DE GRUYTER
SAUR

Isabella Peters

FOLKSONOMIES. INDEXING AND RETRIEVAL IN WEB 2.0

KNOWLEDGE & INFORMATION

DE
G

Knowledge & Information

Knowledge & Information

Studies in Information Science

Knowledge and Information (K&I) is a peer-reviewed information science book series. The scope of information science comprehends representing, providing, searching and finding of relevant knowledge including all activities of information professionals (e.g., indexing and abstracting) and users (e.g., their information behavior). An important research area is information retrieval, the science of search engines and their users. Topics of knowledge representation include metadata as well as methods and tools of knowledge organization systems (folksonomies, nomenclatures, classification systems, thesauri, and ontologies). Informetrics is empirical information science and consists, among others, of the domain-specific metrics (e.g., scientometrics, webometrics), user and usage research, and evaluation of information systems. The sharing and the distribution of internal and external information in organizations are research topics of knowledge management. The information market can be defined as the exchange of digital information on networks, especially the World Wide Web. Further important research areas of information science are information ethics, information law, and information sociology.

De Gruyter · Saur

Isabella Peters

Folksonomies

Indexing and Retrieval in Web 2.0

Translated from German by
Paul Becker

De Gruyter · Saur

D 61

∞ Printed on acid-free paper which falls within the guidelines
of the ANSI to ensure permanence and durability.

ISBN 978-3-598-25179-5
ISSN 1868-842X

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Printed in Germany

Printing: Strauss GmbH, Mörlenbach

Contents

Introduction

We used to rely on philosophers to put the world in order. Now we've got information architects. But they're not doing the work – we are.

--Bruce Sterling, Wired Magazine, 2005.

Five years ago, most librarians, archivists, documentalists and information architects would likely not have dared dream that one of the main areas of their daily work would come to gain the recognition it now enjoys in theory and practice, as well as among internet users, and neither could they have predicted that they themselves – the experts – would not be the ones to blaze its trail, but mainly laymen: I am talking about the indexing of digital information resources via 'tags,' or user-generated descriptors. A study conducted by PEW Research Center (Rainie, 2007) revealed that 28% of internet users have already indexed online content with tags, and 7% stated that they do so several times over the course of a typical day online. But why do folksonomies, the collections of user-generated tags, attract such attention? „One cannot help but wonder whether such enthusiasm for metadata would be the same if people were asked to use only prescribed and standardized vocabularies." (Spiteri, 2005, 85)

Folksonomies are part of a new generation of tools for the retrieval, deployment, representation and production of information, commonly termed 'Web 2.0.' In Web 2.0 it is no longer just journalists, authors, web designers or companies who generate content – every user can do so via numerous online services and through various media, such as photos, videos or text. eMarketer (2007) estimates that by 2011, more than 200 million users worldwide will be contributing to the internet's content. But even today, the growth rates of certain online services are impressive: the social bookmarking service del.icio.us has about 90,000 registered users (Al-Khalifa, Davis, & Gilbert, 2007) and spans around 115m posts linking to anywhere between 30 and 50m URLs – and the data pool grows by 120,000 URLs every day. The photo-sharing service Flickr compiles around 2 billion pictures (Oates, 2007), the German social networking service studiVZ has over 5m registered users and Technorati was able to index more than 70m blogs at the beginning of 2007 (Sifry, 2007).

The heavy growth of user-generated content increases the demand for suitable methods and facilities for the storage and retrieval of said content. In order to meet those demands, companies and computer scientists have developed collaborative information services like social bookmarking, photosharing and videosharing, which enable users to store and publish their own information resources as well as to index these with their own customised tags. Thus the indirect co-operation of users creates a folksonomy for each collaborative information service comprised of each individ-

ual user's tags. Using this folksonomy, all users may then access the resources of the information service in question.

The production of user-generated content, the development of collaborative information services and the usage of folksonomies, as well as the popularity of these three aspects, are mutually dependent. For one thing, this means that the more collaborative information services there are, the more user-generated content, the more tags in folksonomies there will be (see Figure I). At the same time, the growing number of user-generated information resources necessitates ever more collaborative information services in order to store them and folksonomies to index and make them retrievable. Also, a collaborative information service's success, and by the same token its usefulness, increases proportionally to the number of users producing and indexing their own information resources, since this creates more varied access paths as well as greater quantities of different resources.

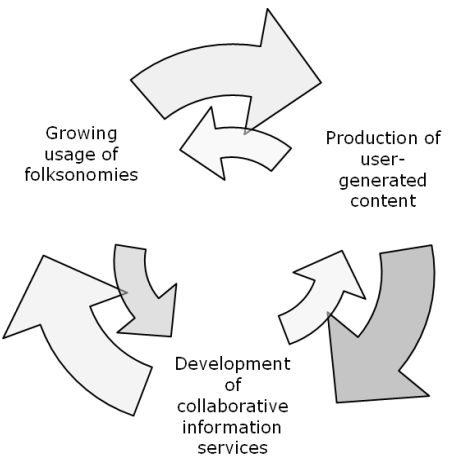

Figure I: The Production of User-generated Content, the Development of Collaborative Information Resources and the Growing Usage of Folksonomies are Mutually Dependent.

