# Handbook of Research on Web 2.0, 3.0, and X.0:
## Technologies, Business, and Social Applications

San Murugesan
*Multimedia University, Malaysia & University of Western Sydney, Australia*

Volume I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Senior Managing Editor: Jamie Snavely
Assistant Managing Editor: Michael Brehm
Publishing Assistant: Sean Woznicki
Typesetter: Carole Coulson, Ricardo Mendoza, Kurt Smith
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference



Library of Congress Cataloging-in-Publication Data

Handbook of research on Web 2.0, 3.0, and X.0 : technologies, business, and social applications / San Murugesan, editor.
p. cm.
Includes bibliographical references and index.
Summary: "This book provides a comprehensive reference source on next generation Web technologies and their applications"--Provided by publisher.
ISBN 978-1-60566-384-5 (hardcover) -- ISBN 978-1-60566-385-2 (ebook) 1. Web 2.0. 2. Social media. I. Murugesan, San.
TK5105.88817.H363 2010
025.042'7--dc22
2009020544

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

# Chapter 20
# Towards Disambiguating Social Tagging Systems

**Antonina Dattolo**
*University of Udine, Italy*

**Silvia Duca**
*University of Bologna, Italy*

**Francesca Tomasi**
*University of Bologna, Italy*

**Fabio Vitali**
*University of Bologna, Italy*

## ABSTRACT

*Social tagging to annotate resources represents one of the innovative aspects introduced with Web 2.0 and the new challenges of the (semantic) Web 3.0. Social tagging, also known as user-generated keywords or folksonomies, implies that keywords, from an arbitrarily large and uncontrolled vocabulary, are used by a large community of readers to describe resources. Despite undeniable success and usefulness of social tagging systems, they also suffer from some drawbacks: the proliferation of social tags, coming as they are from an unrestricted vocabulary leads to ambiguity when determining their intended meaning; the lack of predefined schemas or structures for inserting metadata leads to confusions as to their roles and justification; and the flatness of the structure of the keywords and lack of relationships among them imply difficulties in relating different keywords when they describe the same or similar concepts. So in order to increase precision, in the searches and classifications made possible by folksonomies, some experiences and results from formal classification and subjecting systems are considered, in order to help solve, if not to prevent altogether, the ambiguities that are intrinsic in such systems. Some successful and not so successful approaches as proposed in the scientific literature are discussed, and a few more are introduced here to further help dealing with special cases. In particular, we believe that adding depth and structure to the terms used in folksonomies could help in word sense disambiguation, as well as correctly identifying and classifying proper names, metaphors, and slang words when used as social tags.*

DOI: 10.4018/978-1-60566-384-5.ch020

## INTRODUCTION

The purpose of this chapter is to introduce the reader to the problems of extracting meaningful, organized information from user-generated folksonomies, and to expose a number of limitations in the current approaches that will need to be solved in the immediate future.

In the Web 2.0 era, social tagging is a concept used to refer to the activity of a large number of human readers who associate descriptive terms (often called *tags*) to Web resources they are reading or searching; no rules, restrictions, and not even suggestions are usually offered to readers when generating tags for these resources, in order to maintain the spontaneity and statistically-relevant frequency of use of the terms thought of by real people. The tags actually entered are then analysed through statistical tools to help other users, that use the same terms, to find the same documents. Folksonomies in this context are the classifications of Web resources emerging from the identification of the statistical prominence of some tags over the others.

On the other hand, traditional document classification methods (both on the Web and on printed collections) have preferred stricter and more precise methods for subjecting and classification. Enumerative systems, taxonomies, thesauri and ontologies are generated by dedicated (and human) professionals; they provide construction rules for the classification (at least a controlled vocabulary) and then painstakingly read, digest reflect on the document content and add manually metadata values. These values match both the content of the documents themselves and the expectations and slant of the collection in which the document ends.

Although the manual process usually reaches high quality levels of classification for traditional document collections, it does not scale to the humongous size of the Web, both in terms of costs, time, and expertise of the human personnel required, and as such it cannot be proficiently put into existence for the whole Web.

If the generation of a complete classification system, using a third party army of professionals, is inappropriate and hard to scale, even the alternative approach of author-created metadata falls short of another important issue, namely, the fact that the intended and unintended users of the information are disconnected from the classification process (Mathes, 2004).

On the other hand, social tagging (i.e., reader-created metadata) deals with this limitation: the added value offered from folksonomies is that this operation is entrusted to the mass actions of the readers themselves, that naturally average the extremes and coalesce on a limited numbers of terms that most probably will be the same used by subsequent users searching the same documents. Pioneered by Web social bookmarking services (such as Del.icio.us, http://delicious.com/; Digg, http://digg.com/; Furl, http://www.furl.net/) and photo or videos sharing services (such as Flickr, http://www.flickr.com/), folksonomies contribute to add not just information to resources, but *concretely relevant* information to resources. The list of tags, however unconstrained and subjective, used by individual readers to describe a document, after reaching a critical mass, tend to cluster around particularly frequent terms that become the most meaningful ones that could be used, have been used and will be used to describe that document. Thus final users are not only *connected* to the classification process, but they in fact are *the main actors* of the classification process.

Of course this flexibility comes at a price: social tagging does not handle issues that are easily handled by previous classification methods:

- *Ambiguity*: social tagging does not enforce, or even propose, values from a restricted set of terms (the *controlled vocabulary*), thus in folksonomies we are sure to find the same the ambiguity that we find in natural language (e.g., homonymy, polysemy, synonymy, term variations, and even plain and simple spelling errors).

- *Undistinguished concerns*: social tagging does not enforce, or even propose, a schema for distinguishing the purpose of a metadata value. The tags might be, indifferently, subject descriptors, genres, self-reminders; tangential remarks (such as colours or years, especially for pictures on Flickr); or proper names.
- *Independence of terms*: social tagging does not provide relations to connect and relate different terms: each tag is independent of the others, and no inference is possible. For instance, no exploitation is possible of hierarchies of concepts, as available with taxonomies, and in fact *basic level variation* (whereby terms with different levels of specificity are used on the same resource, e.g., person, actor, celebrity) is a frequent occurrence in folksonomies.

In this chapter we intend therefore to report on a number of ideas, theories, and systems that have been proposed and discussed in literature to deal with these issues, and we intend to provide a few trends in addressing the issues left open by these works.

We first describe the background behind them, by detailing the traditional subjecting and classification approaches as well as the new social approach. We will thus explore numerical and faceted classification schemes, taxonomies, thesauri and ontologies, examining the expressive power and sophistication, and compare them to folksonomies, which are the most recent addition to the set.

The basic idea of most of the above mentioned works is to mix, at least partially, traditional methods and folksonomies in order to generate meaningful and scalable classifications for resources. This in turn corresponds to ways to:

- *Remove ambiguity*. By providing a clear and restricted semantic frame to terms (e.g. a controlled vocabulary) ambiguity disappears and their exact meaning emerges.
- *Add depth*. By associating the terms to a hierarchical semantic frame (e.g., a thesaurus), their specificity level and the relations with other related terms become evident and navigable in order both to perceive and to tackle basic level variation.
- *Add qualification*. By associating specific qualification from a well-known schema (e.g., Dublin Core, http://dublincore.org/) to social tags, we obtain a better and more precise description of their justification, appropriateness and use.
- *Extract ontologies*. A meaningful challenge is the study of folksonomies and their meaning to extract fully developed ontologies that can be used for more than just searching, but even for reasoning and inferences as made possible with the advent of semantic Web technologies.

