotation method used into: pattern-based, machine learning and multi-strategy based.

The role of *Pattern-based annotation* is to find patterns for a defined initial set of entities in a corpus. Thus, when new entities are discovered along with new patterns, the process is repeated until no more entities are discovered or the user stops the process. This process can also use manual rules to find entities in text.

Machine-based annotation uses two methods: probability and induction. Probabilistic annotation tools use statistical models to locate entities within text. Induction tools use either linguistic or structural analysis to perform wrapper induction[74].

---

[74] Wrapper induction is 'a technique for automatically constructing wrappers from labeled examples of a resource's content'. From http://www.cs.washington.edu/homes/weld/wrappers.html [last accessed 28/2/2007]

Finally, *multi-strategy annotation* combines both pattern-based and machine-based methods; however, Reeve and Han claim that until now no system exists that implements the multi-strategy annotation method.

6.4.2 Semantic Annotation Frameworks

Uren et al., on the other hand, have talked about two annotation frameworks: *Annotea* the W3C annotation project (Kahan et al., 2001) and CREAM (Handschuh and Staab, 2003), an annotation framework developed at the university of Karlsruhe.

Annotea (Koivunen, 2005) (Kahan et al., 2001) is a free text annotation tool that associates statements about documents in a collaborative fashion. These statements must have metadata fields such as author, creation time, etc. Annotea uses RDF as the format of the metadata. The types of documents that can be annotated using Annotea are limited to XML and HTML format. The generated metadata can be stored either locally (in the user machine) or on public RDF servers. Examples of tools based on the Annotea framework are Amaya[75] and Annozilla[76].

The **CREAM** (Creating RElational, Annotation-based Metadata) framework (Handschuh et al., 2001; Handschuh and Staab, 2003; Handschuh and Staab, 2003a) allows the creation of relational metadata, metadata that comprises class instances and relationship instances.

The CREAM framework as an annotation framework comprises the following modules that are required for semantic annotation: a *document viewer* to visualize the web page content, an *ontology guide* to help in the annotation process, a *crawler* to search the Semantic Web for an existing annotation for the instance being annotated, an *annotation inference server* for querying annotated documents, and *document management* for managing annotated documents. Furthermore, CREAM is capable of annotating the deep web i.e. databases; therefore when web pages are generated

---

[75] http://www.w3.org/Amaya/ [last accessed 12/2/2007]

[76] A browser based on Mozilla browser to create and view annotations associated with a web page, http://annozilla.mozdev.org/ [last accessed 12/2/2007]

from databases they are automatically annotated. Some examples of tools based on the CREAM framework are OntoMat-Annotizer[77] and S-CREAM (Handschuh et al., 2002).

6.4.3 Semantic Annotation Tools

Uren et al. continued their survey by examining four types of annotation tools: manual, automatic, integrated annotation environments and on-demand annotation. Each type will now be discussed in some detail.

6.4.3.1 Manual Annotation

Manual annotation tools are the most basic ones. They allow users to manually create annotations with or without the support of ontologies. Several annotation tools have been built based on the Annotea framework, among them is the ***Amaya*** browser and editor from W3C (Koivunen, 2005). It can annotate documents with RDF markup without the aid of an ontology.

The **OntoMat-Annotizer** tool (Handschuh et al., 2001) is built using the principles of the CREAM framework. It has a web browser to display the web page being annotated. It also provides a side-bar for displaying the ontology structure for ease of manual annotation. The user can highlight parts of a web page for annotation, then by using drag-and-drop interaction, the user can associate the highlighted instances with a class in the displayed ontology. The tool can only annotate HTML/XML documents. For a good survey about other manual annotation tools see (Heck and Obermark, 1999).

Multimedia annotation falls into the category of manual annotation, thus expanding the range of file types that can be annotated to include video, audio and images. **Meditate**, for character markup, and **SiX**, for trivial screenplay markup, are two tools developed by the University of Southampton that use the OntoMedia[78] ontology to annotate video clips (Jewell et al., 2006).

---

[77] http://annotation.semanticweb.org/ontomat/index.html [last accessed 12/2/2007]

[78] http://ontomedia.ecs.soton.ac.uk/ [last accessed 12/2/2007]

PhotoStuff (Halaschek-Wiener et al., 2005) is a semantic digital image annotation tool. The tool is capable of annotating parts of an image with instance from a pre-defined ontology. When an image is annotated using PhotoStuff, it is then uploaded to a semantic web portal for browsing, searching and managing annotated digital images. For a list of image annotation tools the reader is referred to (W3C, 2006).

6.4.3.2 (Semi)-Automatic Annotation

Automatic or semi-automatic annotation tools during the course of annotation rely on an information extraction engine and a pre-constructed ontology. The information extraction engine can use rules or wrappers written by hand; it can also be supervised, e.g. the system learns from an annotation sample marked by users, or unsupervised, e.g. the system employs strategies to learn how to annotate without human intervention. Next, a sample of (semi-)automatic tools will be summarized.

S-CREAM (Semi-automatic CREAtion of Metadata) (Handschuh et al., 2002) is an extension of OntoMat-Annotizer that is based on the CREAM framework. S-CREAM uses Amilcare[79], an information extraction tool that learns information extraction rules from manual-markup input. When using this tool, the user annotates a set of web pages then feeds them to the tool so it can learn from them and suggest annotations for new web pages.

MnM[80] (Vargas-Vera et al., 2002) provides both automatic and semi-automatic semantic annotation to web pages. MnM integrates a web browser with an ontology editor. It also provides open APIs, such as OKBC[81], to link to ontology servers and for integrating information extraction tools, such as Amilcare.

SemTag (Dill et al., 2003a; Dill et al., 2003b), an application of Seeker[82], is considered one of the largest scale semantic tagging attempts that have been conducted to date. In this exercise, the tool annotated a collection of approximately 264 million web pages and generated approximately 434 million automatically

[79] http://nlp.shef.ac.uk/amilcare/ [last accessed 12/2/2007]

[80] http://kmi.open.ac.uk/projects/akt/MnM/ [last accessed 12/2/2007]

[81] Open Knowledge Base Connectivity, http://www.ai.sri.com/~okbc/ [last accessed 12/2/2007]

[82] A platform for large-scale text analytics

disambiguated semantic tags. SemTag uses a disambiguation algorithm called TBD, for Taxonomy-Based Disambiguation. The algorithm operates by using a vector-space model to assign the correct ontological class or to determine that a concept does not correspond to a class in TAP, hence SemTag uses the TAP ontology to define annotation classes. The TAP ontology is a shallow knowledge base that contains a broad range of lexical and taxonomic information about popular objects such as music, movies, authors, sports, autos, health, etc.

C-PANKOW (Context-driven PANKOW) (Cimiano et al., 2005), is an enhancement of the PANKOW (Pattern-based ANnotation through Knowledge On the Web) annotation technique, where keyword instances to be annotated are put into several linguistic patterns that convey competing semantic meanings. The patterns that are matched most often on the web indicate the meaning of the instance. C-PANKOW uses Google search abstracts to look for the meaning of an instance. It is also an unsupervised automatic annotation tool that avoids the problems of supervised techniques (such as where the document has a similar structure). The tool is considered one of the 'Self Annotating Web' techniques where globally available knowledge is used to annotate web pages.

6.4.3.3 Integrated Annotation Environments

There is also a new trend of annotation tools that helps a document author to semantically annotate their documents as they author it. These tools include a MS Word plug-in called Writing in the Context of Knowledge (**WiCK**[83]). The tool implements a simultaneous authoring of a document and semantic markup (Carr et al., 2004). The tool also helps the user in the process of filling research proposal forms by proposing values for the fields from domain ontologies (e.g. researchers' ontologies).

Semantic Word (Tallis, 2003) is a semantic annotation environment based on MS Word. It helps authors to semantically annotate their documents using predefined annotated templates or/and toolbars that support the creation of semantic descriptors to attach to text regions.

[83] http://wick.ecs.soton.ac.uk/ [last accessed 12/2/2007]

6.4.3.4 On-Demand Annotation

Uren et al. devised this category to talk about systems that are not strictly annotation tools. **Magpie** (DZBOR et al., 2004), is one example of these tools. The tool produces an annotation-like service on demand for users browsing un-annotated web page. It operates from within a browser, where it highlights text strings related to an ontology of the user's choice.

6.4.3.5 Other Annotation Techniques and Tools

There are two approaches to store annotation files: a proxy–based and a browser–based approach (Koivunen et al., 2000). In the proxy-based approach the annotation is stored on a proxy server and when an annotated web page is visited the annotation is merged with the web page from the proxy and then displayed to the user. The browser-based annotation is slightly different in its merging process in that the browser is responsible for merging the web page with its annotation before displaying it to the user. It is also possible to save the annotation on the browser side, however, this is less interesting to the users because of some limitations (e.g. annotation may be updated and/or changed). ComMentor[84] and Yawas[85] are two examples of browser-based annotation tools (Koivunen et al., 2000).

The **Gnowsis**[86] Semantic Desktop (Sauermann, 2005) is an open source framework project led by DFKI[87]. The architecture is based on a Semantic Web server running as a desktop service. The aim of the framework is to provide the glue between desktop applications (email client, browser, office applications, etc.) using semantics derived from ontologies.

Microformats[88] is another semantic markup technique "*designed for humans first and machines second*". It integrates a set of simple open data formats (compound and/or elemental) built upon existing standards with HTML/XHTML/XML files (Figure 6.1).

---

[84] http://dbpubs.stanford.edu:8091/diglib/pub/reports/commentor.html [last accessed 12/2/2007]

[85] http://www.fxpal.com/people/denoue/yawas/ [last accessed 12/2/2007]

[86] http://www.gnowsis.org [last accessed 12/2/2007]

[87] The German Research Centre for AI, http://www.dfki.de/web/ [last accessed 12/2/2007]

[88] http://microformats.org/about/ [last accessed 12/2/2007]

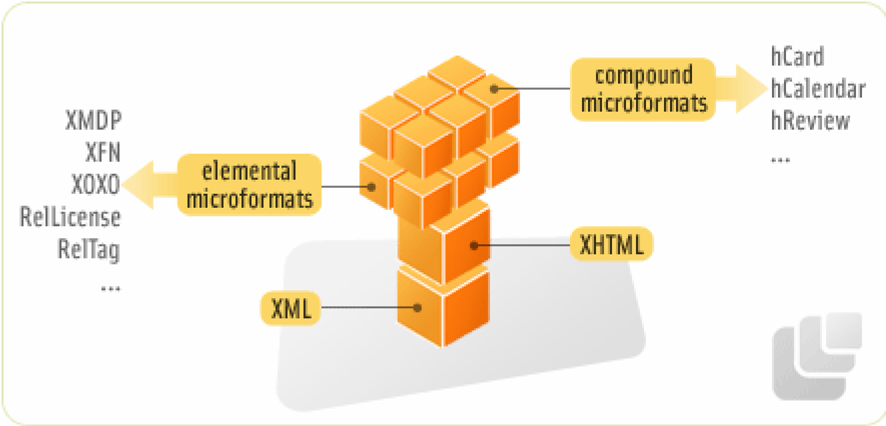


Figure 6.1: Microformats diagram [by Microformats.org]

The list of Microformats data format includes the use of some of the following tags:

- hCard for People and Organizations
- hCalendar for Calendars and Events
- VoteLinks and hReview for Opinions, Ratings and Reviews
- XFN for Social Networks
- rel-license for Licenses

The idea of Microformats is similar to SHOE[89] (Simple HTML Ontology Extension); where the annotation is inserted within the HTML tag; however, Microformats do not use a formal ontology. One application using the Microformats notation is structured blogging[90].

To the best of the researcher's knowledge, it is still not clear how Microformats will advance the development of the Semantic Web. Nonetheless, it seems that the future of this initiative is bright, or as (Khare, 2006) pointed out in his article "*...Microformats may yet take hold in their ecological niche as an appropriately incremental evolution of existing technologies that makes the Web more amenable to automated analysis without infringing on authors' authority to present that data as they wish.*"

[89] http://www.cs.umd.edu/projects/plus/SHOE/ [last accessed 12/2/2007]

[90] http://structuredblogging.org/index.php [last accessed 12/2/2007]

Piggybank[91] (Huynh et al., 2005) is a Firefox plug-in to semantically annotate websites. It also provides screen-scraper functionality (a screen-scraper is a client-side program that extracts specific information from a web page e.g. price, product, and colour from a commerce website). Piggybank converts the information collected by the screen-scraper users add their own tags (i.e. folksonomies) to annotate websites and save these tags in an RDF format. The saved RDF files can be either saved on the user's computer or moved into a collaborative server called a Semantic-Bank.

6.5 Semantic Annotation Tools for eLearning

Few semantic annotation tools exist for annotating learning resources. In a survey paper by (Azouaou et al., 2004) about the different tools for semantic annotation for learning materials, the authors tried to identify some specifications as guidelines for developing semantic annotation tools that fulfil the requirements of educational applications.

They first categorized the three main players in the annotation activity which includes:

- The author of the annotation (the annotator).
- The addressee of the annotation (the user of the annotation).
- The fact that the annotation is semantic or not.

Then, based on the previous characterization, they provided four properties of annotation tools, which are:

- Automatic versus manual annotation.
- Cognitive versus non-cognitive annotation.
- Computational versus non-computational annotation.
- Semantic versus non-semantic annotation.

They also list the requirements for eLearning annotation tools, namely: *usefulness* (which takes into account teaching/learning context); *shareability* (which enables

[91] http://simile.mit.edu/wiki/Piggy_Bank [last accessed 12/2/2007]

teaching/learning actors to communicate through annotation; and *usability*. Then they evaluated the strength and weaknesses of annotation tools based on each category and the requirements they specified.

MemoNote (Azouaou and Desmoulins, 2006a; Azouaou and Desmoulins, 2006b) and AnnForum (Azouaou et al., 2004) were the two evaluated annotation tools dedicated for annotating learning materials. They conclude that the tools which respect most of the requirements are those that are computational, cognitive and semantic; these requirements were reified in MemoNote.

Finally, they pointed out that the problem with general purpose annotation tools was, they usually provide domain independent annotation, thus, do not take into consideration the requirements of special domains.

Another semantic annotation system was produced by (Dehors et al., 2005). Dehors et al. have developed a methodology for semi-automatically extracting annotations from existing pedagogical documents. Their QBLS (Question-Based Learning) system does not require a specific annotation tool; instead it uses MS Word templates that rely on pre-defined layouts which are linked to ontologies to produce semantic annotation. The resultant semantic annotation is then used with the Corese[92] semantic search engine to perform semantic queries. They evaluated their system based on a two-hour exercise session attended by 49 students. The students rated the usability of the conceptual navigation provided by the system with a high score (4.2 out of five). The system also appealed to most teachers that used it for authoring their pedagogical materials; this was because the system relied on well-known software (i.e. MS Word) to produce its output.

6.5.1 Annotation Goals in Education

Annotation in an educational context can be identified as having four goals (Azouaou and Desmoulins, 2005): 1) classifying (organizing into a hierarchy, contextualizing); 2) adding information (reformulating commenting and documenting); 3) planning (scheduling, indirect annotating); and 4) correlating.

[92] http://www-sop.inria.fr/acacia/soft/corese/ [last accessed 13/2/2007]

In this thesis the goal of using semantic annotation is to classify and add information to existing web resources, so they can be retrieved and searched by semantic means, which makes these web resources amenable for machine processing.

6.6 Discussion

From the previous overview of the different aspects of the process of semantic annotation (general and domain specific), several points can be highlighted:

- Most previously mentioned tools rely on either human manual annotations or (semi)-automatic annotation that uses Information Extraction (IE) and Machine Learning (ML) techniques to extract valuable information from a web resource. Both techniques suffer from apparent shortcomings. In the case of manual annotation, the main shortcoming is that it is a human dependent process, which leads to significant effort and sometimes to errors when handled by an incompetent annotator. In the case of (semi)-automatic annotation, the shortcoming can be viewed as a fluctuation in the accuracy and quality of the produced semantic metadata.
- There are few semantic annotation tools dedicated to the eLearning domain, this might be attributed to the sheer interest in the Semantic Web community for building Semantic Web technologies to serve the needs of large industries/ organizations and/or research centres, rather than to education.
- Many of the reviewed semantic annotation tools follow a content-level semantic annotation approach, where the internal pieces of a web resource are linked to ontological terms, i.e. these tools are designed to insert ontology-based markups in web pages (Corcho, 2006). However, this thesis is using a slightly different systematic approach for semantic annotations. The implemented tool has adopted a document-level semantic annotation approach, where an overall description of a web resource is generated without the hassle of performing a content-level interlinking with ontological terms.
- One difference the thesis tool has compared to the Piggybank plug-in is that it uses pre-generated ontologies and deals with a specific domain, while the Piggybank plug-in is open to all and does not comply with any ontologies.

To conclude, the problem of most automatic semantic annotation tools is that they require ‘the man in the middle’ process, which uses extraction technologies. This wastes an extensive amount of processing time in that phase. Moreover, none of the previously mentioned tools have used folksonomies as guides in the process of annotating web resources. So, to test the potential of using people’s metadata (aka folksonomies) in the process of semantic annotation and to check how rich the generated semantic metadata will be; this thesis explores the benefit of using the output of contemporary web services that use tagging as their main assets to create semantic metadata.

6.7 Chapter Summary

The term ‘annotation’ has different interpretations depending on the context that it is used in. Some might think of it as private notes, others as comments or remarks by the author or the visitor of a web page. Despite these different interpretations; annotation, or in particular semantic annotation, is what makes the web amenable for machine processing.

This chapter has discussed in some detail the different platforms, frameworks and tools used for semantic annotations. It also highlighted some important guidelines and requirements that need to be considered when designing an annotation tool for an eLearning domain.

The vision of this thesis is to develop a semantic metadata annotation tool for the use in an educational context. The source of semantic descriptors will come from folksonomy tags; to show the added value of the folksonomy community in the process of semantic annotation. The FolksAnnotation tool will not annotate the *content* of a web resource; instead it will assign *document-level semantic metadata* to a web resource as a whole. The discussion of the folksonomy-based annotation tool along with the design decisions will be the theme of the next chapter.

Chapter 7
Exploring the Value of Folksonomies

While previous chapters have laid out the foundation of the thesis work by signposting the various technologies exploited for building the FolksAnnotation tool, this chapter and the following ones will cover the main contributions of this thesis by discussing the various experiments conducted to justify the thesis hypotheses.

In this chapter, the exploration of the value of folksonomies against automatic indexing mechanism is done by testing ***Hypothesis 1(a)***, which states***:***

> *"Folksonomies, as index keywords, hold more semantic value than keywords automatically extracted by machines*."

The underlying assumption of this hypothesis is that most folksonomy tags are more related to a professional indexer's mindset than keywords extracted using automatic keyword extraction techniques.

The main questions this experiment tries to answer are:

- Do folksonomies only represent a set of keywords that describe what a document is about, or do they go beyond the functionality of index keywords?
- What is the relationship between folksonomy tags, automatically extracted index keywords and keywords assigned by a professional indexer?

- Where are folksonomies positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords?

In order to find out if folksonomies can improve on automatically extracted keywords, it is significant to examine the relationship between them, and between them and professional human indexer keywords. Therefore, this chapter starts by discussing similar works that have compared folksonomy tags to other indexing mechanisms. Then the setup of the experiment and data set selection are explained. Finally, the chapter concludes by reporting and discussing the results of the four phases of the experiment.

7.1 Related Work

Little research has explored the area of folksonomies compared to other indexing mechanisms. Kipp (2006) has examined the differences and similarities between the user keywords (folksonomies), the author and the intermediary (such as librarians) assigned keywords. She used a sample of journal articles tagged in the social bookmarking sites citeulike[93] and connotea[94], which are specialized for academic articles. Her selection of articles was restricted to a set of journals known to include author assigned keywords and to journals indexed in the Information Service for Physics, Electronics, and Computing (INSPEC[95]) database, so that each article selected would have three sets of keywords assigned by three different classes of metadata creators. Her methods of analyses were based on concept clustering via the INSPEC thesaurus, and descriptive statistics. She used these two methods to examine differences in context and term usage between the three classes of metadata creators.

Kipp's findings showed that many users' terms were found to be related to the author and intermediary terms, but were not part of the formal thesauri used by the intermediaries; this was due to the use of broad terms which were not included in the thesaurus or to the use of newer terminology. Kipp then concluded her paper by

[93] http:// citeulike.org/ [last accessed 5/2/2007]

[94] http://connotea.org/ [last accessed 5/2/2007]

[95] A database which provides an intermediary assigned controlled vocabulary for searchers.

saying "*User tagging, with its lower apparent cost of production, could provide the additional access points with less cost, but only if user tagging provides a similar or better search context*."

Similarly, Lin et al. (2006) compared social tagging with controlled vocabularies and title-based automatic indexing. The data set they used was similar to Kipp's data set, with an interest in articles in the medical filed. They concentrated on medical articles in PubMed[96] that have Medical Subject Headings (MeSH) terms and used GATE[97] text-processing engine to extract the indexing keywords. Their results show that there was little overlap among the three indexing methods, with 11% between social tagging and MeSH terms, and 19% between social tagging and automated indexing.

Conversely, Tennis (2006) compared the differences and similarities between social tagging and subject cataloguing using framework analysis. The framework analysis compares the 1) processes, 2) structures, of indexing and 3) the context in which social tagging and subject cataloguing occur. After applying the framework analysis, Tennis has found that social tagging is quite different from subject cataloguing, and there was a superficial similarity in purpose between the two.

7.1.1 Discussion

Apparently, the method that Kipp used does not compare folksonomies to keywords extracted automatically using context-based extraction methods. This extra evaluation method is significant to measure the relationship between automatic machine indexing mechanisms lead by a major search engine like Yahoo compared to human indexing mechanisms, and whether is it possible to replace folksonomies with automatically extracted keywords. As for Tennis's comparison, he did not undertake an in-depth analysis of folksonomy tags; instead he theoretically applied modified rubrics from the library science to compare between social tagging and subject cataloguing, this implies that his work lacks an empirical basis. Finally, Lin et al. is very similar to the experiment described here, differing in the tools and data

[96] http://www.ncbi.nlm.nih.gov/entrez/ [last accessed 5/2/2007]

[97] http://gate.ac.uk/ [last accessed 5/2/2007]

sets used. The tool used for automatic indexing is based on Lucene[98] search engine library that uses full text indexing technique compared to Yahoo TE which uses context-based indexing. Moreover, the data set used for the Lin et al. experiment was chosen to be from the medical field and they used bookmarked articles in connotea. However, in this experiment the data set was taken from the del.icio.us bookmarking service and covered a variety of topics.

7.2 Experiment Setup and Test Data

There are plenty of keyword extraction techniques in the IR literature, most of which are either experimental or proprietary, so they do not have a corresponding freely available product that can be used. Therefore the researcher was limited to what exists in this field such as, SEO keyword analyzer tools, Kea[99]- an open source tool released under the GNU General Public License-, and Yahoo API[100] term extractor. Of these the Yahoo API was the preferred choice.

Kea requires extensive training in a specific domain of interest to come out with reasonable results; SEO tools on the other hand, were biased (i.e. they look for the appearance of popular search terms in a webpage when extracting keywords), besides the IR techniques they are using are very basic (e.g. word frequency/count). The decision to use Yahoo API was made for the following reasons:

- The technique used by Yahoo's API to extract terms is context-based as described in (Kraft et al., 2005), which means it can generate results based on the context of a document; this will lift the burden of training the system to extract the appropriate keywords.
- Also, Yahoo's recent policy of providing web developers with a variety of API's encouraged me to test the quality of their term extraction service.

The experiment was conducted in four phases: in the first phase the researcher exposed a sample of both folksonomy and Yahoo keywords sets to two trained-human indexers who, given a generic classification, evaluated which set held greater

[98] http://lucene.apache.org/java/docs/index.html [last accessed 5/2/2007]

[99] http://www.nzdl.org/Kea/ [last accessed 5/2/2007]

[100] Yahoo API term extractor service was launched on May 2005

semantic value than the other. In the second phase, the researcher used another modified instrument from (Kipp, 2006) to further explore the semantic value of folksonomy tags and the Yahoo keywords. In the third phase, the researcher measured, for a corpus of web literature stored in the del.icio.us bookmarking service, the overlap between the folksonomy set and Yahoo extracted keyword set. In the final phase, one of the human indexers was asked to generate a set of keywords for a sample of websites from our corpus and compare the generated set to the folksonomy set and the Yahoo TE set to measure the degree of overlap. Thus, the analysis of the experiment can be thought of as being in two forms: term comparison (phase 1 and 2) and descriptive statistics (phase 3 and 4).

The rest of this chapter will discuss the comparison system framework used for evaluating phase 3 and 4, the data set and the different phases of the experiment along with the accomplished results.

7.3 The Comparison System Framework

The researcher constructed a system to automatically compare the overlap between the folksonomy, Yahoo TE and human indexer keywords and generate the desired statistics. The system consisted of three distinct components: the Term Extractor, the Folksonomy Extractor and the Comparison Tool as shown in Figure 7.1. The Term Extractor consists of two main components: JTidy[101], an open source Java-based tool to clean up HTML documents and Yahoo Term Extractor[102] (TE), a web service that provides "*a list of significant words or phrases extracted from a larger content*". After removing HTML tags from a website, the result is passed to Yahoo TE to generate the appropriate keywords.

The Folksonomy Extractor that the researcher developed is designed to fetch the keywords (tags) list for a particular website from del.icio.us and then clean-up the list by pruning and grouping tags. Finally, the Comparison Tool role is to syntactically compare the folksonomy list to Yahoo's keywords by counting the number of

[101] http://sourceforge.net/projects/jtidy [last accessed 5/2/2007]

[102] http://developer.yahoo.net/search/content/V1/termExtraction.html [last accessed 5/2/2007]

overlapped keywords between the two sets. The tool then calculates the percentage of overlap between the two sets using the following equation (1):

$$P = \frac{N}{(|F_s| + |K_s|) - N} \times 100 \qquad (1)$$

The above equation can be also expressed using set theory as (2):

$$P = \frac{|F_s \cap K_s|}{|F_s \cup K_s|} \times 100 \qquad (2)$$

Where:

P	Percentage of overlap
N	Number of overlapped keywords
Fs	Folksonomy set
Ks	Keyword set

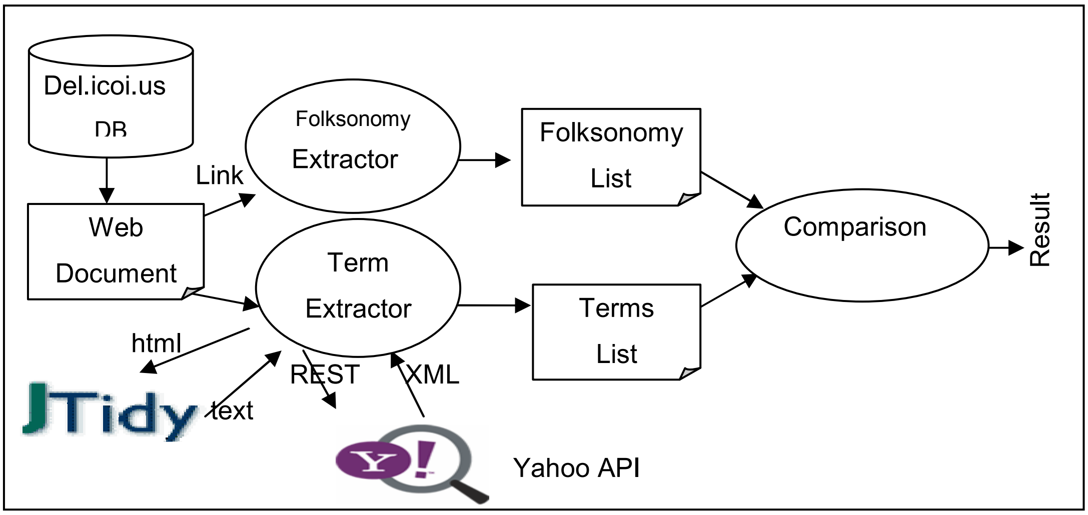


Figure 7.1: The Comparison System Framework.

7.4 Data Selection

The test data used in this experiment was randomly collected from the del.icio.us[103] social bookmarking service. One hundred bookmarked websites spanning various topics from the popular tags webpage were selected, as shown in Table 7.1.

[103] http://del.icio.us/tag/, Data was collected between 24/2 and 27/2 2006

Topic	Number of Web Sites
Software	11
Open source	14
Education	6
Programming	18
Sciences	8
Linux	10
References	13
Development	20
Total	100

Table 7.1: Topics covered in the experiment data set

The selected web resources were chosen based on the following heuristics:

- Bookmarked sites that are of multimedia nature such as audio, video, flash, Word/PDF documents, etc. were avoided, as the Yahoo term extraction service only extracts terms from textual information. By the same token, whole Blog sites were avoided because they usually hold a diversity of topics; the researcher tried to look for web pages with a single theme (e.g. a specific post in a Blog).
- The researcher only chose bookmarked sites with 100 or more participating taggers; this was necessary to ensure there were enough tags describing the website.

7.5 Other General Heuristics

Some other heuristics were used during the experiment lifecycle, to improve the quality of the extraction results which are listed as follows:

1. Most websites that use Google Adsense (an advertisement tool by Google) affected the results of the terms returned by Yahoo extractor. The returned terms from the extractor are filled with the advertisement words provided by Adsense, which will add noise to the extracted terms and at the same time limit the number of returned terms (i.e. Yahoo TE only returns 20 terms). Therefore, in some cases the researcher was forced to manually enter (i.e. copy and paste) the text of a website and place it in a web form that invokes the Yahoo TE service.

2. Yahoo TE is limited to produce only twenty terms, which may consist of one or more words to represent the best candidate for a website (as mentioned on the service website); these terms were put in two forms: a) concatenated to form compound words and b) split out into single words; this action was necessary so that Yahoo TE keywords might match del.icio.us style for single and compound word tags.

7.6 Results

7.6.1 Phase 1

The role of phase one is to determine whether or not folksonomies carry more semantic value than keywords extracted using Yahoo TE. In this phase the phrase ‘semantic value’ means that the tag or keyword used to describe a web resource is relevant to its gist, i.e. the tag or keyword contributes to the description of the resource meaning.

Thus, given the sets of keywords from Yahoo TE and del.icio.us; the two trained indexers[104] were asked to blindly[105] evaluate each keyword from both sets. The indexers were provided with a five-category table to classify the keywords from both sets. The table has the following values: "Strongly relevant" encoded 5, "Relevant" encoded 4, "Undecided" encoded 3, "Irrelevant” encoded 2 and "Strongly irrelevant” encoded 1.

After evaluating 10 websites from the thesis data set, an inter-rater reliability test was conducted for each evaluated web resource to measure the evaluation agreement between the two indexers. This step is essential to measure the consistency among the two indexers.

[104] Two non-professional colleagues were trained during the course of two weeks on the practice of evaluating indexing keywords.

[105] By blindly, the researcher means that both indexers do not know which keyword list belongs to which set (i.e. folksonomy or Yahoo TE).

The inter-rater agreement reliability test that the researcher used to measure the consistency of classifying keywords into categories without any ordering (i.e. nominal data), was the Kappa (k) coefficient, a widely accepted measurement developed by (Cohen, 1960). The value of the resulting Kappa coefficient indicates the degree of agreement between the two raters. For interpreting the meaning of the resulting Kappa value the researcher used (Landis and Koch, 1977) interpretation, where $0 \leq k < 0.2$ means slight agreement, $0.2 \leq k < 0.4$ means fair agreement, $0.4 \leq k < 0.6$ means moderate agreement, $0.6 \leq k < 0.8$ means substantial agreement, and $0.8 \leq k < 1.0$ means almost perfect agreement.

Table 7.2 shows the overall average degree of agreement between the two indexers for the 10 evaluated web resources. The obtained Kappa value for both sets falls in the fair level of agreement, which is considered satisfactory (Bayerl et al., 2003) for the purpose of this experiment. However, the results show that agreement between the indexers about the folksonomy set is slightly lower (0.2005) than their agreement about the Yahoo TE set (0.2162); the difference is statistically significant at $p < 0.001$. The lower Kappa value for the folksonomy set was due to a slight disagreement in evaluating one of the websites in that set, which affected the results accordingly.

	Average Inter-Rater Agreement [Kappa-coefficient value]
Folksonomy	0.2005
Yahoo TE	0.2162

Table 7.2: Average Inter-Rater agreement for the ten evaluated web resources in phase 1

The values summarized in Table 7.3 show the average mode value for each evaluated website from both indexers. For all values except for site 2, 5 and 8, the results for the folksonomy set was higher or equal to Yahoo TE values. By further inspecting the three cases (2, 5 and 8), the researcher has found that what affected the average mode value in these three cases in the folksonomy set was the amount of general tags used to describe these web resources compared to the same Yahoo TE set. In contrast, Yahoo TE extracted more specific keywords (i.e. the same or narrower terms).

The results also show that the folksonomy and Yahoo TE sets scored an equal mode value (4 = relevant) for all sites. The values for the Yahoo TE varied considerably compared to the folksonomy values but the most frequent value in Yahoo TE was still (4) which appeared 3 times compared to 7 times in the folksonomy set.

Moreover, the results show that the folksonomy set has a higher mean and lower standard deviation i.e. 4.15(0.24), this indicates a low variance in the views of the two indexers towards classifying folksonomy tags compared to the values for Yahoo TE, i.e. 3.55(1.01), which indicates a high variance in the views of the two indexers. These results indicate that the folksonomy tags are more relevant to the human indexer's conception than Yahoo TE keywords. Furthermore, the difference between the two means was statistically significant at $p < 0.001$.

Site	**F**	**K**
1	4.5	4
2	4	4.5
3	4	3
4	4	2.5
5	4	4.5
6	4.5	3
7	4	1.5
8	4	4.5
9	4	4
10	4.5	4
Mean	4.15	3.55
SD.	0.24	1.01
Mode	4	4

Table 7.3: The average mode values for each website in both Folksonomy (F) and Yahoo TE (K) set along with the mean, mode and standard deviation for all 10 evaluated websites

The results of this phase gave the researcher the big picture of the semantic relationships held in the folksonomy and Yahoo TE keywords compared to the two indexers views. To better understand the semantics of each classified keyword in the folksonomy and Yahoo TE sets, an in depth analysis is carried out in phase 2.

7.6.2 Phase 2

The role of phase two was to inspect in more detail the semantic categories of the folksonomy set and the Yahoo keywords set compared to the web resource hierarchical listing in the DMOZ[106] directory and to its title keywords (afterwards, these will be called descriptors). Thus, the two indexers were provided with another categorization. The new categorization values were adopted from (Kipp, 2006). Kipp built her scale instrument based on the different relationships in a thesaurus as an indication of closeness of match, into the following categories:

- Same, *encoded 7*- the descriptors and tags or keywords are the same or almost the same (e.g. plurals, spelling variations and acronyms),
- Synonym, *encoded 6*- the descriptors and tags or keywords are synonyms,
- Broader Term (BT), *encoded 5*- the keywords or tags are broader terms of the descriptors,
- Narrower Term (NT), *encoded 4*- the keywords or tags are narrower terms of the descriptors,
- Related Term, *encoded 3*- the keywords or tags are related terms of the descriptors,
- Related, *encoded 2*- there is a relationship (conceptual, etc) but it is not obvious to which category it belongs to,
- Not Related, *encoded 1*- the keywords and tags have no apparent relationship to the descriptors, also used if the descriptors are not represented at all in the keyword and tag lists.

The two indexers applied the modified categorization scale to a sample of 10 bookmarked websites that were chosen from the experiment corpus.

After evaluating the 10 bookmarked websites, an inter-rater reliability test was conducted to evaluate the agreement between the two indexers in their evaluation of each web resource.

[106] http://dmoz.org/ [last accessed 5/2/2007]

Table 7.4 shows the degree of agreement between the two indexers. The agreement between the two indexers resulted in a fair level of agreement with almost equal scores for the folksonomy set (0.2257) and the Yahoo TE set (0.2241). The difference between the two means was statistically significant at $p< 0.001$.

	Average Inter-Rater Agreement [Kappa-coefficient value]
Folksonomy	0.2257
Yahoo TE	0.2241

Table 7.4: Average Inter-Rater agreement for the ten evaluated web resources in phase 2

The values summarized in Table 7.5 show the average mode value for each evaluated website from both indexers. Notice this time for all values, except for site 3, the results for the folksonomy set was higher than Yahoo TE values. By further inspecting site 3, the researcher has found that what caused the decline of the average mode value in this site was the number of tags assigned to it, i.e. 18 tags compared to 28 keywords from Yahoo TE, and also the category of the tags used to describe the website, which fall more in the related category.

The results also show that the folksonomy set scored a higher mode value (5) compared to Yahoo TE (2). However, the results show that the folksonomy set has a higher mean and higher standard deviation i.e. 4.45(1.28), which indicates a high variance in the views of the two indexers towards classifying the folksonomy tags, compared to the values for Yahoo TE, i.e. 2(0.71), which indicates a lower variance in the views of the two indexers. The difference between the two means was statistically significant at $p< 0.001$.

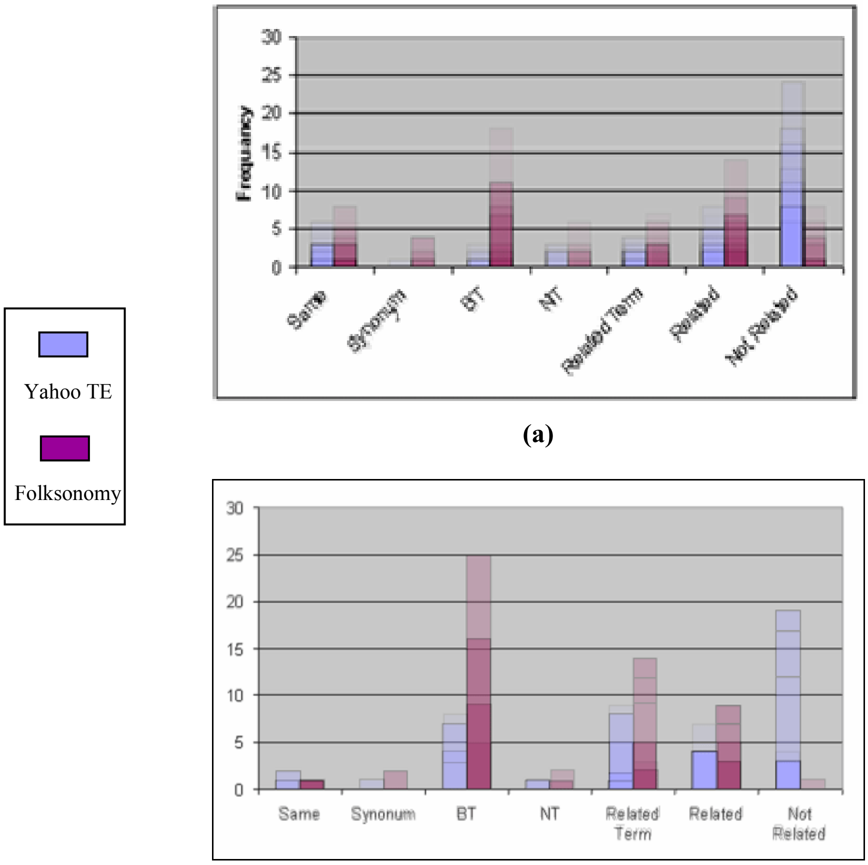
The resultant statistical analysis of this phase stressed the finding of the previous phase and gave more insight in how folksonomies are considered semantically richer than Yahoo TE keywords.

Site	F	K
1	5	1.5
2	5	1
3	1.5	2
4	5	2.5
5	5	2
6	3.5	2
7	5	3
8	6	2
9	3.5	3
10	5	1
Mean	4.45	2
SD.	1.28	0.71
Mode	5	2

Table 7.5: The average mode values for each website in both Folksonomy (F) and Yahoo TE (K) set along with the mean, mode and standard deviation for all 10 evaluated websites

Furthermore, to visualize the results of this phase, a two-column bar graph was generated for each evaluated web resource to reflect the results of each category, i.e. the Gray bars denote the Yahoo keywords frequency and the Dark-Gray bars denote the folksonomy tags frequency.

Figure 7.2 shows the accumulated bar-graph, for both indexers, obtained by juxtaposing each individual bar graph of the 10 evaluated web resources in a layered fashion so that a general conclusion can be drawn easily.



(a)

(b)

Figure 7.2: A visualization of the categorization results for the 10 web resources layered on top of each other shaping a ghost effect, (a) corresponds to the results of the first indexer (b) corresponds to the results of the second indexer.

Comparing the two figures shows that there is almost a good agreement between indexer (a) and indexer (b) in the assignment of Yahoo TE keywords in the 'not-related' category. However, this agreement starts to fluctuate, in order of magnitudes between the two indexers, in the similarity categories (i.e. Same, Synonym, BT, NT, Related Term and Related).

For instance, in Figure 7.2.(a), the folksonomy tags are accumulating more around the 'Broader Term' and 'Related' category, while in Figure 7.2.(b), the folksonomy tags are accumulating more around the 'Broader Term' and 'Related Term' category.

The figure also shows that most of the folksonomy tags fall in the similarity categories compared to a small portion which falls in the 'not related' category. In contrast, most of the Yahoo keywords fall in the 'not related' category compared to a small portion distributed in the similarity categories. Moreover, the figure shows that in all similarity categories the folksonomy set outperforms the Yahoo keyword set.

Finally, the researcher believes that the variance between the two indexers categorization was due to either the different interpretation of the categories meaning or the use of single category with different frequencies, as in the case of indexer (b), thus a further marginal homogeneity analysis using the Stuart and Maxwell test to identify the sources of variability (Bayerl et al., 2003) will be considered for future work.

More in depth analysis of Phase 2

In this section a detailed analysis of both the Yahoo keywords set and the folksonomy set falling in the ‘not related’ and ‘related’ categories is discussed.