The task of folksonomies is to create access paths to information resources using tags. Their perspective on classification systems is an altered one – proponents see this as the crucial difference to knowledge organization systems and documentation languages from library science and professional information services.

To clarify, let's look at some examples from everyday life: suppose you've just moved to a new flat and want to organise your bookshelves systematically. First you must choose the classification system that's right for you – you can arrange the books alphabetically by title or by author, stack them according to size or spine colour, or separate them thematically, by genre. You could create different sections for novels with male and female protagonists, or create an order based on the date of purchase. The same goes for recipes, which can labelled as either hors d'œuvres, main courses or desserts, or classified according to ingredients, such as meat and fish. You could think of many analogous examples for photos, clothes, e-mails or CDs, and would find that they each share a common feature: once a classification

system is settled on, the book, recipe or photo is allowed exactly *one* place on the shelf or in the folder – whatever the taxonomy – where the resource is to be found.

The problem which arises here is that a systematic order of any physical or digital resources always requires a decision on the part of the "guardian of order" as to which of the resource's properties is to be used as a classification criterion. And only one criterion may be selected, since you cannot arrange books alphabetically on a shelf by author *and* title. The guardian must also take care to shape the classification system in such a way as to be able to quickly find the required resources as needed. Here we encounter a problem touching upon the system's usability: an order that appears logical and practical to one user might cause utter confusion in somebody else. A solution for both problems might be the multiple storage of one and the same resource, but that would be neither efficient nor practicable – apart from costing money and space, since most books would have to be bought and stacked at least twice.

Folksonomies take a different approach in the classification and structuring of digital information resources. Instead of choosing a classification criterion and filling it with resources, it is now the resources that are allocated the criteria. Folksonomies turn the classification system from a criteria-centric into a resource-centric approach. This means that multiple storage no longer refers to the resources but to the multiple allocations of the 'folders,' 'drawers' or 'shelves' that are the tags of the folksonomy. Pinned to the information resources are as many tags as are necessary to adequately describe and retrieve them. Thus tags enable the most diverse criteria to be allocated to the resources and in this way guarantee a much broader access to them, which, due to the collaborative construction of the folksonomy, is also independent of the guardian. In the digital world, however, this approach always requires an indexing and retrieval system to render the folksonomy-based classification system manageable. The user may have created numerous access paths to the information resources, but a system will be needed to aggregate the tags and so provide links to the desired resources.

So in order to structure and classify resources, folksonomies sidestep onto a meta-level, which represents the resource via (a whole lot of) tags. This approach is neither new nor especially innovative, however. In the physical world, this meta-level is mainly developed and implemented by libraries, with their methods of knowledge representation. Subject catalogues, classification systems and thesauri are all somewhat antiquated testaments to this endeavour. They are to be distinguished from folksonomies in that they use a controlled vocabulary and a term-based approach, to be applied to the resources by trained professionals. This means that the indexing and retrieval of the resources is only possible using this prescribed vocabulary which, however, represents and finds the resource independently of its representation in language. The user of such a controlled vocabulary must commit the allowed terms to heart before gaining access to the resources. Folksonomies, however, allow users to implement their own terminology for indexing and representing content, as well as for retrieving resources, which at first sight enormously facilitates the information retrieval process. Since the allocation of tags to resources is not bound by a unified set of rules, however, a suitable retrieval system is needed in order to efficiently look for and find the resources. Here, too, librarians were the pioneers with their card indexes, that were later replaced by electronic search and data storage facilities, not least on the internet.

Both collaborative information services and folksonomies mainly serve the user to personally store and arrange digital information resources, which are either self-produced or found online, e.g. videos or photos. The coverage and indexing of *all* resources available on the internet is not the user's professed goal. But little by little it transpires that folksonomies have reached a new feasibility level in indexing web resources, which takes its place next to web catalogues as well as the full text storage and analysis of websites' link-topological properties as practiced by search engines. The low entry barriers for folksonomies make this form of mass indexing possible, since they do not require any specialised knowledge on the part of the user. Added together, the numerous activities of single users then lead to a database of web resources indexed by human hands. Thus the burden of mass indexing is no longer shouldered by single institutions, but carried by the many component parts of the internet community, where each user contributes his share:

> Collaborative tagging is most useful when there is nobody in the 'librarian' role or there is simply too much content for a single authority to classify; both of these traits are true of the web, where collaborative tagging has grown popular (Golder & Huberman, 2005, 198).

> The hierarchical Yahoo directory was developed for the purpose of browsing, but categorization by a limited number of professionals cannot practically deal with the huge number of web pages. [...] Folksonomy seems to be able to deal with the large amount of content (Ohkura, Kiyota, & Nakagawa, 2006).

> As a result, it leads to an emergent categorization of web resources in terms of tags, and creates a different kind of web directory (Choy & Lui, 2006).