Yet, these works are far from covering the whole set of issues that arise in the automatic structuring of purposefully unstructured terms. Some of these issues are still uncovered and hardly discussed in literature. Some of them we will examine further, especially in the case of correct disambiguation and contextualization of proper names (of people, brands, organizations, places, etc.), of identification of metaphors (i.e., exaggerated, offensive, malicious or figurative misrepresentation of concepts through evocative, and yet improper, terms), of individuation and interpretation of slang terms, and of qualification of terms (i.e., the association of the most appropriate qualifying facet to terms that are not meant to contribute to the subject description of a document).

We will try to discuss and detail some ways to address these unmanaged issues; these techniques, whose usefulness we are in the process of proving, come on the other hand with clear limitations themselves, which we will try to describe and justify in our conclusions.

## Background

Descriptive and structured terms used for representing the content of an informational resource is a common approach oriented to organize and manage information on which retrieval operations will be required. Organize information is a practice that associate the work made in libraries, archives, museum, settled to the creation of catalogues, indexes, finding aids, etc., with the treatment of web resources (especially directory).

Experts in library cataloguing commonly assign keywords to books in order to describe the content of data source and aggregate documents regarding the same object. In the same way, collaborative or social tagging, commonly known as folksonomy, is a process that allows users to add different kind of metadata to resources ("anything with a URL", Vander Wal, 2005) and share tags and contents on the Web. But in traditional libraries, cataloguers use *controlled vocabulary* for describing materials and refer to categorization rules based on specific schemes (*classification systems*). Folksonomies, on the contrary, are Web-based systems that allow users to upload their resources, labelling them with arbitrary words, the so-called tags, without referring to a standard classification scheme or a controlled vocabulary for the keywords.

The differences between traditional formal methods of classification and folksonomies are related to the two different approaches to resources description. In the first we can speak about a *top-down* philosophy: we already have a scheme (a defined vocabulary or a classification system) to be adapted to the resources being described. In the second we refer to a *bottom-up* approach: we start from resources, i.e. from the reality, trying to apply descriptors coming from non-controlled terms belonging to our natural language.

The drawbacks of each approach are balanced by the advantages of the other, and viceversa, so that we end up dealing with two complementary ways to associate keywords to resources.

- It is necessary, for this reason, to introduce traditional formal systems in order to understand:
  1) what we mean with subjecting and classification methods based on formal schemes;
  2) how these methods work in order to solve the ambiguity of the natural language and to handle the relationships between concepts;
  3) how and in which sense these systems could be used as linguistic resources;
  4) in which way it is possible integrate *bottom-up* systems with *top-down* one.

## Formal Subjecting and Classification Methods

Formal subjecting and classification methods aims to:

- *regularize the vocabulary*, in order to solve the ambiguity of the language;
- *categorize knowledge*, that is define semantic (i.e. explicitly declared) relationships between terms.

The problem of ambiguity of natural language, in particular, has been long discussed, mostly by librarians and information scientists, and resolved in formal subjecting and classification efforts. The result of this reflection is the production of controlled vocabularies and classification schemes in which terms are organized in structures according to different relationships.

Environments, such as libraries, archives, museums, etc., have to deal with two different kinds of problems when describing a resource:

- *Semantic univocity*. At the first level, it is necessary to define a specific word for a specific concept: each descriptive keyword

has to be unambiguous. This is a complex task, because, in natural languages:
  - a lot of words have more than one meaning (polysemy). This also potentially means that the same word could assume different meanings in function of its grammatical form (e.g. the term is used as noun, verb, adjective or adverb);
  - people speaking about the same concept often do not use the same word, or the same concept could be expressed with different terms (synonymy).

The definition of a vocabulary require the control not only on synonyms, homonyms and homographs (that is polysemy), but also on different forms of the same term (e.g. online; on line; on-line), composed or bound words (e.g. credit card), specific and generic concepts referring to the same content (dachshund is specific, dog is generic but it applies to the same content).

- *Relations among concepts*. Secondarily, it is necessary to place the concept in relation with others at some different levels, exploiting all its characteristics; in this way, it becomes possible to establish different kinds of semantic relationships between accepted concepts. Also, generally the use of a controlled vocabulary resolves the relationships between concepts at the subject level, while classification schemes manage the relationships among classes.

A knowledge organization system may be defined as a structured collection of terms formally defined in a restricted vocabulary. Formal subjecting and classification methods help in creating a knowledge domain for retrieving documents using subject descriptors.

The different kinds of methods, commonly used for traditional libraries, can be grouped in two typologies, respectively based on:

- *vocabulary control*: thesauri and ontologies;
- *categorization and classification*: enumerative and faceted classification schemes. Also, among systems for categorization and classification the taxonomies represent more a theory of knowledge organization than a specific classification method in use.

Next two subsections are focused on these two approaches: the starting point is represented by the systems for vocabulary control, that is the lists of accepted terms and the relationships among them. Then the differences between two different theories of classification and their relation with the taxonomies are discussed.

## Systems for Vocabulary Control

***Indexing languages*** represent the first step for lexical normalization: the necessity of semantic univocity can be solved by a biunivocal relation between a term and a concept, that is one term for each concept and one concept for each term.

Indexing is the act of describing or identifying a document in terms of its subject content (ISO 5963/1985). A subject could be defined as a concept, or a combination of concepts, that represent the content of a document.

A formal index is so a list of accepted words that could be used for describing a resource: subject headings (lists of controlled terms) like the Library of Congress Subject Headings (LCSH) or authority files (controlled term mainly for proper names) are used in the OPAC (*Online Public Access Catalogue*) for disambiguate among the different forms used for expressing a term.

The functional form of a controlled vocabulary are the thesauri.

**Thesauri**. A thesaurus (ISO 2788/1986) defines:

- a consistent (and controlled) terminology;
- the preferred term to be used;

- semantic relationships among the terms.

Its main function is to solve the ambiguity introduced by the use of natural languages, determining the *Preferred Term* (in short, PT) to used in describing a concept; such a choice specifies a relationship between different terms and a concept. This is the first semantic relation present in a controlled vocabulary and named *equivalence* or synonymic relation. Also alternative spellings, acronyms and abbreviation are considered synonymic situations and have to be resolved. In a thesaurus the accepted form, also named *descriptor*, is PT; it may be either a single concept or a bound term (if the concept can be only represented by two or more words). The NP represents the *Non Preferred* term.

Also the choice of the descriptor is a relationship between accepted (that is preferred) and non-accepted terms.

An interesting semantic relation managed by a thesaurus is the *hierarchical* one: it defines a tree of terms representing relations of subordination and up-ordination between the accepted concept and parents-children concepts. More in details, in a thesaurus for a concept we find *Broader Terms* (BT), that is more general terms, and *Narrower Terms* (NT), that is more specific terms.

Some thesauri (following the ISO 2788 Guidelines) design more specifically the simple broader and narrower terms differentiating hierarchical relations in three levels:

- *Generic*: a relation genus/species or class/class member (e.g. house is a member of the class building);
- *Partitive*: a relation whole/part (e.g. nail are part of finger);
- *Instance*: a relation class/instance or a generic topic/named example (e.g. Paris is an instance of a city).