A) Unrelated tags

To explore in greater depth the nature of tags falling in the ‘not related’ category, a further inspection was carried out to analyze the type of tags and keywords found in this category.

Folksonomy tags falling in the ‘not related’ category tend to be either time management tags e.g. ‘todo’, ‘toread’, ‘toblog’, etc., or expression tags e.g. ‘cool’, self-reference tags and sometimes unknown/uncommon abbreviations.

Time management tags, as Kipp said, suggest that the users want to be reminded of the bookmarked resource, but have not yet decided what to do with it. These kinds of tags do not appear in any controlled vocabulary or thesaurus; they are made up for the user’s own needs and do not have any value to anyone except the individual who created them.

Another common type of unrelated tag is the use of expression tags e.g. ‘cool’, ‘awesome’, etc. These reflect what the users think of the bookmarked resource. These tags suggest that the bookmarked web resource might be useful.

Self-reference tags include any tag that has to do with the user's own interest. For instance, delegating content to people e.g 'forchris' or referencing own self e.g. 'mywork' or 'myblog'. These tags usually appear once or twice among all the tags for a given bookmarked web resource.

On the other hand, Yahoo keywords falling in the 'not related' category do not follow a recognized pattern as folksonomy tags do. Most keywords seem to be words that have occurred frequently in the text or in the URL of a web resource; alternatively the position of the word and its style (e.g. heading or sub-title) might be the reason for extracting it. The algorithms that Yahoo TE uses to extract keywords from web sites are obscure which affects further analyses of the extracted keywords.

B) Related tags

This category represents relationships that are ambiguous or difficult to place into the previous similarity categories. These tags often occur when there is a relationship between the tag or keyword and its field of study, or/and a relationship between two fields of study (Kipp, 2006). An example of the first mentioned relationship would be of a web resource talking about open source software which has tags such as 'code' or 'download'. These two tags do not appear explicitly in the DMOZ directory listing nor in the title of the web resource; however, they do describe the field of 'open source' software where someone can download and play with the code. Furthermore, in a web resource that gives examples about FreeBSD, a particular version of the UNIX operating system, del.icio.us users' have tagged the web resource with related tags such as: 'tutorial', 'tips', and 'how-to', these tags were not explicitly mentioned in the web resource; however, they contributed to the description of the web resource by giving it a pedagogical dimension.

Another example of a relationship between two fields of study is a web resource about an open source office application called 'NeoOffice' for the Mac operating system. This web resource is tagged with tags such as 'Microsoft' and 'OpenOffice' to denote the relationship between the 'Mac OS' and 'Microsoft' and between 'NeoOffice' and 'OpenOffice' applications.

7.6.3 Phase 3

As mentioned in the experiment setup, the role of phases three and four was to find the percentage of overlap between the folksonomy set and the keywords generated by Yahoo TE. In this phase and the next one, folksonomy tags, Yahoo TE keywords and the indexer keywords are treated as abstract entities which do not hold any semantic value. This assumption will help see where folksonomies are positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords, and to measure the scope of this overlap.

The overlap measurement used in the comparison framework was interpreted using set theory (Stoll, 1979). The researcher considered the folksonomy set of tags as set F, keywords set from Yahoo TE as set K and keywords set from the indexer as set I, hence:

F = {the set of all tags generated by people for a given URL in del.icio.us}

K = {the set of all automatically extracted keywords for a given URL}

I = {the set of all keywords provided by the indexer}

Using set theory the degree of overlap was described using the following categories:

1. No overlap e.g. $F \neq K$ or $F \cap K = \emptyset$ (i.e. empty set).
2. Partial overlap (this is know as the intersection) e.g. $F \cap K$
3. Complete overlap (also know as containment or inclusion). This can be satisfied if the number of overlapped keywords equals to the folksonomy set (i.e. $F \subset K$) or if the number of overlapped keywords equals to the Yahoo keyword set (i.e. $K \subset F$) or if the number of overlapped keywords equals both folksonomy and keyword set (i.e. $F = K$).

The collected data set (described in the Data Selection subsection) was dispatched to the comparison framework to measure the percentage of overlap between folksonomy tags and Yahoo TE keywords.

After observing the results of 100 websites the researcher can detect that there is a partial overlap ($F \cap K$) between folksonomy tags and keywords extracted using Yahoo TE. The results show that the mean of the overlap was 9.51% with a standard deviation of 4.47%, which indicates a moderate deviation from the sample mean. Also the results show both the maximum and minimum possible overlap with values equal to 21.82% and 1.96% respectively. This indicates that there is neither complete overlap nor no overlap at all, and the most frequent percentage of overlap (i.e. mode) was 12.5%.

Figure 7.3 shows a histogram of the frequency of the results which graphically summarizes and displays the distribution of the percentage of the overlaps using short intervals (2.5 percentages wide). Notice that most of the overlap values (14 values) fall in the interval between 7.5 and 8.75, while the least of the overlap values fall at the ends of the histogram. The shape of the histogram forms the beginning of a normal curve, thus, the researcher believes that with more evaluated websites the histogram will ends up being an approximate normal curve, which can be used as a tool to estimate proportion of overlaps with appropriate margins of errors.

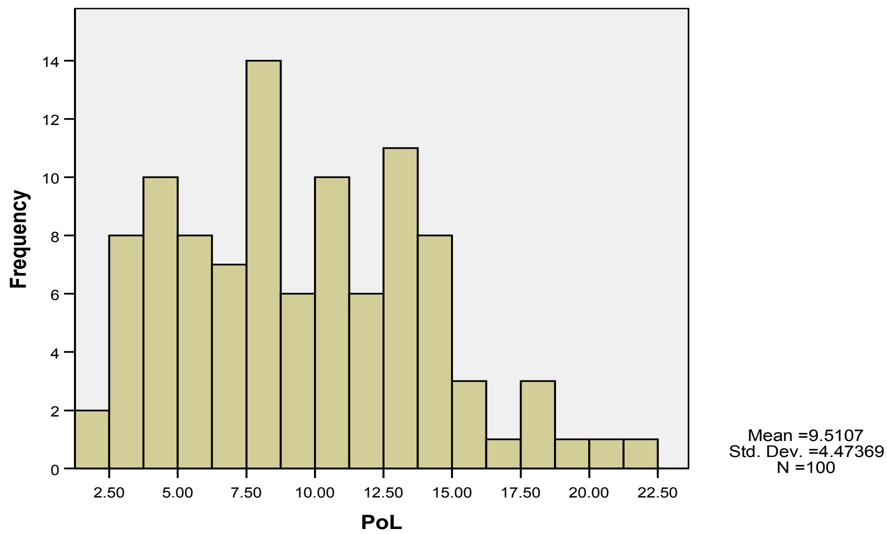


Figure 7.3: Histogram of the Percentage of Overlap (PoL) for 100 websites.

Finally, the results of this phase showed that folksonomy tags can not be replaced by automatically extracted keywords, even if there was a marginal overlap between the two sets. However, to inspect in more depth the position of folksonomies in the spectrum from professionally assigned keywords to context-based machine extracted keywords, phase 4 is carried out to envision the place of folksonomy tags in this spectrum.

7.6.4 Phase 4

The role of phase four is to check the correlation between folksonomy and human keyword assignment, and also between Yahoo TE keywords and the human assignment. This step is necessary to see which technique is most closely related to a cataloguing (indexation) output.

Therefore, tools from library and information science were used to index a sample of 20 websites taken from the thesis data set and to check them against folksonomy and Yahoo TE sets. The assignment of keywords was done using the following guidelines:

1. The use of controlled vocabularies of terms for describing the subject of a website, such as DMOZ (the Open Directory Project) and Yahoo directory.
2. The source code of each website was checked to see if it contains any keywords provided by the website creator.
3. The position (i.e. in titles) and emphasis (such as bold) of words in a website were considered.
4. The indexer was also asked to read the content of the website and generate as many index keywords as possible.

After the end of this process the set of produced keywords for each website was compared using the comparison framework, once with the keywords from the Yahoo TE set and another with the folksonomy set. This step is essential to see whether folksonomies produced the same results as if a human indexer was doing the process.

The results show (see Figure 7.4) that there is partial overlap between the two sets and the indexer set, but this time with higher scores. The folksonomy set was more correlated to the indexer set with a mean of 19.48% and a standard deviation of 5.64%, while Yahoo TE set scored a mean of 11.69% with a standard deviation of 7.06%. Furthermore, the experiment showed one case where there is a complete overlap (inclusion) between the folksonomy set and the indexer set.

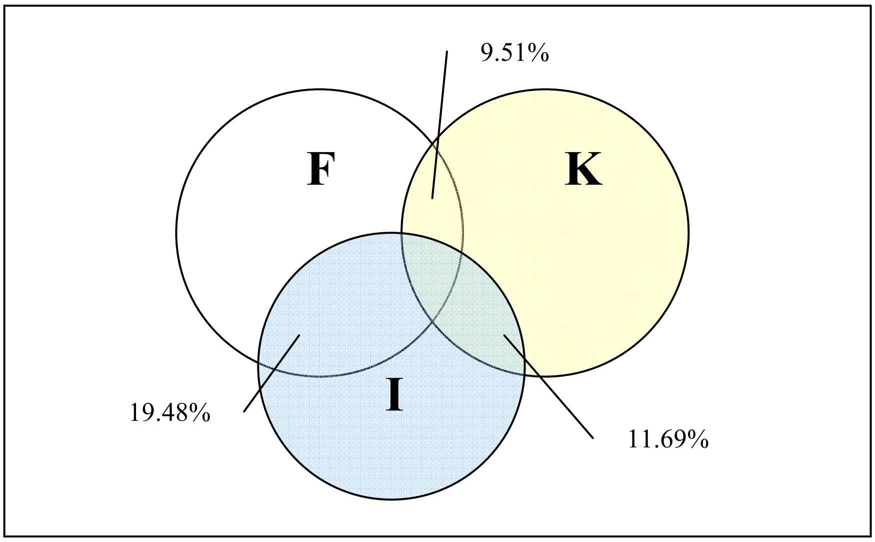


Figure 7.4: A Venn diagram that shows Folksonomy (F), Yahoo TE (K) and the human indexer (I) sets as three distinct circles and highlights the percentage of the overlap between the three sets.

The results of this phase showed that folksonomy tags are more oriented toward the professional indexer keywords. Therefore, this finding positioned the folksonomy tags more closely to the indexer keywords in the spectrum from professionally assigned keywords to context-based machine extracted keywords.

7.7 Discussion

After completing the four phases of this experiment, a number of observations were made. As a first impression, phase 1 was used to evaluate the relevance of the folksonomy tags and Yahoo TE keywords to the human conception. Thus, the results of this phase indicate a significant tendency of the folksonomy tags towards depicting what a human indexer might think of when describing what a web resource is about compared to Yahoo TE keywords.

Another interesting observation was found in phase 2, where some folksonomy tags fall in the ‘Narrower Term’ and ‘Synonym’ categories. These categories were less common than the ‘Broader Term’, ‘Same’ and ‘Related Term’ categories, which implies from the researcher point of view, that this might be attributed to the low number of specialized people who uses the del.icio.us bookmarking service, or it might be due to the varied backgrounds of the del.icio.us users.

In phase 3, the average overlap between the folksonomy set and Yahoo keywords was 9.51%, which implies that there was only a minor intersection between the two sets. Thus, folksonomy tags cannot be replaced completely with keywords generated by machine (in this case Yahoo TE). This finding also opens the door for other potential research directions, for instance in the field of language technology and semantics, which is out of the scope of this research.

In phase 4, the folksonomy tags showed a greater tendency to overlap with the professional indexer produced keywords than with the Yahoo Thus, the results showed that the folksonomy set was more correlated to the indexer set with a mean of 19.48%, while Yahoo TE set scored a mean of 11.69%. This finding also emphasis the researcher claim about the better correlation between folksonomies and professional indexing compared to the correlation between professional indexing and context-based machine extracted keywords.

Finally, it is worth mentioning that the results from this experiment have not been evaluated against a large corpus, especially where this concerns the sample size used by the indexers. This was due to the high effort needed for manual indexing. However, to get a fair judgment the researcher has attempted to choose varied websites topics spanning multiple domains as mentioned in Table 7.1.

7.8 Chapter Summary

In this chapter the researcher has described an experiment consisting of four phases, each phase was designed to explore the semantic value of folksonomies from different perspective.

The first and second phases evaluated the relevance of the folksonomy tags and Yahoo TE generated keywords to the human conception. The evaluation was performed by two trained indexers using an evaluation scale based on the different relationships in a thesaurus as an indication of the closeness of match. The third and fourth experiments were conducted to find the percentage of overlap between the folksonomy tags, keywords generated by Yahoo TE and the human indexer keywords.

The results of phases one and two show that the two human indexers both agreed on the richer semantics of the folksonomy tags compared to Yahoo TE, with $p < 0.001$. The results of phase three showed that the average overlap between the folksonomy set and Yahoo keywords was 9.51%, and the results of phase four showed that the folksonomy set was more correlated to the human indexer set with a mean of 19.48%, while Yahoo TE set scored a mean of 11.69%.

It is clear from the results of this experiment that the folksonomy tags agree more closely with the human generated keywords than those automatically generated. The results also showed that the trained indexers preferred the semantics of folksonomy tags compared to keywords extracted by Yahoo TE. These results were very encouraging, and illustrated the power of folksonomies. The researcher has demonstrated that folksonomies have added new contextual dimensions that are not present in automatic keywords extracted by machines.

This experiment was a first step towards future evaluation techniques on which the researcher is planning to embark. These techniques will measure the semantic value of folksonomies based on knowledge engineering principles and methods, such as Formal Concept Analysis (FCA) and frame-based systems (Stuckenschmidt and Harmelen, 2004). In such techniques concept hierarchies (or 'concepts lattices') are used to define a given term. By using this approach, the intended meaning of a term is addressed instead of finding the exact syntactic match.

So to conclude, folksonomies are very popular and a potential rich source for metadata. The rational of this experiment was based on the motivation of investigating whether folksonomies could be used in semantically annotating web resources. The findings of this experiment were used to justify the use of folksonomies in the process of generating semantic metadata for annotating learning resources as will be described in the next chapter.

Chapter 8
The FolksAnnotation Tool System Architecture

The experiment carried out in the previous chapter (i.e. Chapter 7) showed that folksonomies are potential source of rich metadata. With this in mind, the researcher has implemented a tool, which she named ‘FolksAnnotation’ that utilises folksonomy tags in an attempt to create semantic metadata.

In this chapter a detailed discussion about the FolksAnnotation tool system architecture and its components’ functionalities is presented. Figure 8.1 shows the two main processes used in the FolksAnnotation Tool, namely: the *Normalisation pipeline* which adopts its idea from the classical text normalisation[107] process in IR, and the *Semantic Annotation pipeline* that works as a dictionary lookup process which assigns an ontology instance to a given web resource. Next, a detailed description of the two processes is discussed.

[107] http://en.wikipedia.org/wiki/Text_normalization [last accessed 13/2/2007]

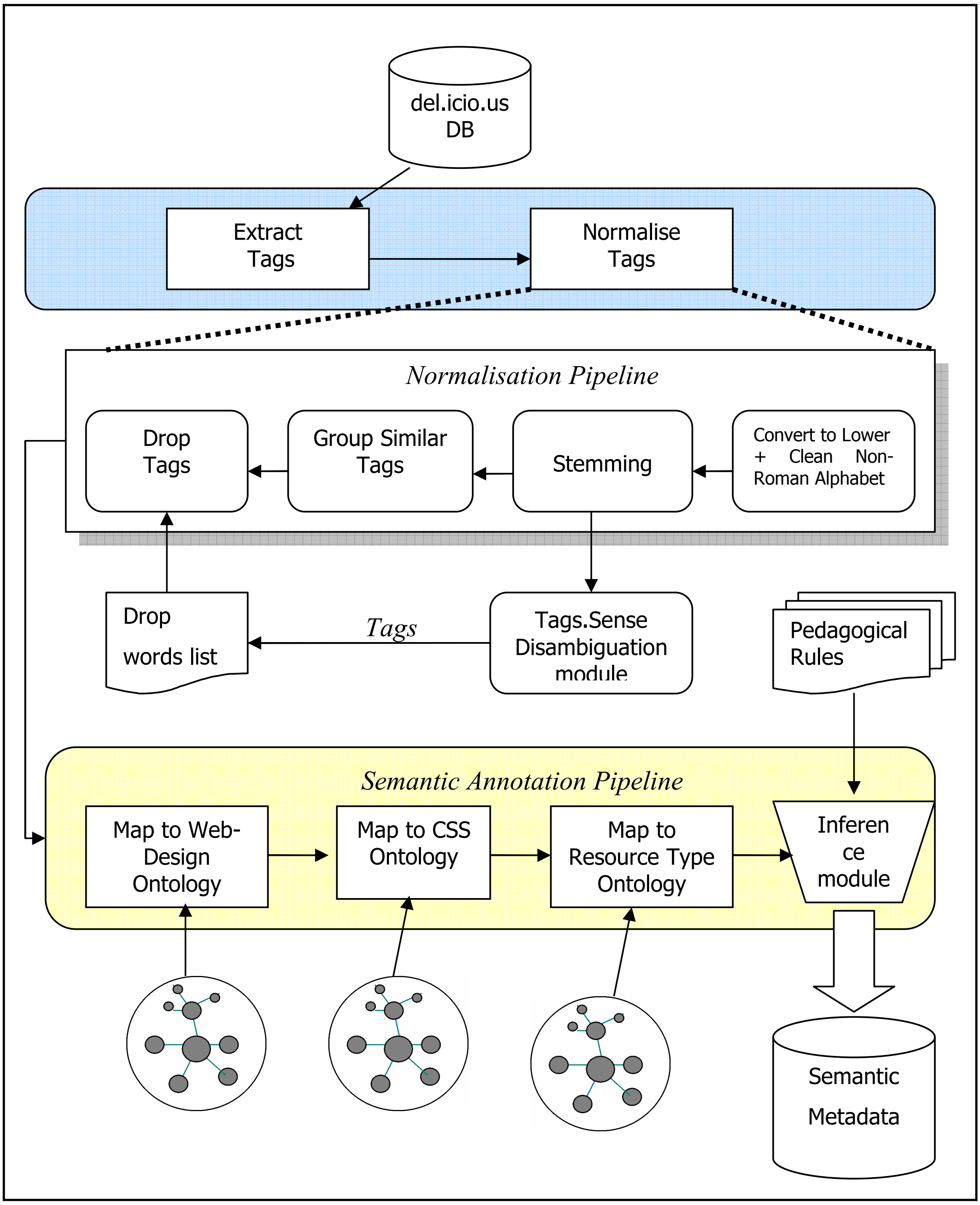


Figure 8.1: Overview of the system illustrating the interplay of the different components

8.1 Tags Extraction and Normalisation

This process starts by fetching a bookmarked web resource from the del.icio.us bookmarking service, then the tag extraction process starts extracting viable information from the web page of the bookmarked web resource, this includes: Title, URL, number of people who bookmarked the resource and the list of all tags assigned to the bookmarked web resource.

The extracted tags are then passed to the normalisation process which applies a series of filters for cleaning the tags. The filters are applied sequentially in the following order:

1. **Lower-case filter**: Tags are converted to lower case so that string manipulation (e.g. comparison) can be applied to them easily,
2. **Non-English filter**: Non-Roman Alphabets are dropped; this step is to insure that only English tags are present when doing the semantic annotation process,
3. **Stemming filter**: Tags are stemmed (e.g. convert plural to singular) using a modified version of the Porter Stemmer[108]. The reason for choosing the Porter Stemmer is two fold: first, the stemmer has been ported into many programming languages including Java which is used as the language to implement the FolksAnnotation Tool; second, other stemmers such as UEA-Lite[109] and Lovins stemmer[110] were aggressive in handling special suffix cases which might lead to nonsense words.
4. **Tags sense Disambiguation filter**: stemmed tags are passed to this module to remove ambiguous tags; further details are presented in section 8.1.1.
5. **Grouping filter:** similar tags and substrings are grouped,
6. Finally, the **removal filter**, where the general concept tags (e.g. programming, web, etc) in the thesis domain of interest and the ambiguous tags are eliminated.

The normalisation process is done automatically and it is potentially useful to clean up the noise in people's tags. Table 8.1and Table 8.2 depict this process by giving an example of tags before and after normalization. Also, Figure 8.2 shows a screen shot of part of the normalization program 'in action' on another web resource.

---

108 http://www.tartarus.org/~martin/PorterStemmer/ [last accessed 13/2/2007]

109 http://www.cmp.uea.ac.uk/Research/stemmer/ [last accessed 13/2/2007]

110 http://www.cs.waikato.ac.nz/~eibe/stemmers/index.html [last accessed 13/2/2007]

123 css	18 gui	7 howto	2 example
56 design	14 html	5 tips	
47 graphs	12 webdev	5 usability	
46 webdesign	10 reference	5 graphing	
28 graph	9 development	3 coding	
27 web	8 cool	3 stats	

Table 8.1: Tags used to annotate a sample web resource[111] stored in the del.icio.us service (before normalization)

123 css	10 reference	5 usability
80 graph	8 cool	3 code
18 gui	7 howto	3 stat
14 html	5 tip	2 example

Table 8.2: Tags after applying the normalization process.

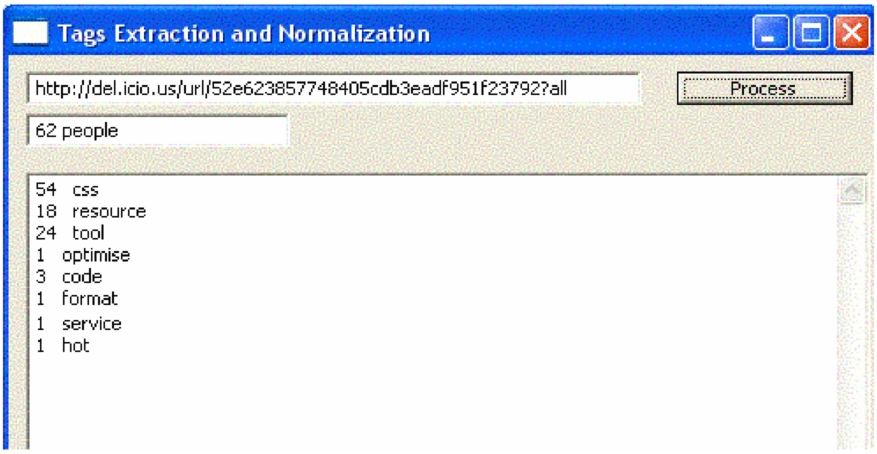


Figure 8.2: A screenshot of the finished normalization process for a bookmarked web resource

At the end of the normalisation phase a list of normalised folksonomy tags are ready to be used in the semantic annotation process; each tag in the list is associated with a number that reflects the tag's frequency occurrence with a given web resource. This number will come in handy when determining the main theme of a web resource

[111] http://apples-to-oranges.com/blog/examples/cssgraphs.html, Date Accessed May 12, 2006 at 10:00 PM GMT

(e.g. Figure 8.13 shows that CSS is the highest tag number which actually reflects the main theme of the bookmarked web resource).

The normalised tag list is then passed to the semantic annotation process, where each normalised folksonomy tag will be mapped to a corresponding ontological term in one of the three ontologies in the system. This process will map ontology instances as descriptors to a web resource.

8.1.1 Tags Sense Disambiguation Module

At the first run of the 'FolksAnnotation' tool without the support of the Tag Sense Disambiguation (TSD) module, the researcher noticed that some inappropriate tags got mapped to the CSS ontology as being part of its instances. These false-positive assignments were due to the multiple meaning of some tags used in different context. For example, the 'property' concept in the CSS ontology has an instance called 'list'. This instance if put in another context might mean 'a collection of things'. These multiple senses of the tag can mislead the annotation procedure since there is no way in the tool to distinguish the semantics of the tags. To overcome the ambiguity of some instances in the CSS ontology, a TSD method was adopted.

TSD is a derivative idea that comes from the Word Sense Disambiguation (WSD) technique, a well-known problem in the natural language processing field. Thus, the goal of a WSD algorithm is "… *to associate the most appropriate meaning or senses to a word w in document d, by exploiting its window of context (or more simply context) C, that is a set of words that precede and follow w. The senses are selected from a predefined set of possibilities, usually known as sense inventory [such as WordNet]*" (Degemmis et al., 2006). In other words, the working idea behind WSD is to enumerate the set of all possible meanings of a word in a given context, and then determine, based on some techniques, which sense is the most appropriate for a given context.

Mihalcea and Moldovan (1999) have classified WSD techniques into five broad categories:

- Dictionary based WSD

- Supervised WSD
- Unsupervised WSD
- Machine learning WSD
- Hybrid methods that combine several techniques with each other.

For my algorithm, the dictionary based WSD has been adopted with some modifications to construct the new TSD algorithm, i.e. make use of ontologies in place of dictionaries to remove the ambiguous tags.

The proposed TSD algorithm is based on constructing a semantic matrix of the concepts' instances for a given domain ontology, in the thesis case this will be the CSS ontology, and then recording the co-occurrence of these instances in the list of tags. Table 8.3 shows an excerpt of a populated semantic matrix built for the 'list' CSS ontology instance. This table is constructed based on the semantic relationships between the 'list' instance and its neighbouring instances in the ontology; see Figure 8.4. The numbers in the columns represent how many times the instance 'list' co-occurred with another instance as will be shown in the next example.

Instances	CSS	Navigation	Menu	li
list	3	1	0	1

Table 8.3: The Semantic Matrix for the 'list' instance, the row headings represents the ambiguous word while the columns headings represent the neighbour instances in the ontology

To illustrate the functionality of the TSD algorithm an example is given below:

- "**Listamatic: Rollover horizontal list**"[112] web resource was bookmarked by 5 people in del.icio.us as follows:

[112] http://css.maxdesign.com.au/listamatic/horizontal02.htm# [last accessed 13/2/2007]

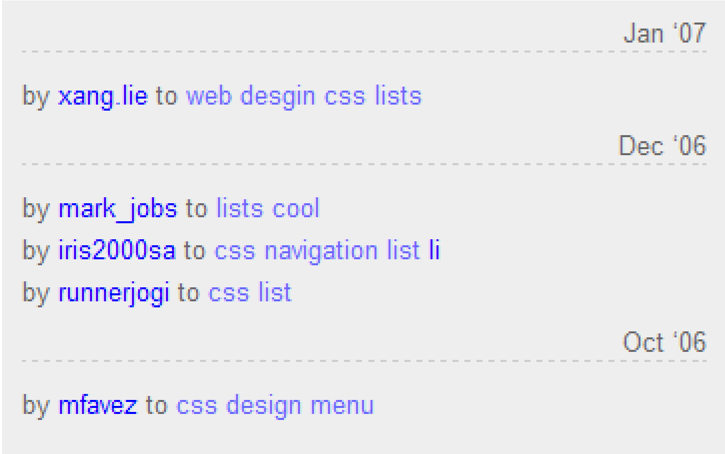


Figure 8.3: A screenshot showing the appearance of the 'list' tag in "Listamatic: Rollover horizontal list" web resource.

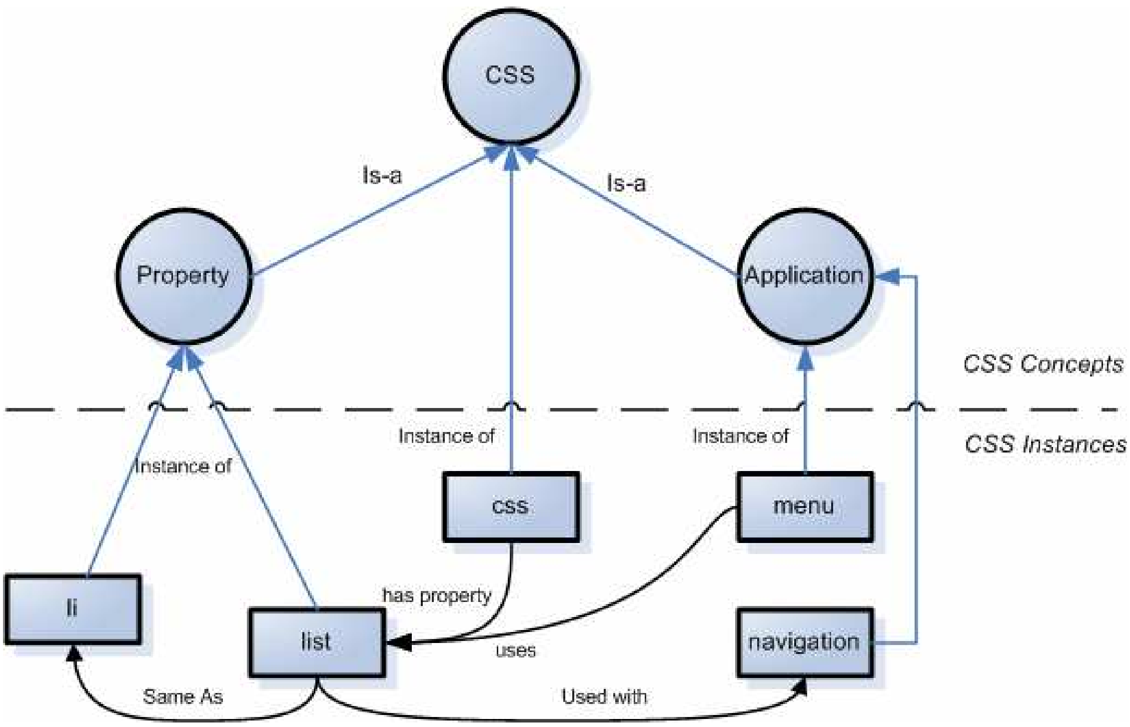


Figure 8.4: A schema that depicts the semantic relationships between the 'list' instance and its neighbouring instances in the CSS ontology.

- Let Tag_a = 'list', be the ambiguous tag whose ambiguity we want to resolve. Based on the semantic matrix given in Table 8.3, the 'list' instance is semantically associated with 'CSS' as 'has property' relationship, with 'menu' as 'uses' relationship, with 'navigation' as 'used with' relationship and with 'li' as 'same as' relationship as shown in Figure 8.4.
- The algorithm starts by traversing the list of all tags in the posts (as one long list), and records the number of times two instances of the semantic matrix co-occurred in the tag's list.

- After finishing recording all the tags co-occurrence, a tag ambiguity index is calculated based on the following formula:

$$Tag\_Index(Tag_a) = \frac{\sum_{i=1} co\_occur(Tag_a, Tag_i)}{All\,Tag_a}$$

Where Tag_a refers to the ambiguous tag, Tag_i refers to the co-occurred instance in the semantic matrix and $All\ Tag_a$ corresponds to the total number of occurrence for Tag_a in the tags' list.

So, given the previous example: Tag_Index('list') = 5/4 = 1.25. The value 1.25 is greater than 0.5 which represents a hypothetical threshold that the researcher has setup to qualify a tag as being from the ontology instances. The value of the threshold was based on a candidate value observed after repetitive trials on the experiment's data set. Therefore, if the *tag-index* for a given instance did not pass the test, it is appended to the list of words that need to be dropped from the tag list.

Another example is given next to show how the algorithm behaves when the tag 'list' is used for a different purpose. The "**CSS - Contents and compatibility**"[113] web resource was bookmarked by 477 people in del.icio.us. Figure 8.5 shows all the appearances of the 'list' tag compiled in one shot.

[113] http://www.quirksmode.org/css/contents.html

Figure 8.5: A screenshot showing the appearance of the 'list' tag in "CSS - Contents and compatibility" web resource.

The tag 'list' has appeared in that bookmarked web resource 5 times; however, the co-occurrence of the tag 'list' with other semantic instances was 0. This yields a *Tag-Index* value of zero, which indicates that the people who have bookmarked this resource were using the tag 'list' to mean 'a collection of things' and not the property 'list'.

Finally, it is worth mentioning that the idea of this algorithm is solely the work of the researcher and originated from her observations of the repetitive patterns spotted while analysing the list of tags in the del.icio.us bookmarking service. This pattern was then discovered to be useful for eliminating ambiguous tags. The researcher, however, does not claim that her finding makes any novel contribution to the field of natural language processing or information retrieval, which is beyond the intention of this thesis.

8.1.2 Related Work in Tags Disambiguation

Recently there has been very little academic research that aims to resolve the ambiguity of tags in the del.icio.us bookmarking system. This might be attributed to the recent appearance of these kinds of Web 2.0 applications. However, Zhang et al. (2006) were among the first researchers tackling this problem. They used a probabilistic generative model called the asymmetric Separable Mixture Model (SMM) model for data co-occurrence to infer the semantics of the folksonomies and to resolve the ambiguity of folksonomy tags. Their model has succeeded in resolving tags ambiguity and in grouping synonym tags together. However, to be able to use their proposed solution, a very large-scale data set needs to be used in order to train the model and infer tags' senses accordingly. This approach is somewhat problematic given a small focused domain such as ours, thus, the researcher thought that using ontological relationships co-occurrences to resolve tags ambiguity per web resource for a focused domain is less time-consuming and less resource-intensive as opposed to the SMM model.

Another similar investigation was carried out by (Kipp and Campbell, 2006). They have used Multi-Dimensional Scaling (MDS) for co-word clustering to visualise the relationships between tags. While their approach helped picture the relationships between tags for a given URL, their approach is still not practical for resolving tags ambiguity because it does not assign weights to tags based on their ambiguity level in a given context, i.e. it only shows the relationships visually.

8.2 Semantic Annotation Pipeline

The semantic annotation process is the backbone process that generates semantic metadata using the three ontologies. The process attempts to match folksonomy terms (after normalisation) from the bookmarked web resource against instances in the ontology (which works as a controlled vocabulary) and only selects those terms that appear in the ontology. This matching procedure is very conservative, i.e. only equivalent instances are matched. This is because the researcher has used string matching to look for instances in the ontology; however a radical improvement for this process is suggested in chapter 12.

```
Lookup the tag instance in the ontology
  If tag instance found then
    Add (URL, Instance, Property) to the metadata file
Get next tag
```

Figure 8.6: Pseudocode for the process of the semantic annotation.

Figure 8.6 demonstrates the process of semantic annotation; this is done by searching for each tag instance within the three ontologies and, if found, creating an RDF statement in the metadata file that associates the web resource URL with the tag instance by a property. The value of the property differs based on the tag value, for example if the tag instance is about an application in the CSS domain, the annotation process will use the 'hasApplication' property to make the RDF statement. The properties values can be obtained from the CSS ontology itself.

8.2.1 Inference Module

The *inference module* is responsible for associating pedagogical semantics (i.e. 'difficulty level' and 'instructional level') to the annotated web resource. These two values are generated from a set of inference rules feed to the inference engine by a separate file.

The pedagogical rules will only function if there is enough information available in the basic semantic descriptors. Figure 8.7 shows the interplay of the reasoning rule pipeline.

Rules are encoded in Turtle-based syntax[114] - Terse RDF Triple Language- supported by the Jena inference engine. This syntax is an extension of the N-Triples/Notation3 format (see chapter 5).

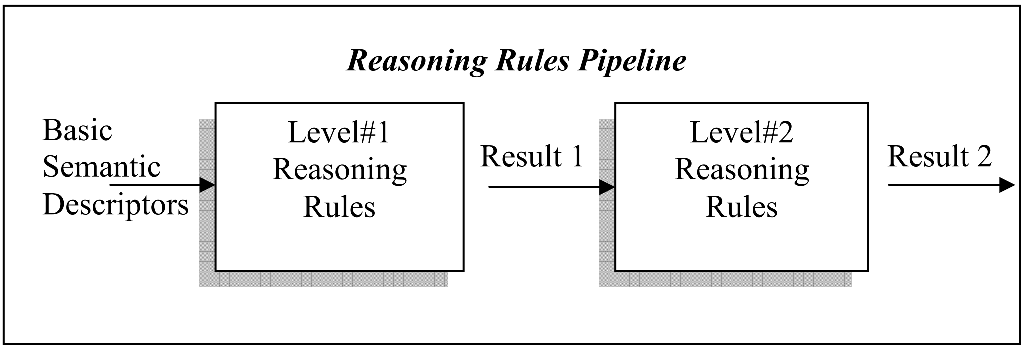


Figure 8.7: The Reasoning rules pipeline

[114] http://en.wikipedia.org/wiki/Turtle_(syntax), http://www.dajobe.org/2004/01/turtle/ [last accessed 13/2/2007]

```
@prefix lom: <http://kmr.nada.kth.se/el/ims/schemas/lom-educational#> .
@prefix sdo: < http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#> .

[rule1e:
(?url sdo:hasProperty sdo:background)  -> (?url lom:difficulty lom:Easy)]

[rule9m:
(?url sdo:hasElement sdo:span) -> (?url lom:difficulty lom:MediumDifficulty)]
```

Figure 8.8: Level 1 reasoning rules excerpt for the difficulty level descriptor

Level 1 reasoning rules are those simple, one predicate rules, that assign a value of instruction level and difficulty level based on the value of the basic semantic descriptors. This results in a rule for each possible ontological instance. Figure 8.8 and Figure 8.9 show two excerpts of the reasoning rules used to determine the difficulty level and instructional level, respectively.

```
@prefix sdo: <http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#> .

[rule1n:
(?url sdo:hasProperty sdo:background) -> (?url sdo:hasInstructionalLevel sdo:novice)]

[rule10m:
(?url sdo:hasElement sdo:span) -> (?url sdo:hasInstructionalLevel sdo:intermediate)]
```

Figure 8.9: Level 1 reasoning rules excerpt for instructional level descriptor

Level 2 reasoning rules are those rules which combine the results of the previous reasoning rules to form a unique instructional level and difficulty level value for a given web resource in whole.

To show how Level 2 rules operate, an example is given. The ‘Nifty Corners’[115] web resource, Figure 8.10, has in its tags list the instances ‘div’ and ‘rounded corner’, these tags have difficulty level of ‘medium’ and ‘difficult’ and instructional level of ‘intermediate’ and ‘advanced’ respectively. Level 2 reasoning rules are then

[115] http://www.html.it/articoli/nifty/index.html [last accessed 31/1/2007]

responsible for propagating the highest values among the difficulty level and instructional level for the basic semantic descriptors, in this case the ‘Nifty Corners’ web resource has a difficulty level of ‘difficult’ and instructional level of ‘advanced’.

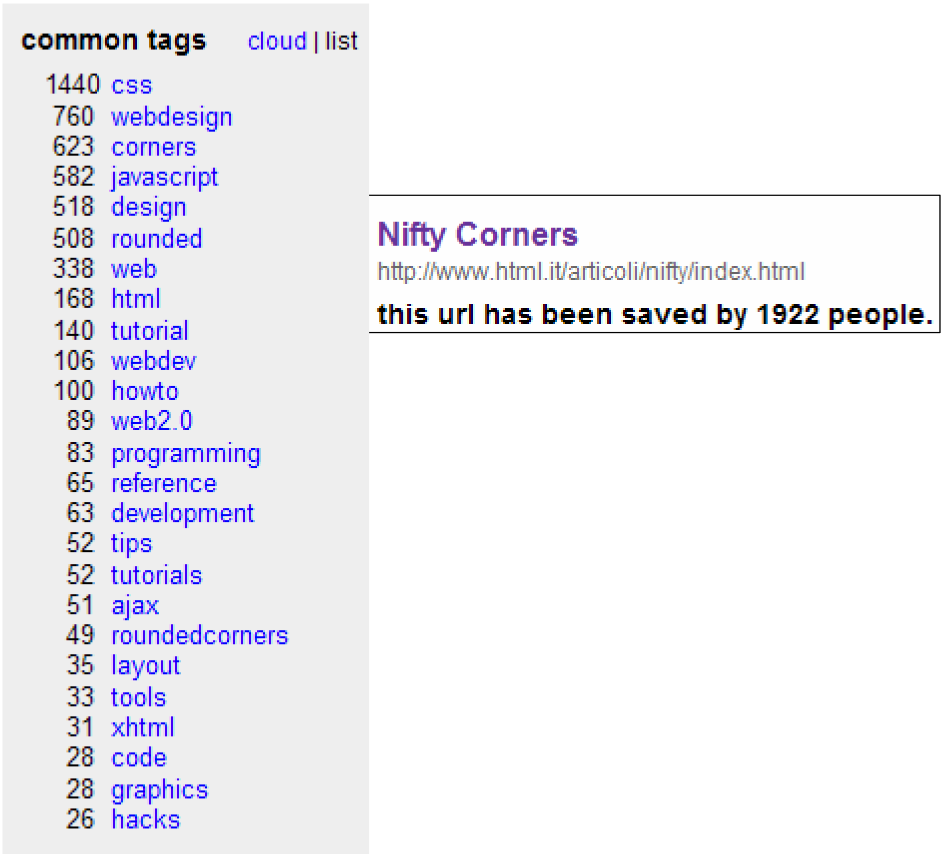


Figure 8.10: The folksonomy list for the ‘Nifty Corners’ web resource (date accessed 31-January-2007 @2:00 PM).

Finally, to simplify feeding the tool with the pedagogical rules, an interface consisting of a two dimensional table where rows represent the CSS instances and columns represent the required difficulty level and instructional level for each instance is shown in Figure 8.11. This editor is used to modify the values stored in the rules file.