It transpires that folksonomies are a method of knowledge representation in the exact same way that libraries' traditional controlled vocabularies are – minus a centralised administration for the vocabulary. Nevertheless, folksonomies aim for an optimised representation and retrievability of information resources. But unlike the established, term-based documentation languages and knowledge organization systems, folksonomies have a few weak spots. Since they forego a restricted terminology, they are confronted by all of language's problems and idiosynchrasies (such as synonyms and homonyms), which come to bear on the process of information retrieval especially. A search for information resources on the subject of 'disco' might lead to fewer relevant hits, since synonymous terms such as 'night club' are omitted from the search results. Similarly, whoever searches for 'club' will be cluttered with useless information if the user is interested in partying and not, in fact, golfing equipment or sports teams. In certain cases, the user's search will even result in no hits at all, which has immense repercussions: „It's impossible to create knowledge from information that cannot be found or retrieved" (Feldman & Sherman, 2001, 4). The goal of knowledge representation – and thereby, of folksonomies – is to provide access to information resources; non-textual resources such as photos and videos are particularly dependent on them. If there is no access path, the resources are deemed unimportant: „The digital consumer is highly pragmatic – their attitude is that if the information is not found immediately, in one place, it is not worth looking for" (Nicholas & Rowlands, 2008). In today's information gathering, which mainly relies on internet search engines, folksonomies are a weak tool for retrieval since they hardly represent a departure from search engines' full text storage. „Users of tagging systems can quickly label (tag) large numbers of objects, but these labels are much

less informative – tags tell us little more than the free-form string that they present," is Heymann and Garcia-Molina's (2006) fitting summary. Thus folksonomies lag far behind the potential of methods for knowledge representation and information retrieval that stem from library science (Peters, 2006).

Why then are folksonomies so popular and successful? Could they offer advantages and access paths to information resources that controlled vocabularies cannot, and that are not visible at first sight? Or could folksonomies still learn from the traditional methods of knowledge representation and the established approaches to information retrieval, and vice versa?

Current State of Research

Outside of information science, (folksonomy) research is mainly carried out in IT. Here folksonomy-based database and retrieval systems are constructed and evaluated in particular (e.g. BibSonomy), and characteristic frequency distributions for tags, users and resources are calculated. Social and media science mainly deal with the network characteristics of folksonomies, user behaviour in folksonomy-based systems and the effect of folksonomies on media consumption, while linguistics investigates the semantic and pragmatic properties of tags as well as their possibilities in the computational processing of language. Library science researches folksonomies with a particular focus on their usefulness for the indexing and retrieval of mostly physical information resources, as well as their effects on the (self-)image of libraries.

Information science's research is located on the intersection of all these disciplines. Its basis is IT, which is complemented by the other disciplines in order to be able to meet the particular demands of information resources adequately. In folksonomy research, information science's task is to adopt a holistic point of view and to apply the results of the above disciplines on the processing, communication and classification of digital information resources, and to develop conceptions for improved methods of knowledge representation and information retrieval.

All these disciplines are at the beginning of their research endeavours, since most folksonomy-based web services have been developed only from 2003 onwards (e.g. del.icio.us) as part of Web 2.0 – the very term 'folksonomy' was coined as late as 2004 (Smith, 2004; Vander Wal, 2004; Mathes, 2004). Characteristically, interest in folksonomies first manifested itself on the internet, and at first did not appear in scientific publications. Only in 2006, Golder & Huberman, Guy & Tonkin and Marlow et al., amongst others, published studies on the subject that are still highly relevant today and can be deemed essential literature. From that point on, the number of publications on folksonomies has grown exponentially, but they are still mainly to be found in the blogosphere[1] or conference literature. Today we can assume that around 1,000 scientific publications on the subject have been published. A research for 'folksonomy' or 'folksonomies' in scientifically oriented databases yields the following picture:

- Web of Science shows 38 results after a title and keyword search,
- Scopus has 198 hits after a title, abstract and keyword search, and 454 results after a search in all text fields,

[1] 'Blogosphere' describes the totality of all blogs and blog entries available on the internet.

- the ACM Portal boasts 435 results after a search in all text fields, and 105 results after a title and abstract search, and
- InfoData finds 29 results after a search in all text fields.

This book summarises the research findings of more than 700 publications.

Monographic publications are rare, however. Exceptions are the popular science books „Tagging. People-powered Metadata“ by Gene Smith (2008), „Everything Is Miscellaneous: The Power of the New Digital Disorder” by David Weinberger (2007) and soon „Understanding Folksonomy. Catalyzing Users to Enrich Information” by Thomas Vander Wal (2009), as well as publications originating in diploma or doctoral theses such as „Social Tagging. Schlagwortvergabe durch User als Hilfsmittel zur Suche im Web“ by Sascha A. Carlin (2007) and „Tagging, Rating, Posting. Studying Forms of User Contribution for Web-based Information Management and Information Retrieval” by Markus Heckner (2009). Smith's book is directed at readers who wish to implement a tagging system within (non-)commercial and in-house company websites, shows advantages and disadvantages of folksonomies and system features and gives advice on the realisation and implementation of such systems. Moreover, he presents three case studies for different sorts of tagging systems: Social Bookmarking, Media Sharing and Personal Information Management. Citing many examples, Weinberger (2007) gives an overview on the different category and classification systems that exist in the real world and discusses their philosophical as well as their (im-)practical backgrounds. Above all, the shortcomings of such systems with regard to digital information are discussed. Hence, Weinberger grapples intensively with the new possibilities for creating metadata in Web 2.0, tags and folksonomies, and their implications for companies and users. Carlin (2007) uses his diploma thesis mainly to provide a literature survey on the subject and does not offer any original ideas or solutions. Heckner (2009) investigates the tagging behaviour of users in a tagging system and beyond platform borders, and additionally creates a model for the categorisation of tags via their functional and linguistic properties. The results flow towards conceptual thoughts on the architecture and implementation of a folksonomy-based online help system.