A specific case of hierarchical relation is the *poly-hierarchy* that refers to the possibility for a term to derive from different classes. This is a particular interesting aspect of vocabulary managing because it solves problems related to the concepts belonging to more than one category (more than one BT may be contemplated for each concept).

Finally, the *associative* relation (defined by *Related Term* RT) includes relations such as cause/effect, agency/instrument; sequence in time or space; characteristic feature.

Vocabulary control, various forms of indexing terms, the use of singular and plural forms in indexing languages, the choice of appropriate terminology, proper names in indexing languages, and the functions of scope notes and definitions are objects of thesaurus implementation.

**Ontologies.** They are part of the W3C standards used in particular for the Semantic Web. An ontology is a collection of *concepts* and *relations* among them, based on *classes,* identified by categories, *properties,* which are different aspects of the class, and *instances* that are the things.

In other words, ontology is a description of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the use of ontology as set-of-concept-definitions.

Generally an ontology is identified by a set of definitions related to a formal vocabulary, since hyponyms and hypernyms, holonyms and meronyms, antonyms are viewed as different relations among concepts.

Ontologies are used to specify standard conceptual vocabularies, provide services for answering queries, publish reusable knowledge bases, and offer services to facilitate interoperability across multiple, heterogeneous systems and databases. The key role of ontologies with respect to database systems is to specify a data modelling representation at a level of abstraction above specific database designs (logical or physical), so that data can be exported, translated, queried, and unified across independently developed systems and services. Common components of ontologies include

*classes*, *individuals*, *attributes*, *relations*, *function terms, restrictions, rules, axioms and events.*

Possible typologies of hierarchies are:

- *is_a* (is a type of) is generic and is a specialization of the concept represented by the class in a wider/narrower sense: all the instances of a subclass are also instances of a superclass.
- *has_a* (has a) is partitive and is a specialization of the concept represented by the property: all the valid instances of a class must provide a value for that property.
- *instance_of* (instance of) is a relationship of belonging of an object (a class-of-one) to a class.

Ontologies are often equated with taxonomic hierarchies of classes, but, in order to specify a conceptualization, it is necessary to state axioms for constraining the possible interpretations of defined terms.

## Classification Systems and Taxonomies

If a thesaurus or a lexical network defines explicit semantic relationships among concepts, the pure classification systems define the membership of a concept (the descriptor) to a category and set the relationship among categories, expressing the link in some kind of notation. We could say that the Aristotelian theory of category is the basis for the major classification schemes in use.

In bibliotheconomy, indexing languages could be distinguished in:

- subjecting that identify the topics related to the document and express them in a controlled vocabulary (see section "Systems for vocabulary control");
- classification that aim to define the field of knowledge the document belongs to.

Bibliographic classification could be divided, in turn, into two macro areas that represent two different approaches in knowledge organization: *top-down* and *bottom-up*.

**Hierarchical-enumerative classification.** This classification uses a *top-down* scheme: knowledge is organized in classes or categories progressively narrower. The most used classification system derives from Melvile Dewey that in 1876 proposed the Dewey Decimal Classification (DDC), today used in libraries. DDC proposes 10 main classes, each divided into 10 divisions (one thousand); each division is divided into 10 subsections (ten thousand), and so on into potential infinity. Each object (e.g. a book) is assigned a number, possibly decimal, and a string of words that identify the subject of the described object (e.g. 800 identifies Literature; 850 Italian Literature; 856 Italian letters; and so on). Other classification schemes are the Universal Decimal Classification (UDC) and the Library of Congress Classification (LCC). These schemes share the use of a verbal description of concepts associated to a notation alongside.

**Faceted classification systems**. An alternative to hierarchical-enumerative classification scheme is represented by the analytical-synthetically classification system, a *bottom-up* scheme that divides a subject into concepts (analytical) and gives rules to use these concepts in constructing a structured subject (synthetically).

This new approach to cataloguing derives from Ranganathan, who, in 1930, proposed the Colon Classification (CC) of documents, based on the concept of *facets*. Faceted classification supports descriptions based on different characteristics of a subject. We can say that the CC is a framework by which any document could be broken down in terms of five facets: personality, matter, energy, space and time (formula PMEST).

- Personality (the something in question, e.g. a person or event in a classification of

history, or an animal in a classification of zoology)
- Matter (what something is made of)
- Energy (how something changes, is processed, evolves)
- Space (where something is)
- Time (when it happens)

This makes it possible to create a heading for composite complex subject without using a determistic list of subjects defined in a hierarchical structure. A kind of poly-hierarchical relationship for each aspect regarding the subject (Quintarelli, 2005). As we show in section "Adding qualification" the faceted mechanism represent the Dublin Core way to organize and describe an resource.

**Taxonomies.** We complete the overview of the subjecting and classification systems with taxonomies. Firstly because of their apparent relation to folksonomy, secondly since the concept is not so clearly defined in literature.

Taxonomies exist at least from 1735, when Linnaeus published his *Sistema Naturae*, a classification of plants and animals. The term taxonomy is used for every kind of system that organizes things in categories. Linnean system, traditional classification schemes, Internet directories, the organization of files and directories in a file system are taxonomic views of the objects organized in categories. The taxonomies represent the classical system of categorization, a concept different from classification (Jacob, 2004). Classification is strictly related to bibliographic enumerative schemes while categorization is less rigorous and it not necessary alludes to a hierarchy in the strict sense (for example, a facet could be defined as a category).

## Folksonomies

The term **folksonomy** is the fusion of folks & taxonomy and has been coined by Thomas Vander Wal in a listserv discussion hosted by the Information Architecture Institute (Smith, 2004); it is the result of personal free tagging of information and objects by members of a (possibly large) community. The tagging is done in a social environment.

A folksonomy allows user communities (rather than taxonomy professionals) to classify Web sites, providing a democratic tagging system that reflects the opinions of the general public.

Tools for the management of a folksonomy are not part of the underlying Web protocols; Web-based communities enable Web users to label and share user-generated contents or to collaboratively label existing contents, such as Web sites, books, works in the scientific and scholarly literatures, and blog entries (Marlow et al., 2006).

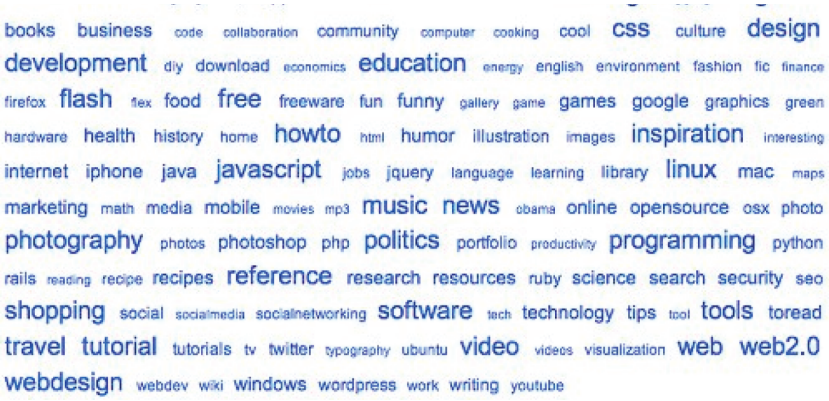
Usually people prefer to use tags, to provide a way to connect items and to propose their meaning in their own understanding.