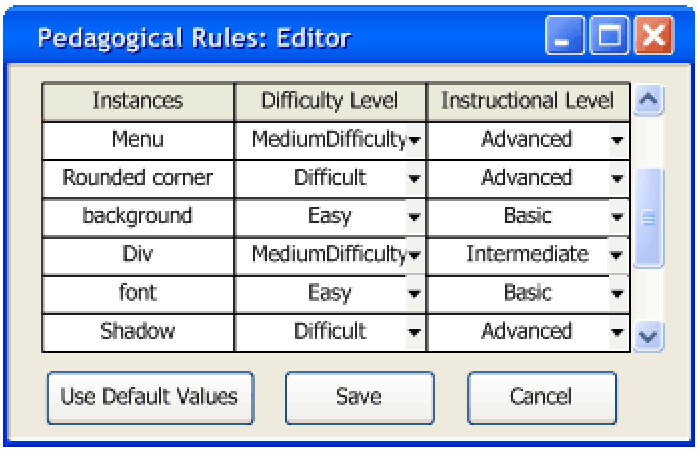


Figure 8.11: The pedagogical rules Editor

8.3 General Heuristic and the Resultant Semantic Metadata

In order to distinguish web resources that talk about CSS as their main theme from those which uses CSS as a supplementary technology; tags' frequencies are used as guides in determining whether to consider a given web resource for the semantic annotation process. Thus, the higher the number associated with a tag, the more likely it represents the main theme of a web resource. Figure 8.12 and Figure 8.13, underline this point graphically, by showing that when the CSS tag is lower in order, as seen in Figure 8.12, the main theme of the web resource is not about CSS, however it uses CSS as one of its technologies, and when the CSS tag is higher in order, as shown in Figure 8.13, the main theme of the web resource is about CSS.

This distinction is necessary to help in eliminating web resources that do not concentrate on CSS as their main topic, and also to help in maximizing the number of tags used in the semantic annotation process.

Figure 8.12: A list of tags for a website about Drag and Drop method, notice the position of the CSS tag in the list.

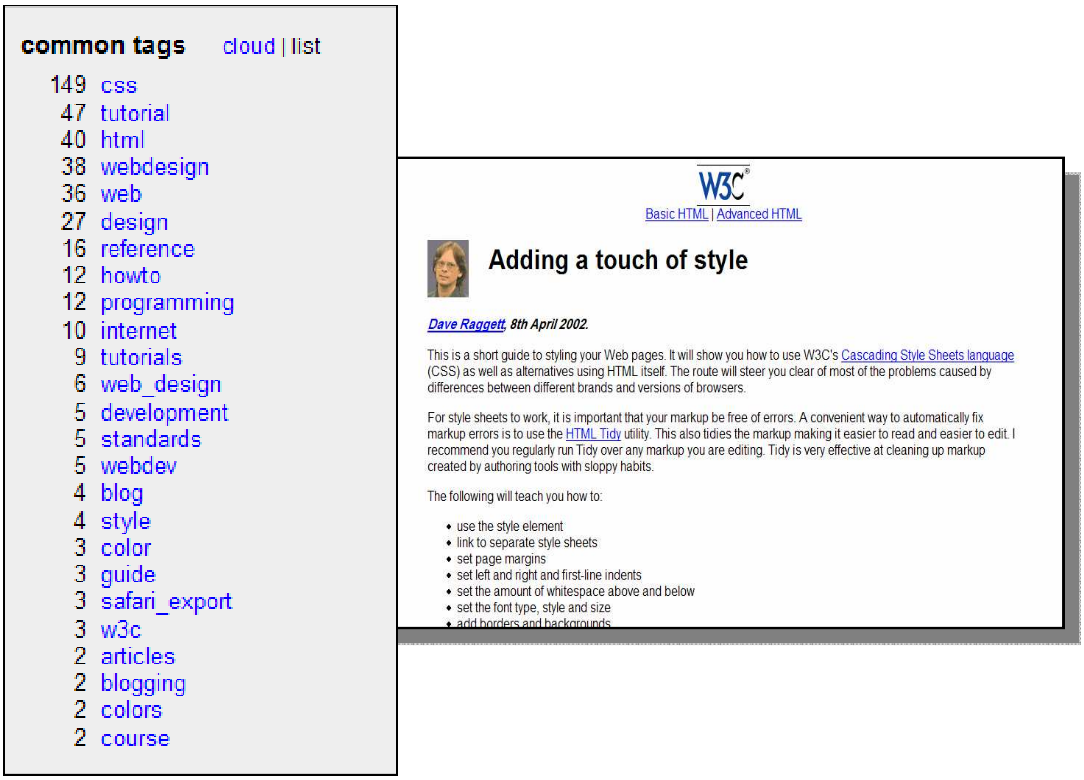


Figure 8.13: A list of tags for a website about CSS, notice the position of the CSS tag in the list.

Figure 8.14 shows the semantic metadata generated automatically for the ‘Nifty Corners’ bookmarked web resource; the original tags list was shown in Figure 8.10.


```
<rdf:RDF
  xmlns:sdo="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#"
  xmlns:wdo=" http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#"
  xmlns:rdo="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#"
  xmlns:rdf=" http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:lom="http://kmr.nada.kth.se/el/ims/schemas/lom-educational#"
  xmlns:dc="http://purl.org/dc/elements/1.1/" >
<rdf:Description rdf:about="http://www.html.it/articoli/nifty/index.html ">
<rdo:hasResourceType
rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#tutorial"/>
 <dc:subject> http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#html</dc:subject>
<rdo:hasResourceType rdf:resource="
http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#template"/>
<dc:subject>http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#ajax </dc:subject>
<sdo:hasTechnique
rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#roundedcorner"/>
<dc:subject>http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#dom</dc:subject>
<dc:subject> http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#css</dc:subject>
<sdo:hasLayout rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#box "/>
<rdo:hasResourceType
rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#article"/>
<sdo:hasInstructionalLevel rdf:resource="
http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#Advanced"/>
<rdo:hasResourceType rdf:resource="
http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#example"/>
<sdo:hasElement rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#div "/>
 <sdo:hasProperty rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#border"/>
<sdo:hasApplication rdf:resource=" http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#menu"/>
 <lom:difficulty rdf:resource=" http://kmr.nada.kth.se/el/ims/schemas/lom-educational#Difficult"/>
 <dc:subject>http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#dhtml </dc:subject>
  <dc:subject>http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#xhtml</dc:subject>
<dc:subject> http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#javascript</dc:subject>
 <sdo:hasApplication rdf:resource=" http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#button"/>
<rdo:hasResourceType rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#guide
"/>
 <rdo:hasResourceType
rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#reference"/>
 <dc:description>A(n) tutorial with title: 'Nifty Corners' suggest the knowledge of the following
topics: css,javascript,html,ajax,dhtml,js,dom,xhtml. This resource is also suggested to be used as:
code, example, article, template, reference, sample, guide, </dc:description>
<sdo:hasSubject rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#effect"/>
 <rdo:hasResourceType rdf:resource="
http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#code"/>
<rdo:hasResourceType
rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#sample "/>
<sdo:hasSubject rdf:resource="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#layout"/>
<dc:subject> http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#js</dc:subject>
  </rdf:Description>
</rdf:RDF>
```


Figure 8.14: The generated RDF Semantic metadata for the ‘Nifty Corners’ web resource.

8.4 Development Tools

The FolksAnnotation tool was built using Java 5 with the support of the Standard Widget Toolkit (SWT)[116] library and used Jena API[117] for ontology manipulation and inference.

As for the search portal, it is a web-based application that provides miscellaneous facets to access the generated semantic metadata. The interface was implemented using Tomcat servlet engine 5.5 that runs JSP pages and used Jena 2 API for ontology manipulation.

8.5 Chapter Summary

This chapter demonstrated the implementation of the FolksAnnotation Tool used to generate the semantic metadata from folksonomies. The chapter also discussed the implementation of the Tags Sense Disambiguation algorithm, which solves the problem of multiple meanings for a given tag. Finally, the chapter concluded with an example of the generated RDF semantic metadata and the general heuristic used to distinguish related CSS web resources from non-related ones.

The next chapter demonstrates the modelling of the three ontologies used in the process of semantic annotation alongside the creation of the semantic metadata.

---

[116] http://www.eclipse.org/swt/ [last accessed 27/2/2007]

[117] http://jena.sourceforge.net/ [last accessed 27/2/2007]

Chapter 9

Domain Ontologies and Semantic Metadata

Chapter 8 indicated that the FolksAnnotation tool has used three ontologies in the semantic annotation pipeline. These three ontologies were necessary for the creation of the semantic metadata. Therefore, this chapter focuses on building the three mentioned ontologies, namely: Web Design ontology, CSS ontology and Resource Type ontology. It further presents the semantic metadata descriptors used to describe CSS resources and elaborates on their functionalities.

9.1 Introduction

Before embarking the process of ontology modelling and building, it is necessary to see whether there exists any ontology suitable for the purpose of the thesis domain. Existing ontologies can be found either in ontology libraries such as the Ontolingua ontology library[118], the DAML ontology library[119], the Protégé ontologies[120] and the SchemaWeb[121], or by using semantic search engines such as Swoogle[122] and OntoSearch[123], or by consulting ontology portals such as ONTHOLOGY[124].

---

118 http://www.ksl.stanford.edu/software/ontolingua/ [last accessed 2/2/2007]

119 http://www.daml.org/ontologies/ [last accessed 2/2/2007]

120 http://protege.stanford.edu/download/ontologies.html [last accessed 2/2/2007]

121 http://www.schemaweb.info/ [last accessed 2/2/2007]

122 http://swoogle.umbc.edu/ [last accessed 2/2/2007]

123 http://www.ontosearch.org/ [last accessed 2/2/2007]

Unfortunately, no matching ontologies were found in these venues that are suitable for the thesis domain of interest, i.e. 'Web Design' and 'CSS', however, the researcher has found one candidate ontology related to the Resource Type domain which will be discussed later in this chapter. As for the Web Design and CSS domains, the researcher decided to construct the ontologies from scratch to serve the thesis purpose. The ontologies will act as conceptual backbones for generating semantic metadata annotations.

9.2 Ontology Building

To build an ontology, it is required to go through different stages, which include: knowledge acquisition, knowledge modelling, knowledge annotation and reuse (Millard et al., 2006). Actually, there is no one correct way to build ontologies and there are several methodologies aimed at designing and building ontologies. Millard et al. (2006) and Noy & McGuinness (2001) iterative approaches are both followed to produce the thesis ontologies.

Therefore, the three domain ontologies that need to be acquired and modelled are: Web Design ontology, CSS ontology, and Resource Type ontology. The two former ontologies' themes were based on observed patterns in peoples' tags in the del.icio.us bookmarking service for our domain of interest (c.f. chapter 4).

However, before identifying the main concepts in each ontology, it is necessary to define the domain and scope of the three ontologies (Azouaou and Desmoulins, 2006a). The domain of the thesis ontologies is to teach CSS in Web design course context. Therefore it is unlikely that the CSS ontology will contain concepts about other Web design domains, such as HTML, JavaScript, etc.

The next logical step after defining the domain is to identify the ontologies scope by asking "*for what are we going to use the ontologies*?" The expected uses of the ontologies are:

- Annotate CSS web resources with fine-grained semantics.

[124] http://www.onthology.org/ [last accessed 2/2/2007]

- Search for CSS resources using the smallest granularity of the domain.
- Provide CSS web resources with pedagogical and technical semantics.

The concepts alongside their instances modelled in the Web Design and CSS ontologies were acquired from multiple sources such as, Wepopedia[125], Dmoz[126] and Yahoo[127] directory where these sources represent a formal controlled vocabulary for Computer Science related topics. The researcher also scanned several Web Design and CSS websites such as W3Schools[128], W3C[129], assorted books, cheat sheets and online courses' curriculum to grasp an idea about the two named domains. Finally, the researcher benefited from the del.icio.us users' tags in the domain of CSS. The reason behind consulting del.icio.us in the ontology modelling is that many terms used within the CSS domain are developing and may not be found in specialized controlled vocabulary.

The modelling of the ontologies involved explicitly representing the acquired knowledge into a formal language. The formal language was expressed using OWL-DL, one flavour of the W3C official ontology language recommendation for modelling ontologies. The researcher chose OWL-DL (see Chapter 5) for its expressiveness and powerful representation. The modelling was done using a well-known ontology editor[130] called Protégé.

Protégé (Noy and McGuinness, 2001) is a free, open source ontology editor and knowledge-base framework developed at the Stanford Medical Informatics. The editor supports the building of ontologies in different languages such as RDFS and OWL using special plug-ins. Protégé also provides the ability to check the consistency, validation and verification of an ontology. This feature was used to check the thesis ontologies for conformance with OWL DL rules.

[125] http:// www.webopedia.com/ [last accessed 3/2/2007]

[126] http://dmoz.org/ [last accessed 3/2/2007]

[127] http://dir.yahoo.com/ [last accessed 3/2/2007]

[128] http://www.w3schools.com/ [last accessed 3/2/2007]

[129] http://www.w3.org/Style/CSS/ [last accessed 3/2/2007]

[130] http://en.wikipedia.org/wiki/Ontology_editor [last accessed 3/2/2007]

The following subsections describe in detail the modelling of the three ontologies along with their concepts and relationships.

9.2.1 Web Design Ontology

The Web design ontology represents an abstract level of the domain of 'Web Design and Development'. It models the concepts in that domain along with their relationships.

The purpose of using the Web Design ontology in this thesis is to place the CSS ontology in the context of its domain. Figure 9.1 shows a hierarchical diagram of the Web Design ontology.

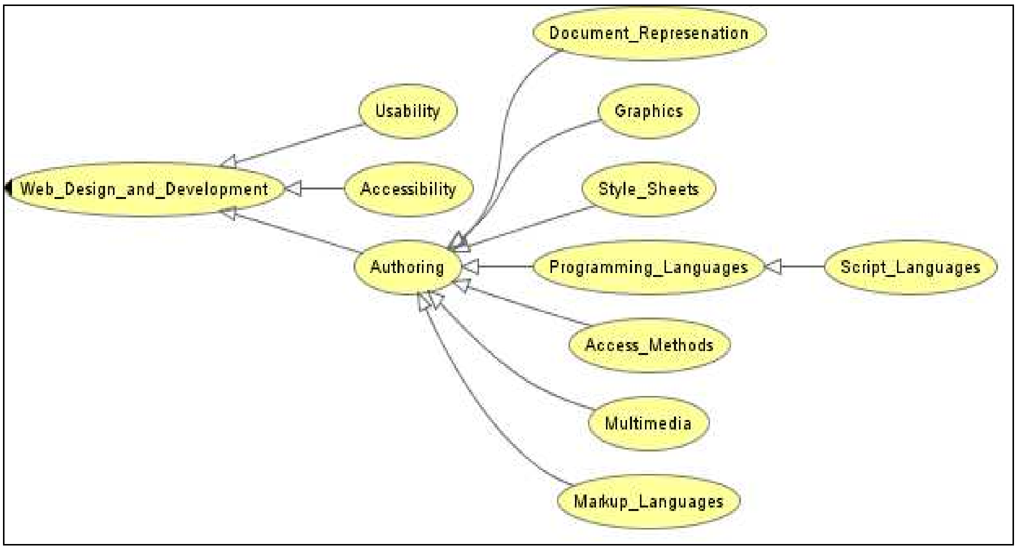


Figure 9.1: 'is-a' diagram showing the hierarchical relationship between the main concepts in the Web Design domain.

Table 9.1 shows the definitions of each concept in the Web Design ontology. In addition, the set of relationships (i.e. properties) that connect the concepts are defined in Table 9.2.

Class	Definition
Usability[131]	Defines the ease-of-use of the user interface. This include learnability, efficiency, etc.
Accessibility	Defines the ease-of-access of the user interface.
Authoring	Describes the main components needed for creating a web resource.
Document_Representation	Describes how an HTML or XML document is represented in a tree structure (e.g. DOM).
Graphics	Describes the software packages used to generate graphic content (e.g. Photoshop).
Style_Sheets	Describes the style of elements in a document marked up using a markup language (e.g. CSS).
Programming_Languages	Describes the programming languages used to develop web applications (e.g. Java).
Script_Languages	Describes the scripting languages used to develop web applications (e.g. JavaScript).
Access_Methods	Describes the techniques used to access a web resource (e.g. AJAX).
Multimedia	Describes the technologies and software used to generate multimedia content (e.g. Flash).
Markup_Languages	Describes the types of markup languages (e.g.HTML).

Table 9.1: Web Design Ontology Concepts

[131] http://www.useit.com/alertbox/20030825.html [last accessed 3/2/2007]

Property	Definition
consistOf	Describes the relationship between the parent class and subclasses (e.g. Authoring consistOf markup languages, document representation, etc).
represented_in	Defines the relationship between document representation and the markup language (e.g. HTML represented_in DOM).
uses	Defines the relationship between Access methods and the technologies supporting it (e.g. AJAX uses XML, CSS, and JavaScript).

Table 9.2: Properties of the Web Design ontology

9.2.2 CSS Ontology

Figure 9.2 shows the CSS ontology. The ontology gives a fine grained listing of the concepts used in the CSS domain.

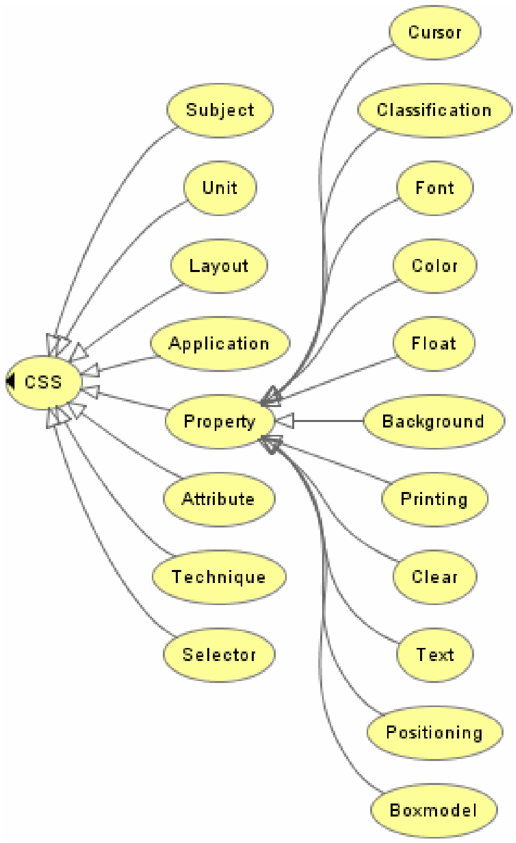


Figure 9.2: 'is-a' diagram showing the hierarchical relationships between the concepts in the CSS domain.

Table 9.3 shows the definition and description of each concept in the CSS ontology. In addition, some of the relationships (aka properties) used in this ontology are defined within the semantic metadata section, i.e. Table 9.7 in the next section. Table 9.4, however, describes the rest of the ontology relationships.

Class	**Definition**
CSS	A thing that represents the main concept in the ontology.
Property	Parent class for the CSS properties.
Printing	The CSS Printing property defines the printing effects of page.
BoxModel	The BoxModel defines the border, margin and padding of an element.
Cursor	The CSS Cursor property defines the pointer shape of the cursor.
Clear	The CSS clear property for controlling flow when using float.
Font	The CSS font property defines the font in text.
Color	The CSS color property defines the color effects of an element.
Positioning	The CSS positioning property describes how to position an element.
Background	The CSS background property defines the background effects of an element.
Box	owl:SameAs BoxModel
Text	The CSS text property defines the appearance of text.
Classification	The CSS classification property specify how and where to display an element.
Attribute	Defines the two main attribute elements {class, id}.
Application	Describes the possible applications of the CSS technology (e.g. Menus, Check Boxes).
Technique	Describes the possible techniques of the CSS technology (e.g. shadow, transparency).

Layout	Describes the layout that a CSS technology provides (e.g. boxes, columns).
Subject	Describes the theme that a CSS domain handles (e.g. navigation, positioning, etc).
Selector	Defines the two main selector elements {div, span}.
Unit	Defines the CSS units (e.g. em, pt, px, etc).

Table 9.3: CSS Ontology Concepts

Property	**Definition**
Related_to	Defines the relationship between related concept instances (e.g. navigation related_to sitemap).
Used_with	Defines the conjunction relationship between two concept instances (e.g. rounded corner used_with button).
Uses	Defines the dependency relationship between two concept instances (e.g. menu uses list).

Table 9.4: Properties of the CSS ontology

9.2.3 Resource Type Ontology

The resource type ontology has a simple taxonomic structure as shown in Figure 9.3. It defines the hierarchy of concepts without specifying any kind of relationship between them. It further models the resource types that go beyond the scope of the common-set provided by LOM. The reason behind using a different vocabulary set is that, different learning resources can come in a variety of forms. For example, suppose a learning resource is of type ‘editor’, a software tool used to create or modify files of a particular type. A possible use of the resource will be to use it as an additional resource in the context of a programming course. This type of resource and others have not been mentioned in the LOM resource type values; therefore, new vocabulary needs to be modelled to represent the new resources types emerging in people’s vocabulary.

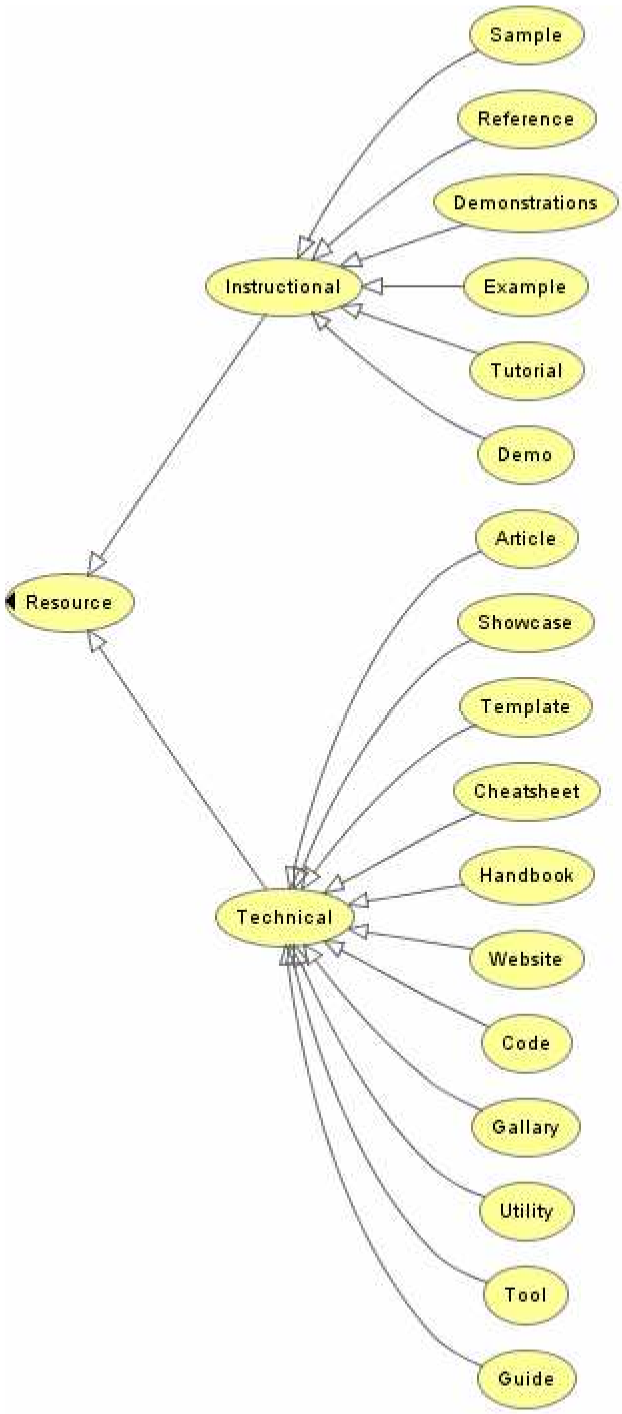


Figure 9.3: Resource Type ontology.

The resource type ontology also overcomes the shortcomings in LOM as mentioned by (Ullrich, 2005); where LOM mixes both the instructional (e.g. tutorial) and technical (e.g. code) as part of the resource type in its listing.

The creation of the resource type ontology was also inspired by examining existing learning resources type vocabularies such as IEEE-LOM, SCORM, Ullrich (2005) and Jovanovic et al. (2006). None of the previously mentioned vocabularies were capable of demonstrating the resource type domain used in this thesis. The problem with LOM and SCORM is that they mix the concepts of instructional and technical together. Furthermore, Ullrich ontology was confined to model the semantic of learning resources used in text-book domain. Finally, Jovanovic et al. (2006) have developed a resources type ontology for their TANGRAM system; a learning web application for the domain of Intelligent Information Systems (IIS) where users (students and teachers) can upload, describe, search or compose a new learning object using components in the system repository. In spite of the comprehensiveness of their resource type ontology, the ontology was developed with a course structure in mind.

9.2.4 Ontology instances

The last step in ontology modelling is to create instances for each concept in the ontologies. Table 9.5 demonstrates sample instances from the three previously mentioned ontologies.

Ontology	Concept	Instances
CSS	Box Model	border, margin, padding.
	Positioning	display, position, visibility, z-index.
	Application	banner, bargraph, button, chart, form, graph, image-map, menu, rating, sitemap, wordpress.
Resource Type	Technical	documentation, article, cheat sheet, code, gallery, guide, eBook, handbook, showcase, template, website, tool, utility.
Web Design	Access Methods	ajax, dhtml.
	Markup Languages	html, xml, xhtml.

Table 9.5: Instances samples from Web Design Ontology, CSS ontology and Resource Type Ontology

9.3 The Semantic Metadata

Learning resources are usually described using standards such as Dublin Core (DC) and IEEE LOM alongside their RDF bindings. The semantic metadata used in this thesis builds on these standards and adds more fine grained semantics to web resources in the thesis domain of interest. Thus an application profile is created by using subset elements of IEEE-LOM, not all elements are used, necessary to support the intended functionalities of the system.

Figure 9.4 shows the RDF graph for the semantic metadata used to annotate web resources in the CSS domain. Note that not all elements in the semantic metadata need to be present at one time. For example, a CSS web resource in the del.icio.us bookmarking service with tags talking only about the application of the resource will have *hasApplication* property but not *hasTechnique* property, since there is no tag that triggers the latter relationship. In other words, the more comprehensive the tag list for a given web resource is, the more likely the semantic descriptors are present.

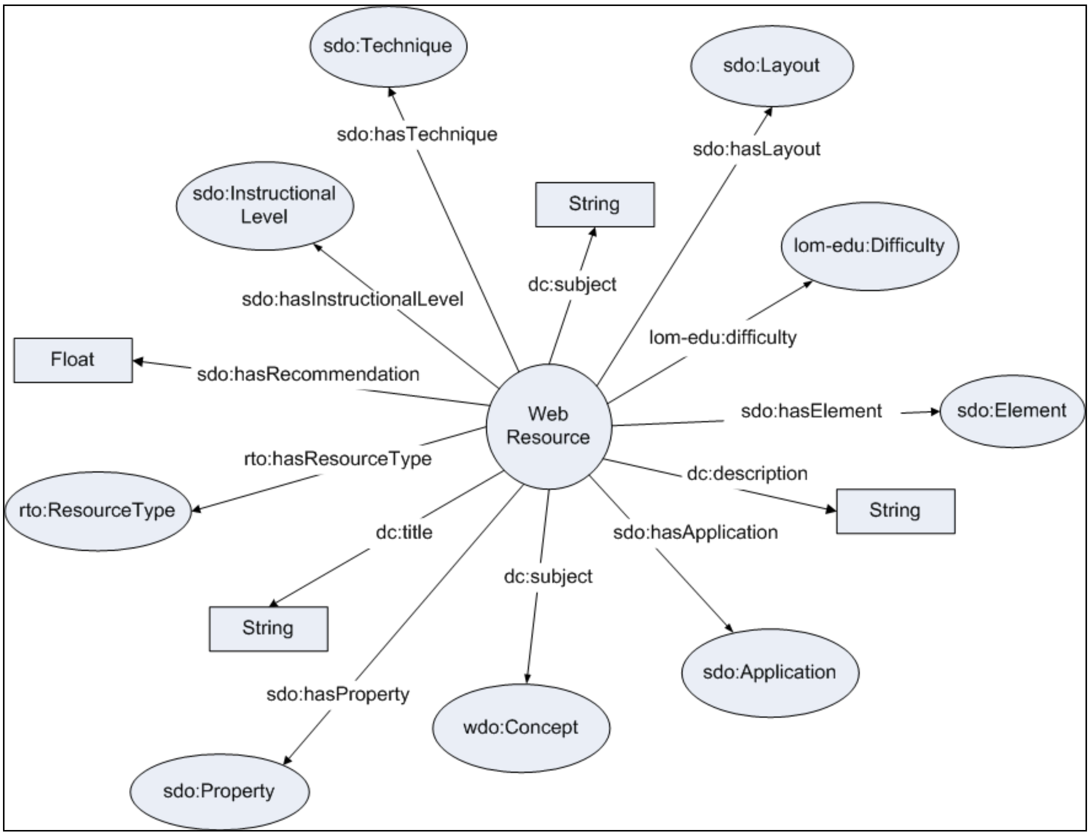


Figure 9.4: An Excerpt of the RDF Graph used to describe a CSS web resource.

The semantic metadata is comprised of a subset of six elements adopted from the categories defined in IEEE-LOM, as summarized in Table 9.6. The approach of using a subset of LOM properties in semantic annotation has also been used in (Brase and Nejdl, 2003; Brase and Painter, 2004; Munoz and Oliveira, 2004).

LOM Descriptor	RDF Binding	Description
1.2- Title	**dc:title** with a literal value	Name given to this learning resource.
1.4- Description	**dc:description** with a literal value	Description of the content of the learning resource.
1.5- Keyword	**dc:subject** with a literal value	A keyword describing a topic in the learning resource. In the case study the value of the keywords will be the folksonomy tags.
5.2-Learning Resource Type	RTO[*]:hasResourceType	Specify the type of learning resource. Example: reference, tutorial, etc. This property links to an entry in the **Resource Type domain** ontology.
5.8- Difficulty	**lom-edu:Difficulty**	How hard it is to work with the learning resource. Example: very easy, easy, medium, difficult, very difficult.
9- Classification	**dc:subject** with a literal value	This property links to an entry in the **Web Design** ontology.

Table 9.6: LOM descriptors used in the CSS Semantic Metadata

[*] Resource Type Ontology Namespace

Table 9.7 shows the extra properties that were used with the generated semantic metadata. Notice that these extra values do not exist in either LOM or DC vocabulary.

Descriptor	RDF Binding	Description
Recommendation	SDO[^]:hasRecommendation with a numeric value	Describes how popular is a web resource based on the number of people who bookmarked it (see section 9.3.2).
Instructional Level	SDO:hasInstructionalLevel	Describes the instructional level of a web resource. Example:novice, intermediate, advanced.

Table 9.7: Extra descriptors used with the CSS semantic metadata

Table 9.8, shows the properties specific to the CSS domain. These properties connect a web resource with a given concept instance.

Property	Description
SDO:hasApplication	Describes the range of applications a CSS technology can be applied to. Example: imagemap, menu, button, etc.
SDO:hasElement	Describes the elements that can be used with CSS technology. Example of attribute elements {class, id}, and example of selector elements {div, span}.
SDO:hasProperty	Describes the properties of the CSS technology. Example: color, background, box model, etc.
SDO:hasTechnique	Describes the range of techniques a CSS technology can give. Example: hover, rollover, shadow, etc.
SDO:hasLayout	Describes the types of page layouts a CSS technology can give. Example: tableless, fluid, fixed, etc.

[^] Subject Domain Ontology Namespace

SDO:hasUnit	Describes the units that a web resource uses. Example: px, pt, etc.

Table 9.8: Specific CSS descriptors with their RDF binding

9.3.1 The Metadata Elements

The researcher cannot stress enough how the variability of the generated semantic metadata is dependent on the quantity and the quality of the assigned folksonomy tags. The number of semantic elements used to describe a web resource heavily relies on the existence of a corresponding tag that populates that element. Accordingly, this affects the completeness and the quality of the generated semantic metadata.

The semantic metadata elements come in two genres: extracted elements and generated elements. The extracted elements are those elements which can be extracted from the list of folksonomy tags. The extracted elements themselves can come in two types: fixed and variable. The fixed elements are those which can be always extracted from a bookmarked web resource, these include: Title and keywords.

The variable elements are those whose presence is not guaranteed, and they come in two types: triggered and inference. The triggered elements are those elements that get activated when a corresponding folksonomy tag was found in the CSS ontology; these include the following elements: Property, Application, Element, Layout, Subject, Unit, Technique and Resource type.

The inference elements are those which get activated when enough tags are satisfied in the triggered elements. The inference elements include: Difficulty level and Instructional level.

The generated semantic elements are those elements which are produced automatically either by using a template or numerically computed. These include: the Description and the Recommendation element.

9.3.2 *The Generated Elements*

In this section the researcher will elaborate more on the process of producing the generated elements.

Automatic generation of the dc:description element

Since the description of a bookmarked web resource provided by the del.icio.us users varies between non-English description, incomplete and self-oriented comments (as shown in Figure 9.5), the system needs to automatically produce the description field for each annotated web resource.

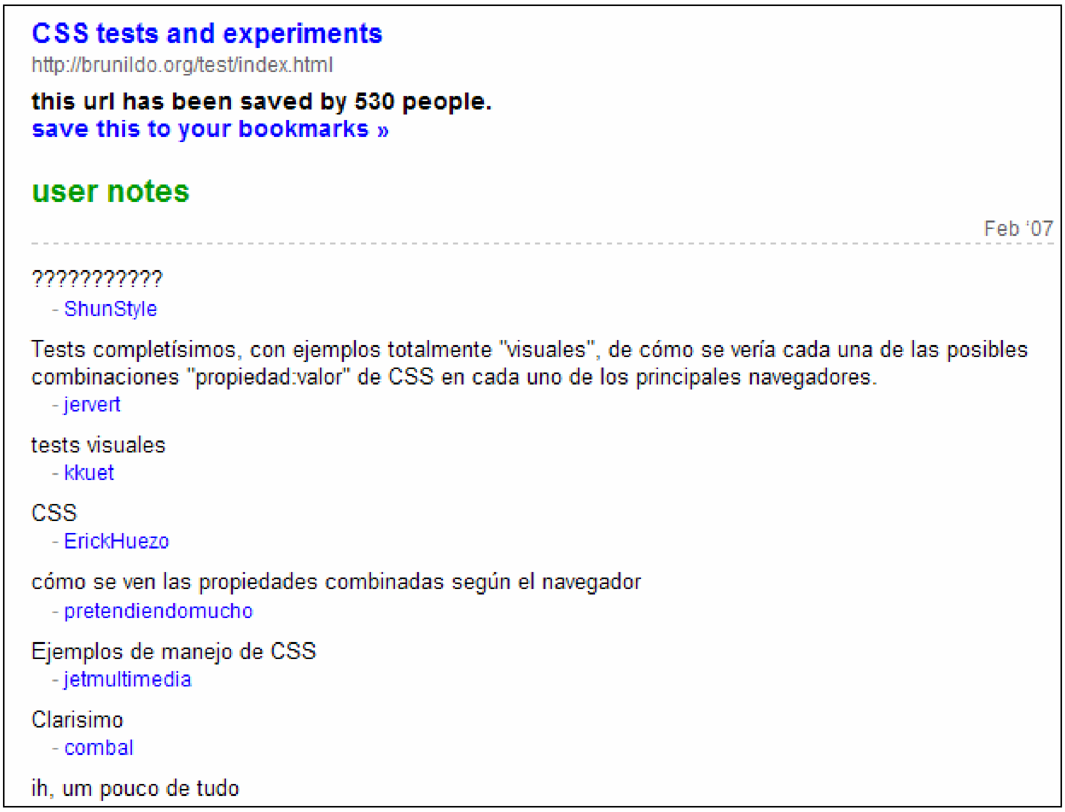


Figure 9.5: A screen shot showing the different inappropriate descriptions applied by the del.icio.us users for the “CSS tests and experiments”[132] web resource

This idea was adopted from (Jovanovic et al., 2006) where they created a template for describing their annotated learning resources. The template slots are filled with instances from the three ontologies. To give an example Figure 9.6 shows a template (a) and its example (b). The angle brackets are replaced with their actually values in

[132] http://brunildo.org/test/index.html [last accessed 4/2/2007]

the corresponding ontology, while the single curly brackets indicates that the enclosed element can have multiple values.

"A(n) <rto:ResourceType> with title: '<dc:title>', suggest the knowledge of the following topics: {<dc:subject>}. This resource is also suggested to be used as: {<rto:ResourceType>}." (a)
"A(n) **reference** with title: **'Daniel Mall: Well Educated CSS'** suggest the knowledge of the following topics: **css, xhtml, ajax, javascript**. This resource is also suggested to be used as: **tool, article, resource, tutorial, code, template**. " (b)

Figure 9.6: Template (a) and example (b) of the dc:description element.

Automatic generation of the recommendation element

The recommendation element holds a numerical value that is computed using two operands. The first operand is the number of people who bookmarked the web resource and the second operand is the number of expression tags in the complete list of tags assigned to a web resource.

The number of people, denoted as P(R), is a computed value that reflects how popular a bookmarked web resource is. This value is a simplified version of how Google's PageRank[133] uses the number of incoming links ('backlinks') to compute a website's popularity. In this case, the incoming links are replaced by the number of people who bookmarked a web resource. The value of P(R) is computed using Equation (1).

P(R) is a function that returns a constant factor based on the number of people who bookmarked a web resource. The researcher derived this entirely empirical function approximately based on a logarithmic scale approach, where the values between 0.1 and 0.5 represent a slow increase in the number of people who are bookmarking a web resource. However, as soon as a web resource gets bookmarking momentum the

[133] PageRank "is a link analysis algorithm which assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set." From http://en.wikipedia.org/wiki/PageRank [last accessed 4/2/2007]

popularity value does not increase significantly. The values in Equation (1) are totally empirical and subject to fine tuning of the relative weighting to optimize the result.

$$P(R) = \begin{cases} 0.1 \quad if \quad x \leq 50 \\ 0.2 \quad if \quad 50 < x \leq 100 \\ 0.3 \quad if \quad 100 < x \leq 500 \\ 0.4 \quad if \quad 500 < x \leq 1000 \\ 0.5 \quad if \quad 1000 < x \leq 3000 \\ 0.6 \quad if \quad 3000 < x \leq 5000 \\ 0.7 \quad if \quad 5000 < x \leq 7000 \\ 0.8 \quad if \quad 7000 < x \leq 10000 \\ 0.9 \quad if \quad x > 10000 \end{cases} \qquad (1)$$

The number of expression tags, denoted as E(R), is computed by counting the number of occurrences expression tags have appeared in the list of tags assigned to a given web resource. A pre-defined list of expression tags was compiled beforehand based on observed expressions usage in the del.icio.us service. The expressions tags' list the researcher has identified so far consists of the following expressions: *cool, interesting, handy, useful, best, fabulous, important, good, hot, awesome, amazing, wow and excellent*.

E(R) is a function that computes the number of occurrences an expression tag T_i has appeared in the list of all tags (L_i). The summation is then divided by U(R), which represents the total number of people who bookmarked the web resource {*R* means Resources, *T* means Tags and *U* means Users}.

$$E(R) = \frac{\sum R(T_i) \in L_i}{U(R)} \qquad (2)$$

Finally, the value of the recommendation element is computed using an empirical equation the researcher has devised:

$$\mathrm{Re}\,commendation = \left[E(R) + P(R)\right] * 5 \qquad (3)$$

The result is calibrated to a 5 point range, as shown in equation (3). Notice that the straight forward approach the researcher has pursued is subject to further evaluation to improve the quality of the ranking provided as will be mentioned in chapter 12.

To test if the proposed equation produces sensible values, the researcher has devised some benchmarking criteria. In this case the best choice was to compare the results of the recommendation equation for a sample of bookmarked web resources in del.icio.us against the same sample using the same model but without the expression operand E(R). This was necessary to examine whether the expression operand would add any value to the ranking of a bookmarked web resource.

Table 9.9 shows real examples pulled from the del.icio.us bookmarking service to illustrate how the recommendation value is computed given different situations.

Web Resource	**No. of People**	**No. of expression tags**	**Recommendation value Without E(R) (A)**	**Recommendation value With E(R) (B)**
Layout Gala[134]	8485	223	4	4.13
Text Inputs on Safari[135]	49	0	0.5	0.5
css Zen Garden Shot[136]	250	5	1.5	1.6
OverZone Software - CSS Tab Designer[137]	1968	51	2.5	2.63

Table 9.9: The result of the computed recommendation value for four examples

From the previous compiled table, the reader can observe that there is a subtle difference between column (A) and column (B), where column (A) represents the

[134] http://blog.html.it/layoutgala/ [last accessed 4/2/2007]

[135] http://www.shauninman.com/post/heap/2006/11/02/text_inputs_on_safari [last accessed 4/2/2007]

[136] http://antenna.readalittle.net/thumblink/zenGarden/ [last accessed 4/2/2007]

[137] http://www.highdots.com/css-tab-designer/ [last accessed 4/2/2007]

recommendation value before taking into account the expression operand and column (B) represents the recommendation value after taking the expression operand into account.

Notice the impact of the expression operand E(R) on the recommendation output for each web resource that has among its tags an expression tag. This slight impact in recommendation value has changed the web resources ranking accordingly.

From this simple evaluation, the researcher can claim that the recommendation value is potentially useful for ranking search results returned from the CSS knowledge base, and for expressing the popularity of a web resource based on the del.icio.us users' recommendations. Finally, it is worth noting that the method for calculating the recommendation element is experimental and it is subject to future modifications.

9.4 Chapter Summary

This chapter started by defining the three ontologies needed for the process of semantic metadata annotation. The ontology building began by acquiring the necessary knowledge required to model the three domains. Then OWL-DL, the formal ontology language, was used to formally represent the concepts and properties of the three ontologies. The chapter concludes with a thorough description of the elements of the semantic metadata and how they were produced.

Chapter 10
Evaluation, Analysis and Discussion

To evaluate the output of the research prototype tool, many evaluation aspects need to be considered. These aspects include the usefulness of the generated semantic metadata, the quality and the representativeness of the metadata semantics.

Barritt and Alderman (2004) determine the usefulness of metadata from two viewpoints: validity, i.e. creating valid metadata for every learning resource, and search-ability, having the search tools in place to use that metadata.