Even though Heckner (2009) provides the first monographic publication on the subject of folksonomies from the perspective of information science, there exists as of yet no scientific work investigating folksonomies as both a method of knowledge representation and as a tool for information retrieval. This book aims to close this gap.

Open Questions in Folksonomy Research

Although the number of writings on the subject of folksonomies increases steadily, some subareas, particularly from an information scientific point of view, remain uncharted.

Unavailable as of yet is a taxonomy of services in Web 2.0 that differentiate between collaborative information services (which use folksonomies and aim to manage information resources) and social software (which also includes resource management, but additionally includes the construction of a knowledge base and the communication with other internet users). Such a taxonomy is sorely needed if the definitional parameters for the observation and usage of folksonomies in Web 2.0 are to be provided and standardised. Moreover, neither a critical analysis of the ap-

plicability of folksonomies within specific collaborative information services, nor an evaluation with regard to the methods of knowledge representation and information retrieval currently in use has so far been carried out.

Summarising statements concerning the applicability, as well as the strengths and weaknesses, of folksonomies in knowledge representation have been published on several occasions, and will be cited here, yet comprehensive solution statements on how to avoid or neutralise the disadvantages are few and far between. A possibility might be the use of semiautomatic processes for cleansing and unifying folksonomies during and after the indexing of information resources (e.g. via natural language processing), as well as the projection of documentation languages and knowledge organization systems on folksonomies.

Here the question arises as to how preexisting knowledge organization systems on can profit with regard to their terminology and to the term relations applied by folksonomies, respectively how this knowledge can reenter the folksonomies.

Collaborative information services and folksonomies are basking in success at the moment; the number of users keeps increasing. Is it conceivable that folksonomies or collaborative information services might one day collapse under the sheer weight of users, tags or resources? No research has yet been expended on this problem, perhaps due to the topic's youth. Answers might be found in network economics and adjacent areas.

In the scientific debate it is almost always assumed that the allocation of tags to resource follows an informetric distribution or power law. Is this correct, or might other forms of distribution apply to the tags?

The area of information retrieval via folksonomies has only entered the scientific debate in the past three years. In what ways resource gathering via folksonomies may be implemented, which retrieval strategies play a role in this process and how a search via folksonomies is different from information retrieval using traditional search tools are therefore questions that have been examined but not yet summarised. A generally accepted algorithm for relevance ranking in folksonomy-based retrieval systems is missing altogether.

There is also no overview of folksonomies' disadvantages in information retrieval with its subareas retrieval effectiveness, search interface design and search result visualisation, as well as in relevance ranking, and no solution statement on how to avoid or neutralise the localised disadvantages. Possible options here would be the use of query tags as indexing tags, a relevance ranking that takes into account resource-inherent properties as well as tags and user activities, the observation of empirical tag distribution on a resource level to increase folksonomies' retrieval effectiveness, or semiautomatic user support during searches that use tools for information retrieval and knowledge representation.

This book will aim to provide conceptual answers to these open questions in folksonomy research, and in so doing comprehensively describe the close relation between the endeavours of knowledge representation and information retrieval. Particular attention will be paid to the reciprocity of these two information scientific research areas and to the effects on folksonomy-based indexing methods and retrieval systems.

Notes on the Book's Structure

Chapter one will provide an introduction to the context of folksonomies. First, a definition of terms will be carried out in order to differentiate 'Web 2.0,' 'social software' and 'collaborative information services,' which are predominantly used synonymously. Folksonomies will be defined as one of the essential components of collaborative information services. Then, several known collaborative information services will be introduced and their functionalities discussed. Here particular attention will be paid to those of the systems' properties that make indexing and research via folksonomies possible. Since collaborative information services are heavily represented online, I will take this opportunity to discuss the totality of information resources, from photos and links up to and including videos. The chapter will close with a summary of all results, which will register the functionalities of information services and their folksonomies as well as their cardinality with regard to the retrieval of information resources.

Chapter two will then provide a short commentary on the basic terms of knowledge representation and information retrieval. I will focus especially on terms which are invaluable for the understanding of the following chapters and of the discussion of folksonomies in knowledge representation and information retrieval. Where possible, folksonomies will already be classified according to their respective contexts as methods of knowledge representation and tools for information retrieval.

Chapters three and four give the book its name, since they deal exclusively with the areas of folksonomies in knowledge representation and folksonomies in information retrieval. Both follow a three-part structure: first the subject area will be introduced and any relevant research that will familiarise the reader with the mode of operation and particularities of folksonomies in the particular subject area cited. The first part then closes with a critical evaluation of folksonomies, discussing its advantages and disadvantages as a tool in knowledge representation and information retrieval. The second part of the chapters is tasked to solve folksonomies' localised problems and disadvantages. These solution statements are of a conceptual nature and have mainly not been implemented in practice or only vaguely investigated, respectively. Where possible, however, available research results will be cited to support the argumentation. Both chapters then close on a forecast that introduces further relevant areas of research and identifies research questions left unanswered.