The three tenets of a folksonomy are the *tags*, words linked to resources; the *resources*, object being tagged, and the *person*, author of the tagging. Each person uses his/her own vocabulary and adds explicit meaning to resources.

The most important Web 2.0 services based on folksonomies are:

- Del.icio.us is a social bookmarking Web service for storing, sharing, and discovering Web bookmarks. The site was founded by Joshua Schachter in late 2003 and acquired by Yahoo! in 2005. It has more than three million users and 100 million bookmarked URLs.
- Flickr is an image and video hosting Website, Web services suite, and online community platform. In addition to being a popular Web site for users to share personal photographs, the service is widely used by bloggers as a photo repository. Its popularity has been fuelled by its organization tools, which allow photos to be tagged and browsed through folksonomic means.
- Furl (File Uniform Resource Locators) is a social bookmarking site that makes easy to save, share, and explore favourite Web

*Figure 1. An example of tag cloud*




pages. Furl enables members to bookmark, annotate, and share Web pages. Topics are used to categorize saved sites, similar to the tagging feature of other social Websites. Additionally, a user may write comments, save clippings, assign each bookmark a rating and keywords (which are given greater weight while searching), and have an option of private or public storage for each topic or item archived.

Folksonomy-based tools enable users to see what are the most used tags relatively to given pages.

## From Folksonomies to Semantic Tags

Folksonomies are strictly related to the concepts of polysemy, synonymy, basic level variation (Golder Huberman, 2005); ambiguity, spaces and multi words (Mathes, 2004) and have to deal with systems of classification and categorization.

- In comparison to traditional subjecting and classification methods, social tagging is flat: no hierarchy of terms is supported, no parent-children and no sibling relationships are contemplated. Folksomomy "is not collaborative, it is not putting things in to categories, it is not related to taxonomy (more like the antithesis of a taxonomy)" (Vander Wal, 2005). Shared social categorization is not conceived for providing hierarchical structures in resources descriptions; but a hierarchy could however emerge. In detail this is possible by associating a semantic frame to terms: thesauri and ontologies in order to give a support in vocabulary control and in relationships managing (synonyms, hierarchies, related terms); facets in order to assign qualification, eventually mapped into metadata schemes (like the Dublin Core). This in particular means to find systems for associating traditional formal *top-down* systems to the new social *bottom-up* one (see background section), trying to combine the two approaches.

In the rest of this section folksonomies are, following this direction, compared to other existing methodologies with the aim to define possible relationships.

## Thesauri/Ontologies and Folksonomies

A lexical network (a thesaurus or an ontology) may be used in the direction of a folksonomy, in order to provide:

- author with a list of controlled words; he/she has to choose a tag in a defined list of accepted words;
- user with a lexical resource useful for comparing used words with the terms accepted in the vocabulary and solving synonymic situations;
- a useful way for determining the hierarchical level of a term and defining the related words.

But to use a restricted formal vocabulary let some problems emerge. Natural languages evolve rapidly and the use of *closed vocabularies* produces situations in which some proper names, slang expressions, metaphorical usage of terms could not appear in them in time to be used when necessary.

In natural languages, and analogously in social tagging, we find:

- variations (masculine/feminine or singular/plural),
- spelling mistakes
- spelling variations

Some NPL (*Natural Language Processing*) techniques could help: in the pre-processing of tags stemming or tokenization are used in order to extract the root of a term, giving lists for alternative/variation spelling in the phase of matching. Moreover it is interesting to maintain the inflected form in order to verify if different grammatical forms are related to different meanings.

But the most meaningful problem is represented by *ambiguity*, intrinsic in a word or determined by the use of the language: neologisms, proper names of contemporary phenomena, and metaphorical uses of lexical units.

Let's give an example: searching for the "Paris Hilton".

In this situation the ambiguity exists in the meaning of the tags; in this case, some considerations are useful:

- All social tagging systems mix the name of hotel chain (Hilton) in the capital of France with the blonde starlet.
- Even if we extend the searching to "Paris Hilton Hotel", the search engines still mix them up without warnings.
- Since the relevance of results is given by frequency, news about the starlet predominate those about the Hotel chain.

## Enumerative Classification Systems and Folksonomies

Hierarchical classification systems present some limitations (Quintarelli, 2005) when they are used both for organizing knowledge and for determining the position of a concept in a hierarchy.

The main issue regards the possibility that an item does not fit exactly inside one and only one category. But we have also to deal with the evolution of language, culture, knowledge, and the update of an existing classification system is an expensive operation. Finally, categories are too rigid and above all static vs. the fluidity and the evolution of language.

One interesting solution could be studied in order to integrate the terms found in folksonomies in existing classification systems. But we have also to deal with the possibility that a taxonomy, or generally a hierarchical subject relationship, emerges from terms used in folksonomies by different users as regard to the same resource described (Kome, 2005, Heymann Garcia-Molina, 2006)

## Facets and Folksonomies

The organization of concepts in an indexing language works on two different levels: on the *semantic level*, in which each concept is considered like a single concept and on the *syntactic level* in which each concept is considered like an element of a combination of concepts.

Two typologies of relationships exist: the *semantic* relationship that is the link among a concept and the more general, the more specific and the similar one (that is synonymic, hierarchical and associative relationships) and *syntactic* relationship that create subject string for represent composite subject. In this latter case we need role specification that is define a category, or a facet, the term belong to (see section "Add qualification"). But in general classification systems do not specify the semantic relationship between the concept and the category. Faceted classifications can specify the role of the tag, but still would not be useful to distinguish among documents where the role of a term would be the same (for instance ambiguities in the terms used for the facet *subject* of the document): they would be able to distinguish a document *about* Paris Hilton from a document *by* Paris Hilton, but still would not be useful to distinguish between a documents about the "Paris Hilton hotel" and the "Paris Hilton person" (since the role of the term would be the same). Some kind of relationship has to be defined.

## Integrate Ontologies and Facets with Folksonomies

Ontologies have a huge potential to improve information organization, management and understanding; knowledge, structured in ontologies, can be processed in a more efficient way allowing more elaborated conclusions.

The complementary features of ontologies and folksonomies justify several works aimed at ontologizing folksonomies: the hope is to take advantage of the combination of the formal, precise and explicit specification of a shared conceptualization provided by ontologies with the usability, flexibility and ease of folksonomies.

Different methodologies and approaches are used in literature. Some works simply extend ontologies in a folksonomy-like approach (Bateman et al., 2006). Other works add *multiple labels* to ontology nodes (Maedche, 2002). Another line of research is concerned with extracting basic semantic relations from folksonomies. Some of them (e.g. Mika, 2005, Van Damme et al., 2007, Specia and Motta, 2007) are based on the association of tags to terms belonging to *lexical databases*. An example of lexical database is WordNet (http://wordnet.princeton.edu); in it "nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations". Relatively to our example (see "Facet and folksonomy"), since WordNet knows that "hotel" is a type of building, both "hotel" and "restaurant" are returned when searching for "building".