Guy et al. (2004) define metadata quality as "*... supports the functional requirements of the system it is designed to support*." Thus, to stipulate the 'functional requirements' of the current work, the researcher has considered that the semantic metadata need to have no errors and the semantic descriptors need to correctly reflect the nature of the described web resource.

Finally, the representativeness of a semantic metadata can be thought of as how well the metadata descriptors describe the semantics of the given domain, in this case the domain of teaching CSS.

Therefore, to evaluate these different aspects, the researcher has implemented a comprehensive evaluation framework, as shown in Figure 10.1, where each evaluation procedure is explained in further detail based on the order they are presented in the diagram (i.e. from left to right using depth-first propagation).

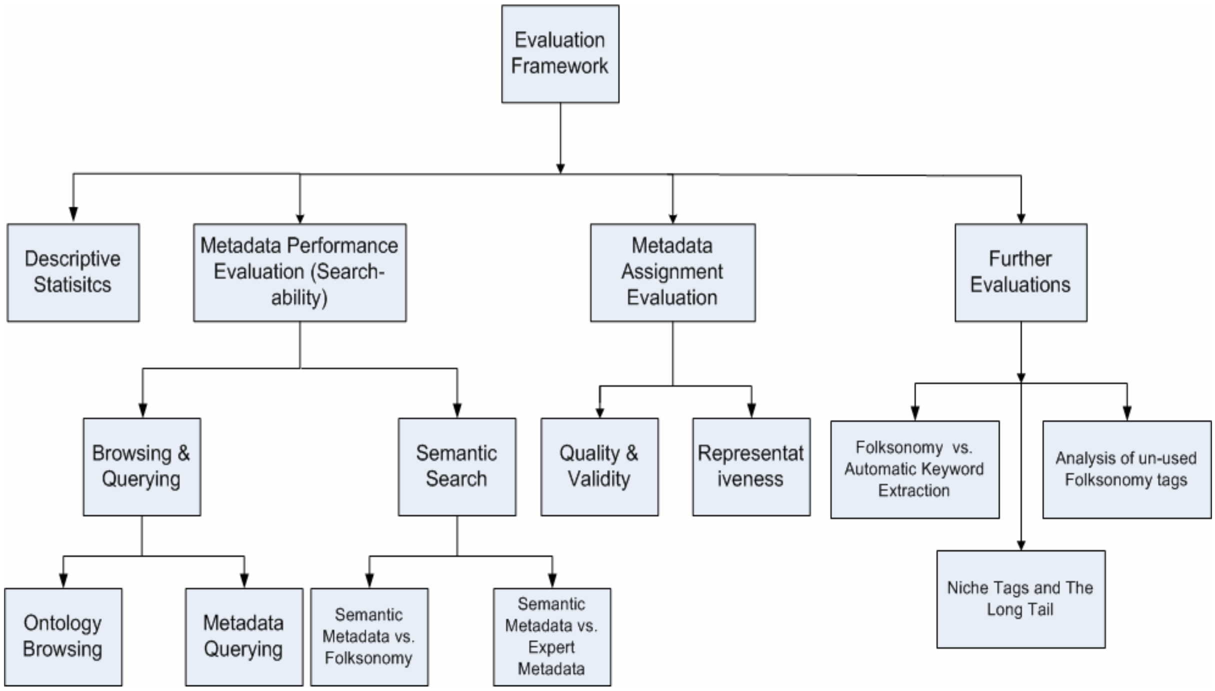


Figure 10.1: The Evaluation Framework

10.1 Descriptive Statistics

The number of URLs the researcher has experimented with was 100, with a total number of 72,458 posts (i.e. people). The maximum number of posts for a given URL was 7776, while the minimum number of posts was 4.
The total number of tags before normalisation was 245,892, and the number of tags after normalisation was 10,900.

The first significance test the researcher has conducted was performed to measure the correlation between the number of people who are tagging a web resource and the number of its assigned tags. The correlation test indicates that there is a significant positive relationship between users tagging a resource and the number of tags assigned to it. The value of the Pearson correlation coefficient R^2 was 0.995 ($p<0.01$). The regression equation for the relationship can be interpreted as:

$$\text{Number of Tags} = 3 * \text{Number of people} + 40$$

The equation implies a positive correlation between the number of people and the number of tags assigned to a web resource so that as the number of users increases the number of tags increases accordingly. However, it is worth noting that this equation is valid for the interval [4, 7776] and the researcher does not claim the validity of the equation given other situations such as other bookmarking services.

To measure the effectiveness of the normalisation pipeline, the researcher also correlated the number of tags for each given URL before normalisation against the number of tags for that URL after normalisation. The correlation value showed a statistical significance relationship with R^2 value of 0.968 ($p<0.01$). Moreover, the correlation value for tags after normalisation versus tags after dropping the general tags also showed a statistical significance relationship with R^2 value of 0.998 ($p<0.01$).

Notice that these findings are only significant for the experiment sample set; therefore, the researcher does not claim that her findings can be generalised to the entire population of the del.icio.us bookmarking service. This is because the content of the del.icio.us bookmarking service is very dynamic and it changes over time. Therefore, generalising the findings might cause false results in subsequent experiments.

10.2 Metadata Searchability Evaluation

The semantic metadata that has been generated using the thesis framework needs to be deployed in a way that is capable of searching for CSS resources at each level of the metadata record. Thus, the semantic metadata can be browsed using hierarchical trees, or it can be searched using the smallest element of the metadata.

Two broad searchability techniques were embraced in this thesis, namely: Browsing & Querying and Semantic Search.

Browsing & Querying includes: ontology browsing and metadata querying (Al-Yahya, 2006), which adds two flexible ways to reach, retrieve and search for annotated learning resources. Since the ontologies are created in a hierarchical taxonomic nature; they can be directly projected to the user as views.

The semantic search technique will exploit the power of semantic relationships between the ontology concepts (see section 10.2.2).

10.2.1 Browsing & Querying

This section illustrates the browsing & querying technique conducted using two mechanisms that show two flexible ways to retrieve and search for annotated learning resources.

Ontology Browsing

Figure 10.2 shows the user interface depicting one of the ontology views (CSS ontology). When a category is selected by clicking on a subcategory link listed on the view, a view-based search is initiated that shows all results returned to the user based on the selection made.

In this search option, the user[*] can retrieve web resources either by browsing the concepts in the CSS ontology, or by browsing the concepts in the resource type ontology. When a concept is selected in either ontology all resources resembling the selected concept are retrieved along with their full description. Figures 10.2, 10.3 and 10.4 show how users can access the web resources by browsing the two named ontologies.

---

[*] A user can be either a teacher or a learner

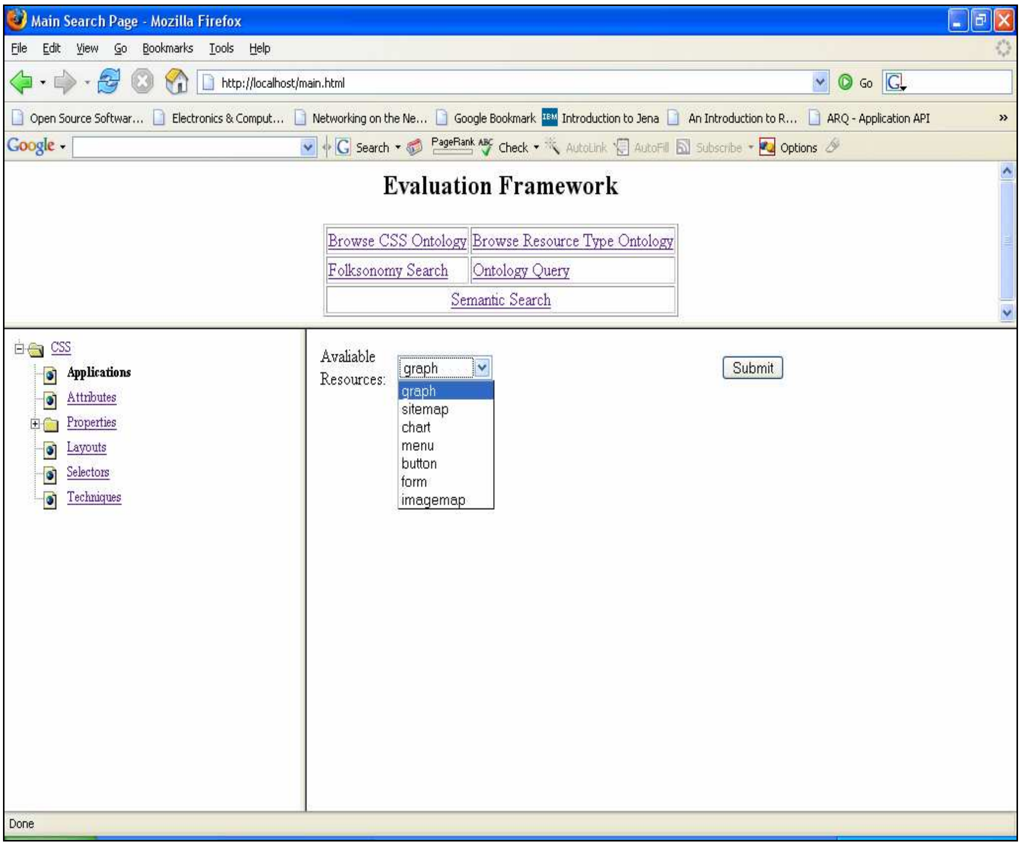


Figure 10.2: Browsing the CSS Ontology

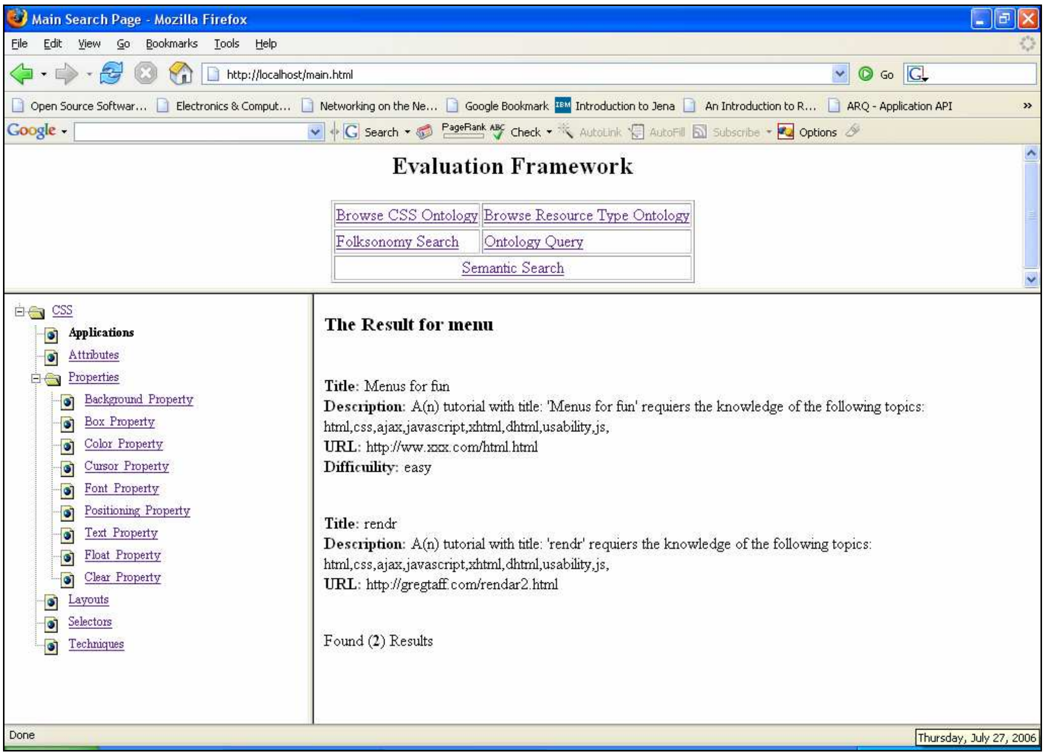


Figure 10.3: Retrieved results after selecting the context "menu" to search the CSS knowledge base

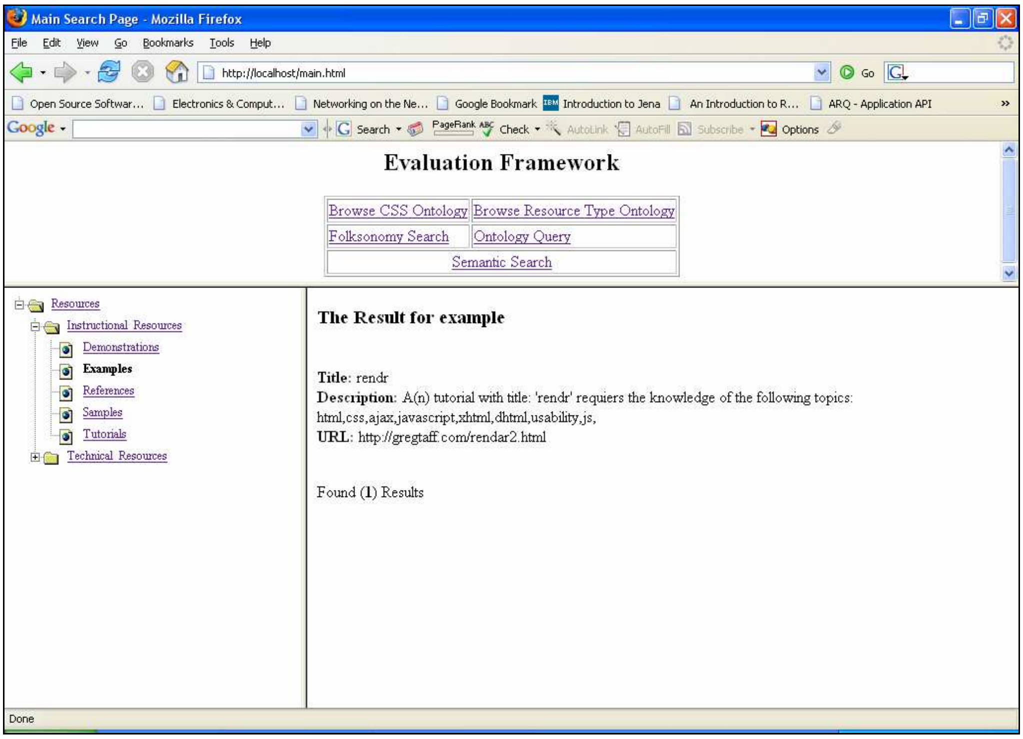


Figure 10.4: Browsing the Resource Type ontology

The browsing algorithm works by reasoning over the data. Thus, when a concept is selected, the search algorithm searches the knowledge base for all resources related to the concept.

One benefit of using the view-based search paradigm is that users can have a grand vision of all concepts provided by the domain and select concepts that represent what they are looking for.

Metadata Querying

To further enhance the experience of searching for CSS resources, a query interface has been implemented, which enables the composition of different queries to access the CSS knowledge base. The user is presented with a set of query filters to choose from, as shown in Figure 10.5. These include query by: resource type, difficulty level, instructional level, subject, technique, attribute, property, layout and/or application.

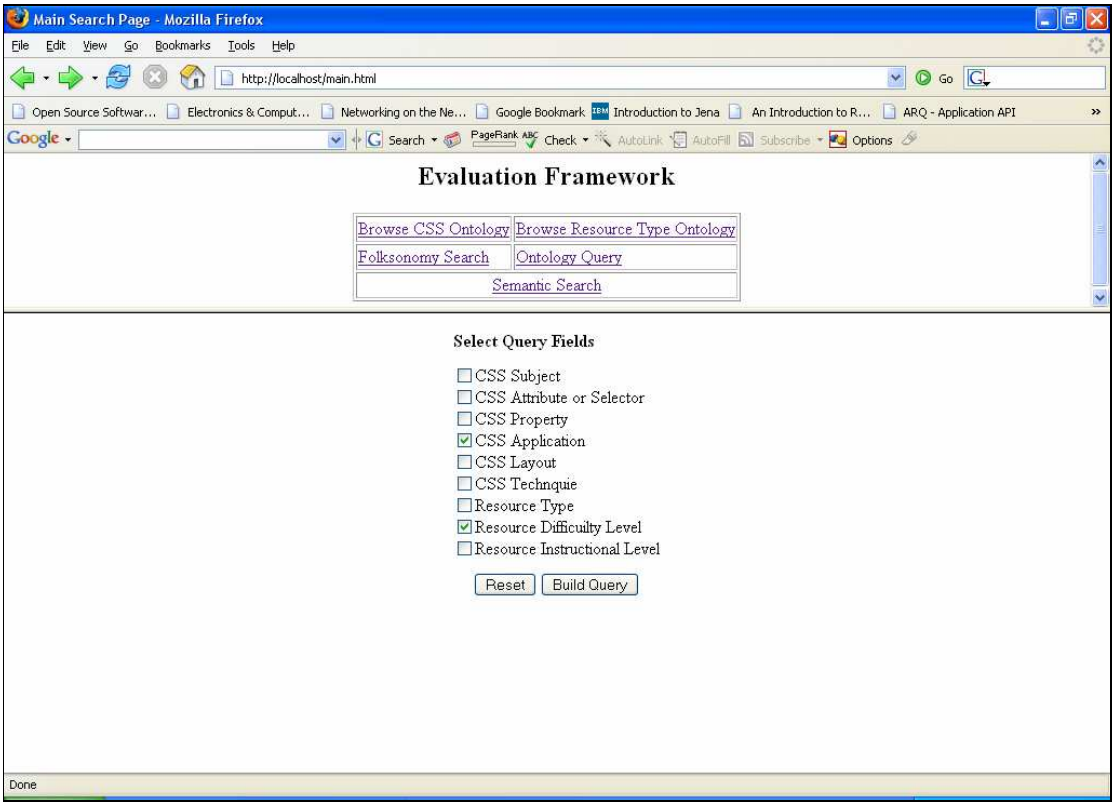


Figure 10.5: Query filters selection

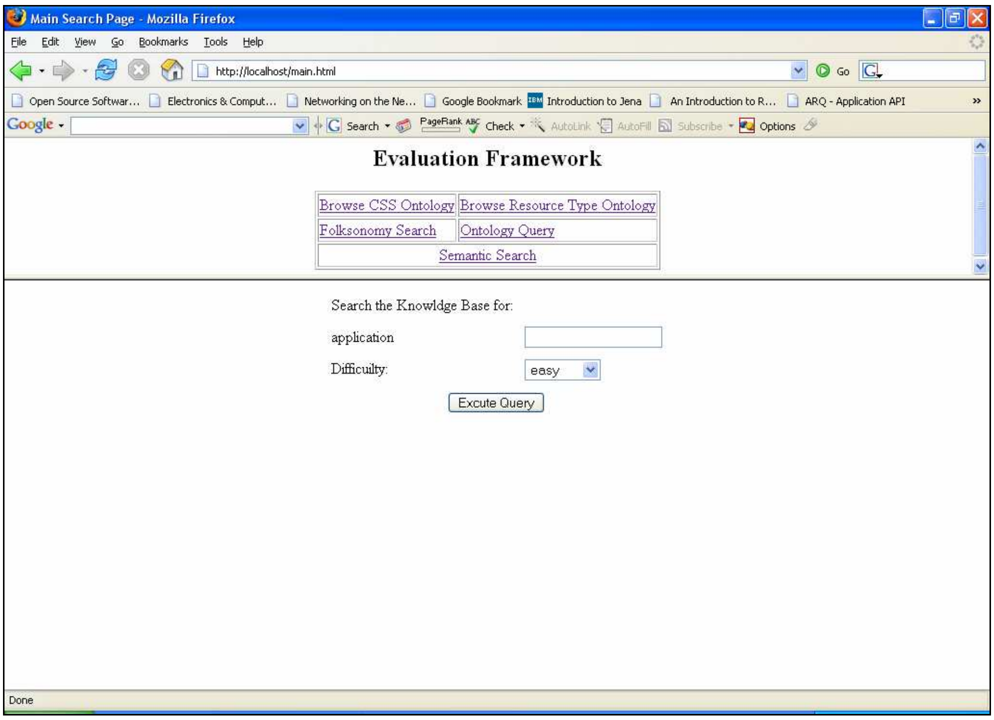


Figure 10.6: Query form builds up

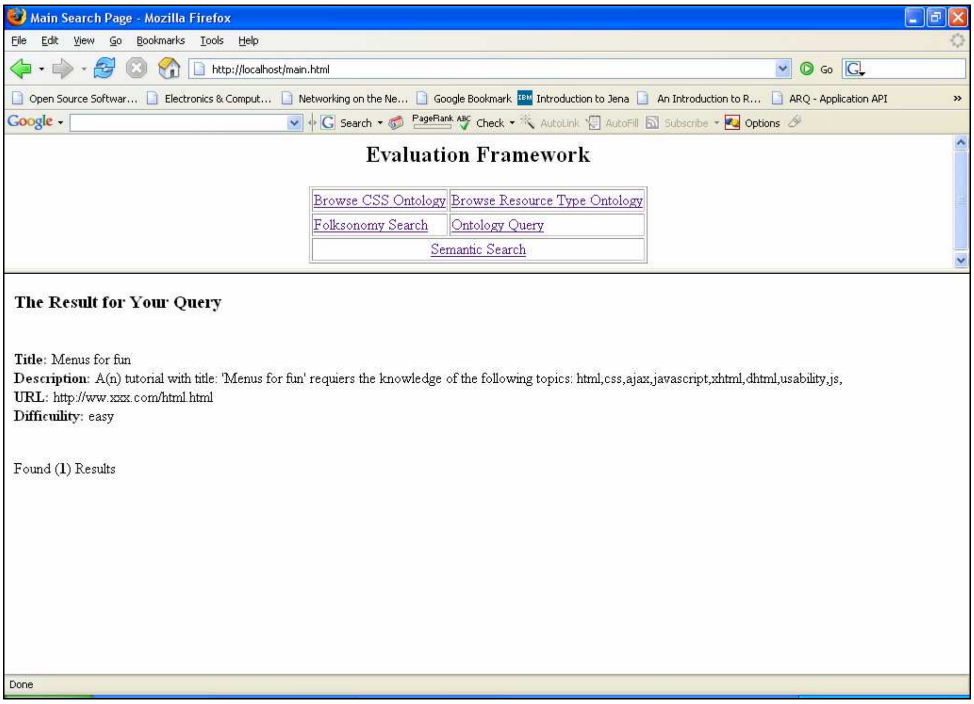


Figure 10.7: Query results after entering "menu" as an application and choosing "easy" as difficulty level

10.2.2 Semantic Search

This section demonstrates the semantic search technique carried out on two scenarios to exploit the power of semantic relationships between the ontology concepts.

Folksonomy vs. Folksonomy Semantic Metadata

To evaluate the performance of the generated semantic metadata, the researcher has embarked on an evaluation procedure adopted from (Vallet et al., 2005), where they compared keywords against semantic topic search using several case studies. However, in this research, the researcher has compared the performance of folksonomy against instances of semantic concepts and evaluated the results using the well-known information retrieval matrices of precision and recall.

The results were assessed in terms of precision, recall and f-measure; the recall is the proportion of all possible correct annotations that were found by the system and are semantically related to the web resource, the precision is the proportion of the retrieved web resources that were found to be correctly related and can be used in the context of the queried semantic concept, and the f-measure evaluates the harmonic mean for the overall performance by equally handling the recall and precision of the results.

Thus, the precision, recall and f-measure are computed as follows:

$$\mathrm{Re}\,call = \frac{\mathrm{Retrieved}}{POSS} * 100\%$$

POSS (Possible) refers to the number of CSS web resources in the knowledge base that contribute to the possible candidate resources, *POSS = Relevant + Semantically-Related. Relevant* refers to those retrieved CSS resources that address the searched instance.

$$\mathrm{Pr}\,ecision = \frac{\mathrm{Re}\,levant}{\mathrm{Re}\,trevied} * 100\%$$

$$F - Measure = 2 * \frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} * 100\%$$

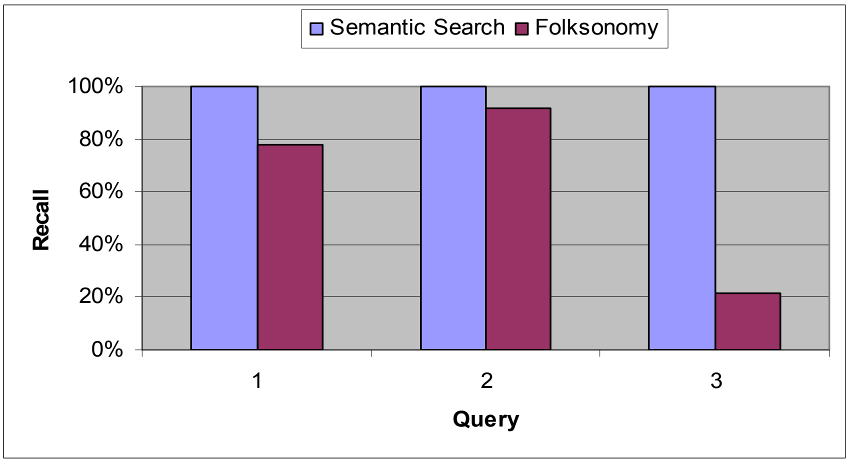
The evaluation framework allows for the search of CSS instances in two ways. In the first approach a keyword query is made to search for folksonomy tags in the corpus, while in the second approach a semantic search on the CSS ontology for the same instance is made.

A knowledge base of 100 semantic metadata items was built for the purpose of semantic search, and the retrieval algorithm performance was tested with three examples to compare the results of the two approaches (i.e. folksonomy-only search vs. semantic search). Moreover, the relevance of the retrieved web resources was based on a manual evaluation, where the researcher has used a manual metric that ranks all retrieved web resources for each query, on a scale from 0 to 5 ('0' indicates not relevant and '5' indicates strongly relevant). Thus, the measurements are subjective and limited; however, they are indicative of the degree of performance for ontology-based folksonomy search over folksonomy search alone.

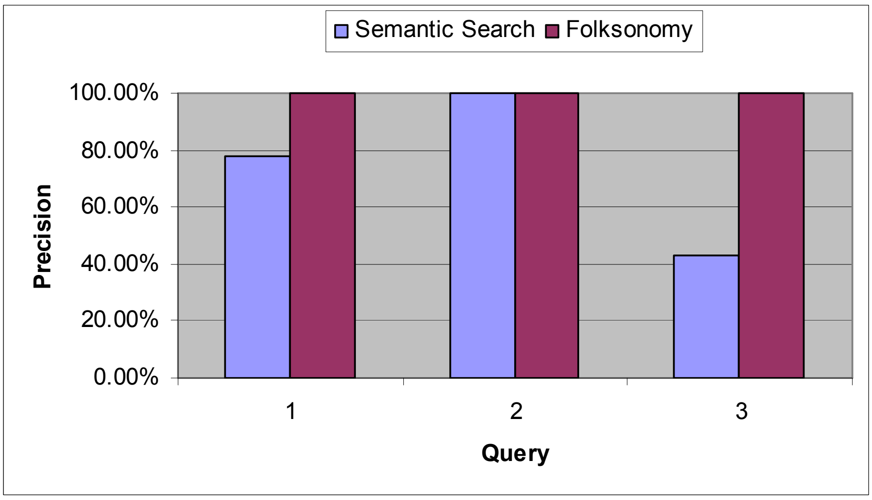
The semantic search has been designed to return semantically-related resources that have an explicit relationship between the retrieved resources. Even if the resource is not totally relevant to the concept the user has searched for, the retrieved resource can be helpful in one way or another. Next, the observed results of the three case studies are reported showing the different levels of the tool performance.

				Retrieved Documents			Recall		Precision		F-Measure	
Query No.	**Semantic Relationship**	**CSS Concept**	**Concept Instance**	**Correct Results**	**Semantically Related**	**Semantically Relevant**	**Semantic Search**	**Folksonomy Search**	**Semantic Search**	**Folksonomy search**	**Semantic Search**	**Folksonomy search**
Qa	Related_to	Subject	Positioning	7	2	0	100 %	77.7 %	77.7 %	100 %	87.5 %	87.5 %
Qb	Uses	Application	Menu	11	1	1	100 %	91.6 %	100 %	100 %	100 %	95.6 %
Qc	Used_with	Technique	Rounded Corner	3	11	3	100 %	21.4 %	42.8 %	100 %	59.9 %	35.2 %

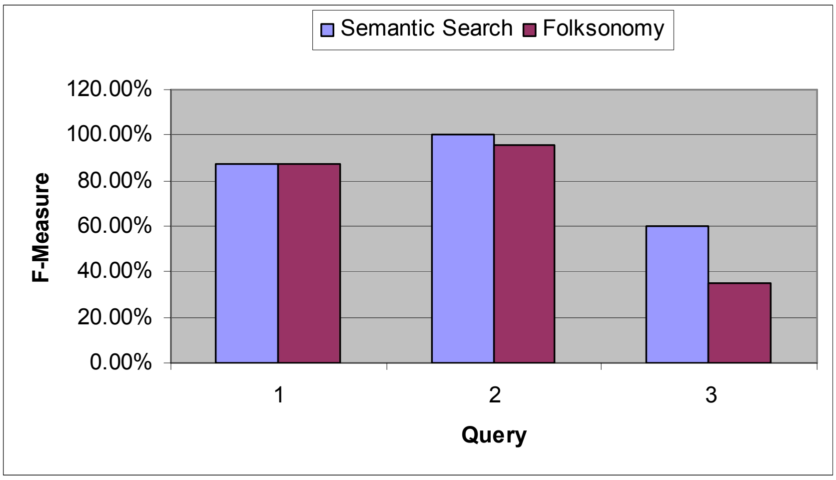
Table 10.1: A summarisation of the three query results showing the used semantic relationships, the concepts, their instances, the number of retrieved documents and the values of precision, recall and F-measure.



(A)



(B)



(C)

Figure 10.8: The Recall (A), Precision (B) and F-Measure (C) of ontology-based folksonomy search against folksonomy search alone. The performance of both techniques is shown for three different queries 1, 2 and 3

Query 1. "Resources about the Positioning subject."

In this example, the semantic search uses the 'Related_to' semantic relationship to retrieve further web resources from the CSS knowledge base. Apparently there were seven resources with explicit folksonomy tags describing them, and there were 2 semantically related resources that are connected to the subject 'Positioning' with a 'Related_to' relationship. However, from these 2 returned web resources, zero were relevant to the subject being searched for. The manual ranking of the 9 retrieved web resources have a mean relevance value of 3.5 out of 5 in the semantic search, versus 5 in the folksonomy search.

The recall for the semantic search was at its maximum level (100%) compared to folksonomy search recall which was 77.78%,; the reason behind this difference was because the semantic search results were boosted by the number of retrieved resources from the semantic relationship. On the other hand, the precision for the folksonomy search outperforms the precision of the semantic search, i.e. 100% versus 77.78%; this was due to the fact that all retrieved web resources from the folksonomy search are actually relevant resources, while in the semantic search only 7 resources were relevant.

Query 2. "Resources about CSS as a Menu application."

In this example the semantic search outperforms folksonomy search in recall (100% versus 91.60%) and was similar to it in precision (both 100%). The equal precision value can be attributed to the low number of semantically-related instances in CSS knowledge base, which in this case was just one instance. Thus, the manual ranking of the 12 retrieved web resources for the semantic search has a mean relevance value of 4.8 out of 5 compared to 5 in the folksonomy search.

Query 3. "Resources about Rounded corner technique."

In this example, the CSS knowledge base has only a few instances of rounded corner technique resources (3 instances), therefore, when querying for web resources that refer to this technique, all resources both relevant and semantically related to the query are retrieved. This causes precision to drop to lower values compared to the

previous queries. Although the total recall of folksonomy search is low (21.42%), it still has a good precision (100%) compared to the semantic search (42.85%).
Moreover, the manual ranking of the 14 retrieved web resources has resulted in a mean relevance value of 1.7 out of 5 in the semantic search, versus 5 in the folksonomy search.

Discussion

It can be seen from the evaluation results of the previous case studies, the importance of the added value of the semantic search as opposed to the folksonomy search alone (as the researcher foresaw), the semantic search exploited the power of semantic relationships represented in the CSS ontology to retrieve more results with varied relevance. This variation in relevant web resources affected the precision of the semantic search in favour of the folksonomy search. However, the recall for the semantic search was always better than the folksonomy search, which makes sense given the power of the ontology relationships.

Moreover, the harmonic mean, where the researcher has weighted the recall and precision as being of equal importance, showed that for the given three queries the semantic search over weighted the folksonomy search in both query 2 and 3, and equals it in query 1. The result indicates that the semantic search gave in better results than the folksonomy search alone; this finding is similar to other results obtained from research dealing with retrieving learning objects using an ontological approach such as (Lee et al., 2006).

In summary, the thesis prototypical tool achieved better recall in all three queries with respect to folksonomy-based search, and one important point to re-iterate is that the semantic search and the folksonomy search both use folksonomy tags as their basic assets, and the only difference between the two is that semantic search uses the ontologies and their associated metadata to allow for more results.

Expert vs. Folksonomy Semantic Metadata

To compare the quality and performance of the generated semantic metadata against an expert annotation, the researcher has randomly chosen 10 CSS web resources from the data set of 100 and asked an expert in the Web design and CSS domain to annotate them using the three ontologies. The human expert used the three ontologies to aid him in the process of annotation. The performance of the tool was then compared against the human annotator. The researcher had annotated all web resources beforehand to be considered as benchmark standard to which the two parties were compared. The researcher has again used IR measurements of precision, recall and f-measure as follows:

$$\mathrm{Re}\,call = \frac{\mathrm{Retrieved}}{\mathrm{Relevant}}$$

Relevant: means all web resources marked relevant by the researcher.

$$\mathrm{Pr}\,ecision = \frac{\mathrm{Relevent}}{\mathrm{Retrieved}}$$

$$F - Measure = 2 * \frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} * 100\%$$

		Recall		**Precision**		**F-Measure**	
Concept	**Instance**	**Expert**	**Folks.**	**Expert**	**Folks.**	**Expert**	**Folks.**
Property	background	100%	60 %	100 %	100%	100%	75.00%
Application	menu	100%	125 %	100%	100%	100%	111.11%
Layout	Box	100%	20%	100%	100%	100%	33.33%
Technique	overlay	100%	50%	100%	100 %	100%	66.67%
Subject	navigation	33%	167%	100%	60%	49.62%	88.28%
Resource type	Code	11%	111%	100%	90%	19.82%	99.40%
Resource type	article	0%	175%	0%	57%	0%	85.99%
Resource type	example	80%	80%	100%	100%	88.89%	88.89%

Table 10.2: A summarisation of the results of precision, recall and F-measure for the eight queries issued on the annotated web resources by both the human expert and the FolksAnnotation tool.

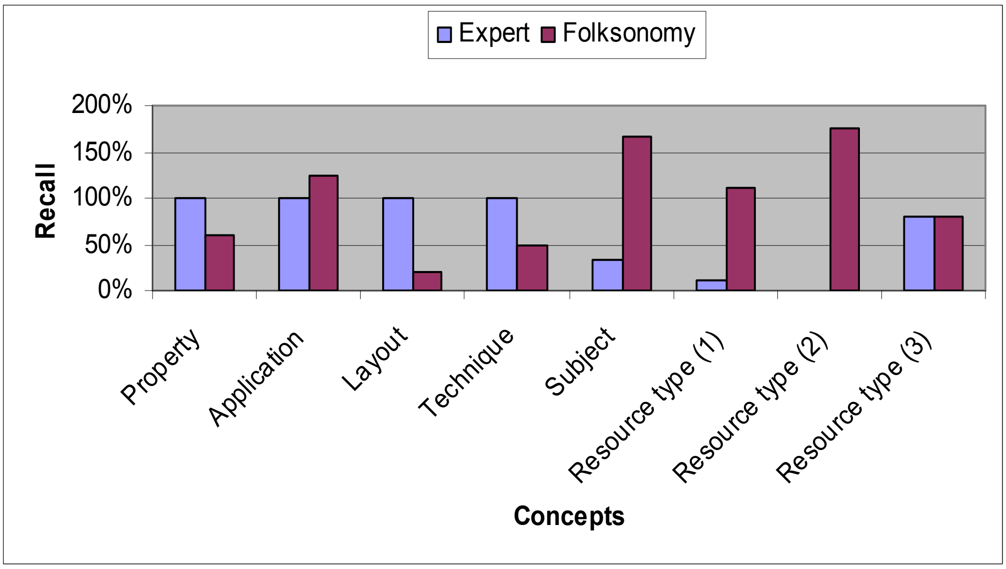
Table 10.2 summarises the results after running the eight search queries against the folksonomy metadata and the human expert metadata. The table shows interesting cases obtained from the case study. As an example, in the instances: background, box and overlay, the recall score for the human expert surpasses the folksonomy score, this was because the human expert has correctly annotated and covered all the

potential web resources. In contrast, the folksonomy metadata group failed to cover all the potential web resources which affected their recall score accordingly.
However, the precision for both parties was at its highest value (i.e. 100%), which is because all retrieved web resources from both parties were relevant.

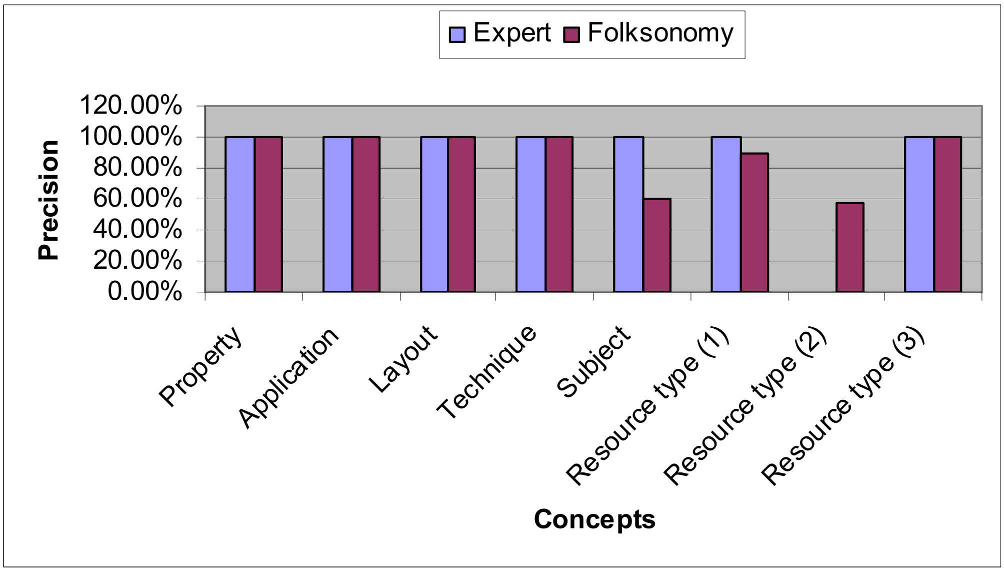
Another interesting observation can be found in the instances: menu, navigation, code and article, where the recall score for the folksonomy metadata group outperforms the human expert recall score. This phenomenon can be attributed to the use of extra contextual dimensions in the folksonomy metadata group to boost the results; i.e. this happens when more than one instance for a given concept is used to populate the field. To illustrate the effect of the added contextual dimension an example is given. ‘Navigation’ is an instance of the Subject concept; del.icio.us users use this instance with web resources talking about ‘sitemaps’ or ‘menus’. Apparently, del.icio.us users use the ‘Navigation’ instance in web resources that do not directly address ‘sitemaps’ or ‘menus’; they add it to web resources talking about buttons or lists to extrapolate its indirect use in ‘Navigation’ as a Subject.

As a result, the precision of the folksonomy group has fallen down dramatically against the human expert precision score, except for the ‘menu’ and ‘article’ instances, where in the former it equals the precision score of the human expert and in the latter the human expert did not annotate any of the web resources as being of type ‘article’. Moreover, the recall results of the ‘Navigation’ and ‘Code’ instances showed that the human expert has missed annotating some web resources from the data set as having these descriptors, this illustrated that even a human expert can not compete with the aggregate knowledge of crowds.

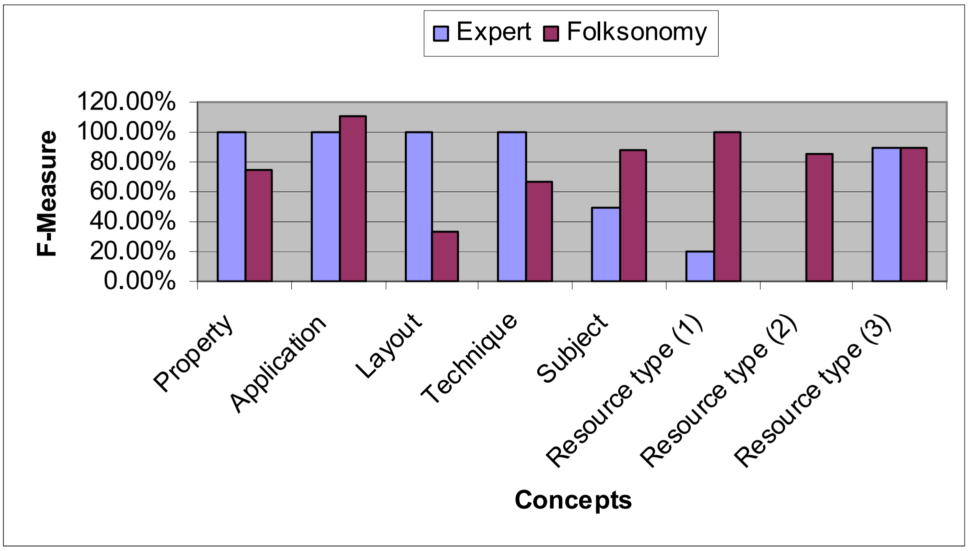
A third interesting observation can be found in the ‘example’ instance, where the human expert and the folksonomy group both have an equal recall score of 80%; this was because both parties have missed annotating a web resource as being an ‘example’. However, both parties’ precision score was at its highest value, this was due to the same reason of having both parties correctly annotating the web resource as being an ‘example’ instance resource type.