The third chapter is wholly dedicated to folksonomies and their task as a method for knowledge representation. First the term 'folksonomy' will be more closely defined and the difficulty concerning the competing terms for this idea discussed. Then follow a description of the three intrinsic parts of a folksonomy ('users,' 'tags' and 'resources'), and an observation on the cognitive effort that indexing via folksonomies demands of users. The differentiation of folksonomies into three variants ('broad,' 'narrow' and 'extended narrow folksonomy'), as well as illustrations on collective intelligence in groups of people are necessary to explain the formation of different tag distributions on a resource level. While tag distributions arise mainly through the statistical accumulation of user activity after indexing, the section 'Tagging Behaviour' will address actual user behaviour during indexing. The properties of the tags themselves and a possible allocation to different sorts of tags will be examined after. While the three component parts of a folksonomy and their interconnections can be exploited for different recommender systems, here light will be shed on the proposal of tags during indexing. On the basis of the preceding observations,

advantages and disadvantages of folksonomies will be deduced and summarised. This critical evaluation forms the basis of the chapter's latter sections, in which folksonomies are compared with traditional methods of knowledge representation, and where semiautomatic strategies meant to increase folksonomies' effectiveness in indexing are presented. Then methods that will minimise disadvantages and optimise advantages are developed and presented. The goal is an improved form of knowledge representation that will feed on folksonomies' strengths without however forgetting about the virtues of established indexing methods (see 'tag gardening'). The chapter ends on a forecast on the further possibilities of knowledge representation via folksonomies and research questions left unanswered.

The fourth chapter will discuss folksonomies as tools for information retrieval. First, the relation between knowledge representation and information retrieval will be examined, with a particular focus on folksonomies as access vocabulary during research. After that, the different retrieval strategies ('searching,' 'browsing' and 'retrieving') that can be implemented via folksonomies are illustrated. In the same context, the differentiation between active and passive information retrieval will be discussed. Active information retrieval distinguishes itself through its use of tags as information filters and thus as search terms during retrieval Information filtering and collaborative filtering stand for passive information retrieval, which uses tags to formulate search requests or as the basis of a recommender system. As opposed to the active retrieval strategies discussed in the preceding section, the user is 'delivered' possible relevant resources by the retrieval or tagging system. The recommender systems (collaborative filtering) making use of folksonomies and their three-part structure of users, tags and resources will be discussed in particular detail. An explanation will be given as to the role folksonomies play in resource gathering, and what sorts of information might further be retrieved in this way. A whole section will deal with folksonomies' retrieval effectiveness, since it can elucidate whether folksonomies as a method of knowledge representation can also serve as a tool for information gathering. What follows is a discussion of several of folksonomies' visualisation techniques, especially tag clouds, as well as their usefulness as facilities for efficient information retrieval. After this comes a summary of folksonomies' disadvantages in information retrieval as localised in the preceding sections. This will form the basis of the four following subject areas that deal with neutralising these disadvantages. The first suggestion is to supplement the tags added to the resources during indexing by more tags drawn from search requests. Relevance ranking makes the decision whether a resource is relevant for a search or not easier for the user. Folksonomies offer query-, user- and resource-specific properties that may be exploited for relevance ranking and which are introduced in this section as ranking algorithms. Users' indexing activities may be statistically analysed and equally exploited for information retrieval. Popular tags, that is tags with a high indexing frequency, are in this context considered 'power tags' and localised in a fictitious retrieval system, where they can be used as tools for restricting the number of search results. The conception and procedure of this approach will be explained in this section. Chapter three will have already discussed tag gardening for knowledge representation, so at this point the concept will be projected onto information retrieval. Research questions still unanswered, as well as possible further uses for folksonomies in information retrieval, will be considered in the forecast.

In the conclusion, the points made in the preceding chapters will be examined, and the book' contribution to folksonomy research considered.

Bibliography

Al-Khalifa, H. S., Davis, H. C., & Gilbert, L. (2007). Creating Structure from Disorder: Using Folksonomies to Create Semantic Metadata. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies, Barcelona, Spain* .

Carlin, S. A. (2007). *Social Tagging. Schlagwortvergabe durch User als Hilfsmittel zur Suche im Web - Ansatz, Modelle, Realisierungen.* Boizenburg: Hülsbusch.

Choy, S., & Lui, A. K. (2006). Web Information Retrieval in Collaborative Tagging Systems. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong* (pp. 352–355).

eMarketer (2007). *UGC Users Outnumber Creators: Are Enough People Recording Their Cats?* from http://www.emarketer.com/Article.aspx?id=1005081.

Feldman, S., & Sherman, C. (2001). *The High Cost of Not Finding Information,* from http://web.archive.org/web/20070320063457/http://www.viapoint.com/doc/IDC+on+The+High+Cost+Of+Not+Finding+Information.pdf.

Golder, S., & Huberman, B. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science, 32*(2), 198–208.

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying Up Tags? *D-Lib Magazine, 12*(1).

Heckner, M. (2009). *Tagging, Rating, Posting. Studying Forms of User Contribution for Web-based Information Management and Information Retrieval.* Boizenburg: Hülsbusch.