The use of linguistics and lexical resources help in solving ambiguities but we have to deal with problems related to the use of natural language, since they do not deal with pseudo subclasses (e.g., hotels are not subclasses of Amenities, because there exists hotels that are not amenable); they do not really deal with has-a hierarchies (e.g., they do not know that Paris is the location of the hotel); they do not really deal with instance-of hierarchies, i.e. with individuals and proper names (e.g., they do not know what thing is Paris, nor what thing is Paris Hilton).

Other works (e.g. Echarte et al., 2007) define, on the one side, generic ontology structures in order to represent any folksonomy and, on the other side, an algorithm to obtain an ontology containing the tagged information from the folksonomy itself. The main advantage of this approach is that user annotates using a folksonomic approach, but the system stores such information in an ontology; in this way, two typical problems of folksonomy

are overcome: "tag variability (for example, blog, blogs, blogging) and tags defined in terms of the objective of the tag and not on the content (for example, toread, whilist, etc.)".

Another line of research is concerned with adding more ontology-like features to social tagging, e.g., allowing users to add a specify hierarchies in tags through the use of hierarchical and faceted metadata structures (Yee et al., 2003), which can be added to user generated content. An example in this direction is offered in (Quintarelli et al., 2006) by the ">" character (e.g. Business > hotel > Hilton or France > Paris > Hilton). Basic issues, like polysemy, homonymy and base level variations, are solved in this way, using contextualization and user-added semantic value.

Unfortunately some open issues emerge:

- there is no distinction between the different types of hierarchies. This also means that at each level of the hierarchy, the relationships between concepts are not semantically expressed;
- multiple hierarchies may exist to identify the same terminal values: a concept may belong to different classes (poly-hierarchy);
- there is no identification of pseudo-hierarchies (e.g., showbiz > California > Paris Hilton) and of bogus hierarchies (weapons of mass destruction > blonde > Paris Hilton);
- there is no reference vocabulary for instances (such as proper names) which count for more than 25% of all the tags of documents.
- ontologies do not solve words ambiguity and are not updated on natural language evolution, neither on metaphorical uses of lexical units.

Next section discusses in more detail current open issues and defines some possible research lines for ontology emerging from folksonomies.

## FUTURE RESEARCH DIRECTIONS

**Word Sense Disambiguation** is a problem well recognized and addressed in computational linguistic (Yngve, 1995). But while in computational linguistic the disambiguation can be performed on the neighbouring sentences and words, in folksonomic tags we have almost no context around.

For this reason, the above-mentioned limits impose new and innovative approaches. We are currently experimenting with a few of them.

### Clustering

In order to study the tags behaviour, it is important to do a statistical analysis of tags in order to identify groups, or more appropriately clusters, of related tags. **Clustering** is the classification of objects into different groups, sometimes even overlapping, so that the data in each group (ideally) share some common trait, often expressed as proximity according to some defined measure of distance. In particular, semantic clustering is the clustering of objects based on proximity in their meaning. Through clustering it is possible to determine similarity in meaning based on the contexts according to which the documents are tagged, i.e., by examining not only the individual tag, but also all the tags that are associated to the document, and all the tags that are associated to all documents that include the individual tag.

The distance between tags is then computed by considering the relationships that compose the context of use of these tags. This technique allows us to differentiate each context of use of an ambiguous tag. For instance, "apple" is clustered differently to refer either to a fruit or a company, and is disambiguated by considering whether it is appearing near tags such as "computer" rather than "pie". Consider for instance a theoretical tagging of a document by different users (Table 1)

The fact that these terms all refer to the same document does allow us to infer that their semantic distance is limited, and that in some way at least

*Table 1. An example of document tagging*

| *Tag* | *User* | *Document* |
|---|---|---|
| Kids | Joe Green | Document A |
| Cartoon | Joe Green | Document A |
| Aladdin | Joe Green | Document A |
| Disney | Mary Violet | Document A |
| Cartoon | Mary Violet | Document A |
| Movie | Hugh Orange | Document A |
| Kids | Hugh Orange | Document A |

one meaning of both "Aladdin" and "Disney" belongs to the same neighbourhood of at least one meaning of the word "cartoon", given the fact that this term appears in both tag sets where they appear, i.e. we can infer that they are *clustered* together because of some (unspecified) semantic justification involving "cartoon". A reasonable expectation is also that the *other* meanings of these words are clustered differently, and therefore have different distances between them.

There exists different approaches to tags clustering. Motta and Specia (2007) in their paper show a specific analysis based on tags co-occurrence, in order to find "similarity" of tags. They use two smoothing heuristics to avoid having a high number of these very similar clusters. For every two clusters:

- if one cluster contains the other, that is, if the larger cluster contains all the tags of the smaller one, remove the smaller cluster;
- if clusters differ within a small margin, that is, the number of different tags in the smaller cluster represents less than a percentage of the number of tags in the smaller and larger clusters, add the distinct words from the smaller to the larger cluster and remove the smaller.

That would be an important classification of tags, in a specific prospective, and generates a set of clusters resulting from distinct seeds that are similar to each other. Another possible algorithm that we have considered would be based-on a fuzzy approach. Clustering is *hard* if it produces an exact partition of the data set, as in the case of the Motta and Specia approach, and it is termed *fuzzy* if it produces a fuzzy set covering the data set, whereby objects belong to clusters with a certain degree of truthness expressed as a number between 0 and 1.

In order to talk about disambiguation of polysemic terms we prefer to rely on fuzzy clustering, since hard clustering does not allow any ambiguity, and forces to resolve it automatically by selecting only the *best* cluster for each term and excluding all the others. Fuzzy clustering, on the other hand, allows terms to belong to multiple clusters with different degrees of certainty, and can take semantic ambiguity in consideration.

## Identifying Proper Names

Another approach to disambiguation is to provide a way by which (at least a few) users add structure and depth to social tags. This can be obtained by providing a syntactically simple mechanism to qualify the terms used. As mentioned, a similar mechanism has been proposed (Quintarelli et al., 2005), but limited to expressing *is_a* relations (i.e., the BT/NT generic hierarchies between terms) as pairs of generic/specific tags such as *feline* > *cat* (see section "Thesauri" and "Ontologies").

We intend to concentrate on a different hierarchy, the ***instance_of* relationship** (Fisher, 1998), which connects an instance to a **category**, i.e., a proper name to a common name or an individual to its category. Rather than requiring the author of the tag to identify the immediately broader term of each relevant term, we only expect a categorical term (and, in fact, just about *any* reasonable categorical term) for each proper name (be it of individuals, organizations, places, etc.), such as *person:Paris Hilton* as opposed to *hotel:Paris Hilton*, or *fruit:apple* as opposed to *company:apple*, at the same time expecting any degree of variability in the categorical term, i.e., allowing for variations such as *socialite:Paris Hilton*, *heiress:Paris Hilton*, *inn:Paris Hilton*, *destination:Paris Hilton*, or really any other category, generic or specific, that the mind of the reader comes up with in the spur of the moment.

Such social tags would be exactly composed of exactly two parts, the *category* and the *proper name*. In fact, the relationship *instance_of* only matters for proper names, and the tag author needs only answer the simple questions "Is this a proper name? And if so, what is its category?"