Expert
Folksonomy
Recall
200%
150%
100%
50%
0%
Property
Application
Layout
Technique
Subject
Resource type (1)
Resource type (2)
Resource type (3)
Concepts

(A)


Expert
Folksonomy
Precision
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Property
Application
Layout
Technique
Subject
Resource type (1)
Resource type (2)
Resource type (3)
Concepts

(B)



(C)

Figure 10.9: The Recall (A), Precision (B) and F-Measure (C) of folksonomy search results against the human expert search results. The performance of both techniques is shown for eight different queries.

Another interpretation of the results is shown in Figure 10.9. The figure shows the results obtained after evaluating the search performance of the human expert assignment against the folksonomy assignment.

The recall results, Figure 10.9(A), show that the folksonomy results were better than the expert results in almost half of the queries. However, the precision results, Figure 10.9(B), show that the human expert outperforms the folksonomy results in two cases and equals them in the rest, except for one case (i.e. Resource type (2)) where the human expert did not assign a value to the web resource; this action affected the human expert precision and recall scores accordingly.

Finally, in order to thoroughly compare human expert and folksonomy performance it is necessary to observe differences of f-measure, since it is an aggregate measure of both precision and recall. Thus in Figure 10.9(C), the researcher can assume that in this particular case study, folksonomy search has performed better in most of the cases compared to the human expert search results. This outcome can be attributed to the high values of the recall in the folksonomy results which leveraged the f-measure results accordingly.

10.2.2.1 In-depth Manual Qualitative and Quantitative Evaluation

A manual evaluation was also carried out to compare the difference between folksonomy assignment and its human expert counterpart. The difference is measured quantitatively, by counting the number of instances assigned to each descriptor, and qualitatively by thoroughly examining the quality of folksonomy assignment against the human expert as a gold-standard.

For the *property* concept, the human expert tends to be more precise and diverse in using instances from the CSS ontology for this concept. On the contrary, folksonomy tags were few and only cover one, two or three sub-concepts in this category. Among the ten folksonomy annotated web resources, two of them did not have the property descriptor field filled.

Moreover, for the ten web resources, the folksonomy tags assigned to the *property* field were quantitatively less than the human expert assignment; i.e. roughly 4.5:2 ratio in favour of the human expert.

Qualitatively speaking, evaluating the content of the *property* descriptor for the folksonomy group showed that most folksonomy tags were sub-sets of the human expert assignment, except for three web resources where folksonomy tags covered different instances for this field. The researcher found that both the human expert descriptors assignment and the folksonomy assignment; of the three web resources, both were valid in terms of web resource ‘about-ness’.

For the *element* (i.e. *attribute)* concept, the human expert did not populate this field for the ten web resources, however, the folksonomy metadata group have one web resource that has an attribute field populated. By further inspecting this web resource the researcher has found that the folksonomy assignment was correct.

For the *selector* concept, the human expert has indicated that seven web resources used an instance of the *selector* concept compared to four assignments from the folksonomy metadata group. Both parties used the same instance (i.e. div) for the annotated web resources.

For the *unit* concept, the human expert has showed that eight web resources demonstrated the use of one or more instances of the unit concept. On the other hand, the folksonomy metadata group did not annotate any of the ten web resources with this piece of information.

For the *application* concept, the human expert has assigned a single instance for each of the ten web resources; on the contrary, the folksonomy metadata group has seven of its web resources application field filled with one or more instances from the application concept. These instances if compared to the human expert assignment are considered either the same as the human expert assignment or they added more contextual applications that have not been stated explicitly in the annotated web resource. To give an example, the "**Super Easy Blendy Backgrounds**"[138] was assigned an application instance of 'Textbox background' by the human expert, while in the folksonomy metadata for the same web resource the folksonomy metadata group has assigned 'Menu' and 'Button' instances, which are both extra valid applications for the designated web resource.

For the *layout* concept, the human expert has assigned a single instance for each of the ten web resources, in contrast; the folksonomy metadata group has only three web resources with layout instances assigned to them. Again, as observed in the *application* concept, the number of assigned instances in the folksonomy metadata group was equal or more than their human expert counterpart assignment. These more instances added new potential dimension to the layout of an annotated web resource.

For the *technique* concept, the human expert has assigned one or more instances for only eight web resources; on the contrary, the folksonomy metadata group has seven web resources annotated with instances from the *technique* concept. However, the total number of assigned instances by the human expert was 10 compared to 15 in the folksonomy metadata group. This again shows that del.icio.us users are adding more contextual dimensions to the annotated web resources. Not surprisingly, the human expert assignment has agreed with the del.icio.us users' assignment for four web resources.

[138] http://www.alistapart.com/articles/supereasyblendys [last accessed 9/2/2007]

For the *subject* concept, both the human expert and the folksonomy metadata group have assigned instances to the *subject* field for the ten web resources. However, the number of assigned instances for each web resource was different between the two parties. For the human expert, the rate of assignment was one or two instances as a maximum for the *subject* field. On the contrary, the folksonomy metadata group assigned more than one *subject* instance to each web resource. The total number of assignments for the human expert was 12 as against 19 for the folksonomy group. From all the ten annotated web resources for both parties, only four of them were not matching, this returns us again to the notion of added contextual dimensions in people's tags. In other words, the added instances in people's tags were not wrong, they just highlighted a potential subject for the annotated web resource.

For the *resource type* concept, the power of collective intelligence was shown in action. For the ten web resources, the folksonomy group has assigned a total of 81 resource type instances compared to 12 assignments by the human expert, which constitute a ratio of roughly 1:7 instances assignments for the folksonomy group. Notice that all the human expert instances assignments were sub-sets of the folksonomy group assignment; this demonstrates the power of aggregated intelligence. As for the quality and validity of the folksonomy group assignments, the researcher thinks that they were both technically and pedagogically acceptable.

Discussion

The quantitative results of the precision and recall statistics showed that the folksonomy results were better than the human expert results in most of the queries, and this is what the f-measure has also justified. However, the researcher does not claim that these observations can be generalised; instead these observations represent interesting cases that might happen when attempting to analyse larger data sets. They also show the level of extra information that a typical indexer might not spot.

As for the qualitative manual evaluation of both assignments, the researcher has found that although the human expert is more precise than folksonomy users when annotating a web resource, due to the existence of a predefined template to fill, folksonomy tags have added potential contextual dimensions to most of the web resources. These contextual dimensions can be attributed to the suggested possible applications applied to a given web resource that the taggers have in mind or to the different perspectives a tagger might think of when tagging a web resource.

Moreover, when tagging web resources with elements from the resource type ontology, the human expert tends to annotate resources with pedagogical instances and forgets about the technical aspect of the resource. This observation demonstrates the power of aggregating group intelligence against an individual human subject-matter expert mindset.

10.3 Metadata Assignment Evaluation

The metadata assignment evaluation stage is necessary to evaluate the *quality, validity and representativeness* of the generated semantic metadata record. To verify these requirements, the researcher used a blend of quantitative and qualitative evaluation techniques.

So, to evaluate the previous requirements a set of questions need to be answered, which are:

- Are the semantics of the metadata elements ***clear and unambiguous***?
- How ***well*** does the metadata describe the web resource?
- How ***accurate*** is the generated metadata about the web resource?

To answer these questions, a questionnaire was designed and distributed to a group of subject domain experts to rate the appropriateness of the descriptors and metadata assigned; this evaluation technique was adopted from (Al-Yahya, 2006). The questionnaire measured how well the user believes the metadata predicts the actual contents of the web resource.

To validate the suitability of the questionnaire, a pilot test was conducted before the questionnaire was used with the target population. First, two colleagues were asked to read, revise and evaluate the questionnaire, then they filled it out in front of the researcher so that she could spot any deficiencies or difficulties encountered while answering the questionnaire. Both colleagues had previous experience with teaching

Web Design and a good knowledge of metadata. The comments of these reviewers were used to revise the questionnaire.

The validity of the questionnaire was also enhanced by distributing it to samples other than the subject population of the current study. These samples were reached via some mailing lists that the researcher is subscribed to, such as the Systers[139] mailing list.

The questionnaire was then distributed to two target populations: Web designers and experts in the field of learning technologies and metadata (called the specialists group in the subsequent discussion). The Web designers' community was reached using mailing lists that resides at Yahoo Groups and other focused groups such as css-discuss[140]. The total response from the Web designers group was 29 respondents and the total response from the specialists group was 22 respondents.

The professional roles of the respondents within the web designers' community are shown in Figure 10.10. The figure indicates that 80% of the respondents were web designers or/and programmers while 12% of the respondents preferred to choose 'others' to indicate that they are either amateur web programmer/designer or high school teachers. The rest of the responses (8%) were divided equally between being professors or researchers and being postgraduates or undergraduates.

[139] http://www.systers.org/ [last accessed 7/2/2007]

[140] http://css-discuss.org/ [last accessed 7/2/2007]

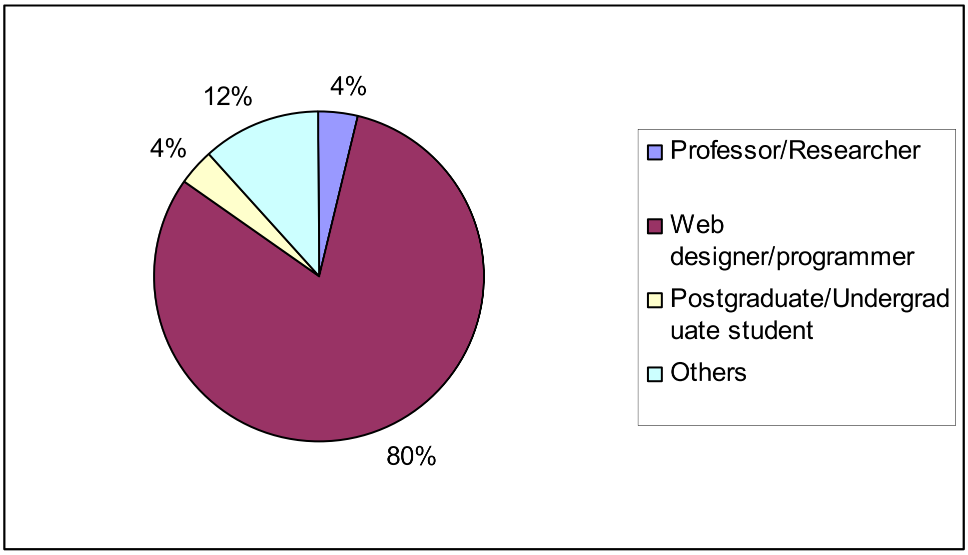


Figure 10.10: Distribution of the Web designers' group based on their professional role

Table 10.3 shows the expertise statistics for the Web designers' community respondents. The respondents were asked to rate their expertise on a scale from 1, which indicates 'poor' expertise, to 5, which indicates 'excellent' expertise. The table shows that the overall respondents' knowledge in the fields of Web programming, Web design and Cascading Style Sheets (CSS) are all above average. However, the respondents' expertise in teaching web design was below midpoint.

Expertise	Response Average
Web programming in general	3.35
Knowledge of web design domain in general	3.77
Knowledge of Cascading Style Sheets (CSS)	3.78
Experience in teaching or tutoring a web design course	2.23

Table 10.3: Averages of the expertise level for the web designers' group

On the other hand, the professional role for the specialists' group respondents is shown in Figure 10.11. The figure indicates that 36% of the respondents were either professors or researchers, 27% were postgraduates or undergraduates, 18% of the respondents preferred to choose 'others' to indicate their specific professional role such as IT support, learning technology consultant, learning technology adviser or development team leader. 14% of the respondents were cataloguers and 5% were Web designers or/and programmers.

Notice the diverse range of expertise in the specialists' group population; this is useful to evaluate the generated metadata from different perspectives.

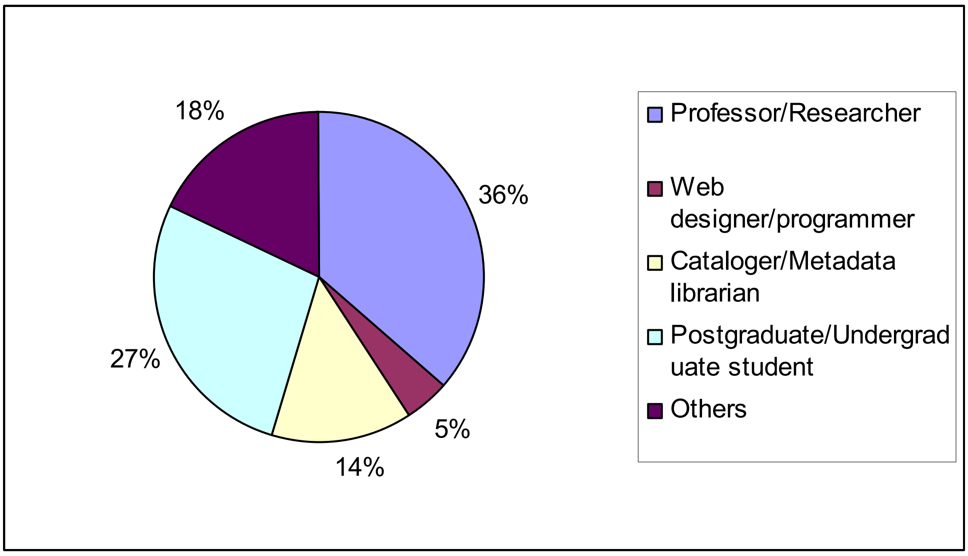


Figure 10.11: Distribution of the specialists group based on their professional role

Table 10.4 shows the expertise statistics for the specialists' group respondents. The table shows that the overall respondents' knowledge in the fields of Web programming, Web design and CSS are all above average. However, the respondents' expertise in teaching web design was below midpoint.

Expertise	Response Average
Web programming in general	3.4
Knowledge of web design domain in general	3.67
Knowledge of Cascading Style Sheets (CSS)	3.20
Experience in teaching or tutoring a web design course	2.4

Table 10.4: Averages of the expertise level for the specialists group

By comparing the level of expertise in both groups, the researcher can observe that the Web designers group is more knowledgeable in the field of Web design and CSS;

however, the specialists group is slightly better than the Web designers group in the experience of teaching Web design.

10.3.1 Metadata Representativeness

Two questions in the questionnaire were designed to capture the respondents view on the representativeness of the metadata elements. The first question concerns the usefulness of the metadata descriptors used to describe a CSS web resource and the second question concerns the required metadata fields needed to search for CSS web resources.

The respondents were asked to rate (based on a scale from 1 to 5 where 1 represents 'useless' and 5 represents 'very useful') how useful was each metadata element used to describe and search for web resources in the domain of teaching CSS.

Table 10.5 shows the mean and standard deviation of each metadata descriptor used for the purpose of describing CSS web resources. The overall statistics for the Web designers' responses show that the mean of all the metadata elements are above average, except for the 'Difficulty level' element which is slightly below midpoint. However, the standard deviation for all elements is quite high, which indicates the varied view between respondents. By comparing the means of all elements, it is apparent that the 'Description' and 'Title' elements are among the most useful descriptors.

Metadata Element	Mean	Standard Deviation
Title	3.94	1.44
URL	3.75	1.25
Description	4	1.13
Keywords	3.25	1.45
Difficulty level	2.8	1.2
Instructional level	3.2	1.47
Resource type	3.38	1.31
Recommendation	3.14	1.16

Table 10.5: Overall score of the usefulness of the metadata descriptors used to describe a CSS web resource rated by the web designers group

Table 10.6, on the other hand, shows the result of the mean and standard deviation for each metadata element used for the purpose of searching for CSS web resources. The overall statistics of the Web designers' responses show that the mean of the metadata elements are all above average, except for the 'Difficulty level' element again, which is slightly below midpoint. However, the standard deviation for most elements is quite high, which indicates the varied view between respondents, except for two elements which are 'Selector' and 'Difficulty level'. This indicates some consistency in the respondents rating towards these two elements. By comparing the means of all elements, it is apparent that most elements are equally likely useful descriptors for retrieving/searching for a CSS web resource, except for 'Difficulty level' which did not appeal to the community of Web designers.

Metadata Element	Mean	Standard Deviation
Property	4.27	1.10
Element (i.e. Attribute)	3.36	1.62
Selector	4.54	0.68
Units	3.36	1.56
Application	3.27	1.5
Layout	3.54	1.21
Technique	3.72	1.01
Subject	3.72	1.19
Resource type	3.45	1.04
Difficulty level	2.8	0.87
Instructional level	3.27	1.35

Table 10.6: Overall score of the usefulness of the metadata elements used to search for a CSS web resource rated by the web designers group

As for the specialists group, Table 10.7 shows the mean and standard deviation of each metadata descriptor for this group. The overall statistics of the responses show that the mean of the metadata elements are all above average. However, the standard deviation for most elements is quite high, except for the 'Keywords' and 'Recommendation' elements, which indicates the varied view between respondents. By comparing the means of all elements, it is apparent that the 'Description', 'Title' and 'Keywords' elements are among the most useful descriptors.

Metadata Element	Mean	Standard Deviation
Title	3.8	1.46
URL	3.42	1.3
Description	3.81	1.4
Keywords	4.42	0.8
Difficulty level	3.33	1.23
Instructional level	3.33	1.23
Resource type	3.25	1.3
Recommendation	3.5	0.8

Table 10.7: Overall score of the usefulness of the metadata descriptors used to describe a CSS web resource rated by the specialists group

Table 10.8 on the other hand, shows the result of the mean and standard deviation for each metadata element used for the purpose of searching for CSS web resources. The overall statistics of the specialists group responses show that the mean of the metadata elements are all above average, except for the 'Element', 'Selector' and 'Unit' elements, which is slightly below midpoint. However, the standard deviation for most elements is moderate, which indicates some consistency in the respondents rating towards these elements. By comparing the means of all elements, it is apparent that the 'Subject', 'Technique' and 'Property' elements are among the most useful descriptors for retrieving/searching for a CSS web resource.

Metadata Element	Mean	Standard Deviation
Property	3.42	1.08
Element (i.e. Attribute)	2.9	1.3
Selector	2.5	1.12
Units	2.72	0.65
Application	3.27	0.65
Layout	3.09	0.7
Technique	3.45	0.82
Subject	3.81	1.07
Resource type	3.6	1.12
Difficulty level	3.36	1.3
Instructional level	3.45	1.3

Table 10.8: Overall score of the usefulness of the metadata elements used to search for a CSS web resource rated by the specialists group

Discussion

It can be noticed from the previous statistics that the overall rating for most metadata elements in both groups was fairly acceptable. Both Web designers and specialists group agreed on the usefulness of the metadata descriptors given the varied views between the respondents. Unsurprisingly, the respondents did not under-estimate the importance of the provided metadata descriptors for either searching or describing CSS web resources.

Furthermore, one question in the questionnaire was asked to see if there were any other metadata elements that might be useful for describing or searching for web resources from a CSS knowledge base.

The respondents from the specialists group have answered with the following suggestions:

- "Relationships; indicate relationship between resources. In other words in learning objects whether a certain asset needs to be viewed before the next, etc.,
- Author,

- Creation date,
- CSS version,
- Accessibility,
- Objective of each resource".

As can be seen from the specialists group answers most suggestions are either applicable to produce given the set of folksonomy tags such as CSS version, or inapplicable because it requires human intervention such as determining the objective of a resource or its author.

On the other hand, the Web designers group have suggested the following additions:

- "Language, 'hacks', in a way of referencing the specific (ab)use of different browsers' understanding or not of certain techniques which can often be on a very minute scale of computing,
- Browser compatibility,
- 'Hacks' - CSS settings tweaks needed to compensate for differences in CSS implementations,
- Browser support (e.g. "this selector/property/technique works in IE6+, Firefox 1.0+, and Safari 1.0+.")".

The Web designers responses were more technical than those of the specialists group, however, they pointed out some important descriptors that can be used to enhance the description of a CSS web resource such as browser compatibility and language.

10.3.2 Metadata Quality and Validity

The questionnaire was designed to also include a question about the quality and validity of a random sample of three CSS web resources. These three automatically generated semantic metadata records were selected based on their coverage of the various aspects of the CSS metadata descriptors. Therefore, the three metadata records were exposed to both groups (Web designers and specialists) to rate them based on a metric used to evaluate the quality and validity of metadata elements by (Greenberg, 2005). The evaluation is based on a three-tier scale, which is:

- **“Good**: Good metadata accurately represented the resource and would facilitate accurate resource discovery. A good metadata element does not require any revision.
- **Fair**: Fair metadata would be somewhat useful for resource discovery of the resource being represented. In this case, a revision(s) would generally improve the quality of the metadata element.
- **Reject**: Reject (poor quality) metadata was inaccurate. In this case, the metadata element required substantial revision to be useful for resource discovery.”

The results of the quality and validity for each metadata element of the three resources are discussed next.

Metadata Elements Assessment

The quality and validity for the ‘Title’, ‘Description’, ‘Keywords’, ‘Resource Type’, ‘Property’, ‘Selector’, ‘Element’, ‘Technique’, ‘Application, ‘Subject’, ‘Layout’, ‘Difficulty level’ and ‘Instructional level’ metadata elements were evaluated for each metadata record produced by the thesis prototype tool.

10.3.2.1 Title Metadata

A web resource title is automatically extracted from the del.icio.us web resource post page. It was anticipated that the quality and validity for this element will be high, that is because when posting a web resource link to del.icio.us, del.icio.us uses the exact web resource title that is explicitly provided in the HTML code of that web resource. Table 10.9 shows the Web designers group evaluation of the title element of the three web resources. Notice that 65% of the web designers agreed that the title provided by the web resource itself was enough; hence ‘Good’, while 28% think that it is ‘Fair’ and 8% did not like the assignment.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good	56%	88%	50%	65%
Fair	33%	12%	38%	28%
Reject	11%	-	12%	8%

Table 10.9: Overall evaluation of the Title element for the Web designers group

On the other hand, Table 10.10 shows the specialists group evaluation of the title element of the three web resources. 52% of the specialists group agreed that the title provided by the web resource itself was 'Good', while 30% think that it is 'Fair' and 18% did not like the assignment.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good	56%	67%	34%	52%
Fair	22%	33%	33%	30%
Reject	22%	-	33%	18%

Table 10.10: Overall evaluation of the Title element for the specialists group

The difference between the two groups' ratings is not that noticeable, more than 80% of both sets of respondents agreed on the acceptable assignment of the title element, while less than 20% did not accept the automatic assignment. This good agreement between both groups indicates that the del.icio.us title extraction process was successful in assigning bookmarked web resources with appropriate title.

10.3.2.2 Description Metadata

The description metadata element was generated automatically from a pre-defined template as mentioned in chapter 9. Table 10.11 shows that the Web designers group did not like the way the description of a web resource was reported. Thus, an average of 56% of the evaluations deemed that the description needs to be rejected. However, an average of 23% and 11% of the Web designers reported that the description is 'Fair' and 'Good' respectively. In other words, although the description element is not accepted by most Web designers as shown in the table, some of them thought that the description might be good enough to describe the pedagogical aspect of a web resource.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good	22%	0%	12%	11%
Fair	22%	50%	25%	32%
Reject	56%	50%	62%	56%

Table 10.11: Overall evaluation of the Description element for the Web designers group

By the same token, Table 10.12 shows that the specialists group rarely liked the way the description of a web resource was reported. On average 29% rated this element as being ‘Good’, 38% thought it was ‘Fair’ while 33% rejected this element.
However, the element is considered fairly acceptable when summing the overall rating of the acceptability region i.e. ‘Good and Fair’. This yields a total of 67% acceptance compared to 33% rejection.

Evaluation ↓	**Resource #1**	**Resource #2**	**Resource #3**	**Average Evaluation**
Good	12%	50%	25%	29%
Fair	50%	25%	38%	38%
Reject	38%	25%	37%	33%

Table 10.12: Overall evaluation of the Description element for the specialists group

10.3.2.3 Keyword Element

The keywords assigned to the annotated web resources are extracted from the folksonomy tags that have been assigned to the three ontologies. Table 10.13 summarises the average evaluation results for the keyword element. The results show that on average 44% of the Web designers still do not like the keywords assigned to the three web resources. However, 40% think that the keywords were ‘Good’ while 15% think that they were ‘Fair’. Notice that most of the keywords used in populating the template of the description element gained better evaluations from the Web designers group; this might means that these keywords are good enough to be used alone without forcing them in any template. Also, the small difference between evaluating the keyword element as being ‘Good and Fair’ against being ‘Rejected’, is some how oriented toward accepting its value (40%+15% = 55%) rather than rejecting it completely (44%).

Evaluation ↓	**Resource #1**	**Resource #2**	**Resource #3**	**Average Evaluation**
Good	33%	38%	50%	40%
Fair	22%	12%	12%	15%
Reject	44%	50%	38%	44%

Table 10.13: Overall evaluation of the Keyword element for the Web Designers group

Despite the specialists group moderate acceptance of the description element, they showed more positive rating for the keyword element as shown in Table 10.14. They rated the keyword element for the three web resources as either being 'Fair' or 'Good', thus, they did not reject it. The overall rating of the keyword element was believed to be 'Fair' or 'Good' with 58% and 42% acceptance, respectively. This implies that the specialists group valued the keywords elements more than the Web designers did.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good	38%	50%	38%	42%
Fair	62%	50%	62%	58%
Reject	-	-	-	-

Table 10.14: Overall evaluation of the Keyword element for the Web specialists group

10.3.2.4 Variable Elements

These are metadata elements that might or might not appear in every CSS web resource, thus from the three evaluated web resources, the reader will find that some elements have appeared only once in the three web resources.

10.3.2..4.1 Resource Type Element

Table 10.15 shows the Web designers evaluation for the resource type element. It seems from the results that there was some consistency between the two groups in evaluating the resource type element as being 'Fair'. Most of the Web designers respondents evaluation falls in the acceptance region (32%+40%=72%) rather than rejecting the element entirely (28%). This indicates that all suggested resource types provided by folksonomy tags are acceptable and valid for describing the pedagogical functionality of a bookmarked web resource.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good	34%	25%	38%	32%
Fair	33%	38%	50%	40%
Reject	33%	37%	12%	28%

Table 10.15: Overall evaluation of the resource type element for the Web Designers group

Along the same line, the specialists group has also rated the resource type element as being 'Fair' with an overall acceptance of 91% (i.e. 41%+50%) compared to 9% rejection, see Table 10.16. This result emphasis on the added contextual dimension present in people's tags.

Evaluation ↓	**Resource #1**	**Resource #2**	**Resource #3**	**Average Evaluation**
Good	12%	62%	50%	41%
Fair	63%	38%	50%	50%
Reject	25%	-	-	9%

Table 10.16: Overall evaluation of the resource type element for the specialists group

10.3.2..4.2 Subject Element

Table 10.17 shows the Web designers evaluation for the subject element. On average 52% of the Web designers' group marked this element as being 'Fair', while only 20% rejected it. Moreover, 80% (28% + 52%) of Web designers think that this element was acceptably assigned.

Evaluation ↓	**Resource #1**	**Resource #2**	**Resource #3**	**Average Evaluation**
Good	22%	38%	25%	28%
Fair	56%	50%	50%	52%
Reject	22%	12%	25%	20%

Table 10.17: Overall evaluation of the Subject element for the Web Designers group

In the same vein, 54% of the specialists group has marked the subject element as being 'Good', 29% as being 'Fair' and only 17% rejected its assignment.

Evaluation ↓	**Resource #1**	**Resource #2**	**Resource #3**	**Average Evaluation**
Good	88%	63%	12%	54%
Fair	12%	25%	50%	29%
Reject	-	12%	38%	17%

Table 10.18: Overall evaluation of the Subject element for the specialists group

10.3.2..4.3 Application Element

Table 10.20 and Table 10.21, consecutively, show the Web designers and specialists group evaluations for the Application element. Notice that this element did not appear in the first resource (Resource#1), thus, the Application element has been only evaluated for Resource#2 and #3.

The Web designers group has accepted the element assignment with an overall rating of being ‘Fair’ (50%) and ‘Good’ 19%, however, 31% of Web designers did reject the element assignment.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		-	38%	19%
Fair		50%	50%	50%
Reject		50%	12%	31%

Table 10.19: Overall evaluation of the Application element for the Web Designers group

The specialists group, on the other hand, did not seem to rate the application element assignment as being ‘Good’. However, on average most of the group (69%) agreed that the element was fairly assigned, while 31% of them thought it was not acceptable which is the same as the Web designers’ group decision.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		-	-	-
Fair		62%	75%	69%
Reject		38%	25%	31%

Table 10.20: Overall evaluation of the Application element for the specialists group

10.3.2..4.4 Technique Element

Table 10.21 shows the Web designers evaluation for the Technique element. Again this element has appeared only in Resource#2 and #3. On average 75% of the Web designers’ group marked this element as being ‘Good’ compared to 19% who ranked it as being ‘Fair’. However, small portion of the Web designers’ group (i.e. 6%) has rejected the assignment.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		88%	62%	75%
Fair		12%	25%	19%
Reject		0%	12%	6%

Table 10.21: Overall evaluation of the Technique element for the Web Designers group

Similarly, on average 44% of the specialists group has rated this element as being 'Good', 50% as being 'Fair' and only 6% thought it was not acceptable. For a second time, the same portion of Web designers and the specialists group have the same opinion about rejecting this element assignment, and it is small compared to the number of respondents who accepted it.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		50%	38%	44%
Fair		38%	62%	50%
Reject		12%	-	6%

Table 10.22: Overall evaluation of the Technique element for the specialists group

10.3.2..4.5 Property Element

Table 10.23 shows the Web designers evaluation for the Property element. On average 50% of the respondents rated this element as being 'Fair', 32% rated it as being 'Good' and 18%' rejected it.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		25%	38%	32%
Fair		50%	50%	50%
Reject		25%	12%	18%

Table 10.23: Overall evaluation of the Property element for the Web Designers group

Likewise, Table 10.24 shows the specialists group evaluation for the Property element. On average 50% rated this element assignment as being 'Good', 38% rated it as being 'Fair' while 12% rejected its value.

When it comes to accepting the element as it is or accepting it after providing some minor changes, the views of the Web designers' group and the specialists group were opposite to each other, i.e. 50% of Web designers thought it was ‘Fair’ while the same proportion of the specialists group though it was ‘Good’.

Evaluation ↓	Resource #1	Resource #2	Resource #3	Average Evaluation
Good		38%	62%	50%
Fair		50%	25%	38%
Reject		12%	12%	12%

Table 10.24: Overall evaluation of the Property element for the specialists group

10.3.2..4.6 Element, Layout and Selector Elements

The Element, Layout and Selector descriptors have appeared once in each of the three resources; therefore, a single table was compiled to display their ratings. Table 10.25 summarizes the Web designers’ evaluation of the three elements. By looking at the largest value in each column, the researcher can observe that the Element descriptor assignment was fairly accepted with 56% of Web designers voting for it, the Layout descriptor was also fairly accepted with 50% of the Web designers’ group accepting it. The Selector descriptor was highly rejected with 50% votes from the Web designers group. However, when comparing the Selector element acceptance region (i.e. Fair and Good) it can be seen that its sum (12% + 38% = 50%) equals the percentage of people who rejected this element. This indicates the varied views of the Web designers’ group toward this element assignment.

Evaluation ↓	Resource #1 (Element)	Resource #2 (Layout)	Resource #3 (Selector)
Good	22%	25%	12%
Fair	56%	50%	38%
Reject	22%	25%	50%
Largest Value	56%	50%	50%

Table 10.25: Overall evaluation of the Element, Layout and Selector descriptors for the Web Designers group

Contrary to the Web designers’ group rating, Table 10.26 shows that the largest proportion of the specialists group has rated the three elements as being ‘Fair’, and the rejection of the three elements was notably small compared to the total acceptance of these elements.

Evaluation ↓	Resource #1 (Element)	Resource #2 (Layout)	Resource #3 (Selector)
Good	25%	25%	13%
Fair	38%	62%	62%
Reject	37%	13%	25%
Largest Value	38%	62%	62%

Table 10.26: Overall evaluation of the Element, Layout and Selector descriptors for the specialists group

10.3.3 Discussion

The results of the quality and validity for each metadata element of the three web resources were assessed by two expert groups (i.e. Web designers and experts from the field of learning technologies and metadata). For the three annotated web resources both the Web designers group and the specialists group agreed in giving the following metadata descriptors: 'Title', 'Resource type', 'Subject', 'Application', 'Technique', 'Property', 'Attribute' and 'Layout'; either a 'Good' or 'Fair' rating. However, the two groups diverge in their opinion of the rest of the metadata descriptors which are: 'Description', 'Keywords' and 'Selector'. In the specialists group they rate these elements as 'Fair', 'Good' and 'Fair' respectively; while, the web designers group has rated them as 'Reject'.

From the previous evaluation, the researcher can observe that the specialists group ranked the metadata descriptors higher than Web designers group; this might be because the specialists group knows the importance of using metadata to describe learning resources. Moreover, the Web designers group was dissatisfied with the automatic description generated by the tool, and also the keywords assigned, this was contrary to what the specialists group thinks! However, both groups agreed on the quality and validity of the metadata descriptors assigned by the thesis tool, i.e. on average both groups rated the metadata elements as 'Fair'; these results justify the researcher's claim about the quality and validity of the folksonomic metadata generated using folksonomy tags.

10.4 Further Evaluations

In this section an extended evaluation is carried out to examine the discarded folksonomy tags (unused tags) left out by the semantic annotation pipeline. Furthermore, a significance test was performed to measure the correlation between folksonomy tags assignment and Yahoo TE keywords assignment for a sample of web resources. Finally, a closer look at the assigned tags that fall in the Long Tail is carried out, to determine whether these tags are considered fine-grained or not.

10.4.1 Analysis of Unused Tags

To evaluate the unused folksonomy tags a classification scheme from (Sen et al., 2006), which was adopted and modified from (Golder and Huberman, 2006) to categorise folksonomy words, was used. Sen et al. have classified folksonomy tags into three groups:

- (P)ersonal tags: "have an intended audience of the tag applier themselves. They are often used to organize a user's own resources (self-reference, task organization, time management).
- (S)ubjective tags: express people opinions related to a web resource, and
- (F)actual tags: identify 'facts' about the described web resource such as people, places, or concepts."

To use the Sen et al. (2006) classification, the researcher also modified the classification with some additional heuristics, which are:

- The number of times the tag has been used is an indicator of its agreed meaning; therefore, lower tags occurrence indicates personal use.
- Compound tags and vague abbreviations are considered personal, since no one knows what do they mean, or why they were formed in this shape. And their tag occurrence is at its minimum.
- Misspelled tags are not counted in the classifications.

Thus, a sample of 100 randomly selected bookmarked web resources[141] with a total of 72,458 posts and 5250 unused tags were manually inspected and classified based on the previous assumptions.

Figure 10.12 shows the overall distribution of the three categories: 71% were personal tags, 21% were factual tags and 8% were subjective tags.

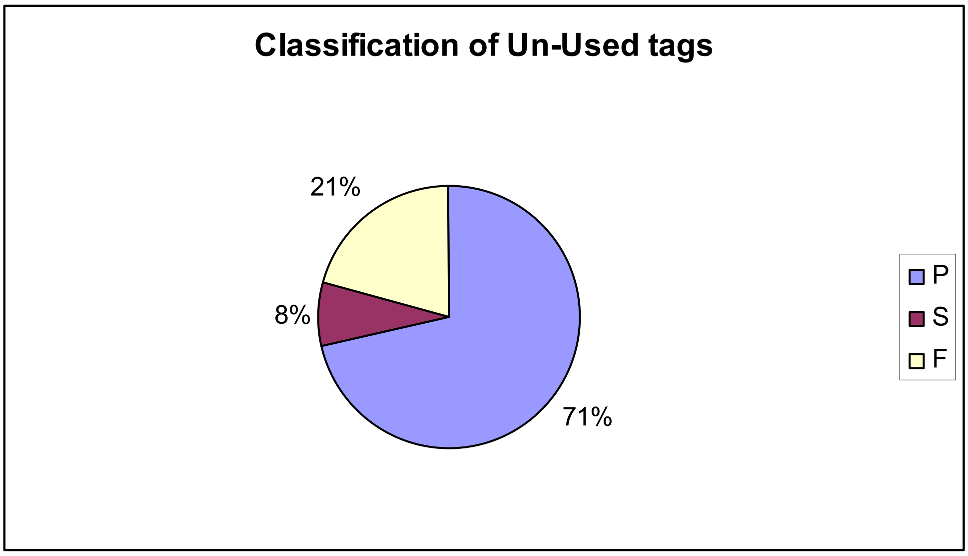


Figure 10.12: Classification of unused tags

As has been reported in chapter 4, tags can be either abbreviations, acronyms, complete words (singular/plural variations) or compound words. The manual inspection of the unused tags stressed the previously mentioned classification and helped in revealing more concise pattern in people's compound tags, despite the dynamic nature of the compound tags. The devised general pattern can be formulated as follows:

$$[P^*,S^*,F^*]_{3!}$$

Where P(ersonal), S(ubjective) and F(actual) each of which can be either an abbreviation, acronym or complete word (singular and/or plural variations). (3!) indicates the possible number of permutations between the three categories (hence 6 possible orders) and (*) indicates that there are zero or more possible occurrences of the category.

[141] Between 11-November and 20-December 2006

To illustrate this pattern some actual examples from the thesis data set are presented in the following form ([*pattern*], *example*) as shown in Table 10.27:

Pattern	**Example**
[F,F,F]	bugscrossbrowser
[F,P]	cssotherproject
[F,F]	csslib
[F,S]	csswelldone
[S,F,F]	nicesiteportal
[S,F]	goodexample
[P,P/F,P]	personaltoolbarfolder
[P,F,S]	nagtyexerciseevaluation

Table 10.27: Examples of patterns in people tags

Spelling errors constituted around 6% of the entire unused tags. By examining the types of errors that people have generated, the researcher found that some users inserted an extra character by mistake when typing the tag e.g. (termplate), or switched the places of characters e.g. (hmtl), or even missed a character e.g. (tutoria). These misspelled tags are usually used only by the person who created them and they do not gain much attention from other del.icio.us users.

Next, a detailed analysis of the three main categories is carried out.

(P)ersonal tags

These are tags that have an intended audience. They are often used to organise a user's own resources, and can be roughly classified into: self-reference tags, task and time management tags and others (Kipp, 2006).

Self-reference tags classification includes any tag that has to do with the users' own interest. Examples of these include dates e.g. (January, monthly and night), names e.g (tojack) and own reference e.g. (mylink, mysite and myblog). These tags usually appear among all tags for a given bookmarked web resource.

On the other hand, the most frequent task and time management tags were 'howto', 'tip' 'toread', 'work', 'todo' and their varieties such as 'readlater', 'todescribe', ‘tostudy’, etc. These tags tend to function as reminders and to-do lists to manage someone’s own future activities.

Other foreign tags were also spotted. These tags use English characters in their scripture; thus they can not be removed in the normalisation pipeline. Such examples include the frequently used Spanish tag ‘herramienta’ which means ‘tool’ and the Portuguese tag ‘Artigo’ which means ‘Article’[142].

Also there are some occurrences of prepositions in the tag list such as for, with, and, one and in. These preposition tags might be inserted unintentionally by a user who is thinking that the del.icio.us service deals with sentences/phrases as whole tags. A quick examination of the tag list that contains these prepositions justifies the researcher assumption.

Pedagogical tags have also appeared in users’ posts, among these tags are ‘learn(ing)’ and ‘Teach(ing)’. These tags can be used to emphasis the pedagogical nature of a web resource.

Finally, compound tags with minimum tag occurrence are considered personal tags, since no other del.icio.us user has used them. These tags constitute around 35% of all personal tags. By the same token, abbreviations are considered personal tags since no one knows what is their intended meaning other than the person who created them. These tags constitute around 6% of all personal tags.

(S)ubjective tags

These are tags that express people’s opinions on the bookmarked web resource. Although these tags constitute a small portion of the unused tags (i.e. 8%), an in depth inspection was carried out to analyse them.

Two classifications were observed: either the subjective tags were compound or informal. The compound tags consisted of a subjective qualifier with either a factual

[142] Google translation service was used to translate foreign tags.

or personal tag, e.g. beautifulsite, goodfor. Informal subjective tags include words that are produced by the user's own vocabulary such as fuckie, Kool or kickass.

(F)actual tags

These are tags which identify 'facts' about the described web resource such as people, places, or concepts. A more specific rough classification can be: web resource title/URL/author, synonyms (either near or far), rights/language, compound tags, generic, acronyms, spelling variation and other areas of application or usage.

Usually del.icio.us users use the title, the author or words that appear in the URL to bookmark a web resource. This pattern might be used because it is easy to remember a bookmark category by its title, author or URL. To give an example, most users who bookmarked articles from the A-List-Apart (alistapart.com) website have used tags such as: ala, alistapart and zeldman (a popular author in the website).

Another notable category was the use of synonym tags (with their two types near and far). Near synonyms mean that the average person can predict/use the tag (casual vocabulary), and far synonyms mean that only the elite user can predict/use the tag. As an example, the tag 'library' is an instance in the resource type ontology, which means 'a collection of things'. On the one hand, 'Database' and 'collection' are two near tags that can be used as synonyms by the average del.icio.us user; this is evident when more than one user uses these tags. On the other hand, the tag 'Grid' is considered a far synonym because the average user can not predict it as a straightforward synonym; this was evident from the number of users who used this tag; it was used only by one user. The problem of synonyms can be potentially solved by using thesauri such as WordNet[143].

The rights tags are used to indicate the privilege to use a web resource, e.g. (free, opensource, freedom, etc.) These tags constitute useful information to populate the rights element in a typical metadata record. By the same token, the language tags indicate the language of a web resource. The language tag comes in different forms (complete words, abbreviations, spelling error, etc.), e.g. English, langen, en, inglish,

---

[143] http://wordnet.princeton.edu/ [last accessed 9/2/2007]

etc. Also this type of tag can be useful to populate the language element in a typical metadata record.