Heymann, P., & Garcia-Molina, H. (2006). *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems: InfoLab Technical Report,* from http://dbpubs.stanford.edu/pub/2006-10.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can Social Bookmarking Improve Web Search? In *Proceedings of the International Conference on Web Search and Web Data Mining, Palo Alto, California, USA* (pp. 195–206).

Mathes, A. (2004). *Folksonomies - Cooperative Classification and Communication Through Shared Metadata,* from www.adammathes.com/academic/computer-mediated-communication/ folksonomies.html.

Marlow, C., Naaman, M., d. boyd, & Davis, H. (2006). HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *Proceedings of the 17th Conference on Hypertext and Hypermedia, Odense, Denmark* (pp. 31–40).

Nicholas, D., & Rowlands, I. (2008). In Praise of Google. *Library & Information Update, December*, 44–45.

Oates, G. (2007). *Heiliger Bimbam!,* from http://blog.flickr.net/de/2007/11/13/heiliger-bimbam/.

Ohkura, T., Kiyota, Y., & Nakagawa, H. (2006). Browsing System for Weblog Articles based on Automated Folksonomy. In *Proceedings of the Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland*.

Peters, I. (2006). Against Folksonomies: Indexing Blogs and Podcasts for Corporate Knowledge Management. In *Preparing for Information 2.0. Proceedings of Online Information Conference, London, Great Britain* (pp. 93–97). London: Learned Information Europe Ltd.

Rainie, L. (2007). *28% of Online Americans Have Used the Internet to Tag Content. Forget Dewey and is Decimals, Internet Users are Revolutionizing the Way We*

Classify Information – and make Sense of It, from http://www.pewinternet.org/~/media//Files/Reports/2007/ PIP_Tagging.pdf.pdf.

Sifry, D. (2007). *The State of the Live Web, April 2007,* from http://technorati.com/weblog/2007/04/328.html.

Smith, G. (2004). *Folksonomy: Social Classification,* from http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

Smith, G. (2008). *Tagging. People-powered Metadata for the Social Web.* Berkeley: New Riders.

Spiteri, L. (2005). *Controlled Vocabulary and Folksonomies,* from http://www.termsciences.fr/IMG/pdf/Folksonomies.pdf.

Sterling, B. (2005). Order Out of Chaos. *Wired Magazine, 13*(4), from http://www.wired.com/wired/archive/13.04/view.html?pg=4.

Vander Wal, T. (2004). *Feed On This,* from http://www.vanderwal.net/random/entrysel.php?blog=1562.

Vander Wal, T. (2009). *Understanding Folksonomy: Catalyzing Users to Enrich Information.* Beijing [u.a.]: O'Reilly Media [to appear].

Weinberger, D. (2007). *Everything is Miscellaneous. The Power of the New Digital Disorder.* New York, NY: Times Books [u.a.].

All online sources cited here and over the course of this book have last been accessed on March 25th, 2009.

Chapter 1
Collaborative Information Services

This chapter will provide a systematic description of collaborative information services and their particular tagging and search functionalities. Before describing the web services, however, it is absolutely necessary to differentiate the various concepts within this subject area from one another.

Web 2.0 vs Social Software vs Collaborative Information Services

The context that houses collaborative information services is termed '**Web 2.0**.' Web 2.0 spans all activities and technical requirements that allow users of the World Wide Web to self-publish content, in the form of profiles, bookmarks, photos, videos, posts etc. and to make it accessible to other users, as well as to communicate with them. Furthermore, 'Web 2.0' describes a state of affairs in which the internet distinguishes itself through the continuous development and combination ('mash-up') of known online services and incorporates users in this development process via feedback loops (Lange, 2006). This state is also called 'perpetual beta' (Cox, Clough, & Marlow, 2008).

The term 'Web 2.0' was created as the name for a conference, held by O'Reilly Media Inc.[2], in which the latest developments online were to be discussed (Notess, 2006a). Since then, the suitability of this term has been the subject of ample and contentious debate (Dvorak, 2006; Shaw, 2005; Notess, 2006a; Millard & Ross, 2006). Nevertheless, Web 2.0 has established itself as a concise buzzword in daily language (Sixtus, 2006; Braun & Weber, 2006; Röttgers, 2007; Schachner & Tochtermann, 2008, 23ff; Gissing & Tochtermann, 2007). O'Reilly (2005) himself presented the ideas behind the concept of Web 2.0 in a meme map (see Figure 1.1), in which both innovations in users' online behaviour as well as the continuing development of the technical basis were illustrated. The most frequently cited idea is the 'architecture of participation,' which puts the main emphasis on user-generated content as well as online communication. Ankolekar, Krötzsch and Vrandecic (2007) concentrate on software development: "It is notable that the term Web 2.0 was actually not introduced to refer to a vision, but to characterise the current state of the art in web engineering" (Ankolekar, Krötzsch, & Vrandecic, 2007).

[2] http://www.oreilly.com.