Among the advantages of this approach:

- The *instance_of* social tag has always exactly two levels, and never more. Therefore the categorical term can be chosen from any level of a multi-level *is_a* hierarchy of terms (such as WordNet).
- The *instance_of* social tag easily deals with the fact that no vocabulary of proper names exists, but only of categories. Proper names constitute a hearty percentage of tags in real life folksonomies. A method for devising a meaningful measure of such percentage is under way within our research team, but our initial considerations for sites such as del.icio.us suggests that over 20% of tags are proper names.
- All inferences and experiments in ontology building are always performed on the categories only, and never on the proper names, which are by definition open and are simply rewritten as non-controlled vocabulary.

Also note that a tag separator that is explicitly different from space would allow for spaces to be available in tags, and thus for first name/family name pairs (as well as for city names such as San Francisco and New York) to be recognizable as such and to be considered as single tags rather than as two separate ones.

One of the most interesting key points of proposing a richer syntax for disambiguation in folksonomies is that it is not necessary for all users to adopt it: in fact, it suffices for a few, and actually even just one author to use the syntax, to disambiguate all other associations of the same tag to the same document, even if they keep on relying on the unsophisticated syntax.

## Adding Qualification

**Qualification** can be used to conceptualize the tags of a folkonomy, and to let a real fully-fledged ontology to emerge from the concepts described therewith. The simple addition of a tag in a list is not sufficiently eloquent to determine if it describes facts about the document or about *the content* of the document.

Tags in folksonomies, in fact, are used to describe the subject of the content of the document (i.e., what the document talks about), as well as incidentals about the characteristics of the document, its intended or perceived uses, and the relevance to the author of the tagging.

Consider for instance the list of tags *"DVD release date", "kids", "cartoon", "Disney" "Aladdin"* and *"Christmas presents"*. A human could immediately and reasonably infer that the document associated to this list of tags *talks about* the "DVD release date" for the movie *titled* "Aladdin", which is of *type* "cartoon", *produced* (or *authored*) by "Disney" and that the author of

the tags is interested in it *in relation to* making "Christmas presents" *aimed at* some "kids" (their own, possibly).

In order to qualify correctly the justification and meaning of these tags, a possible solution may be to populate some faceted classification properties such as Dublin Core. For instance, the mentioned tags could populate properties such as, respectively, *dc:subject, dc:audience*, *dc:content-type, dc:creator, dc:title, dc:relation*, and so on. In fact, it is not even necessary to use the Dublin Core properties correctly (in our case, "kids" for *dc:audience* is a bit of a stretch, Disney and Aladdin may be the *dc:author* and *dc:title* of the movie, but most surely not of the document talking about the DVD release date, and "cartoon" for *dc:content-type* is technically wrong) as long as reasonably distinguished qualifiers are used.

Enticing users to qualify their tags can be done in at least two different ways:

- By using a positional organization of the tags, in a similar way to a Colon Classification (see section "Faceted classification systems") on which are based Dublin Core facets.
- By providing them with a specific list-like selector with terms from a controlled vocabulary for at least a few of the facets of the Dublin Core schema.

Faceted qualifications not only allow the association of tags to their category, but they also provide relationships that enable the correct generation of metadata property statements. Metadata plays a role very important in both cases, and also the use of the RDF standard (*Resource Description Framework*, http://www.w3.org/RDF), based upon the idea of making "statements" about Web resources in the form of subject-predicate-object expressions, makes it is possible to associate a computable form of the correct role of each tag of every document.

## Disambiguating Slang Words

Dealing with folksonomies a big problem is to contextualize the tags according to the document they are associated to. This implies (as explained) describing the semantic distance of the tags in relation with other tags used by the different users for the same resource or by the same users for different resources. **Contextualization** also means defining the role of the tag as regard to its specific scope of use in terms of categories, and facets. To correctly assign the terms to their category it is possible to use linguistic resources to associate at least approximately the terms to their context.

Some existing linguistic resources include WordNet (see also section "Integrate ontologies and facets with folksonomies"), a large lexical database of the English language. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (*synsets*), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated through an API or directly within the browser. WordNet, though, provides definitions only for terms belonging to the "official" language, which is often a limited, bowdlerized, averaged view of the multifaceted, multi-localized and ever-evolving language that is really used by people for folksonomic tags.

A large number of tags - again, as per proper names, we cannot provide reliable figures yet, but we notice a visible incidence - does not in fact belong to the view of the English language proposed by WordNet, either because the word simply does not exist in the official language (e.g., *fanfic*) or because the official definition provided does not really match the meaning intended in current or local usage of the word (e.g., *douche bag*). These terms, which we cumulatively call ***slang*** (even if there are subtle distinctions that should be made in using such term) cannot be satisfyingly

catered for by traditional linguistic resources, both because of their often irreverent tone, and because of their frenetic creation and evolution. There are therefore two additional resources we are considering, although way less sophisticated technically than WordNet, that may give hints so as to disambiguate and provide some meaning to terms unreliably described by WordNet:

- *Urban Dictionary* (http://www.urbandictionary.com) is a dictionary of slang with definitions provided by users. For each term it is possible to have different definitions ordered according to credibility or just simply coolness. All slang terms we have encountered so far in folksonomies (except for foreign words) are present in Urban Dictionary with more or less credible definitions. One disadvantage of Urban Dictionary is the level of noise that is present: a large number of terms are really extremely limited in scope (even down to usage within a single US High School) and many definitions are clearly nothing but jokes, exercises in low-level humour, or personal offences, with limited usefulness except possibly for the self-esteem of their compilers.
- *Wikipedia* (http://en.wikipedia.org) is the well-known largest multilingual online encyclopaedia, built collaboratively using Wiki software. Wikipedia articles have been written by volunteers around the world, and nearly all of its content can be edited by anyone with access to the Internet. While much better guarded against humorous exploitation of its definitions, the encyclopaedic rather than linguistic purpose of Wikipedia makes concrete disambiguation of tags quite difficult manually, and impossible automatically: almost every categorical word in English has multiple pages related to it (including people, places, books, records and movies with that term as name or title), and often is associated to a *disambiguation* page to (manually) guide the reader to the actual meaning sought. On the other hand, Wikipedia does provide adequate light to all public personas, all large corporations, all main brands, or all major places whose proper name is used in folksonomic tags, as most of them have a page on Wikipedia, so it is a relevant source of information for disambiguation of such tags.

## CONCLUSION

Metadata represents one of the most popular ways for retrieving relevant information in search engines. The conceptual basis of social tagging is that users' information associated to documents via folksonomies provides a good and reliable set of descriptors of the documents themselves, i.e., social tags are really representative of the aims and content of the documents. The analysis of this "data on data" is fundamental in the new frontiers of the Web, as it aims at establishing a collective knowledge and allowing a global collaboration environment for the production and the description of electronic resources. However, the polysemy of natural language requires us to not get rid of controlled vocabularies already, especially whenever it is necessary to convey meaning through concepts rather than potentially ambiguous natural language words.

In this paper we have presented a collection of works and efforts to bring together formal classification methods and social classification efforts. The path towards joining in a single all-encompassing environment these radically different approaches is still long. We have listed a few of the still unanswered issues (proper names, slang, facets) and proposed a few possible ways to approach them (cluster analysis, syntactical extensions to tags, and socially generated linguistic resources). Of course, the realization and concrete usefulness of these approaches are, as of

now, fully undemonstrated, but we are confident that they will at least be considered interesting initial steps.