Compound tags made up a good share of all tags in the factual category. These are different from the compound tags mentioned in the personal category, in that more than one del.icio.us user has used it. The use of compound tags might be attributed to people who are trying to preserve the maximum amount of facts about a web resource in one tag. It also shows that people are mixing generic tags with more specific ones to qualify them, e.g. 'cssarticle'.

Generic, acronyms and spelling variation tags, although there were not many of them, yet, they constitute a noise in the tag lists.

Sometimes factual tags refer to other potential areas of the application or usage of a web resource. For example in a web resource that talks about the 'shadow' technique, one tagger have used the tag 'dreamweaver', this might indicate that the person who assigned the tag was thinking of using this technique in Dreamweaver[144], and by inspecting the content of the web resource, it did not mention any thing about the software package or its usage in CSS.

Discussion

The researcher claims that it is the spelling errors, different patterns of compound tags and the use of non-English tags which made the unused tags outweigh the used ones. This finding is also verified by (Guy and Tonkin, 2006) who have mentioned the problem of folksonomies which include typographical errors and spelling variations, and this was also evident in the thesis inspected controlled domain.

From the in-depth analysis of del.icio.us users' tags, the researcher can envision that the normalisation pipeline did not process tags in a typical way. This deficiency of the normalisation process is due to the aggressive process conduct in each step in the pipeline. In the future work section a more detailed suggestion to overcome the normalisation process deficiency will be discussed.

[144] Dreamweaver is a software application used for creating websites

Another observation is that in the compound tags, users tend to mix and match different tags forms (plural and singular), e.g. inspirationscss (plural+singular), inspirationcss (singular+singular), or one of them containing spelling errors, e.g. tuotorialcss. These acts make it very difficult to build the best normalisation process. This raises the issue of the need for more language processing techniques, which will be discussed in the future work chapter.

The rights and language tags are two good sources of more information about the resource; however, their inconsistent appearance has made it difficult to capture them in a general form.

Another interesting finding is that, no matter how large is the normalised tag cloud, the number of assigned tags is limited by the total number of instances in the ontologies.

The unused tags uncover an important finding in this thesis which states: not all tags are useful for semantic annotation due to the variations in people's vocabulary and their background knowledge, also the normalisation process has played an important role in reducing the amount of noise in people's tags. However, from the thesis sample roughly 50% of all normalised tags can be used in the process of semantic metadata generation. Finally, it is worth mentioning that the process of analysing the unused tags depends solely on the subjective view of the researcher; however, the analysis appears indicative and reflects the meta-noise[145] present in people's tags.

10.4.2 Folksonomy vs. Automatic Keyword Extraction Assignment

One of the evaluation procedures the researcher has carried out using the FolksAnnotation tool was to compare the number of folksonomy tags assigned to the ontologies against the Yahoo TE keywords assignment for the same web resource.
For the purpose of this evaluation a set of 30 web resources wase randomly selected from the del.icio.us bookmarking service, and for each web resource two sets of

[145] Meta-noise refers to inaccurate or irrelevant metadata. from http://en.wikipedia.org/wiki/Meta_noise [last accessed 9/2/2007]

keywords (namely, folksonomy tags and Yahoo TE keywords) were prepared to be passed through the semantic annotation pipeline.

The results of this experiment showed that the number of assigned keywords from the folksonomy set is much higher than the Yahoo TE set with mean and standard deviation of 14.17(8.25) and 4.24(2.47), respectively. The difference between the means was statistically significant at $p < 0.001$. Thus, the results demonstrated that folksonomy tags are more useful in generating semantic metadata than automatically extracted context-based keywords.

10.4.3 Discussion

While running the previous experiment the researcher has noticed the following observations:

- The outstanding performance of Yahoo TE was evident in some cases where the number of people bookmarking a web resource was not that high, i.e. less than 50, however, once a bookmarked web resource gets momentum, Yahoo TE cannot beat people's collective intelligence, this observation was also stated by (McFedries, 2006).
- Another interesting observation is that, even if Yahoo TE extracts more information from web resources tagged by a small number of people, it was evident that people tags' tend to get a better ontology assignment than Yahoo TE keywords.
- There were situations where the keywords extracted by Yahoo TE are misleading; i.e. Yahoo TE keywords were ambiguous and got wrongly assigned to the ontology.
- Some of people's tags gave the main gist of a web resource and/or describe a tacit knowledge that does not appear explicitly in the described web resource. These folksonomy tags added potential contextual dimensions to a web resource and unveil an important difference between human vs. machine knowledge extraction.
- The Yahoo TE service clogs when a web resource does not contain text. Thus, most non-textual web resources (e.g. image, flash, java applets, etc.) when passed to Yahoo TE will not produce 'meaningful' keywords. However, people dive into a web resource, despite its non-textual nature, and

go beyond its content, i.e. dig into the web resource source file, to extract more meaningful aspects of the resource.

10.4.4 Niche Tags and The Long Tail

The Long Tail (Figure 10.13), as defined in Wikipedia[146]:

> "… *was first coined by Chris Anderson [October 2004] in Wired magazine article to describe certain business and economic models such as Amazon.com or Netflix. The long tail is the colloquial name for a long-known feature of statistical distributions … In these distributions a high-frequency or high-amplitude population is followed by a low-frequency or low-amplitude population which gradually 'tails off.'* "

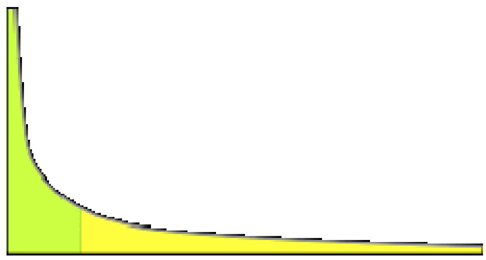


Figure 10.13: The long tail, colored in yellow [Wikipedia, 2007]

Wu et al. (2006) also gave a concise definition of the Long Tail which "… *describes the mass of users who search for documents using a variety of low-frequency keywords that would have been underserved by controlled vocabularies*." A final definition was given by Grimes and Torres (2006) states that the "*Long Tail is a theory that in a statistical distribution the accumulated minority can be more important than the simple majority*."

In Hypothsis#2, the researcher claimed that "***Fine-grained metadata values come from The Long Tail***", thus, some folksonomy tags (niche-tags) from the CSS ontology create a fine-grained index for a web resource. To verify this claim the

146 http://en.wikipedia.org/wiki/The_Long_Tail [last accessed 10/2/2007]

researcher has analysed the list of tags used to semantically annotate web resources in the experiment data set.

One observation the researcher has found when compiling the list of tags used to create the semantic metadata was: the distribution of all tags that are used to semantically annotate a web resource have always formed a long tail shape, as shown in Figure 10.14. Notice that the tags: 'list' (1 time), 'menu' (2 times), 'button' (9 times) and 'rollover' (10 times), are niche tags from the CSS ontology and at the same time fall in 'The Long Tail' region.

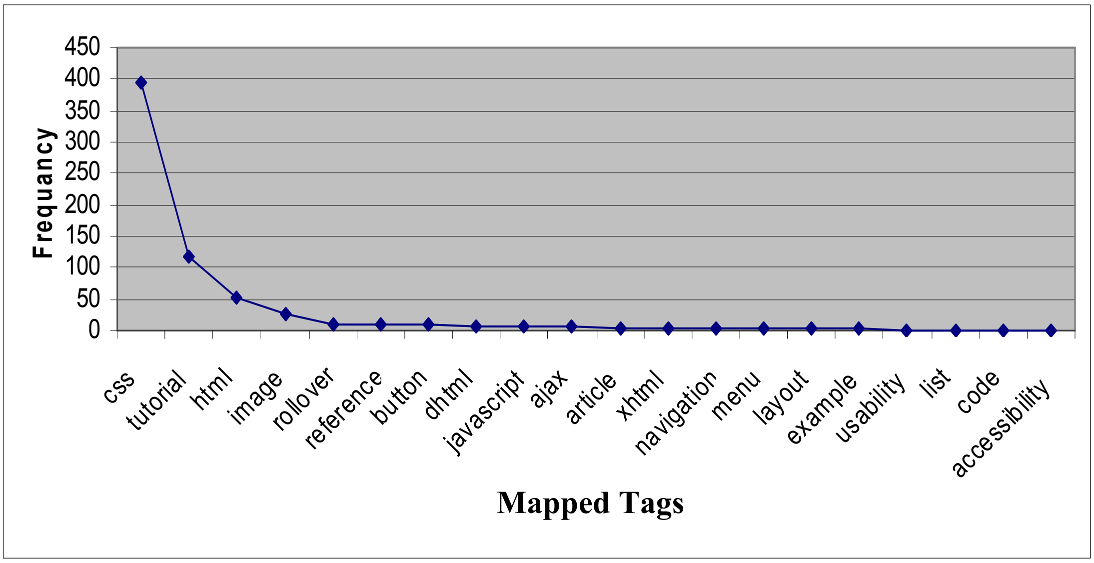


Figure 10.14: The Long Tail shape for the mapped tags used to semantically annotate the "*What Are CSS Sprites? > A Quick Example: Button Rollovers*"[147] web resource

The researcher has examined the graph for each tag list to determine the tags that fall in 'The Long Tail' portion, and found that from the 100 annotated web resources 80% have one or more niche-tags.

The average value of niche-tags for all web resources was 16% with a standard deviation of 11.77%. This implies that on average 16% of the mapped tags in the tag list for each web resource will be a niche-tag. This finding verifies the researcher's claim about the source of the fine-grained metadata values.

[147] http://www.informit.com/articles/article.asp?p=447210&rl=1 [last accessed 28/2/2007]

10.5 Evaluation Summary

After going through the various evaluations stages, a summarisation of the results is presented as follows:

- The searchability of the generated semantic metadata was demonstrated using different techniques. Each technique, be it browsing or semantic searching, adds flexible ways to reach, retrieve and search for annotated learning resources.
- The search for folksonomy semantic metadata preformed better than the search by folksonomy tags alone.
- The quality and validity of the generated semantic metadata was acceptable.
- The representativeness of the semantic metadata descriptors was generally useful.
- Folksonomy tags hold both formal and informal terms; however the majority of tags represent informal terms.
- Folksonomy tags are more useful in generating semantic metadata than context-based keywords.
- Fine grained metadata descriptors come from The Long Tail.

10.6 Chapter Summary

In this chapter, the researcher has presented a comprehensive framework to evaluate the usefulness, the quality, the searchability and the representativeness of the generated semantic metadata.

For each stage of the evaluation the researcher has shown a detailed analysis of the outcomes and accompanied each stage with a thorough discussion. The researcher has used a blend of qualitative and quantitative methods to measure the quality of the generated folksonomic semantic metadata; be it designing a questionnaire to collect experts' opinion on the generated folksonomic metadata or evaluating the metadata retrieval performance against human expert metadata using IR measurement.

Finally, the researcher has found that not all normalised tags can be used in the process of semantic metadata generation and also fine grained metadata values in the thesis case study came from The Long Tail.

Chapter 11
Related Work

Automatic metadata generation is a well-known research area in the fields of library and information science. This thesis work extends these research areas by exploiting other features from multiple research domains including: the Semantic Web, eLearning and Web technologies.

Basically, automatic metadata generation techniques can be categorized into two types: techniques to generate Standard Metadata and techniques to generate Semantic Metadata. The researcher, however, will add a third technique that deals with the generation of Folksonomic Metadata. Therefore, in this chapter, a discussion about the three genres and how they relate to the thesis current work will be highlighted.

11.1 Standard Metadata Techniques

Most metadata assignment techniques follow one of two approaches: extraction or classification (Paynter, 2005). In the extraction approach, metadata is extracted from documents using techniques such as natural language processing. This approach is appropriate for uncontrolled metadata fields such as title, description and creator. On the other hand, the classification approach relies on controlled vocabulary to assign metadata to documents.

There are many metadata extraction and generation tools used in an educational context, among them is the DC-dot program[148] that extracts Dublin core from the author's META tags in an HTML document. Another application is the Automatic Metadata Generation framework (AMG) developed by Erik Duval and his team (Cardinaels et al., 2005). This prototype framework operates by extracting metadata from the content in two learning management systems namely: SIDWeb and Toledo-Blackboard. Despite the fact that the extracted metadata is compliant with IEEE-LOM it only provides values to a limited number of elements in the LOM. However, for those fields where metadata can be automatically generated, the accuracy of the data is assured, when compared to a manual assignment.

The previously mentioned systems generate LOM metadata, primarily for human consumption. The scope of their research is different from the thesis main research theme and so this thesis system produces its metadata in a format suitable for machine processing by semantic web tools.

11.2 Semantic Metadata Techniques

In a developing field such as the Semantic Web, it is impossible to complete a comprehensive survey of the new tools and new versions of existing tools that are used to generate semantic metadata, due to the rapid progress in this area. However, most of these tools are created to generate general purpose semantic metadata, without taking into consideration the requirements of the educational field e.g. (Reeve and Han, 2005; Uren et al., 2005).

Luckily, some examples from the semantic field do exist that adhere to the requirements of the educational domain. The most recent example is an ongoing project carried out in the laboratories of Advanced Research in Intelligent Educational Systems (ARIES), Canada (Brooks and McCalla, 2006). This project is replacing the standard metadata (i.e. IEEE-LOM) with more flexible ecological approach based on semantics. The approach sees metadata as the process of reasoning over observed interactions of users with a learning object for a particular purpose.

[148] http://www.ukoln.ac.uk/metadata/dcdot/ [last accessed 6/2/2007]

Another example is the TANGRAM system (Jovanovic et al., 2006). TANGRAM is a learning web application for the domain of Intelligent Information Systems (IIS) where users (students and teachers) can upload, describe, search or compose a new learning object using components in the system repository. The system provides a solution for automatic metadata generation of learning objects (LO) components. Thus, each generated semantic metadata item/object attached to a LO allows the TANGRAM system to assemble these objects into new LOs personalized to the users' goals, preferences and learning style.

Despite the similarity between the purpose and outcome of the TANGRAM system and the thesis system, the FolksAnnotation tool does not rely on any algorithms to generate or extract metadata from web resources (that are equivalent to LO). Moreover, the thesis system uses a freely accessible web service (i.e. del.icio.us) for generating semantic metadata, while the TANGRAM system operates from within a learning management system. Finally, my system adds an extra layer of semantics to existing human generated metadata (i.e. folksonomies) which opens the doors for a wide range of intelligent applications.

11.3 Folksonomic metadata techniques

To the best of the researcher's knowledge, utilising folksonomies in the process of creating semantic metadata for eLearning applications is not yet a well-researched area. However, there is a nascent prototypical tool called 'CommonFolks' (Bateman et al., 2006) that is being developed in the laboratories of Advanced Research in Intelligent Educational Systems (ARIES), Canada, to create ontological metadata (i.e. semantic metadata) from people's tags to annotate learning resources to be used in adaptive eLearning systems.

The system's goal is to employ collaborative tagging in order to make metadata creation fast, easy and machine consumable using the English language ontology (WordNet). The approach works by appending WordNet with tags in a 'is-a' relationship. The tool is still in its early stages and no evaluation results have yet been reported.

11.4 Discussion

From the previous overview it is apparent that the researcher's work differs from prior research in automatic metadata generation in number of ways. First, the thesis tool relies on keywords generated by people to create the semantic metadata; while this approach is novel the researcher has not encountered any research dealing with such an approach. Secondly, the thesis system has not used any kind of algorithms to extract keywords from web resources; thus, it uses people's tags to generate new values for the elements of the semantic metadata. Finally, the thesis tool generates semantic metadata rather than standard metadata.

Table 11.1 summarises the different tools discussed in this chapter by comparing them based on various characteristics.

The Tool	**Type of Generated Metadata**	**Metadata Format**	**Ontology used**	**Metadata compliance**	**Resource Type**
DC-dot	Standard	XML/RDF	N/A	Dublin core	Web page
AMG	Standard	XML	N/A	IEEE-LOM	Learning Object
ARIES Project	Semantic	RDF	Proprietary	N/A	Learning Object
TANGRAM	Semantic	RDF	ALOCoMCS[149] & ALOCoMCT[150]	IEEE-LOM Application profile	Learning Object (PowerPoint)
CommonFolks	Semantic	N/A	WordNet	N/A	Web resource
FolksAnnotation	Semantic	RDF	Proprietary	IEEE-LOM Application profile	Web resource

Table 11.1: A Summary of automatic metadata generation in the eLearning domain

[149] Abstract Learning Object Content Model Content Structure (ALOCoMCS) ontology

[150] Abstract Learning Object Content Model Content Type (ALOCoMCT) ontology

11.5 Chapter Summary

This chapter presented an overview of the different types of automatic metadata generation techniques. It discussed each tool, compared and contrasted their functionalities with the thesis tool. Finally, the chapter concludes by summarizing the main characteristics for each related work.

Chapter 12
Conclusion and Further Work

This thesis has presented a novel approach to produce semantic metadata using free and unstructured human created keywords. In this chapter the researcher reflects on the achieved objectives and findings. It then concludes with planned follow-up work and further research.

12.1 Research Justification

The successful results returned by the evaluation framework confirmed that folksonomy tags are capable of creating acceptable semantic metadata. It has also confirmed that the FolksAnnotation tool system architecture was an appropriate step in the right direction, in that it normalised the folksonomy tags before using them in the process of semantic annotation.

Overall, the experiments' results have justified the feasibility of the research endeavour to achieve the following objectives:

1. To show that folksonomies, as index keywords, hold more semantic value than keywords automatically extracted by machines. Thus, based on the early experiment conducted in this research, which compares folksonomy tags to context-based automatic keywords extraction, the results demonstrated that folksonomy tags hold more semantic value than context-based automatically created keywords. This conclusion was derived based on quantitative evaluation procedures, using overlap measures, and qualitative evaluation procedures, based on experts' opinions.

2. To generate semantic metadata that annotates web resources for educational purposes using folksonomy tags.
3. To demonstrate that searching by folksonomies mapped to ontologies yields more retrieved web resources than searching by folksonomies alone. This objective is significant and illustrates the power and benefits of using Semantic Web technologies in the eLearning domain.
4. To demonstrate that folksonomy annotations cover more contextual dimensions than a human expert does. In section 10.2.2 the results of comparing folksonomies annotation against subject-matter expert annotation showed that the folksonomy tags provided some of the semantic metadata fields with extra contextual dimensions that were not proposed by the human expert. This finding emphasised the power of collective intelligence.
5. To illustrate that fine-grained metadata elements come from the Long Tail. This objective was reached by analysing the folksonomy tags that were used in the semantic annotation process. Plotting a distribution graph of the tags used to annotate each web resource has formed a Long Tail shape, and by further examining the tags falling in the Long Tail portion, the researcher discovered that most fine-grained[151] tags reside in that portion.

12.2 Research Findings

Social bookmarking services have the potential to become a mining source for learning resources. In this research, the researcher has showed how she successfully managed to convert folksonomy tags into useful semantic metadata. She has used a comprehensive framework to evaluate and demonstrate the usefulness, the quality and the representativeness of the generated semantic metadata.

Based on the research justifications the researcher can summarise the research findings into six points:

1. Folksonomy tags demonstrated that they are a “good enough” source for creating semantic metadata. This might be attributed to the latent (implicit) semantics embedded in the tags used to describe web resources. The observed latent

[151] Tags that show the greatest differentiation and discrimination for a web resource.

semantics helped the researcher to build the appropriate ontologies that captured folksonomy semantics and converted folksonomy tags to semantic metadata.

2. Analysis of tagging behaviours showed that folksonomy tags include both formal metadata, such as CSS domain specific tags, and informal metadata such as self reference tags. However, the majority of folksonomy tags were from the informal group where they do not adhere to a formal ontology.
3. Although folksonomy tags are neither perfect nor complete, they have added potential contextual dimensions to the generated metadata, as has been demonstrated in the evaluation framework.
4. Semantic Web technologies, i.e. ontologies, have enriched the way that learning resources in a given domain can be retrieved. Also they provided a flexible mechanism to share the meaning of a given domain compared to standard metadata.
5. Folksonomy tags showed the power of aggregating people's intelligence which helped in producing meaningful metadata. This was done without requiring their consensus in choosing the tags.
6. Folksonomy tags are better than automatically generated keywords.

12.3 Further work

Even if the prototype tool presented in this thesis has achieved its intended goals, there are many potential extensions that can enhance the tool's performance and output. The extensions are divided into two parts, namely: tool enhancements and further evaluations.

12.3.1 Tool Enhancements

Although the tool has successfully met the thesis goals, further enhancements need to be carried out to improve its performance and output. These enhancements include: expanding the semantic metadata and ontologies, improving the normalisation pipeline and the semantic annotation pipeline.

12.3.2 Metadata Descriptors Expansion and Enhancement

The completeness of the semantic metadata is not guaranteed since the existence of the appropriate descriptors depends highly on tags provided by the people who have tagged the web resource. Even if this is considered a major shortcoming of the tool, however, the benefit of having cheap human generated metadata may mitigate this shortcoming.

From the evaluation phase, the researcher has found that some metadata descriptors such as the 'Description' field needs to have a major redesign. The researcher also has discovered that there are missing descriptors that might be considered important to enrich the generated semantic metadata; these descriptors are author, technical requirements, language and rights. One potential enhancement to the tool is to incorporate sophisticated data mining and information extraction techniques (as described in chapter 6) before processing the folksonomy tags to help in broadening the range of possible metadata values. The future enhancement of the tool will take some of these issues into consideration beside the other suggested improvements to the current prototype.

12.3.3 Ontologies Expansion

The researcher has found during the process of semantic annotation, that the CSS ontology appears to have only grasped the minimum requirements of the domain; however, various expansions can be added to the current ontologies to reflect the state-of-the-art development in the field of CSS and Web design. This observation was reached due to the fact that CSS domain does not stand alone. The blurred boundaries between the CSS domain and other domains and applications such as Blogging software, XHTML, etc. caused a decline in the number of the folksonomy tags used in the process of semantic annotation.

The same comments apply to the Web design domain. The momentum witnessed in Web 2.0 applications has pushed the Web design community to experience new methodologies and techniques and to develop new vocabularies for this domain. This implies continued update of the CSS and Web design ontologies to reflect the advancement in these domains.

12.3.4 Improving the Normalisation Pipeline

The normalisation pipeline is a focal process in the semantic annotation tool. While this process has helped in minimising the amount of noise in people's tags, the process has shown some shortcomings in its performance.

One major shortcoming was attributed to the aggressive nature of the normalisation process, which caused some tags to lose their value. To give an example, a lot of compound tags that use special characters to separate them (e.g. _,+,/) lost their meaning when these connectors are eliminated. Unfortunately, the normalisation pipeline did not take into consideration the process of splitting compound words that use special characters.

Another shortcoming is that tags lost their context when grouped, but having some tags near each other may reveal hidden information about the tag's meaning. This observation was witnessed when developing the TSD algorithm. This shortcoming does not have an obvious solution, since it requires the reengineering of the whole normalisation pipeline. However, a potential solution might include the introduction of a Natural Language Processing (NLP) pipeline to process tags in their entirety for each post.

To enhance the normalisation process the following improvements are suggested:

- **Normalisation processes modularisation**: To be able to produce normalised tags without losing much of their semantics, the normalisation processes need to be modularised such that each module handles a specific and unique task. Figure 12.1 shows the processes after modularisation. The process of converting to lower case and removing non-Roman Alphabets was split into two different processes with a decomposition process between them, as will be explained in the next point.
- **Tags Decomposition process**: This is a new process introduced to the normalisation pipeline and adopted from (Tonkin, 2006). Tonkin proposed a process for tags decomposition using a splitter that utilises a wordlist to identify candidate terms. This process generates a class of possible decompositions that are then dispatched to a part-of-speech or grammatical tagger to markup each possible term by its part of speech. Finally, a naïve

Bayesian classifier is used to classify the tokenised terms into two groups: known and unknown terms. This decomposition process will help in eliminating the number of compound terms, as has been witnessed in the thesis evaluation phase. It will also include a routine to split compound tags, which uses special characters to separate them.

- **Tags Sense Disambiguation (TSD) reengineering:** It was clear from chapter 8 that TSD module was just a fast and superficial solution to overcome the problem of tags ambiguity in the process of semantic annotation. However, the researcher believes that the issue of tags semantics needs more profound research in the field of concepts relations and semantic distances as proposed by (Zhang et al., 2006).

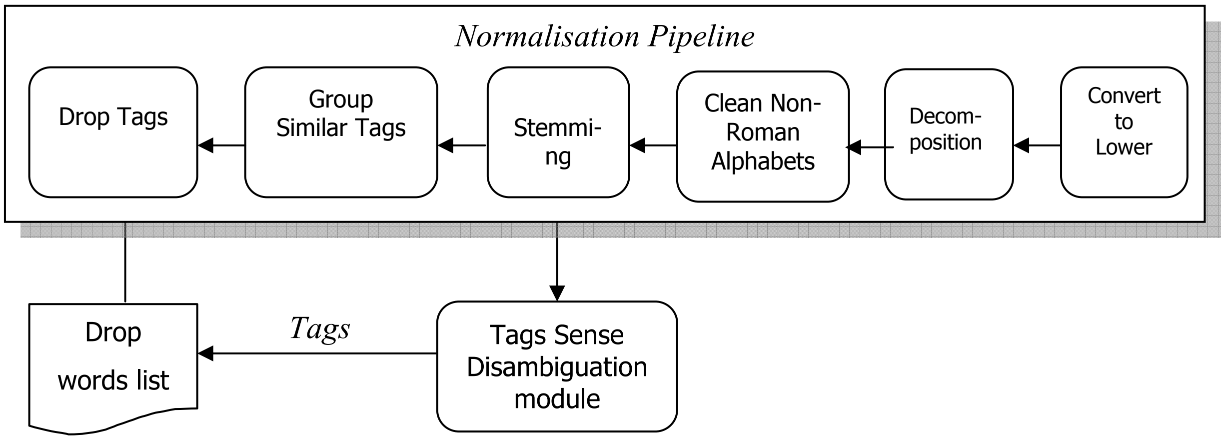


Figure 12.1: The enhanced normalisation pipeline

12.3.5 Improving the Semantic Annotation Pipeline

As a consequence of enhancing the normalisation process, the semantic annotation pipeline can be further enhanced by introducing a WordNet module to capture the semantics of the unused tags that remain after the main ontologies annotation. This proposed solution is questionable, however it provides a possibility to measure to what extent the WordNet thesauri will improve the process of semantic annotation.

Moreover, the semantic annotation process itself needs to be more flexible to accept a number of ontologies to perform the process of semantic annotation. In the current system architecture the invocation of ontologies is hardwired in the source code of the FolksAnnotation tool. However, when introducing a flexible module to plug-in any ontology in the tool, this will broaden the potential usage of it.

Figure 12.2 demonstrates the proposed potential improvements to the semantic annotation pipeline.

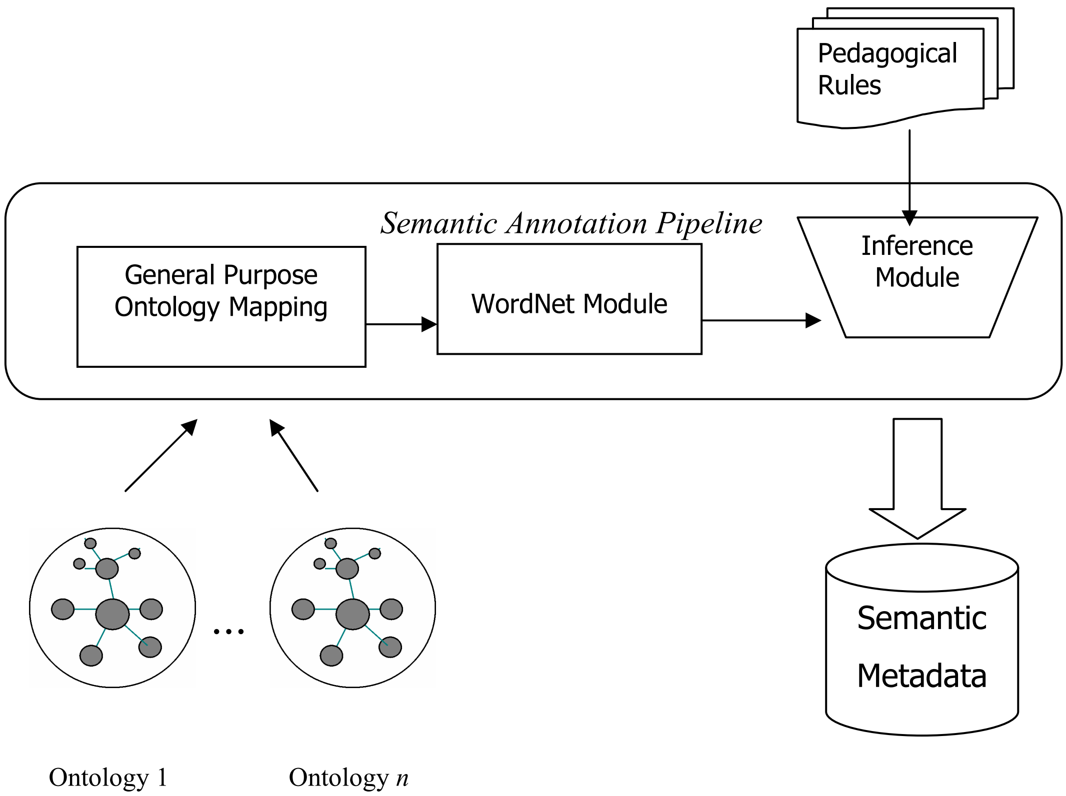


Figure 12.2: The enhanced semantic annotation pipeline

12.3.6 Further Evaluation Factors

The comprehensive evaluation framework performed in this thesis has only focused on the resultant semantic metadata and did not compare the performance of the system and the quality of the semantic annotation to other peer systems. Therefore, the researcher is planning to conduct a comparative study to compare the thesis tool performance against other automatic metadata generation tools such as (Cardinaels et al., 2005) and automatic semantic annotation tools such as (Cimiano et al., 2005).

In addition, an ontology assessment procedure is proposed so that the ontology content is assessed from several different points of view. These include different classes of users (i.e. teacher and student) and different applications need (e.g. introductory Web design).

Finally, the researcher could explore the possibility of using the tool with other social bookmarking services or even other web applications that use folksonomy tags as part of their architecture. This step is essential to derive a general conclusion for the usefulness of folksonomy tags in various and different contexts.

12.4 Future Research Directions

There are several future research directions the FolkAnnotation tool can contribute to; one possible route is to envision the potential applications of the tool. Among the possible applications is to integrate the tool with systems that provide personalised or adaptive content such as recommender systems or Adaptive Hypermedia Systems (AHS). Another possible application is to convert the tool into a Web Service to provide an interoperable and unified access method to its service. A discussion about both proposed applications follows.

12.4.1 Personalisation, Adaptation and Recommender Systems

Before discussing the idea of integrating the FolksAnnotation tool into personalised or recommender systems; a brief definition of the terms personalisation and adaptation are highlighted first. According to Merriam-Webster dictionary the word personalisation means: "*to make personal or individual; specifically: to mark as the property of a particular*[152]*"*. Adaptation means: "*adjustment to environmental conditions: as* ***a****: adjustment of a sense organ to the intensity or quality of stimulation* ***b****: modification of an organism or its parts that makes it more fit for existence under the conditions of its environment*[153]". Based on these definitions, the researcher can think of personalisation as tailoring the output of a system to adapt to the user preferences that are stored in a profile. For instance, based on the thesis case

[152] http://209.161.33.50/dictionary/personalization [last accessed 28/2/2007]

[153] http://209.161.33.50/dictionary/Adaptation [last accessed 28/2/2007]

study, users of an adaptive hypermedia system can receive links personalised to their difficulty level or instructional level from the output of the FolksAnnotation tool.

Adaptive hypermedia systems (AHS) offer "… *an alternative to traditional 'one-size-fits-all' hypermedia and Web systems by adapting to the goals, interests, and knowledge of individual users represented in the individual user models.*" (Chen and Magoulas., 2005).

Adaptation in AHS is based on three approaches: rules, algorithmic methods and hybrid rule-based/algorithmic methods for adaptation (Stefani et al., 2006). Rule based AHS usually use hard-coded rules to supply the user with the necessary adaptation. This was a major drawback in the rule based approach, which has been replaced or/and complemented with algorithmic methods. Algorithmic methods use machine learning techniques such as data mining for discovering user behaviour (Stefani et al., 2006).

Moreover, the three common types of adaptation in most adaptive hypermedia systems are: content, layout and navigation adaptations (Stefani et al., 2006). Content adaptation is a candidate to exploit the output of the FolksAnnotation tool.

Adaptive Educational Hypermedia Systems (AEHS) are a derivative type of AHS that inherit their major features. AEHS consist of the following components: document space, user model, observations and adaptation components (Henze and Nejdl, 2003). A complete discussion of these components is beyond the scope of this thesis; however, the researcher will briefly elaborate on the user model component and describe its role in utilising the output of the FolksAnnotation tool to suggest appropriate links to AEHS users.

The user model is responsible for storing, describing and inferring information, knowledge, preferences about an individual user (Henze and Nejdl, 2003). There are two approaches to creating a user model; one is to use collaborative filtering techniques, and the other is to use content-based filtering (Chen and Magoulas., 2005). Collaborative filtering techniques try to match an individual profile to similar profiles. On the other hand, content-based filtering operates by extracting features from content that the user liked or used in the past (Chen and Magoulas., 2005). The researcher believes that content-based filtering is a key approach to exploit the output of the FolksAnnotation tool.

To employ the FolksAnnotation tool in an AEHS that uses content-based filtering, the tool can be used to suggest learning resources that match individual preferences.

Recommender systems can be considered parallel to adaptive hypermedia systems, since both try to personalise the output to the user needs. However, recommender systems are designed in the first place for the domain of e-commerce; however, they can be applied to the eLearning domain after adjusting them to the needs of the domain, i.e. incorporating pedagogical theories. Discussing the similarities and differences between recommender systems and AHS is beyond the scope of this thesis; however the reader is referred to (Stefani et al., 2006).

A recommender system is a program that collects the behaviour of users of a particular system to find trends and make recommendations based on their profiles (Adomavicius and Tuzhilin, 2005). The recommendation algorithm is built using different mathematical and statistical models to derive future guesses of people's interest. There are different recommendation strategies such as user-to-user correlation, item-to-item correlation, item-to-user correlation, etc. (Parsons et al., 2004). However, a complete discussion of the algorithms used by recommender systems is beyond the scope of this thesis.

Recommender systems are beneficially used as alternative to search algorithms, where users are recommended resources they might have not discovered by themselves. Examples of recommendation systems on the Web include: Amazon.com, Last.fm, eBay, to name but a few.

By integrating the FolksAnnotation tool as a backend database for providing potential learning resources in an Adaptive eLearning system that utilises a recommendation engine e.g. (Kristofic, 2005; Lemire et al., 2005), the recommender engine can operate by using the descriptors provided by the semantic metadata (e.g. resource type, difficulty level, etc.) to match a user's profile preferences with the appropriate learning resources. This technique can be considered a simplified version of the item-to-user correlation recommendation strategy.

12.4.2 Web Services

W3C (2004) defines a Web Service as "…*a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.*"

Web services are the approach of choice to support cross-services co-operation between different agents and services. Imagine the following scenario;

> Sue is teaching a course on Cascading Style Sheets (CSS) as part of the web development course in her institute. In her daily quest for finding suitable learning resources to support her curriculum, she uses the del.icio.us bookmarking service to hunt for resources instead of spending her time Googling.
>
> Sue believes that del.icio.us contains links to massive amounts of useful materials that can be used in an educational context, and will be of great help to her.
>
> There is no semantic metadata in del.icio.us to describe the educational purpose of these materials, but for Sue this lack of metadata is not a major problem, because she can use her agent to consult the appropriate Web service to generate this missing information. So, she runs her agent, which is provided with Sue's preferences, to communicate with the FolksAnnotation Web service, which works as an interface to the del.icio.us bookmarking service, to convert peoples' tags into more structured and meaningful metadata records.

By utilising the communication between Sue's agent and the Web service, she removes the hurdle of visiting the designated bookmarked website or even going through all the tags that people have generated to know what the site is about. Moreover, she can use the returned links, harvested by her agents from del.icio.us, in her course database portal.

12.5 Conclusion

This thesis presents a tool called FolksAnnotation for creating semantic metadata from folksonomy tags for the use in an educational context. The novelty of this work resides in the integration of the Semantic Web technologies with the free form metadata generated by people, to produce structured metadata with explicit semantics.

The FolksAnnotation tool applies an organisational schema to people's tags in a specific domain of interest (i.e. teaching CSS). Thus, the folksonomy tags in this thesis are modelled not as text keywords but as RDF resources that comply with pre-defined ontologies. This technique provides two benefits:

Benefit 1: While the folksonomy approach retrieves documents using 'bags of words', the property-value pairs approach enables more advanced search, such as question answering; reasoning as well as document retrieval. Thus, the thesis approach provides a property-value relationship that is semantically rich and allows for more 'intelligent' search that is not provided by mere folksonomy search.

Benefit 2: Typical semantic annotation tools (as discussed in chapter 6) depend on an intermediate process called Information Extraction (IE) to extract the main concepts from a document before relating them to the designated ontologies. The IE process is a very complex phase in the semantic annotation lifecycle, and encompasses many advanced techniques from the natural language processing domain. Moreover, the processing time required to accomplish the IE task is significant. Therefore, instead of using the IE process as an intermediate phase for extracting knowledge from documents, why not rely on peoples' generated metadata? Thus, by using folksonomies as knowledge artefacts in the process of semantic annotation, the researcher believes that she has used a potential cheap source of metadata generated by people's collective intelligence.

The evaluation of the FolksAnnotation tool has required a phase to test the value of folksonomy tags compared to keywords extracted using context-based keyword extraction technique (chapter 7). The different experiments carried out in that evaluation phase demonstrated the superior value of the folksonomy tags compared to context-based keywords. Moreover, the subsequent evaluation phases conducted in the comprehensive evolution framework (chapter 10) revealed the advantages of the folksonomy tags over other metadata generation techniques.

To conclude, the field of folksonomy research is still nascent and this thesis represents a first step towards more profound research in the educational domain. The researcher has contributed to the endeavour of automatically generating semantic metadata using folksonomy tags. Finally, this research has achieved its intended goals; however, there are still many open research questions that need to be solved, such as:

- What if folksonomy tags span multiple domains?
- Can the tool be used with other social bookmarking services?
- Can we devise a general approach to use the prototype tool with different tagging systems beyond the social bookmarking domain?

Such questions are hoped to shape future research directions for anyone interested in the field of generating semantic metadata from folksonomy tags for educational purposes.

Appendix A. Metadata Questionnaire

Audience and Purpose

The questionnaire was then distributed to two target populations: Web designers and experts in the field of learning technologies and metadata (called the specialists group in the subsequent discussion). The Web designers' community was reached using mailing lists that resides at Yahoo Groups and other focused groups such as css-discuss. The specialists group was reached using focused mailing list such as CETIS Metadata and Digital Repository Special Interest Group and the Learning Societies Lab (LSL) mailing list at the University of Southampton. The total response from the Web designers group was 29 respondents and the total response from the specialists group was 22 respondents.

The purpose of this questionnaire was to measure the usefulness (i.e. validity and appropriateness) of the Cascading Style Sheets (CSS) metadata elements for the purpose of describing learning resource in the domain of CSS.

Metadata Descriptions Instructions:

Metadata are defined as data about data, and in the educational field metadata are used to describe learning resources so they can be easily searched, retrieved and shared. The following questionnaire consists of three parts: ***part 1***, asks about your background in the field of CSS. ***Part 2***, lists all the metadata elements used to describe CSS web resources. The following lists the elements of the CSS metadata record and their intended use; and, for a full listing of CSS properties, units and attributes please look at CSS cheat-sheet.

1. **Title**: the title of the web resource.
2. **URL** (Uniform Resource Locator): a web resource URL.
3. **Description**: a short sentence describing the web resource in terms of proposed background knowledge and possible educational use.

4. **Keywords**: an essential or definitive word that can be used for indexing a web resource, for later search and retrieval.
5. **Difficulty level:** describes how difficult is a web resource (e.g. Easy, Medium Difficulty or Difficult) – notice that this value is set subjectively and can be changed accordingly.
6. **Instructional level**: describes the level of instruction on which a web resource can be used (e.g. Novice, Intermediate or Advanced) – notice that this value is set subjectively and can be changed accordingly.
7. **Resource type**: describes the instructional type of a web resource (e.g. tutorial, code, example, etc.)
8. **Recommendation**: is a number from 1 to 5 that is used to rate the popularity of a web resource.
9. **Property**: defines the properties of CSS (e.g. background, font, colour etc.) that are being used in the described web resource.
10. **Element i.e. Attribute**: defines the attributes of CSS (e.g. class and id) that are being used in the described web resource.
11. **Selector**: defines the CSS selectors (e.g. div and span) that are being used in the described web resource.
12. **Units**: defines the CSS units (e.g. em, pt, px, bolder, lighter or larger) that are being used in the described web resource.
13. **Application**: defines the possible function of a web resource in the domain of CSS (e.g. menu, sitemap, form etc.).
14. **Layout**: describes what layout technique a web resource is promoting (e.g. tableless, fluid, column, etc.).
15. **Technique**: describes what technique a web resource is promoting (e.g. Rollover, hover, image replacement, fade, etc.).
16. **Subject**: gives the main theme of a web resource in the CSS domain. Possible Subject values include {box model, layout, navigation, positioning, effect, and typography}.

Finally ***Part 3***, shows you an example of three generated metadata records and asks you to evaluate their usefulness.

Part 1: About You

Before you start the questionnaire, can you tell me please what your professional role is?