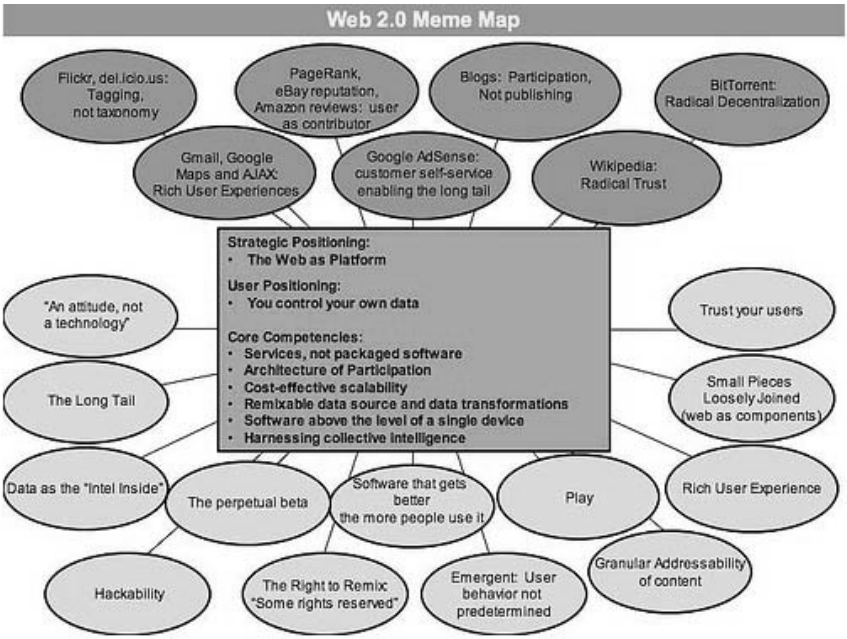

Figure 1.1: Meme Map of the Concept 'Web 2.0' after O'Reilly. Source: O'Reilly (2005, Fig. 1).

To declare the existence of a second version of the World Wide Web presupposes the existence of a previous version (Tredinnick, 2006; Notess, 2006a). Cormode and Krishnamurthy (2008) extensively discuss the differences between 'Web 1.0' and 'Web 2.0' (see also O'Reilly, 2005). They localize these differences on the three pivots of:

1. technology,
2. structure,
3. social aspects.

Online services and websites in 'Web 1.0' are rather static on a technological and structural level, may not be adjusted by users' sites and mainly serve as sources of information, while Web 2.0 services especially focalize communication and exchange of resources between users. They demand little technological know-how on the part of the user, since Web 2.0 services separate content, navigation elements and technology from each other (Tredinnick, 2006), adapt to the user's needs with regard to the design, structure, processing and creation of their content, in other words: because they put the user at the center of the online service:

> The key to Web 2.0 is harnessing the ways in which users use information to add value to information (either through direct or indirect user-participation) in creating the information sources they use. In other words, Web 2.0 reflects collective use over time, rather than reflecting an organization's preferred view of itself. Web 2.0 is built out of real use and need, not idealized use and need (Tredinnick, 2006, 232).

That is why Danowski und Heller (2006) also speak of a change in the World Wide Web from an object-centric to a person-centric network (Danowski & Heller, 2006; Schmitz et al., 2007).

This book will proceed from a definition of the term 'Web 2.0' that is represented in Figure 1.2. Here the term 'Web 2.0' is composed of the three areas of technology, respectively AJAX and RSS, licences and social software. From a technological perspective facilitates the creation and publication of content, which is why Coates (2003) calls Web 2.0 "the Mass Amateurisation of (Nearly) Everything": "Updating a website on a daily basis is no longer an activity that only a trained professional (or a passionate hobbyist) can accomplish. It's now open to pretty much everyone, cost-free and practically effortlessly" (Coates, 2003). For one thing, the technology of Web 2.0 tools recedes so far into the background behind an intuitive user interface that even the most technically unversed user may create and publish information resources in this way.

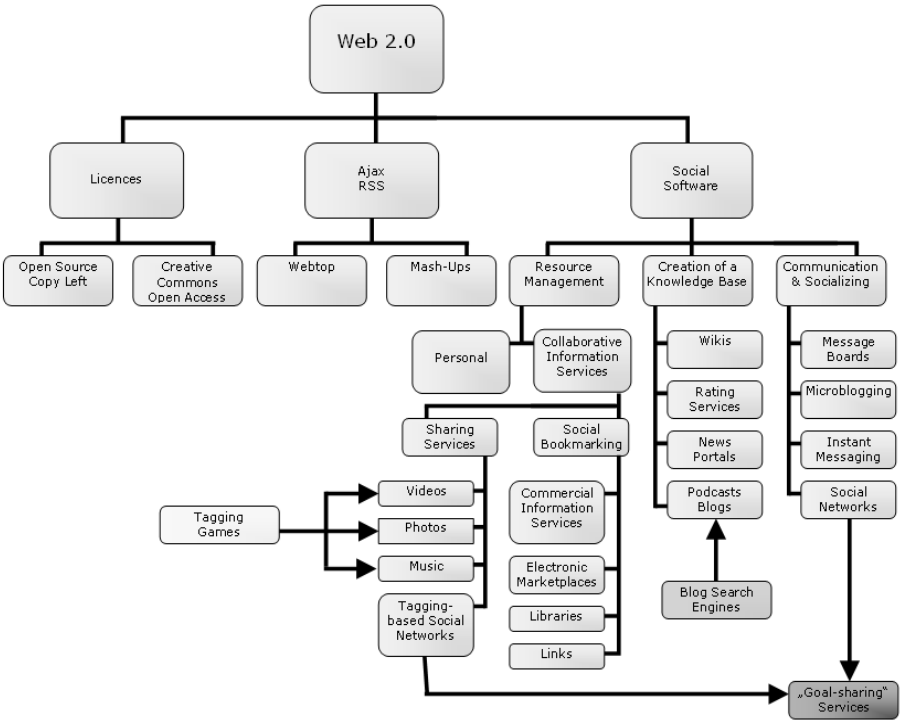

Figure 1.2: Classification of Services, Technologies and Licences in Web 2.0.