We also need to discuss some intrinsic limitations in what we are proposing, that makes solutions harder to implement and exploit. In particular:

- As already mentioned, both Urban Dictionary and Wikipedia are not designed to be used as linguistic resources in automatic engines, but rather as interactive reference tools for humans. Thus, besides the obvious problems of reliability, noise and information overload that their use imply, accessing definition features of the terms (even the simple distinction between common names and proper names) is difficult, error-prone and heavily dependent on NLP algorithms to work on their definitions.
- Clustering algorithms, and in fact any algorithm that attributes relevance to items by considering information available outside of the items themselves, is open to malicious attacks by determined individuals and organizations planning to take advantage of the algorithm. The practice of *edit wars*, *spamdexing*, or *Googlebombing* are clear examples of these kinds of exploitations, and are impossible to deal with in an automatic way (i.e., by the algorithm itself), since any kind of prevention becomes automatically part of the algorithm and as such is open to (possibly different kinds of) further exploitation. Only manual operations on clearly identified attacks can be considered adequate responses to these practices, and they require massive manpower for even a starting and limitedly successful Web service.

It is hard to see a simple solution to these problems, but on the other hand they are shared with a large number of other (and fairly successful) services, which we would never think of giving up to. As such, these problems will make all these services float together or sink together, and solutions found for one will work for all the others.

## ACKNOWLEDGMENT

The authors would like to thank the colleagues and students that have contributed and are contributing to this research. In particular, a big *thank you* goes to Giovanni Rossi, as well as to the *folksonomy folks* (Nicola Di Matteo, Ferdinando Tracuzzi, Barbara Angius and Natalino Mondella) of the department of Computer Science, for their ongoing work and early contributions to these activities. The author would also like to acknowledge the European Thematic Network Project Acume 2 (http://acume2.web.cs.unibo.it/), within which a part of the activities here described are being delivered.

## REFERENCES

Agirre, E., & Edmonds, P. (2006). *Word sense disambiguation: algorithms and applications*. Dordrecht: Springer.

Au Yeung, C. M., Gibbins, N., & Shadbolt, N. (2007). Mutual Contextualization in Tripartite Graphs of Folksonomies. In *The 6th International Semantic Web Conference (ISWC 2007), LNCS (4825/2008)* (pp. 966-970). Berlin-Heidelberg: Springer-Verlag.

Baruzzo, A., Dattolo, A., Pudota, N., & Tasso, C. (2009). Recommending New Tags Using Domain-Ontologies. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, September 15-18, Milan, Italy, vol. 3, pp. 409-412. ISBN 978-0-7695-3801-3, doi http://doi.ieeecomputersociety.org/10.1109/WI-IAT.2009.313.

Bateman, S., Brooks, C., & McCalla, G. (2006). Collaborative tagging approaches for ontological metadata in adaptive e-learning systems. In *Proceedings of the 4th International Workshop on Applications of Semantic Web Technologies for E-Learning (SWEL'06)*. (Lecture Notes in Learning and Teaching, (pp. 3-12). Dublin: National College of Ireland.

Casoto, P., Dattolo, A., Omero, P., Pudota, N., & Tasso, C. (2008). *Accessing, Analyzing, and Extracting Information from User Generated Contents*. Chapter XXVII of this handbook.

Christiaens, S. (2006). Metadata Mmechanisms: From Oontology to Ffolksonomy… and Bback. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *On the Move to Meaningful Internet Systems 2006: OMT 2006 Workshops*, (Vol. 4278, (pp. 199-207). Berlin-Heidelberg: Springer-Verlag.

Dattolo, A., Tasso, C., Farzan, R., Kleanthous, S., Bueno Vallejo, D., & Vassileva, J. (Eds.). (2009). *Proceedings of International Workshop on Adaptation and Personalization for Web 2.0* (AP- WEB 2.0 2009), Trento, Italy, June 22, 2009, CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-485.

Echarte, F., Astrain, J., Cordoba, A., & Villadangos, J. (2007). Ontology of Ffolksonomy: A new Mmodeling Mmethod. *Semantic Aauthoring, Aannotation, and Kknowledge Mmarkup (SAAKM), K-CAP 2007*. Retrieved on September 14, 2008, from http://ceur-ws.org/Vol-289/p08.pdfhttp://ceur-ws.org/Vol-289/p08.pdf

Farrell, S., Lau, T., & Nusser, S. (2007). Building Communities with People-Tags. In C. Baranauskas, P. Palanque, J. Abascal, & S.D.J. Barbosa (Eds), *Proceedings of Human-Computer Interaction - INTERACT 2007, 11th IFIP TC 13 International Conference* (pp. 357-360). Berlin-Heidelberg: Springer-Verlag.

Fisher, D. H. (1998). From thesauri towards ontologies? In W. Mustafa el Hadi, J. Maniez &, S. Pollitt (Eds.), *Structures and relations in knowledge organization: Proceedings of the 5th International ISKO Conference* (pp. 18-30). Würzburg: Ergon.

Golder, A. S., & Huberman, B. A. (2005). The structure of collaborative tagging. *Information Dynamics Lab*. Retrieved on June 10, 2008, from http://arxiv.org/ftp/cs/papers//0508/0508082.pdfhttp://arxiv.org/ftp/cs/papers//0508/0508082.pdf.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198–208. doi:10.1177/0165551506062337

Gordon-Murnane, L. (2006). Social bookmarking, folksonomies, and Web 2.0 tools. *Searcher Mag Database Prof*, *14*(6), 26–38.

Heymann, P., & Garcia-Molina, H. (2006). Collaborative Ccreation of Ccommunal Hhierarchical Ttaxonomies in Ssocial Ttagging Ssystems. (Technical. Report. InfoLab). Retrieved on June 10, 2008, from http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=2006-10&format=pdf& compression=&name=2006-10.pdfhttp://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=2006-10& format=pdf& compression=&name=2006-10.pdf.

Hotho, A., Jaschke, R., Schmitz, C., & Summe, G. (2006). BibSonomy: A Social Bookmark and Publication Sharing System. In *Proceedings of the Conceptual Structures Tool Interoperability. Workshop at the 14th International Conference on Conceptual Structures*, July. Retrieved June 30, 2008, from http://www.kde.cs.uni-kassel.de/jaeschke/paper/hotho06bibsonomy.pdf.

ISO 2788. (1986). *Guidelines for the establishment and development of monolingual thesauri* (2nd ed.). Genevea: International Organization for Standardization.

ISO 5963. (1985). *Documentation: mMethods for examining documents, determining their subjects, and selecting indexing terms*. Genevea: International Organization for Standardization.

Jacob, E. K. (2004, Winter). Classification and categorization: Aa difference that makes a difference. [from http://sils.unc edu/~fu/IR/fulltext/jacob_classification_and_categorization.pdfhttp://sils.unc. edu/~fu/IR/fulltext/jacob_classification_and_categorization.pdf.]. *Library Trends*, *52*(3), 515–540. Retrieved on June 10, 2008.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kome, S. H. (2005). *Hierarchical Ssubject Rrelationships in Ffolksonomies*, Master's thesis, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Maedche, A. (2002). Emergent semantics for ontologies – support by an explicit lexical layer and ontology learning. *IEEE Intelligent Systems - Trends & Controversies, 17(1)*, 78–86.

Manning, C., & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, Ttagging Ppaper, Ttaxonomy, Fflickr, Aacademic Aarticle, TtorRead. In [New York: ACM Press.]. *Proceedings of Hypertext, 2006*, 31–39.