1. Professor/Researcher
2. Web designer/programmer
3. Cataloger/Metadata librarian
4. Postgraduate/Undergraduate student
5. Others; specify ________________________

How would you rate your self in the following areas, on a scale from 1 to 5 where 1 indicates novice and 5 indicates expert?

	1	2	3	4	5
Web programming in general					
Knowledge of web design in general					
Knowledge of Cascading Style Sheets (CSS)					
Experience in teaching or tutoring a web design course					

Part 2: Metadata Descriptors Evaluation

1. On a scale from 1 to 5 (1 represents 'useless' and 5 represents 'very useful'), please evaluate how useful are the following metadata elements to the learner/instructor for **describing** a web resource in the domain of teaching CSS:

General Metadata Elements	**1**	**2**	**3**	**4**	**5**
Title					
URL					
Description					
Keywords					
Difficulty level					
Instructional level					
Resource type					
Recommendation					

2. Please evaluate how useful are the following metadata descriptors for the use by a learner/instructor to **retrieve/search** a CSS knowledge base:

CSS Specific Metadata Descriptors	**1**	**2**	**3**	**4**	**5**
Property					
Element (i.e. Attribute)					
Selector					
Units					
Application					
Layout					
Technique					
Subject					
Resource type					
Difficulty level					
Instructional level					

3. Do you think there are other metadata descriptors/elements (general or specific to CSS) that might be useful for the retrieval of web resources from a CSS knowledge base?

Part 3: Annotation Evaluation

For the following three annotated web resources please rate the quality and validity of the provided metadata elements; hence, you need to visit the designated web resource to make a fair judgment. Notice that the evolution is done based on a three-tier scale adopted from Greenberg, J. (2004)[154], which are:

- **Good**. Good metadata accurately represented the resource and would facilitate accurate resource discovery. A good metadata element does not require any revision.
- **Fair**. Fair metadata would be somewhat useful for resource discovery of the resource being represented. In this case, a revision(s) would generally improve the quality of the metadata element.
- **Reject**. Reject (poor quality) metadata was inaccurate. In this case, the metadata element required substantial revision to be useful for resource discovery.

[154] Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging, 6(4): [28 manuscript pages.]

Resource 1:

Title	Daniel Mall: Well Educated CSS
URL:	http://www.danielmall.com/archives/2006/05/17/well_educated_css .php
Description	A(n) reference with title: 'Daniel Mall: Well Educated CSS' suggest the knowledge of the following topics: css, xhtml, ajax, javascript,. This resource is also suggested to be used as: tool, article, resource, tutorial, code, template.
Keywords	css, xhtml, ajax, javascript, reference, tool, article, resource, tutorial, code, template, id, class, layout
Resource type	Reference, tool, article, resource, tutorial, code, template.
Element	Id, class
Subject	Layout

⇒ Please **tick** the appropriate box to evaluate each metadata element for the previously given web resource.

Metadata Element	**Good**	**Fair**	**Reject**
Title			
Description			
Keywords			
Resource type			
Element			
Subject			

Resource 2:

Title	Cornershop - Rounded Graphics for CSS Box Corners
URL	http://wigflip.com/cornershop/
Description	A(n) generator with title: 'Cornershop - Rounded Graphics for CSS Box Corners' suggest the knowledge of the following topics: xhtml,css,html,dhtml,. This resource is also suggested to be used as: tutorial, template.
Keywords	xhtml, css, html, dhtml, background, background-image, border, layout, box, rounded corner, tabbed menu, template, tutorial, generator
Resource type	Template, tutorial, generator
Property	Background, border
Subject	Layout
Layout	Box
Technquie	Rounded corner
Application	Menu, tab

⇒ Please **tick** the appropriate box to evaluate each metadata element for the previously given web resource.

Metadata Element	**Good**	**Fair**	**Reject**
Title			
Description			
Keywords			
Resource type			
Property			
Subject			
Layout			
Technquie			
Application			

Resource 3:

Title	A Cool CSS Effect - Dashboard [Updated] » Dustin Bachrach Blog
URL	http://dbachrach.com/blog/2006/10/a-cool-css-effect-dashboard/
Description	A(n) tutorial with title: 'A Cool CSS Effect - Dashboard [Updated] Â» Dustin Bachrach Blog' suggest the knowledge of the following topics: css, javascript, ajax, html, usability, xhtml, dhtml,. This resource is also suggested to be used as: code, reference, resource, guide.
Keywords	css, javascript, ajax, html, usability, xhtml, dhtml, code, reference, resource, guide, background, effect, hover, fade, transparent, dim, overlay, popup, div.
Resource type	Tutorial, code, reference, resource, guide.
Property	Background, opacity
Subject	Effect
Technquie	Hover, fade, transparent, dim, overlay
Application	Popup
Selector	div

⇒ Please **tick** the appropriate box to evaluate each metadata element for the previously given web resource.

Metadata Element	**Good**	**Fair**	**Reject**
Title			
Description			
Keywords			
Resource type			
Property			
Subject			
Technquie			
Application			
Selector			

Appendix B. CSS Ontology

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:p1="http://www.owl-ontologies.com/assert.owl#"
  xml:base="http://www.ecs.soton.ac.uk/~hsak04r/css_ontology">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="Boxmodel">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Property"/>
    </rdfs:subClassOf>
    <owl:disjointWith>
      <owl:Class rdf:ID="Classification"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Printing"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Text"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Font"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Cursor"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Positioning"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Color"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Background"/>
    </owl:disjointWith>
  </owl:Class>
  <owl:Class rdf:ID="Position_Type">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Positioning"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Layout">
```

```xml
  <owl:disjointWith>
    <owl:Class rdf:ID="Subject"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Technique"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Property"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Application"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Attribute"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="CSS"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:ID="Selector"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Technique">
  <owl:disjointWith>
    <owl:Class rdf:about="#Subject"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Attribute"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Layout"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#CSS"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:cardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
      >1</owl:cardinality>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="hasTechnique"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Application"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Selector"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Property"/>
  </owl:disjointWith>
```

```
</owl:Class>
<owl:Class rdf:about="#Printing">
  <owl:disjointWith>
    <owl:Class rdf:about="#Text"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Background"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cursor"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Font"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Classification"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Boxmodel"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Property"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Positioning"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Font">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Property"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Positioning"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Classification"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Text"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Printing"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cursor"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Boxmodel"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Background"/>
```

```
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Property">
  <owl:disjointWith>
    <owl:Class rdf:about="#Attribute"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Subject"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#CSS"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Application"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Layout"/>
  <owl:disjointWith rdf:resource="#Technique"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Selector"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:ID="Clear">
  <rdfs:subClassOf rdf:resource="#Property"/>
</owl:Class>
<owl:Class rdf:about="#Positioning">
  <owl:disjointWith rdf:resource="#Printing"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cursor"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Boxmodel"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Classification"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Background"/>
  </owl:disjointWith>
  <rdfs:subClassOf rdf:resource="#Property"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Text"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Font"/>
</owl:Class>
<owl:Class rdf:about="#Classification">
  <owl:disjointWith>
    <owl:Class rdf:about="#Background"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Positioning"/>
  <owl:disjointWith>
```

```
      <owl:Class rdf:about="#Text"/>
    </owl:disjointWith>
    <owl:disjointWith rdf:resource="#Boxmodel"/>
    <owl:disjointWith>
      <owl:Class rdf:about="#Cursor"/>
    </owl:disjointWith>
    <owl:disjointWith rdf:resource="#Printing"/>
    <owl:disjointWith rdf:resource="#Font"/>
    <rdfs:subClassOf rdf:resource="#Property"/>
    <owl:disjointWith>
      <owl:Class rdf:about="#Color"/>
    </owl:disjointWith>
  </owl:Class>
  <owl:Class rdf:about="#Application">
    <owl:disjointWith>
      <owl:Class rdf:about="#Selector"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:about="#Subject"/>
    </owl:disjointWith>
    <owl:disjointWith rdf:resource="#Technique"/>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#CSS"/>
    </rdfs:subClassOf>
    <owl:disjointWith>
      <owl:Class rdf:about="#Attribute"/>
    </owl:disjointWith>
    <owl:disjointWith rdf:resource="#Layout"/>
    <owl:disjointWith rdf:resource="#Property"/>
  </owl:Class>
  <owl:Class rdf:about="#Attribute">
    <owl:disjointWith rdf:resource="#Layout"/>
    <owl:disjointWith rdf:resource="#Technique"/>
    <owl:disjointWith rdf:resource="#Application"/>
    <owl:disjointWith>
      <owl:Class rdf:about="#Subject"/>
    </owl:disjointWith>
    <owl:disjointWith rdf:resource="#Property"/>
    <owl:disjointWith>
      <owl:Class rdf:about="#Selector"/>
    </owl:disjointWith>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#CSS"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#Background">
    <owl:disjointWith rdf:resource="#Printing"/>
    <owl:disjointWith rdf:resource="#Font"/>
    <owl:disjointWith>
      <owl:Class rdf:about="#Text"/>
    </owl:disjointWith>
```

```
  <rdfs:subClassOf rdf:resource="#Property"/>
  <owl:disjointWith rdf:resource="#Boxmodel"/>
  <owl:disjointWith rdf:resource="#Classification"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Positioning"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cursor"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Text">
  <owl:disjointWith rdf:resource="#Background"/>
  <owl:disjointWith rdf:resource="#Positioning"/>
  <owl:disjointWith rdf:resource="#Printing"/>
  <owl:disjointWith rdf:resource="#Classification"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Boxmodel"/>
  <rdfs:subClassOf rdf:resource="#Property"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cursor"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Font"/>
</owl:Class>
<owl:Class rdf:ID="BackgroundProperty">
  <rdfs:subClassOf rdf:resource="#Background"/>
</owl:Class>
<owl:Class rdf:about="#Selector">
  <owl:disjointWith rdf:resource="#Property"/>
  <owl:disjointWith rdf:resource="#Layout"/>
  <owl:disjointWith rdf:resource="#Technique"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Subject"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Attribute"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#CSS"/>
  </rdfs:subClassOf>
  <owl:disjointWith rdf:resource="#Application"/>
</owl:Class>
<owl:Class rdf:about="#Cursor">
  <owl:disjointWith rdf:resource="#Positioning"/>
  <rdfs:subClassOf rdf:resource="#Property"/>
  <owl:disjointWith rdf:resource="#Background"/>
  <owl:disjointWith rdf:resource="#Printing"/>
  <owl:disjointWith rdf:resource="#Classification"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Color"/>
  </owl:disjointWith>
```

```
    <owl:disjointWith rdf:resource="#Text"/>
    <owl:disjointWith rdf:resource="#Boxmodel"/>
    <owl:disjointWith rdf:resource="#Font"/>
  </owl:Class>
  <owl:Class rdf:about="#Subject">
    <owl:disjointWith rdf:resource="#Layout"/>
    <owl:disjointWith rdf:resource="#Selector"/>
    <owl:disjointWith rdf:resource="#Technique"/>
    <owl:disjointWith rdf:resource="#Property"/>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#CSS"/>
    </rdfs:subClassOf>
    <owl:disjointWith rdf:resource="#Attribute"/>
    <owl:disjointWith rdf:resource="#Application"/>
  </owl:Class>
  <owl:Class rdf:ID="Float">
    <rdfs:subClassOf rdf:resource="#Property"/>
  </owl:Class>
  <owl:Class rdf:ID="Unit">
    <rdfs:subClassOf>
      <owl:Class rdf:about="#CSS"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="#Color">
    <owl:disjointWith rdf:resource="#Printing"/>
    <owl:disjointWith rdf:resource="#Text"/>
    <owl:disjointWith rdf:resource="#Boxmodel"/>
    <rdfs:subClassOf rdf:resource="#Property"/>
    <owl:disjointWith rdf:resource="#Classification"/>
    <owl:disjointWith rdf:resource="#Cursor"/>
    <owl:disjointWith rdf:resource="#Font"/>
    <owl:disjointWith rdf:resource="#Background"/>
    <owl:disjointWith rdf:resource="#Positioning"/>
  </owl:Class>
  <owl:Class rdf:about="#CSS">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >css</rdfs:label>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >This is the Class of the Scripting Language Cascading Style
Sheet</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:ID="Type">
    <rdfs:subClassOf rdf:resource="#Application"/>
  </owl:Class>
  <owl:ObjectProperty rdf:ID="hasPositionType">
    <rdfs:domain rdf:resource="#Positioning"/>
    <rdfs:range rdf:resource="#Position_Type"/>
    <owl:inverseOf>
      <owl:ObjectProperty rdf:ID="isPositionTypeOf"/>
    </owl:inverseOf>
  </owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:about="#isPositionTypeOf">
  <rdfs:domain rdf:resource="#Position_Type"/>
  <rdfs:range rdf:resource="#Positioning"/>
  <owl:inverseOf rdf:resource="#hasPositionType"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="used_with">
  <rdfs:range rdf:resource="#CSS"/>
  <rdfs:domain rdf:resource="#CSS"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasElement">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Attribute"/>
        <owl:Class rdf:about="#Selector"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasSubProperty">
  <owl:inverseOf>
    <owl:ObjectProperty rdf:ID="isSubPropertyOf"/>
  </owl:inverseOf>
  <rdfs:range rdf:resource="#Property"/>
  <rdfs:domain rdf:resource="#Property"/>
  <rdfs:subPropertyOf>
    <owl:ObjectProperty rdf:ID="hasProperty"/>
  </rdfs:subPropertyOf>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="related_to">
  <rdfs:domain>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Subject"/>
        <owl:Class rdf:about="#Technique"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:domain>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Application"/>
        <owl:Class rdf:about="#Property"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="isPropertyOf">
  <rdfs:range rdf:resource="#CSS"/>
  <rdfs:domain rdf:resource="#Property"/>
```

```
  <owl:inverseOf>
    <owl:ObjectProperty rdf:about="#hasProperty"/>
  </owl:inverseOf>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasUnit">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="#Unit"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#hasTechnique">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="#Technique"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#hasProperty">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="#Property"/>
  <owl:inverseOf rdf:resource="#isPropertyOf"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasApplication">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="#Application"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasApplicationType">
  <rdfs:range rdf:resource="#Type"/>
  <rdfs:domain rdf:resource="#Application"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="#isSubPropertyOf">
  <rdfs:domain rdf:resource="#Property"/>
  <rdfs:subPropertyOf rdf:resource="#isPropertyOf"/>
  <rdfs:range rdf:resource="#Property"/>
  <owl:inverseOf rdf:resource="#hasSubProperty"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasLayout">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="#Layout"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasSubject">
  <rdfs:range rdf:resource="#Subject"/>
  <rdfs:domain rdf:resource="#CSS"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="uses">
  <rdfs:range rdf:resource="#CSS"/>
  <rdfs:domain rdf:resource="#CSS"/>
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="hasInstructionalLevel">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range>
    <owl:DataRange>
      <owl:oneOf rdf:parseType="Resource">
        <rdf:rest rdf:parseType="Resource">
          <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >intermediate</rdf:first>
```

```xml
          <rdf:rest rdf:parseType="Resource">
            <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
            >novice</rdf:first>
            <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
          </rdf:rest>
        </rdf:rest>
        <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >advanced</rdf:first>
      </owl:oneOf>
    </owl:DataRange>
  </rdfs:range>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="hasRecommendation">
  <rdfs:domain rdf:resource="#CSS"/>
  <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#float"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="hasTechnquieFor">
  <rdfs:range>
    <owl:DataRange>
      <owl:oneOf rdf:parseType="Resource">
        <rdf:rest rdf:parseType="Resource">
          <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >text</rdf:first>
          <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
        </rdf:rest>
        <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >image</rdf:first>
      </owl:oneOf>
    </owl:DataRange>
  </rdfs:range>
  <rdfs:domain rdf:resource="#Technique"/>
</owl:DatatypeProperty>
<Type rdf:ID="dropdown"/>
<Unit rdf:ID="bolder"/>
<Application rdf:ID="imagemap"/>
<Technique rdf:ID="hover"/>
<Unit rdf:ID="lighter"/>
<Classification rdf:ID="list">
  <isPropertyOf>
    <CSS rdf:ID="css">
      <hasLayout>
        <Layout rdf:ID="multicolumn"/>
      </hasLayout>
      <hasProperty>
        <Boxmodel rdf:ID="margin">
          <isPropertyOf rdf:resource="#css"/>
        </Boxmodel>
      </hasProperty>
      <hasLayout>
        <Layout rdf:ID="table"/>
      </hasLayout>
```

```
<hasUnit>
 <Unit rdf:ID="pt"/>
</hasUnit>
<hasTechnique>
 <Technique rdf:ID="overlay"/>
</hasTechnique>
<hasLayout>
 <Layout rdf:ID="tableless">
  <uses>
   <Selector rdf:ID="div"/>
  </uses>
 </Layout>
</hasLayout>
<hasLayout>
 <Layout rdf:ID="box"/>
</hasLayout>
<hasProperty>
 <Classification rdf:ID="clip">
  <isPropertyOf rdf:resource="#css"/>
 </Classification>
</hasProperty>
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>css</rdfs:label>
<hasSubject>
 <Subject rdf:ID="effect"/>
</hasSubject>
<hasUnit>
 <Unit rdf:ID="pixel"/>
</hasUnit>
<hasProperty>
 <Text rdf:ID="textshadow">
  <isPropertyOf rdf:resource="#css"/>
 </Text>
</hasProperty>
<hasTechnique>
 <Technique rdf:ID="imagereplacement"/>
</hasTechnique>
<hasApplication>
 <Application rdf:ID="menu">
  <uses>
   <Classification rdf:ID="li">
    <isPropertyOf rdf:resource="#css"/>
   </Classification>
  </uses>
  <hasApplicationType rdf:resource="#dropdown"/>
  <hasApplicationType>
   <Type rdf:ID="popup"/>
  </hasApplicationType>
  <hasApplicationType>
   <Type rdf:ID="vertical"/>
  </hasApplicationType>
```

```
    <hasApplicationType>
      <Type rdf:ID="tab"/>
    </hasApplicationType>
    <hasApplicationType>
      <Type rdf:ID="horizontal"/>
    </hasApplicationType>
    <uses rdf:resource="#list"/>
  </Application>
</hasApplication>
<hasProperty>
  <Boxmodel rdf:ID="padding">
    <isPropertyOf rdf:resource="#css"/>
  </Boxmodel>
</hasProperty>
<rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
>CaseCading Style Sheets</rdfs:comment>
<hasApplication>
  <Application rdf:ID="chart">
    <owl:sameAs>
      <Application rdf:ID="graph">
        <hasApplicationType>
          <Type rdf:ID="bar"/>
        </hasApplicationType>
      </Application>
    </owl:sameAs>
  </Application>
</hasApplication>
<hasSubject>
  <Subject rdf:ID="navigation">
    <related_to>
      <Application rdf:ID="sitemap"/>
    </related_to>
    <related_to rdf:resource="#menu"/>
  </Subject>
</hasSubject>
<hasApplication rdf:resource="#graph"/>
<hasProperty>
  <Positioning rdf:ID="zindex">
    <isPropertyOf rdf:resource="#css"/>
  </Positioning>
</hasProperty>
<hasProperty>
  <Property rdf:ID="opacity">
    <isPropertyOf rdf:resource="#css"/>
  </Property>
</hasProperty>
<hasSubject>
  <Subject rdf:ID="positioning">
    <related_to>
      <Position_Type rdf:ID="relative">
        <isPositionTypeOf>
```

```
          <Positioning rdf:ID="position">
            <hasPositionType>
              <Position_Type rdf:ID="static">
                <isPositionTypeOf rdf:resource="#position"/>
              </Position_Type>
            </hasPositionType>
            <hasPositionType rdf:resource="#relative"/>
            <isPropertyOf rdf:resource="#css"/>
            <hasPositionType>
              <Position_Type rdf:ID="absolute">
                <isPositionTypeOf rdf:resource="#position"/>
              </Position_Type>
            </hasPositionType>
          </Positioning>
        </isPositionTypeOf>
      </Position_Type>
    </related_to>
    <related_to rdf:resource="#static"/>
    <related_to rdf:resource="#position"/>
    <related_to>
      <Positioning rdf:ID="display">
        <isPropertyOf rdf:resource="#css"/>
      </Positioning>
    </related_to>
    <related_to rdf:resource="#absolute"/>
    <related_to>
      <Positioning rdf:ID="visibility">
        <isPropertyOf rdf:resource="#css"/>
      </Positioning>
    </related_to>
    <related_to rdf:resource="#zindex"/>
  </Subject>
</hasSubject>
<hasTechnique rdf:resource="#hover"/>
<hasTechnique>
  <Technique rdf:ID="grouping"/>
</hasTechnique>
<hasTechnique>
  <Technique rdf:ID="roundedcorner">
    <used_with>
      <Application rdf:ID="button"/>
    </used_with>
    <used_with rdf:resource="#menu"/>
  </Technique>
</hasTechnique>
<hasSubject>
  <Subject rdf:ID="typography">
    <related_to>
      <Font rdf:ID="font">
        <isPropertyOf rdf:resource="#css"/>
      </Font>
```

```
      </related_to>
      <related_to>
        <Text rdf:ID="text">
          <isPropertyOf rdf:resource="#css"/>
        </Text>
      </related_to>
    </Subject>
  </hasSubject>
  <hasTechnique>
    <Technique rdf:ID="replacement">
      <hasTechnquieFor
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >image</hasTechnquieFor>
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >image or text replacment</rdfs:comment>
      <hasTechnquieFor
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >text</hasTechnquieFor>
    </Technique>
  </hasTechnique>
  <hasTechnique>
    <Technique rdf:ID="preload">
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >image preload</rdfs:comment>
      <hasTechnquieFor
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >image</hasTechnquieFor>
    </Technique>
  </hasTechnique>
  <hasSubject>
    <Subject rdf:ID="layout"/>
  </hasSubject>
  <hasLayout>
    <Layout rdf:ID="centering"/>
  </hasLayout>
  <hasLayout>
    <Layout rdf:ID="column"/>
  </hasLayout>
  <hasUnit>
    <Unit rdf:ID="px"/>
  </hasUnit>
  <hasLayout>
    <Layout rdf:ID="liquid"/>
  </hasLayout>
  <hasApplication rdf:resource="#sitemap"/>
  <hasApplication>
    <Application rdf:ID="banner"/>
  </hasApplication>
  <hasProperty>
    <Background rdf:ID="background">
      <isPropertyOf rdf:resource="#css"/>
```

```
    <hasSubProperty>
      <BackgroundProperty rdf:ID="image">
        <isSubPropertyOf rdf:resource="#background"/>
      </BackgroundProperty>
    </hasSubProperty>
  </Background>
</hasProperty>
<hasElement rdf:resource="#div"/>
<hasLayout>
  <Layout rdf:ID="center"/>
</hasLayout>
<hasUnit rdf:resource="#lighter"/>
<hasTechnique>
  <Technique rdf:ID="curve"/>
</hasTechnique>
<hasProperty rdf:resource="#font"/>
<hasUnit>
  <Unit rdf:ID="larger"/>
</hasUnit>
<hasProperty>
  <Boxmodel rdf:ID="border">
    <isPropertyOf rdf:resource="#css"/>
  </Boxmodel>
</hasProperty>
<hasElement>
  <Attribute rdf:ID="class"/>
</hasElement>
<hasTechnique>
  <Technique rdf:ID="shadow"/>
</hasTechnique>
<hasSubject>
  <Subject rdf:ID="boxmodel"/>
</hasSubject>
<hasProperty rdf:resource="#list"/>
<hasApplication>
  <Application rdf:ID="rating"/>
</hasApplication>
<hasProperty rdf:resource="#text"/>
<hasLayout>
  <Layout rdf:ID="fixed"/>
</hasLayout>
<hasProperty>
  <Float rdf:ID="float">
    <isPropertyOf rdf:resource="#css"/>
  </Float>
</hasProperty>
<hasElement>
  <Attribute rdf:ID="id"/>
</hasElement>
<hasApplication>
  <Application rdf:ID="bargraph"/>
```

```
</hasApplication>
<hasProperty>
  <Color rdf:ID="color">
    <isPropertyOf rdf:resource="#css"/>
  </Color>
</hasProperty>
<hasApplication>
  <Application rdf:ID="form"/>
</hasApplication>
<hasTechnique>
  <Technique rdf:ID="slidingdoor">
    <related_to rdf:resource="#menu"/>
  </Technique>
</hasTechnique>
<hasTechnique>
  <Technique rdf:ID="transparent"/>
</hasTechnique>
<hasSubject>
  <Subject rdf:ID="compatibility"/>
</hasSubject>
<hasProperty rdf:resource="#li"/>
<hasTechnique>
  <Technique rdf:ID="transparency"/>
</hasTechnique>
<hasTechnique>
  <Technique rdf:ID="dim"/>
</hasTechnique>
<hasLayout>
  <Layout rdf:ID="fluid"/>
</hasLayout>
<hasElement>
  <Selector rdf:ID="span"/>
</hasElement>
<hasProperty>
  <Font rdf:ID="fontsize">
    <isPropertyOf rdf:resource="#css"/>
  </Font>
</hasProperty>
<hasTechnique>
  <Technique rdf:ID="rollover"/>
</hasTechnique>
<hasSubject>
  <Subject rdf:ID="print"/>
</hasSubject>
<hasUnit rdf:resource="#bolder"/>
<hasUnit>
  <Unit rdf:ID="em"/>
</hasUnit>
<hasProperty>
  <Cursor rdf:ID="cursor">
    <isPropertyOf rdf:resource="#css"/>
```

```
        </Cursor>
      </hasProperty>
      <hasProperty>
        <Classification rdf:ID="ul">
          <owl:sameAs rdf:resource="#list"/>
          <isPropertyOf rdf:resource="#css"/>
        </Classification>
      </hasProperty>
      <hasProperty rdf:resource="#display"/>
      <hasApplication rdf:resource="#imagemap"/>
      <hasTechnique>
        <Technique rdf:ID="fade"/>
      </hasTechnique>
      <hasProperty rdf:resource="#visibility"/>
      <hasProperty>
        <Clear rdf:ID="clear">
          <isPropertyOf rdf:resource="#css"/>
        </Clear>
      </hasProperty>
      <hasProperty rdf:resource="#position"/>
      <hasApplication rdf:resource="#button"/>
    </CSS>
  </isPropertyOf>
  <owl:sameAs rdf:resource="#li"/>
  <owl:sameAs rdf:resource="#ul"/>
</Classification>
<Background rdf:ID="backgroundimage"/>
<owl:DataRange>
  <owl:oneOf rdf:parseType="Resource">
    <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
    <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >easy</rdf:first>
  </owl:oneOf>
</owl:DataRange>
<owl:DataRange>
  <owl:oneOf rdf:parseType="Resource">
    <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >basic</rdf:first>
    <rdf:rest rdf:parseType="Resource">
      <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >advance</rdf:first>
      <rdf:rest rdf:parseType="Resource">
        <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >intermediate</rdf:first>
        <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
      </rdf:rest>
    </rdf:rest>
  </owl:oneOf>
</owl:DataRange>
<owl:DataRange>
  <owl:oneOf rdf:parseType="Resource">
```

```
      <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >easy</rdf:first>
      <rdf:rest rdf:parseType="Resource">
        <rdf:rest rdf:parseType="Resource">
          <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
          <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
          >medium</rdf:first>
        </rdf:rest>
        <rdf:first rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
        >hard</rdf:first>
      </rdf:rest>
    </owl:oneOf>
  </owl:DataRange>
  <Application rdf:ID="wordpress"/>
  <Type rdf:ID="flyout"/>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 2.2, Build 311) http://protege.stanford.edu -->
```

Appendix C. Web Design Ontology

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns="http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:p1="http://www.owl-ontologies.com/assert.owl#"
  xml:base="http://www.ecs.soton.ac.uk/~hsak04r/web_design.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="Style_Sheets">
    <owl:disjointWith>
      <owl:Class rdf:ID="Programming_Languages"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Markup_Languages"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Access_Methods"/>
    </owl:disjointWith>
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Authoring"/>
    </rdfs:subClassOf>
    <owl:disjointWith>
      <owl:Class rdf:ID="Document_Represenation"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Graphics"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Multimedia"/>
    </owl:disjointWith>
  </owl:Class>
  <owl:Class rdf:ID="Web_Design_and_Development">
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >The Web Design and Development ontology was created based on Yahoo
Directory listing and Dmoz.org directory.</rdfs:comment>
  </owl:Class>
  <owl:Class rdf:about="#Multimedia">
    <owl:disjointWith rdf:resource="#Style_Sheets"/>
    <owl:disjointWith>
```

```xml
    <owl:Class rdf:about="#Document_Represenation"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Access_Methods"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Graphics"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Markup_Languages"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Programming_Languages"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Authoring"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="#Graphics">
  <owl:disjointWith>
    <owl:Class rdf:about="#Markup_Languages"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Access_Methods"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Multimedia"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Programming_Languages"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Style_Sheets"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Authoring"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Document_Represenation"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:ID="Usability">
  <owl:disjointWith>
    <owl:Class rdf:about="#Authoring"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Accessibility"/>
  </owl:disjointWith>
  <rdfs:subClassOf rdf:resource="#Web_Design_and_Development"/>
</owl:Class>
<owl:Class rdf:ID="Scripting_Languages">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Programming_Languages"/>
  </rdfs:subClassOf>
</owl:Class>
```

```xml
<owl:Class rdf:about="#Document_Represenation">
  <owl:disjointWith>
    <owl:Class rdf:about="#Programming_Languages"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Authoring"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Markup_Languages"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Multimedia"/>
  <owl:disjointWith rdf:resource="#Graphics"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Access_Methods"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Style_Sheets"/>
</owl:Class>
<owl:Class rdf:about="#Access_Methods">
  <owl:disjointWith>
    <owl:Class rdf:about="#Programming_Languages"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Markup_Languages"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Authoring"/>
  </rdfs:subClassOf>
  <owl:disjointWith rdf:resource="#Graphics"/>
  <owl:disjointWith rdf:resource="#Multimedia"/>
  <owl:disjointWith rdf:resource="#Document_Represenation"/>
  <owl:disjointWith rdf:resource="#Style_Sheets"/>
</owl:Class>
<owl:Class rdf:about="#Authoring">
  <rdfs:subClassOf rdf:resource="#Web_Design_and_Development"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Accessibility"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Usability"/>
</owl:Class>
<owl:Class rdf:about="#Markup_Languages">
  <owl:disjointWith rdf:resource="#Graphics"/>
  <owl:disjointWith rdf:resource="#Multimedia"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Programming_Languages"/>
  </owl:disjointWith>
  <rdfs:subClassOf rdf:resource="#Authoring"/>
  <owl:disjointWith rdf:resource="#Document_Represenation"/>
  <owl:disjointWith rdf:resource="#Style_Sheets"/>
  <owl:disjointWith rdf:resource="#Access_Methods"/>
</owl:Class>
<owl:Class rdf:about="#Programming_Languages">
```

```
  <owl:disjointWith rdf:resource="#Markup_Languages"/>
  <owl:disjointWith rdf:resource="#Style_Sheets"/>
  <rdfs:subClassOf rdf:resource="#Authoring"/>
  <owl:disjointWith rdf:resource="#Graphics"/>
  <owl:disjointWith rdf:resource="#Document_Represenation"/>
  <owl:disjointWith rdf:resource="#Access_Methods"/>
  <owl:disjointWith rdf:resource="#Multimedia"/>
</owl:Class>
<owl:Class rdf:about="#Accessibility">
  <owl:disjointWith rdf:resource="#Usability"/>
  <owl:disjointWith rdf:resource="#Authoring"/>
  <rdfs:subClassOf rdf:resource="#Web_Design_and_Development"/>
</owl:Class>
<owl:ObjectProperty rdf:ID="consistOf">
  <rdfs:range rdf:resource="#Web_Design_and_Development"/>
  <rdfs:domain rdf:resource="#Web_Design_and_Development"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="represented_in">
  <rdfs:domain rdf:resource="#Markup_Languages"/>
  <rdfs:range rdf:resource="#Document_Represenation"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="uses">
  <rdfs:domain rdf:resource="#Access_Methods"/>
  <rdfs:range>
    <owl:Class>
      <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#Style_Sheets"/>
        <owl:Class rdf:about="#Markup_Languages"/>
        <owl:Class rdf:about="#Programming_Languages"/>
        <owl:Class rdf:about="#Document_Represenation"/>
      </owl:unionOf>
    </owl:Class>
  </rdfs:range>
</owl:ObjectProperty>
<Style_Sheets rdf:ID="jsss">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >JavaScript Style Sheets</rdfs:comment>
</Style_Sheets>
<Scripting_Languages rdf:ID="js">
  <owl:sameAs>
    <Scripting_Languages rdf:ID="javascript"/>
  </owl:sameAs>
</Scripting_Languages>
<Usability rdf:ID="usability"/>
<Accessibility rdf:ID="accessibility"/>
<Document_Represenation rdf:ID="sax"/>
<Access_Methods rdf:ID="ajax">
  <uses>
    <Style_Sheets rdf:ID="css">
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >Cascading Style Sheets</rdfs:comment>
```

```
    </Style_Sheets>
  </uses>
  <uses rdf:resource="#javascript"/>
  <uses>
    <Markup_Languages rdf:ID="xhtml">
      <represented_in>
        <Document_Represenation rdf:ID="dom"/>
      </represented_in>
    </Markup_Languages>
  </uses>
  <uses>
    <Markup_Languages rdf:ID="xml">
      <represented_in rdf:resource="#sax"/>
      <represented_in rdf:resource="#dom"/>
    </Markup_Languages>
  </uses>
  <uses>
    <Markup_Languages rdf:ID="html">
      <represented_in rdf:resource="#dom"/>
    </Markup_Languages>
  </uses>
</Access_Methods>
<Access_Methods rdf:ID="dhtml">
  <uses rdf:resource="#dom"/>
  <uses rdf:resource="#javascript"/>
  <uses rdf:resource="#html"/>
  <uses rdf:resource="#xhtml"/>
  <uses rdf:resource="#css"/>
  <uses rdf:resource="#xml"/>
</Access_Methods>
<Authoring rdf:ID="authoring">
  <consistOf rdf:resource="#html"/>
  <consistOf rdf:resource="#javascript"/>
  <consistOf rdf:resource="#ajax"/>
  <consistOf>
    <Style_Sheets rdf:ID="xsl">
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
      >Extensible Stylesheet Language</rdfs:comment>
    </Style_Sheets>
  </consistOf>
  <consistOf rdf:resource="#js"/>
  <consistOf rdf:resource="#dom"/>
  <consistOf rdf:resource="#jsss"/>
  <consistOf rdf:resource="#xml"/>
  <consistOf rdf:resource="#sax"/>
  <consistOf rdf:resource="#xhtml"/>
  <consistOf rdf:resource="#css"/>
  <consistOf rdf:resource="#dhtml"/>
</Authoring>
<Web_Design_and_Development rdf:ID="web_design_and_development">
  <consistOf rdf:resource="#accessibility"/>
```

```
    <consistOf rdf:resource="#usability"/>
    <consistOf rdf:resource="#authoring"/>
  </Web_Design_and_Development>
  <Graphics rdf:ID="photoshop"/>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 2.2, Build 311) http://protege.stanford.edu -->
```

Appendix D. Resource Type Ontology

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:p1="http://www.owl-ontologies.com/assert.owl#"
  xml:base="http://www.ecs.soton.ac.uk/~hsak04r/resource_type.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="Instructional">
    <owl:disjointWith>
      <owl:Class rdf:ID="Technical"/>
    </owl:disjointWith>
    <rdfs:subClassOf>
      <owl:Class rdf:ID="Resource"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Showcase">
    <owl:disjointWith>
      <owl:Class rdf:ID="Handbook"/>
    </owl:disjointWith>
    <rdfs:subClassOf>
      <owl:Class rdf:about="#Technical"/>
    </rdfs:subClassOf>
    <owl:disjointWith>
      <owl:Class rdf:ID="Cheatsheet"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Gallary"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Article"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Guide"/>
    </owl:disjointWith>
    <owl:disjointWith>
      <owl:Class rdf:ID="Website"/>
    </owl:disjointWith>
    <owl:disjointWith>
```

```
    <owl:Class rdf:ID="Utility"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Template"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Code"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Tool"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Handbook">
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Code"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Website"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Guide"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Utility"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tool"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Guide">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tool"/>
```

```
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Website"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Code"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Utility"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Showcase"/>
</owl:Class>
<owl:Class rdf:ID="Reference">
  <rdfs:subClassOf rdf:resource="#Instructional"/>
  <owl:disjointWith>
    <owl:Class rdf:ID="Tutorial"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Example"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Demonstrations"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Sample"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="Demo"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Code">
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <owl:disjointWith>
```

```
    <owl:Class rdf:about="#Utility"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tool"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Website"/>
  </owl:disjointWith>
</owl:Class>
<owl:Class rdf:about="#Utility">
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Website"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tool"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Showcase"/>
</owl:Class>
<owl:Class rdf:about="#Tool">
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith>
```

```
    <owl:Class rdf:about="#Website"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Utility"/>
</owl:Class>
<owl:Class rdf:about="#Demo">
  <owl:disjointWith>
    <owl:Class rdf:about="#Sample"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Reference"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Demonstrations"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tutorial"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Example"/>
  </owl:disjointWith>
  <rdfs:subClassOf rdf:resource="#Instructional"/>
</owl:Class>
<owl:Class rdf:about="#Example">
  <rdfs:subClassOf rdf:resource="#Instructional"/>
  <owl:disjointWith rdf:resource="#Reference"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Demonstrations"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Sample"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Demo"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tutorial"/>
  </owl:disjointWith>
```

```
</owl:Class>
<owl:Class rdf:ID="Howto">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Article"/>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Tutorial"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="#Website">
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith rdf:resource="#Utility"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Cheatsheet"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Tool"/>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith rdf:resource="#Code"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="#Cheatsheet">
  <owl:disjointWith>
    <owl:Class rdf:about="#Gallary"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith rdf:resource="#Website"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith rdf:resource="#Utility"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
```

```
  <owl:disjointWith rdf:resource="#Tool"/>
</owl:Class>
<owl:Class rdf:about="#Sample">
  <owl:disjointWith rdf:resource="#Demo"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Tutorial"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:about="#Demonstrations"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Example"/>
  <owl:disjointWith rdf:resource="#Reference"/>
  <rdfs:subClassOf rdf:resource="#Instructional"/>
  <owl:sameAs>
    <Example rdf:ID="example"/>
  </owl:sameAs>
</owl:Class>
<owl:Class rdf:about="#Gallary">
  <owl:disjointWith>
    <owl:Class rdf:about="#Article"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith rdf:resource="#Tool"/>
  <owl:disjointWith rdf:resource="#Utility"/>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith rdf:resource="#Website"/>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith rdf:resource="#Cheatsheet"/>
</owl:Class>
<owl:Class rdf:about="#Demonstrations">
  <owl:disjointWith>
    <owl:Class rdf:about="#Tutorial"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Reference"/>
  <owl:disjointWith rdf:resource="#Example"/>
  <rdfs:subClassOf rdf:resource="#Instructional"/>
  <owl:disjointWith rdf:resource="#Demo"/>
  <owl:disjointWith rdf:resource="#Sample"/>
</owl:Class>
<owl:Class rdf:about="#Tutorial">
  <owl:disjointWith rdf:resource="#Demo"/>
  <owl:disjointWith rdf:resource="#Demonstrations"/>
  <owl:disjointWith rdf:resource="#Sample"/>
  <owl:disjointWith rdf:resource="#Example"/>
```

```
  <rdfs:subClassOf rdf:resource="#Instructional"/>
  <owl:disjointWith rdf:resource="#Reference"/>
</owl:Class>
<owl:Class rdf:about="#Article">
  <owl:disjointWith rdf:resource="#Tool"/>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith rdf:resource="#Utility"/>
  <owl:disjointWith>
    <owl:Class rdf:about="#Template"/>
  </owl:disjointWith>
  <owl:disjointWith rdf:resource="#Cheatsheet"/>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith rdf:resource="#Website"/>
  <owl:disjointWith rdf:resource="#Gallary"/>
  <owl:disjointWith rdf:resource="#Guide"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:about="#Template">
  <owl:disjointWith rdf:resource="#Tool"/>
  <owl:disjointWith rdf:resource="#Showcase"/>
  <owl:disjointWith rdf:resource="#Utility"/>
  <owl:disjointWith rdf:resource="#Guide"/>
  <owl:disjointWith rdf:resource="#Handbook"/>
  <owl:disjointWith rdf:resource="#Code"/>
  <owl:disjointWith rdf:resource="#Website"/>
  <owl:disjointWith rdf:resource="#Article"/>
  <rdfs:subClassOf>
    <owl:Class rdf:about="#Technical"/>
  </rdfs:subClassOf>
  <owl:disjointWith rdf:resource="#Cheatsheet"/>
  <owl:disjointWith rdf:resource="#Gallary"/>
</owl:Class>
<owl:Class rdf:about="#Technical">
  <rdfs:subClassOf rdf:resource="#Resource"/>
  <owl:disjointWith rdf:resource="#Instructional"/>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasResourceType">
  <rdfs:range rdf:resource="#Resource"/>
  <rdfs:domain rdf:resource="#Resource"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasSubResourceType">
  <rdfs:subPropertyOf rdf:resource="#hasResourceType"/>
</owl:ObjectProperty>
<Code rdf:ID="code"/>
<Website rdf:ID="website"/>
<Resource rdf:ID="resource">
  <hasResourceType>
    <Template rdf:ID="template"/>
```

```
</hasResourceType>
<hasResourceType>
  <Website rdf:ID="portal"/>
</hasResourceType>
<hasResourceType>
  <Utility rdf:ID="utility"/>
</hasResourceType>
<hasResourceType rdf:resource="#website"/>
<hasResourceType>
  <Sample rdf:ID="sample"/>
</hasResourceType>
<hasResourceType>
  <Tool rdf:ID="checker"/>
</hasResourceType>
<hasResourceType>
  <Showcase rdf:ID="showcase"/>
</hasResourceType>
<hasResourceType rdf:resource="#example"/>
<hasResourceType>
  <Cheatsheet rdf:ID="cheatsheet"/>
</hasResourceType>
<hasResourceType rdf:resource="#resource"/>
<hasResourceType>
  <Tool rdf:ID="tool"/>
</hasResourceType>
<hasResourceType>
  <Demonstrations rdf:ID="demonstration"/>
</hasResourceType>
<hasResourceType rdf:resource="#code"/>
<hasResourceType>
  <Gallary rdf:ID="gallery"/>
</hasResourceType>
<hasResourceType>
  <Tool rdf:ID="editor"/>
</hasResourceType>
<hasResourceType>
  <Reference rdf:ID="reference"/>
</hasResourceType>
<hasResourceType>
  <Tutorial rdf:ID="tutorial">
    <hasSubResourceType>
      <Howto rdf:ID="howto"/>
    </hasSubResourceType>
  </Tutorial>
</hasResourceType>
<hasResourceType>
  <Tool rdf:ID="generator"/>
</hasResourceType>
<hasResourceType>
  <Handbook rdf:ID="ebook"/>
</hasResourceType>
```

```
    <hasResourceType>
      <Handbook rdf:ID="handbook"/>
    </hasResourceType>
    <hasResourceType>
      <Tool rdf:ID="viewer"/>
    </hasResourceType>
    <hasResourceType>
      <Website rdf:ID="blog"/>
    </hasResourceType>
    <hasResourceType>
      <Tool rdf:ID="optimizer"/>
    </hasResourceType>
    <hasResourceType>
      <Guide rdf:ID="guide"/>
    </hasResourceType>
    <hasResourceType>
      <Tool rdf:ID="validator"/>
    </hasResourceType>
    <hasResourceType>
      <Demo rdf:ID="demo"/>
    </hasResourceType>
    <hasResourceType>
      <Technical rdf:ID="documentation"/>
    </hasResourceType>
    <hasResourceType>
      <Article rdf:ID="article">
        <hasSubResourceType rdf:resource="#howto"/>
      </Article>
    </hasResourceType>
  </Resource>
  <Utility rdf:ID="bookmarklet"/>
  <Tool rdf:ID="calculator"/>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 2.2, Build 311) http://protege.stanford.edu -->
```

References

Adomavicius, G. and A. Tuzhilin (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions." IEEE Transactions on Knowledge and Data Engineering **17**(6): pp. 734-749.

Al-Khalifa, H. S. and H. Davis (2006d). *A Tool for Semantic Web Annotation based on Folksonomies and Domain Ontologies*. KnowledgeWeb PhD Symposium 2006 (KWEPSY2006), in the 3ed European Semantic Web conference, Budva, Montenegro.

Al-Khalifa, H. S. and H. Davis (2006e). *Towards Semantic Metadata for Learning Resources Using Folksonomies and Domain Ontologies*. The 5th International Semantic Web Conference (ISWC'06) | Doctoral Consortium, Athens, USA.

Al-Khalifa, H. S. and H. Davis (2006f). *Semantic Metadata for Learning Resources Using Folksonomies and Domain Ontologies*. The First IEEE International Conference on Digital Information Management (ICDIM), Banglore, India.

Al-Khalifa, H. S. and H. Davis (2006g). *Measuring the Semantic Value of Folksonomies.* In Proceedings of the Second International IEEE Conference on Innovations in Information Technology, Dubai, UAEIEEE Computer Society.

Al-Khalifa, H. S. and H. Davis (2006h). *FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies*. In Proceedings of the Second International IEEE Conference on Innovations in Information Technology, Dubai, UAEIEEE Computer Society.

Al-Khalifa, H. S. and H. Davis (2006i). *Delicious Learning Resources*. In Proceedings of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CCISS'06), OnlineSpringer.

Al-Khalifa, H. S. and H. C. Davis (2005). *AraCore: An Arabic Learning Object Metadata for Indexing Learning Resources*. In Proceedings of The First Online Metadata and Semantics Research Conference (MTSR'05), Online,pp. 38-43, Rinton Press.

Al-Khalifa, H. S. and H. C. Davis (2006a). *Harnessing the Wisdom of Crowds: How to Semantically Annotate Web Resource Using Folksonomies*. A Poster In Proceedings of IADIS Web Applications and Research 2006 (WAR2006), Online.

Al-Khalifa, H. S. and H. C. Davis (2006b). *Folksonomies versus Automatic Keyword Extraction: An Empirical Study.* In Proceedings of IADIS Web Applications and Research 2006 (WAR2006), Online.

Al-Khalifa, H. S. and H. C. Davis (2006c). *The Evolution of Metadata from Standards to Semantics in E-Learning Applications*. In Proceedings of Hypertext'06, Odense, DenmarkACM Press.

Al-Khalifa, H. S. and H. C. Davis (2006j). "Folksonomies versus Automatic Keyword Extraction: An Empirical Study." IADIS INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND INFORMATION SYSTEMS (IJCSIS) **1**(2): pp. 132-143.

Al-Khalifa, H. S. and H. C. Davis (2007a). "EXPLORING THE VALUE OF FOLKSONOMIES FOR CREATING SEMANTIC METADATA." International Journal on Semantic Web and Information Systems (IJSWIS) **3**(1): pp. 13-39.

Al-Khalifa, H. S. and H. C. Davis (2007c). *Replacing the Monolithic LOM: A Folksonomic Approach*. In Proceedings of The 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), Niigata, JapanIEEE Computer Socirty [In press].

Al-Khalifa, H. S., H. C. Davis and L. Gilbert (2007b). *CREATING STRUCTURE FROM DISORDER: USING FOLKSONOMIES TO CREATE SEMANTIC METADATA*. In Proceedings of the 3rd International Conference on Web Information Systems and Technologies (WEBIST), Barcelona, SpainSpringer.

Al-Yahya, M. (2006). A New Model for Identifying and Describing Question/Answer Resource Semsntics for Distribuated Access. Nottingham, University of Nottingham. PhD**:** 211.

Antoniou, G. and F. van Harmelen (2004). A Semantic Web Primer, The MIT Press.

Arun, S. (2004). "Metadata management: past, present and future." Decision Support Systems **37**(1): 151.

Aurnhammer, M., P. Hanappe and L. Steels (2006). *Augmenting Navigation for Collaborative Tagging with Emergent Semantics*. 5th International Semantic Web Conference (ISWC 2006), Athens, GA, USALecture Notes in Computer Science, Springer.

Azouaou, F., W. Chen and C. Desmoulins (2004). *Semantic Annotation Tools for Learning Material*. International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL), Winston Salem, USA.

Azouaou, F. and C. Desmoulins (2005). *Semantic Annotation for the Teacher: Models for a Computerized Memory Tool*. Workshop on Applications of Semantic Web Technologies for e-Learning (SW-EL@ AIED'05), Amsterdam, The Netherlands.

Azouaou, F. and C. Desmoulins (2006a). *Using and Modeling Context with Ontology in e-learning: the Case of Personal Annotations*. Workshop on Applications of Semantic Web Technologies for e-Learning (SW-EL@A'06), Dublin, Ireland.

Azouaou, F. and C. Desmoulins (2006b). "Teachers' document annotating: Models for a digital memory tool." International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL) **16**(1/2): pp. 18-34.

Barritt, C. and F. Alderman (2004). Creating a Reusable Learning Objects Strategy. San Diego, Pfeiffer.

Bateman, S., C. Brooks and G. McCalla (2006). *Collaborative Tagging Approaches for Ontological Metadata in Adaptive ELearning Systems*. in Proceedings of the Fourth International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL 2006) in conjunction with 2006

International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006), Dublin, Ireland,3-12.

Bayerl, P. S., H. Lüngen, U. Gut and K. I. Paul (2003). *Methodology for Reliable Schema Development and Evaluation of Manual Annotations*. Workshop Notes for the Workshop on Knowledge Markup and Semantic Annotation, Second International Conference on Knowledge Capture (K-CAP 2003), Sanibel, Florida, USA,pp. 17-23, ACM Press.

Bayerl, P. S., H. G. Lungen, U. and K. I. Paul (2003). *Methodology for reliable schema development and evaluation of manual annotations*. in Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (KCAP), SanibelIsland, Florida,USAACM Press.

Bechhofer, S., L. Carr, C. Goble, S. Kampa and T. Miles-Board (2002). *The Semantics of Semantic Annotation*. In Proceedings of First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, Irvine, California,1151-1167.

Begelman, G., P. Keller and F. Smadja (2006). *Automated Tag Clustering: Improving search and exploration in the tag space*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The Semantic Web." Scientific American **284**(5): 34-43.

Bielenberg, K. and M. Zacher (2005). Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation. Digital Media. Bremen, Germeny, University Bremen. Master of Science in Digital Media**:** 120.

Bosworth, A. (2006). *Social Bookmarking Vs Spam*. Retrieved February 17, 2007 from http://www.sourcelabs.com/blogs/ajb/2006/01/social_bookmarking_vs_spam.html.

Brase, J. and W. Nejdl (2003). Ontologies and Metadata for eLearning. Handbook on Ontologies. S. Staab and R. Studer, Springer Verlag**:** 579--598.

Brase, J. and M. Painter (2004). "Inferring Metadata for a Semantic Web Peer-to-Peer Environment." Educational Technology & Society **7 (2)**: 61-67.

Brasher, A. and P. McAndrew (2004). *Human-generated learning object metadata*, Agia Napa, Cyprus,723, Springer-Verlag.

Brooks, C. and G. McCalla (2006). "Towards flexible learning object metadata." International Journal of Continuing Engineering Education and Life-Long Learning **16**(No.1/2): pp. 50 - 63.

Bruijn, J. d. and D. Fensel (2005). Ontology Definitions. Encyclopedia of Library and Information Science, Marcel Dekker, Inc.

Brunton, R. (2006). *Social Bookmarking versus Search Engines*. Retrieved February 17, 2007 from http://weblog.brunton.org.uk/archives/2006/06/social_bookmark_1.html.

Campbell, D. G. (2006). *A phenomenological framework for the relationship between the Semantic Web and user-centered tagging systems*. Proceedings of the 17th SIG Classification Research Workshop, Austin, Texas.

Cardinaels, K., M. Meire and E. Duval (2005). *Automating Metadata Generation: the Simple Indexing Interface*. Proceedings of the 14th international conference on World Wide Web, Chiba, Japan,pp. 548 - 556, ACM Press.

Carr, L., T. Miles-Board, A. Woukeu, G. Wills and W. Hall (2004). *The Case for Explicit Knowledge in Documents*. In Proceedings of ACM Symposium on Document Engineering, Milwaukee, Wisconsin.,pp. 90-98.

Cattuto, C., V. Loreto and L. Pietronero (2006). Collaborative Tagging and Semiotic Dynamics.

CEN (2003). *Internationalisation of the IEEE Learning Object Metadata*. EUROPEAN COMMITTEE FOR STANDARDIZATION, Retrieved February 18, 2007 from ftp://cenftp1.cenorm.be/PUBLIC/CWAs/e-Europe/WS-LT/cwa14643-00-2003-Jan.pdf.

Chatzinotas, S. and D. Sampson (2004). *EMAP: Design and implementation of educational metadata application profiles*. Proceedings - IEEE International Conference on Advanced Learning Technologies (ICALT 2004), Joensuu, Finland,876, IEEE Computer Society.

Chen, Y., Sherry and G. D. Magoulas. (2005). Adaptable and Adaptive Hypermedia Systems, Hershey: IRM Press.

Choy, S.-O. and A. K. Lui (2006). *Web Information Retrieval in Collaborative Tagging Systems*. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006), Hong Kong,pp. 352-355, IEEE Computer Society.

Christiaens, S. (2006). *Metadata Mechanisms: From Ontology to Folksonomy. and Back*. On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Montpellier, France,pp. 199-207, Lecture Notes in Computer Science, Springer.

Cimiano, P., G. Ladwig and S. Staab (2005). *Gimme' The Context: Context-driven Automatic Semantic Annotation with C-PANKOW*. The Fourteenth International World Wide Web Conference (WWW2005), Chiba, Japan,332-342, ACM Press.

Cohen, J. (1960). "A coefficient of agreement of nominal data." Educational and Psychological Measurements **20**(1): pp. 37-46.

Corcho, O. (2006). "Ontology based document annotation: trends and open research problems." Int. J. Metadata, semantics and Ontologies **Vol. 1**(1): pp.47-57.

De Nicola, A., M. Missikoff and F. Schiappelli (2004). *Towards an Ontological Support for eLearning Courses*. OTM Workshops, Sanibel - Florida, USA,pp. 773-777, Lecture Notes in Computer Science, Springer.

Degemmis, M., P. Lops and G. Semeraro (2006). WordNet-Based Word Sense Disambiguation for Learning User Profiles.

Dehors, S. and C. Faron-Zucker (2006). *Reusing Learning Resources based on Semantic Web Technologies*. Sixth International Conference on Advanced Learning Technologies (ICALT 2006), Kerkrade, The Netherlands,pp. 859-863, IEEE Computer Society.

Dehors, S., C. Faron-Zucker, J.-P. Stromboni and A. Giboin (2005). *Semi-automated Semantic Annotation of Learning Resources by Identifying Layout Features*. International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL'05), Amsterdam, Netherlands.

Dennis, B. M. (2006). *Foragr: Collaboratively Tagged Photographs and Social Information Visualization*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Dill, S., N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins, J. A. Tomlin and J. Y. Zienberer (2003a). "A Case for Automated Large Scale Semantic Annotation." Journal of Semantic Web **1**(1): 115-132.

Dill, S., N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin and J. Y. Zien (2003b). *SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation*. In Proceedings International WWW2003 Conference, Budapest, HungaryACM Press.

Ding, Y. (2005). "Study of Design Issues on an Automated Semantic Annotation System." AIS SIGSEMIS Bulletin **2**: 45-51.

Downes, S. (2004). "Resource Profiles." Journal of Interactive Media in Education (Special Issue on the Educational Semantic Web) **5**.

Dubinko, M., R. Kumar, J. Magnani, J. Novak, P. Raghavan and A. Tomkins (2006). *Visualizing tags over time*. Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland,pp. 193 - 202, ACM Press.

Duval, E. (2004). *We're on the road to.* Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano, Switzerland,pp. 3-5, AACE.

Duval, E. and W. Hodgins (2003). *A LOM research agenda*. Proceedings of 12th International World Wide Web (WWW'03), Budapest, Hungary.

Duval, E., W. Hodgins, S. Sutton and S. L. Weibel (2002). "Metadata Principles and Practicalities." D-Lib Magazine(Vol. 8, No. 4).

Duval, E., N. Smith and M. V. Coillie (2006). *Application Profiles for Learning*. Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies, ICALT 2006, Kerkrade, The Netherlands,pp. 242-246, IEEE Computer Society.

DZBOR, M., E. MOTTA and J. B. DOMINGUE (2004). *Opening Up Magpie via Semantic Services*. In Proc. of the 3rd Intl. Semantic Web Conference, JapanSpringer-Verlag.

Euzenat, J. (2002). "Eight questions about Semantic Web annotations." Intelligent Systems, IEEE **17**(2): 55 - 62.

Farrell, S. and T. Lau (2006). *Fringe Contacts: People-Tagging for the Enterprise*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Gasevic, D., J. Jovanovi and V. Devedzic (2004). *Enhancing Learning Object Content on the Semantic Web*. in Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04), Joensuu, Finland.

Golder, S. A. and B. A. Huberman (2005). *The Structure of Collaborative Tagging Systems.* Information Dynamics Lab, HP Labs.

Golder, S. A. and B. A. Huberman (2006). "The Structure of Collaborative Tagging Systems." Journal of Information Science **32**(2): pp. 198-208.

Gómez-Pérez, A. and D. Manzano-Macho (2004). "An overview of methods and tools for ontology learning from texts." The Knowledge Engineering Review **19** (3): 187-212.

Greenberg, J. (2005). "Metadata extraction and harvesting: A comparison of two automatic metadata generation applications." Journal of Internet Cataloging **6**(4): 59-82.

Greenberg, J., S. Sutton and D. G. Campbell (2003). "Metadata: a fundamental component of the semantic web." Bulletin of the American Society for Information Science **29**(4): 16.

Grimes, J. and T. Torres (2006). *Power to the People: Social Bookmarking*. Retrieved February 28, From http://earhart.umd.edu/movabletype/blogs/sschreib/lbsc708y06/pdf/3_may.pdf.

Gruber, T. (2005). "Ontology of Folksonomy: A Mash-up of Apples and Oranges." AIS SIGSEMIS Bulletin **2**(3&4).

Guy, M., A. Powell and M. Day (2004). "Improving the Quality of Metadata in Eprint Archives." Ariadne **January**(38).

Guy, M. and E. Tonkin (2006). "Folksonomies: Tidying up Tags?" D-Lib Magazine **V 12**(1).

Haase, K. (2004). *Context for semantic metadata*. Proceedings of the 12th ACM International Conference on Multimedia, New York, USA,204, ACM Press.

Halaschek-Wiener, C., J. Golbeck, A. Schain, M. Grove, B. Parsia and J. Hendler (2005). *PhotoStuff - An Image Annotation Tool for the Semantic Web*. The 4th International Semantic Web Conference, Poster Paper, Galway, Ireland.

Hammond, T., T. Hannay, B. Lund and J. Scott (2005). "Social Bookmarking Tools (I): A General Review." D-Lib Magazine **11**(4).

Han, P., Z. Wang, Z. Li, B. Kramer and F. Yang (2006). *Substitution or Complement: An Empirical Analysis on the Impact of Collaborative Tagging on Web Search*. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006), Hong Kong,pp. 757-760, IEEE Computer Society.

Handschuh, S. and S. Staab (2003). "CREAM: CREAting Metadata for the Semantic Web." Computer Networks **42**(5): pp. 579-598.

Handschuh, S. and S. Staab (2003a). Annotation for the Semantic Web, IOS Press.

Handschuh, S. and S. Staab (2003b). "Annotation of the Shallow and the Deep Web." Frontiers in Artificial Intelligence and Applications **Volume(96)**.

Handschuh, S., S. Staab and F. Ciravegna (2002). *S-CREAM - Semi-automatic CREAtion of Metadata*. In Proc. of the European Conference on Knowledge Acquisition and Management - EKAW-2002, Madrid, Spain Springer.

Handschuh, S., S. Staab and A. Maedche (2001). *CREAM: Creating relational metadata with a component-based, ontology-driven annotation framework*. K-CAP'01, Victoria, British Columbia, CanadaACM Press.

Harmelen, F. v. (2004). "The Semantic Web: What, Why, How, and When." IEEE Distributed Systems Online **vol. 05** (no. 3).

Heck, R. M. and S. M. L. C. H. Obermark (1999). A survey of web annotation systems. Technical report. Grinnell, Iowa, Grinnell College.

Henze, N. and W. Nejdl (2003). *Logically Characterizing Adaptive Educational Hypermedia Systems*. In Proceeding of the International Workshop of Adaptive Hypermedia and Adaptive Web-based Systems (AH 2003), Budapest, Hungary,pp. 15--28.

Hotho, A., R. Jäschke, C. Schmitz and G. Stumme (2006a). *Information Retrieval in Folksonomies: Search and Ranking*. Proceedings of the 3rd European Semantic Web Conference, Budva, MontenegroLNCS, Springer.

Hotho, A., R. Jäschke, C. Schmitz and G. Stumme (2006b). *Trend Detection in Folksonomies*. First International Conference on Semantics and Digital Media Technologies, SAMT 2006, Athens, Greece, pp. 56-70, Lecture Notes in Computer Science, Springer.

Hotho, A., R. Jäschke, C. Schmitz and G. Stumme (2006c). *Emergent Semantics in BibSonomy*. Proc. Workshop on Applications of Semantic Technologies, Informatik 2006Lecture Notes in Informatics, Springer.

Huynh, D., S. Mazzocchi and D. Karger (2005). *Piggy Bank: Experience the Semantic Web Inside Your Web Browser*. International Semantic Web Conference (ISWC2005), Galway, Ireland.

Jewell, M. O., K. F. Lawrence, A. Prugel-Bennett and m. c. schraefel (2006). *Annotation of Multimedia Using OntoMedia*. First International Workshop on Semantic Web Annotations for Multimedia (SWAMM), Edinburgh, UK.

John, A. and D. Seligmann (2006). *Collaborative Tagging and Expertise in the Enterprise*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Jovanovic, J., D. Gašsevic and V. Devedzic (2006). "Ontology-Based Automatic Annotation of Learning Content." International Journal on Semantic Web and Information Systems (IJSWIS) **2**(2): 91-119.

Kahan, J., M.-R. Koivunen, E. Prud'Hommeaux and R. R. Swick (2001). *Annotea: An Open RDF Infrastructure for Shared Web Annotations*. in Proc. of the WWW10 International Conference, Hong KongACM Press.

Kalfoglou, Y. (2001). Exploring Ontologies. Handbook of Software Engineering and Knowledge Engineering, World Scientific Publishing. 1**:** pp. 863-887.

Khare, R. (2006). Microformats: The Next (Small) Thing on the Semantic Web? Internet Computing, IEEE. 10**:** 68- 75.

Kipp, M. E. (2006). *Exploring the context of user, creator and intermediate tagging*. IA Summit 2006, Vancouver, Canada.

Kipp, M. E. I. and D. G. Campbell (2006). *Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices*. Proceedings Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas (US).

Kiryakov, A., B. Popov, I. Terziev, D. Manov and D. Ognyanoff (2003). *Semantic Annotation, Indexing and Retrieval*. In proccedings of the Second International Semantic Web Conference (ISWC'2003), Florida, USA,pp. 484-499.

Koivunen, M.-R. (2005). *Annotea and Semantic Web Supported Collaboration*. Invited talk at Workshop on User Aspects of the Semantic Web (UserSWeb), at Europen Semantic Web Conference (ESWC 2005), Herakliion, Greece.

Koivunen, M.-R., D. Brickley, J. Kahan, E. P. Hommeaux and R. R. Swick. (2000). *The W3C Collaborative Web Annotation Project. or how to have fun while building an RDF infrastructure*. W3C, Retrieved February 18, 2007 from http://www.w3.org/2000/Talks/www9-annotations/.

Kraft, R., F. Maghoul and C. C. Chang (2005). *Y!Q: Contextual Search at the Point of Inspiration*. The ACM Conference on Information and Knowledge Management (CIKM'05), Bremen, Germany,pp. 816-823, ACM Press.

Kristofic, A. (2005). *Recommender system for adaptive hypermedia applications.* In IIT.SRC 2005- Informatics and Information Technology Student Research Conference, Bratislava,pp. 229-234.

Kroski, E. (2005). *The Hive Mind: Folksonomies and User-Based Tagging*. Retrieved February 18, 2007 from http://infotangle.blogsome.com/category/folksonomies.

Lambiotte, R. and M. Ausloos (2005). *Collaborative tagging as a tripartite network*. Retrieved February 18, 2007 from http://arxiv.org/abs/cs/0512090.

Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics **33**(1): pp. 159-174.

Lassila, O. (1998). "Web metadata: a matter of semantics." IEEE Internet Computing **2**(4): 30.

Lawrence, K. F. and m. c. Schraefel (2006). *Freedom and Restraint: Tags, Vocabularies and Ontologies*. In Proceedings of 2nd IEEE International

Conference on Information & Communication Technologies: From Theory to Applications, Damascus, Syria.
Lee, M.-C., K. H. Tsai and T. I. Wang (2006). *An Ontological Approach for Semantic-Aware Learning Object Retrieval*. Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT2006), Kerkrade, The Netherlands,208-210, IEEE Computer Society.
Lee, M.-C., K. H. Tsai and T. I. Wang (2006). *An Ontological Approach for Semantic-Aware Learning Object Retrieval*. Sixth International Conference on Advanced Learning Technologies (ICALT 2006),pp. 208-210, IEEE Computer Society.
Lemire, D., H. Boley, S. McGrath and M. Ball (2005). "Collaborative Filtering and Inference Rules for Context-Aware Learning Object Recommendation." International Journal of Interactive Technology and Smart Education **2**(3).
Lin, X., J. E. Beaudoin, Y. Bui and K. Desai (2006). *Exploring characteristics of social classification*. Proceedings of the 17th SIG Classification Research Workshop, Austin, Texas.
Liu, H., P. Maes and G. Davenport (2006). "Unraveling the taste fabric of social networks." International Journal on Semantic Web and Information Systems **2**(1): 42-71.
LTSC (2002). *IEEE P1484.12.1-2002 Draft Standard for Learning Object Metadata, Learning Object Metadata Working Group of the IEEE Learning Technology Standards Committee (2002)*. Retrieved February 18, 2007 from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdfIEEE standard department.
LTSO (2004). *Learning Technology Standards Observatory*. Retrieved February 18, 2007 from http://www.cen-ltso.net/Users/main.aspx?put=194.
Lytras, M. and A. Naeve (2006). "Semantic e-learning: synthesising fantasies." British Journal of Educational Technology **37**(3): pp. 479-491.
Marlow, C., M. Naaman, d. boyd and M. Davis (2006). *Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead*. WWW 2006, Edinburgh, UKACM Press.
Martino, K. (2005). *Google vs. del.icio.us*. Retrieved February 17, 2007 from http://www.paradox1x.org/weblog/kmartino/archives/004166.shtml.
Mathes, A. (2004) Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Online: Online: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html. Date Accessed February 28, 2006
Mathes, A. (2004). *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Retrieved February 18, 2007 from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.
McFedries, P. (2006). Folk Wisdom. IEEE Specturm. Feb.**:** p.80.
Menchen, E. (2005). *Feedback, Motivation and Collectivity in a Social Bookmarking System*. Kairosnews Computers and Writing Online Conference, Retrieved February 18, 2007 from http://kairosnews.org/node/4338.
Michlmayr, E. (2005). *A Case Study on Emergent Semantics in Communities*. Workshop on Semantic Network Analysis, International Semantic Web Conference (ISWC2005), Galway, Ireland.
Mihalcea, R. and D. Moldovan (1999). *A Method for Word Sense Disambiguation of Unrestricted Text*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), Maryland, NY, USA.

Mika, P. (2005). *Ontologies are us: A unified model of social networks and semantics.* In Proceedings of the Fourth International Semantic Web Conference (ISWC 2005), Galway, Ireland,122-136, Lecture Notes in Computer Science.

Millard, D., F. Tao, K. Doody, A. Woukeu and H. Davis (2006). "The Knowledge Life Cycle for e-learning." International Journal of Continuing Engineering Education and Lifelong Learning: Special Issue on Application of Semantic Web Technologies in E-learning **16**(1/2): 110-121.

Millen, D., J. Feinberg and B. Kerr (2005). Social Bookmarking in the Enterprise. ACM Queue. vol. 3.

Mohan, P. and C. Brooks (2003). *Learning Objects on the Semantic Web*. In Proc. of the IEEE International Conf. on Adv. Learning Technologies, Athens, Greece,pp. 195-199.

Motelet, O. and N. A. Baloian (2006). *Hybrid System for Generating Learning Object Metadata*. Sixth International Conference on Advanced Learning Technologies (ICALT 2006), Kerkrade, The Netherlands,pp. 563-567, IEEE Computer Society.

Motive (2005). *The Motive Internet Glossary*. Retrieved February 18, 2007 from http://www.motive.co.nz/glossary/folksonomy.php.

Munoz, L. S. and J. P. M. d. Oliveira (2004). Adaptive Web-Based Courseware Development Using Metadata Standards and Ontologies. CAiSE 2004. A. Persson and J. Stirna. Berlin Heidelberg, Springer-Verlag. **3084:** 414-428.

Najjar, J., E. Duval, S. Ternier and F. Neven (2003a). *Towards Interoperable Learning Object Repositories: the ARIADNE Experience*. Proceedings of the IADIS International Conference WWW/Internet2003, Algarve, Portugal,pp. 219-226.

Najjar, J., S. Ternier and E. Duval (2003b). *The Actual Use of Metadata in ARIADNE: An empirical Analysis*. 3rd Annual Ariadne Conference, Leuven, Belgium, pp. 1-6.

Najjar, J., M. Wolpers and E. Duval (2006). *Attention Metadata: Collection and Management*. WWW2006 workshop on Logging Traces of Web Activity: The Mechanics of Data Collection, Edinburgh, ScotlandACM press.

Navigli, R., P. Velardi and A. Gangemi (2003). "Ontology Learning and Its Application to Automated Terminology Translation." IEEE Intelligent Systems **18**(1): pp. 22-31.

Nilsson, M., M. Palmér and A. Naeve (2002). *Semantic Web Metadata for e-Learning - Some Architectural Guidelines*. Proceedings of the 11th international conference on World Wide Web (WWW '02), Honolulu, Hawaii, USAACM Press.

NISO (2004). *Metadata types and functions*. Retrieved February 18, 2007 from http://www.loc.gov/catworkshop/readings/metadatabasics/types.html.

Niwa, S., T. Doi and S. Honiden (2006). *Web Page Recommender System based on Folksonomy Mining*. Third International Conference on Information Technology: New Generations (ITNG 2006), Las Vegas, USA,pp. 388-393, IEEE Computer Society.

Noy, N. F. and D. L. McGuinness (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Retrieved February 10, 2007 from http://smi.stanford.edu/smi-web/reports/SMI-2001-0880.pdf.

Ohmukai, I., M. Hamasaki and H. Takeda (2005). *A Proposal of Community-based Folksonomy with RDF Metadata*. The 4th International Semantic Web Conference (ISWC2005), Galway, Ireland.

Paolillo, J. C. and S. Penumarthy (2007). *The Social Structure of Tagging Internet Video on del.icio.us*. 40th Annual Hawaii International Conference on System Sciences (HICSS 2007), Waikoloa, Big Island, Hawaii,p.85, IEEE Computer Society.

Parsons, J., P. Ralph and G. K. (2004). *Using viewing time to infer user preference in recommender systems*. AAAI Workshop in Semantic Web Personalization, San Jose, California, USA.

Paynter, G. W. (2005). *Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources*. JCDL'05: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Tools & techniques: applying machine learning to collection development, Denver, CO, USA,pp. 291-300, ACM Press.

Peterson, E. (2006). "Beneath the Metadata: Some Philosophical Problems with Folksonomy." D-Lib Magazine **12** (11).

Qin, J. and N. Hernández. (2006). "Building interoperable vocabulary and structures for learning objects." Journal of the American Society for Information Science and Technology **57**(2): pp. 280-292.

Quintarelli, E. (2005). *Folksonomies: power to the people*. ISKO Italy-UniMIB meeting, Milan, ItlayRetrieved February 18, 2007 from http://www.iskoi.org/doc/folksonomies.htm.

Recker, M. M. and D. A. Wiley (2001). "A non-authoritative educational metadata ontology for filtering and recommending learning objects." Interactive Learning Environments **9**(3): 255.

Redeker, G. H. J. (2003). *An educational taxonomy for learning objects*. Proceedings of The 3rd IEEE International Conference on Advanced Learning Technologies (ICALT2003), Athens, Greece,pp. 250-251, IEEE Computer Society.

Reeve, L. and H. Han (2005). *Survey of semantic annotation platforms*. Proceedings of the 2005 ACM symposium on Applied computing, Santa Fe, New Mexico, USAACM Press.

Robson, R. (2000). *Report on Learning Technology Standards*. The World Conference on Educational Multimedia, Hypermedia and Telecommunications,971-976.

Russell, T. (2006). *cloudalicious: folksonomy over time*. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA,pp. 364-364, ACM Press.

Sampson, D. (2004). *The evolution of educational metadata: From standards to application profiles*. Proceedings - IEEE International Conference on Advanced Learning Technologies (ICALT 2004), Joensuu, Finland,1072, IEEE Computer Society.

Sauermann, L. (2005). *The Gnowsis Semantic Desktop for Information Integration*. 1st Workshop on Intelligent Office Appliances (IOA 2005): Knowledge-Appliances in the Office of the Future, Kaiserslautern, Germeny.

Scerri, S., C. Abela and M. Montebello (2005). *semantExplorer: A Semantic Web Browser*. IADIS International Conference WWW/Internet 2005, Lisbon, Portugal.

Sen, S., S. Lam, A. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, M. Harper and J. Riedl (2006). *tagging, communities, vocabulary, evolution*. CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work,190, ACM.

Shen, K. and L. Wu (2005). *Folksonomy as a Complex Network*. Retrieved February 18, 2007 from http://arxiv.org/abs/cs/0509072.

Sheth, A., C. Ramakrishnan and C. Thomas (2005). "Semantics for the Semantic Web: The Implicit, the Formal and the Powerful." International Journal on Semantic Web & Information Systems **1**(1): 1-18.

Shirky, C. (2005). *Ontology is Overrated: Categories, Links, and Tags.* Retrieved February 18, 2007 from http://shirky.com/writings/ontology_overrated.html.

Sieck, S. (2005). connotea and citeulike: "folksonomies" emerge within scholarly communities. E. Insight.

Sinha, R. (2005). *A cognitive analysis of tagging*. Retrieved February 18, 2007 from http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html.

Stefani, A., B. Vassiliadis and M. Xenos (2006). *Behavioral patterns in hypermedia systems: a short study of e-commerce vs. e-learning practices*. International Workshop on Adaptive and Personalized Semantic Web, pp. 57-64, Springer Verlag.

Stock, E. (2006). *User-Created Content: Selfish, Social, Selfless*. Retrieved February 18, 2007 from http://onotech.blogspot.com/2006/01/user-created-content-selfish-social.html.

Stojanovi, L., S. Staab and R. Studer (2001). *eLearning in the Semantic Web*. In Proceedings of the WWWNet Conference, Orlando, USA.

Stoll, R. R. (1979). Set Theory and Logic. Mineola, N.Y., Dover Publications.

Stratakis, M., V. Christophides, K. Keenoy and A. Magkanaraki (2003). E-Learning Standards**:** 45.

Stuckenschmidt, H. and F. v. Harmelen (2004). Information Sharing on the Semantic Web, Springer.

Stutzman, F. (2006). *Del.icio.us Spam, Social Technology and Trust*. Retrieved February 17, 2007 from http://chimprawk.blogspot.com/2006/09/delicious-spam-social-technology-and.html.

Szekely, B. and E. Torres (2005). Ranking Bookmarks and Bistros: Intelligent Community and Folksonomy Development. Harvard University. *Unpublished*

Tallis, M. (2003). *Semantic Word Processing for Content Authors*. In Workshop Notes of the Knowledge Markup and Semantic Annotation Workshop (SEMANNOT 2003), Second International Conference on Knowledge Capture (K-CAP 2003), Sanibel, Florida, USA.

Tennis, J. T. (2006). *Social Tagging and the Next Steps for Indexing*. Proceedings 17th SIG/CR Classification Research Workshop, Austin, Texas.

Thompson, C. and R. Mooney (1999). *Automatic construction of semantic lexicons for learning natural language interfaces*. Proceedings of 16th National Conference on Artificial Intelligence (AAAI'99), Orlando, FL, USA,pp. 487–493.

Tonkin, E. (2006). *Searching The Long Tail: Hidden Structure in Social Tagging*. Proceedings 17th SIG/CR Classification Research Workshop, Austin, Texas.

Tosic, M. and V. Milicevic (2006). *The Semantics of Collaborative Tagging System*. In Proceedings of the 2nd Workshop on Scripting for the Semantic Web, Retrieved February 23, 2007 from http://www.semanticscripting.org/SFSW2006/Paper6.pdf.

Trant, J. and B. Wyman (2006). *Investigating social tagging and folksonomy in art museums with steve.museum*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Trevino, E. M. (2006). Social Bookmarks: Personal Organization and Collective Discovery on the Web. Arts in Communication. Chicago, Illinois, University of Illinois. Master of Arts in Communication**:** 93.

Ullrich, C. (2005). "The learning-resource-type is dead, long live the learning-resource-type!" Learning Objects and Learning Designs **1**(1): 7-15.

Uren, V., P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta and F. Ciravegna (2005). "Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art." Journal of Web Semantics **4**(1): 34.

Vallet, D., M. Fernandez and P. Castells (2005). *An Ontology-Based Information Retrieval Model*. Second European Semantic Web Conference, ESWC 2005, Crete, Greece,455-470, Lecture Notes in Computer Science, Springer.

van Heijst, G., A. T. Schreiber and B. J. Wielinga (1997). "Using explicit ontologies in KBS development." **46**(2-3): 183.

Vander Wal, T. (2005). *Explaining and Showing Broad and Narrow Folksonomies*. Retrieved February 18, 2007 from http://www.personalinfocloud.com/2005/02/explaining_and_.html.

Vander Wal, T. (2007). *Folksonomy*. Retrieved February 18, 2007 from http://vanderwal.net/folksonomy.html.

Vargas-Vera, M., E. Motta, J. Domingue, M. Lanzoni, A. Stutt and F. Ciravegna (2002). MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup. The 13th International Conference on Knowledge Engineering and Management (EKAW 2002). A. Gomez-Perez, Springer Verlag.

Vaughan, L. and M. Thelwall (2004). "Search engine coverage bias: evidence and possible causes." Information Processing & Management **40**(4): 707.

Verbert, K., D. Gašević, J. Jovanović and E. Duval (2005). Ontology-based learning content repurposing. Special interest tracks and posters of the 14th international conference on World Wide Web. Chiba, Japan, ACM Press.

Veres, C. (2006). *The Language of Folksonomies: What Tags Reveal About User Classification*. Natural Language Processing and Information Systems, 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006, Klagenfurt, Austria,58-69, Springer Berlin / Heidelberg.

Veres, C. (2006b). *Concept Modeling by the Masses: Folksonomy Structure and Interoperability*. Conceptual Modeling - ER 2006, the 25th International Conference on Conceptual Modeling, Tucson, AZ, USA,325-338, Springer Berlin / Heidelberg.

Voss, J. (2007). *Tagging, Folksonomy & Co - Renaissance of Manual Indexing?* Retrieved February 18, 2007 from http://arxiv.org/abs/cs/0701072.

Voss, J. (2007). *Tagging, Folksonomy & Co - Renaissance of Manual Indexing?* Retrieved February 11, 2007 from http://arxiv.org/abs/cs/0701072.

Voss, J. (January 26 2007). *Tagging, Folksonomy & Co - Renaissance of Manual Indexing?* Retrieved February 18, 2007 from http://arxiv.org/abs/cs/0701072.

W3C (2001). *Extensible Markup Language (XML)*, Retrieved February 11, 2007 from http://www.w3.org/XML/.

W3C (2002). *Resource Description Framework (RDF)*, Retrieved February 11, 2007 from http://www.w3.org/RDF/.

W3C (2004). *OWL Web Ontology Language (OWL)*, Retrieved February 11, 2007 from http://www.w3.org/TR/owl-ref/.

W3C (2004). *Web Services Architecture*. Retrieved February 27, 2007 from http://www.w3.org/TR/ws-arch/.

W3C (2006). *Image Annotation on the Semantic Web: Tools Overview*. Retrieved February 12, 2007 from http://www.w3.org/2001/sw/BestPractices/MM/resources/Tools.html.

Wash, R. and E. Rader (2006). *Collaborative Filtering with del.icio.us*. Conference on human factors in computing systems (CHI2006), Montreal, CanadaACM Press.

WeBreakStuff (2005). *Tagging and social bookmarks*. Retrieved February 18, 2007 from http://webreakstuff.com/blog/2005/10/tagging-and-social-bookmarks/.

Wikipedia (2006) Folksonomy. Online: http://en.wikipedia.org/wiki/Folksonomy. Date Accessed March 26, 2006

Wikipedia (2007). *Folksonomy*. Retrieved February 18, 2007 from http://en.wikipedia.org/wiki/Folksonomy.

Wiley, D. A. (2002). "Connecting Learning Objects to Instructional Design Theory: A Definition, a Metaphor, and a Taxonomy." In: David A. Wiley (Ed.): The Instructional Use of Learning Objects and Learning Designs: pp. 3-23.

Woukeu, A., G. Wills, G. Conole, L. Carr, S. Kampa and W. Hall (2003). *Ontological Hypermedia in Education: A framework for building web-based educational portals*. Proceedings of ED-MEDIA 2003-World Conference on Educational Multimedia, Hypermedia & Telecommunications 2003, Honolulu, Hawaii, USA.

Wu, H., M. Zubair and K. Maly (2006). *Harvesting social knowledge from folksonomies*. Proceedings of the seventeenth conference on Hypertext and hypermedia, Odense, Denmark,pp. 111-114, ACM Press.

WU, S.-H. and W.-L. HSU (2002). *SOAT: a semi-automatic domain ontology acquisition tool from Chinese corpus*. Proceedings of the 19th international conference on Computational linguistics, Taipei, Taiwan,pp.1-5, Association for Computational Linguistics.

Wu, X., L. Zhang and Y. Yu (2006). *Exploring Social Annotations for the Semantic Web*. World Wide Web Conference (WWW2006), Edinburgh, Scotland.ACM Press.

Xu, Z., Y. Fu, J. Mao and D. Su (2006). *Towards the Semantic Web: Collaborative Tag Suggestions*. in Proceedings Collaborative Web Tagging Workshop in 15th International World Wide Web Conference (WWW2006), Edinburgh, Scotland.

Yin, Z., Z. Xu and A. El Saddik (2003). *Study of metadata for advanced multimedia learning objects*. Canadian Conference on Electrical and Computer Engineering (IEEE CCECE 2003), Ottawa Univ., Ont., Canada,pp. 1099-1102, IEEE Computer Society.

Zhang, L., X. Wu and Y. Yu (2006). "Emergent Semantics from Folksonomies: A Quantitative Study." Journal on Data Semantics VI **4090**: 168-186.

Zhihong, Z. and Z. Mingtian (2003). *A Distributed Description Logic Approach to Semantic Annotation*. Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Application and Technologies,pp. 219-221, IEEE.