Mathes, A. (2004, December). Folksonomies - Ccooperative Cclassification and Ccommunication Tthrough Sshared Mmetadata. December 2004. Retrieved on June 10, 2008, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdfhttp://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf.

Mika, P. (2005). Ontologies Aare Uus: A Uunified Mmodel of Ssocial Nnetworks and Ssemantics. In Y. Gil, E. Motta, V. R. Benjamins, & M. Musen (Eds.), *Proceedings of the 4th International Semantic Web Conference (ISWC2005)* (pp. 522-536). Berlin-Heidelberg: Springer-Verlag.

Mitkov, R. (2003). *The Oxford Handbook of Computational Linguistics*. Oxford/New York: Oxford University Press.

Morrison, P. J. (2008). Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. *Information Processing & Management*, *44*(4), 1562–1579. doi:10.1016/j.ipm.2007.12.010

Ohmukai, I., Hamasaki, M., & Takeda, H. (2005). A proposal of Community-based folksonomy with RDF metadata. In *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*.

Parameswaran, M., & Whinston, A. B. (2007). Research issues in social computing. *Journal of the Association for Information Systems*, *8*(6), 336–350.

Peckham, A. (2005). *Urban Dictionary: Fularious Street Slang Defined*. Kansas City: Andrews McMeel.

Quintarelli, E. (2005). Folksonomies: pPower to the people. *Proceedings of ISKO Italy-UniMIB mMeeting*. Retrieved on June 10, 2008, from http://www.iskoi.org/doc/folksonomies.htmhttp://www.iskoi.org/doc/folksonomies.htm.

Quintarelli, E., Resmini, A., & Rosati, L. (2006). *FaceTag: Integrating Bbottom-up and Ttop-down Cclassification in a Ssocial Ttagging Ssystem*. Paper presented at the EuroIA Conference, Berlin-Heidelberg, DEGermany.

Schmitz, P. (2006), Inducing ontology from flickr tags. In *Collaborative Web Tagging workshop. Proceeding of the 15th International World Wide Web Conference*. Retrieved June 30, 2008, from http://www.ibiblio.org/www_tagging/2006/22.pdf.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, *34*(1), 15–29. doi:10.1177/0165551506078083

Smith, G. (2004). Folksonomy: sSocial classification. *Atomiq (August 3, 2004)*. Retrieved on June 10, 2008, from http://atomiq.org/archives/2004/08/folksonomy_social_classification.htmlhttp://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

Specia, L., & Motta, E. (2007). Integrating folksonomies with the Semantic Web. *Proceedings of the ESWC 2007, Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, (pp. 624-639). Retrieved on June 10, 2008, from http://www.eswc2007.org/pdf/eswc07-specia.pdfhttp://www.eswc2007.org/pdf/eswc07-specia.pdf.

Spiteri, L. F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*, *26*(3), 13–25.

Spyns, P., De Moor, A., Vandenbussche, J., & Meersman, R. (2006). From folksologies to ontologies: how the twain meet. In R. Meersman, Z. Tari et al. (Eds.), *OTM 2006, LNCS 4275* (pp. 738-755). Berlin-Heidelberg: Springer-Verlag.

Taylor, A. G. (2004). *The organization of information*. Westport/London: Libraries Unlimited.

Van Damme, C., Hepp, M., & Siorpaes, K. (2007). FolksOntology: An Iintegrated Aapproach for tTurning fFolksonomies into oOntologies. *Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, (pp. 71-84). Retrieved on June 10, 2008, from http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdfhttp://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdf.

Vander Wal, T. (2005). Folksonomy Ddefinition and Wwikipedia. *Off the Top (November 2, 2005)*. Retrieved in June 2008, from http://vanderwal.net/random/category.php?cat=153http://vanderwal.net/random/category.php?cat=153.

Veres, C. (2006). The Language of Folksonomies: What Tags Reveal About User Classification. *LNCS (3999/2006). Natural Language Processing and Information Systems* (pp. 58-69). Berlin-Heidelberg: Springer-Verlag.

Weinberger, D. (2007). *Everything is miscellaneous: the power of the new digital disorder*. New York: Times Books.

Wright, A. (2008). *Glut: Mastering Information Through the Ages*. New York: Cornell University Press.

Wu, X., Zhang, L., & Yu, Y. (2006). Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web* (pp. 417-426).

Yee, K. P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image searching and browsing. *Proceeding of ACM CHI 2003*, (pp. 401-408). Retrieved on June 10, 2008, from http://flamenco.berkeley.edu/papers/flamenco-chi03.pdfhttp://flamenco.berkeley.edu/papers/flamenco-chi03.pdf.

## ADDITIONAL READINGS

Yngve, V. H. (1995). Syntax and the problem of multiple meaning. In W. N. Locke and D. A. Booth (Eds.), *Machine Translation of Languages* (pp. 208-26). New York: John Wiley and Sons.

Zhang, L., Wu, X., & Yu, Y. (2006). Emergent Semantics from Folksonomies: A Quantitative Study. *Journal on Data Semantics VI: Special Issue on Emergent Semantics. LNCS(4090/2006)* (pp. 168-186). Berlin-Heidelberg: Springer-Verlag.

## KEY TERMS AND DEFINITIONS

C**ategorization:** The basic cognitive process of arranging into classes or categories. The word classification identifies especially the system used in libraries for describe, with a specific notation, the content of a book. Categorization is a more theoretical theory

**Folksonomies:** Folksonomy is 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 (usually shared and open to others). Folksonomy is created from the act of tagging by the person consuming the information.

**Metadata:** Data that describes other data. The term may refer to detailed compilations such as data dictionaries and repositories that provide a substantial amount of information about each data element. It may also refer to any descriptive item about data, such as a title field in a media file, a field of key words in a written article or the content in a meta tag in an HTML page

**Ontologies:** Definition (computer_science): An ontology is a collection of concepts and relations among them, based on the principles of classes, identified by categories, properties that are different aspects of the class and instances that are the things

**Tags:** A tag is a generic term for a language element descriptor. The set of tags for a document or other unit of information is sometimes referred to as markup, a term that dates to pre-computer days when writers and copy editors marked up document elements with copy editing symbols or shorthand

**Taxonomies:** Taxonomy is the science of classification according to a pre-determined system, with the resulting catalogue used to provide a conceptual framework for discussion, analysis, or information retrieval. In theory, the development of a good taxonomy takes into account the importance of separating elements of a group (taxon) into subgroups (taxa) that are mutually exclusive, unambiguous, and taken together, include all possibilities

**Thesaurus:** A thesaurus is the vocabulary of an indexing language, that is a controlled list of accepted terms. The role of a thesaurus is to specify a preferred term (descriptor) to be use in indexing and to establish relationships between concepts at different levels: define synonyms, specify hierarchies, individuate related terms

**Web 2.0:** Web 2.0 is the popular term for advanced Internet technology and applications including blogs, wikis, RSS and social bookmarking. The expression was originally coined by O'Reilly Media and MediaLive International in 2004, following a conference dealing with next-generation Web concepts and issues

**Web 3.0:** Web 3.0 is defined as the creation of high-quality content and services produced by gifted individuals using Web 2.0 technology as an enabling platform. Web 3.0 refers to specific technologies that should be able to create the Semantic Web.