Then, in the sense of 'perpetual beta,' the technology is made accessible to users for processing in so-called 'mash-ups' or for corrections (van Veen, 2006; Zang, Rosson, & Nasser, 2008; Lackie & Terrio, 2007). Both processes are implemented via the so-called 'Application Programming Interfaces' (short 'API'). Weiss (2005) explains the concept of APIs with the example of the photosharing service Flickr:

> The API, or Application Programming Interface, is a set of technical documentation which tells a developer how to interact with a software engine. [...] Flickr published its API so that Web developers anywhere could write their own applications to leverage the Flickr photo and tag database (Weiss, 2005, 22).

Zang, Rosson and Nasser (2008) write about interviews conducted with programmers of mash-ups, in which they found out that 77.8% of all mash-ups consist of map materials, 40.7% of photos and 14.8% of news items. The most commonly used APIs are Google Maps'[3] (96.4%), Flickr's (39.3%) und Amazon's (14.3%).

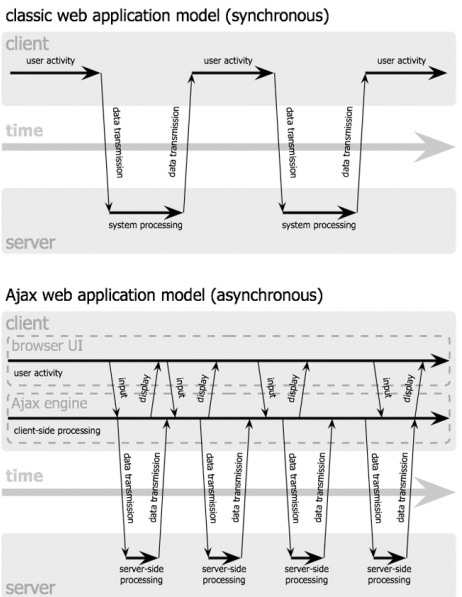

Figure 1.3: The AJAX Concept. Source: Garrett (2005, Fig. 2).

Another important technological component of Web 2.0 is its capacity for asynchronous data transmission using AJAX (short for 'Asynchronous JavaScript and XML'; Garrett, 2005; Crane, Pascarello, & James, 2006; Clark, 2006). AJAX does not signify a new programming language in this context, but describes instead a novel connection of familiar programming techniques, as Garrett (2005) emphasizes:

> Ajax isn't a technology. It's really several technologies, each flourishing in its own right, coming together in powerful new ways. Ajax incorporates:
> standards-based presentation using XHTML and CSS;
> dynamic display and interaction using the Document Object Model;
> data interchange and manipulation using XML and XSLT;
> asynchronous data retrieval using XMLHttpRequest;
> and JavaScript binding everything together (Garrett, 2005).

[3] http://maps.google.com.

This connection affects users' interaction with online services in particular and conveys the impression of a desktop application that processes the user's requests almost without delay (see also Figure 1.3):

> The Ajax engine allows the user's interaction with the application to happen asynchronously – independent of communication with the server. So the user is never staring at a blank browser and an hourglass icon, waiting around for the server to do something (Garrett, 2005).

AJAX transforms the desktop into a 'Webtop' (Lange, 2006) that grants users the decentralized storage of resources as well as access to these resources from every computer with internet connection (Notess, 2006b). Storage does not require any hard drive capacity, and searching and installing software updates is no longer necessary since changes to the program are implemented online. But the greatest advantage of these online applications is surely that multiple users can access and edit the same resource. The different versions are often saved during the process itself, so that a return to earlier incarnations of the resource may always be loaded.

So-called RSS[4] or Atom Feeds (Wusteman, 2004; Hammersley, 2005) allow for a new form of information distribution and platform-independent exchange of data:

> RSS is a way of syndicating web content through the use of content feeds, which consists of XML marked-up files. RSS feeds usually combine either the lead paragraph, or a summary of an article published on the web or on a blog, and a hyperlink back to its resource (Tredinnick, 2006, 230).

They abridge the information to be transmitted to its minimum and provide for it to be sent directly from its creator to the users, where the latter may decide for themselves which feeds they want to receive at all and what information in particular. Wusteman (2004) explains the difference between RSS and Atom:

> All versions of RSS and Atom are written in XML. The main distinction between them is whether or not they are based on the W3C [World Wide Web Consortium, Anm. d. A.] standard RDF (Resource Description Framework). RSS 1.0 is based on RDF; all the others are not (Wusteman, 2004, 408).

Feeds are received via a feed reader (e.g. Bloglines[5] or NewsGator[6]), which bundles all incoming feeds, not unlike an e-mail program, and displays them for the user. Selecting one such message will then lead directly to the original website, where the full text may be accessed. RSS or Atom feeds may also be integrated into other websites, so that with each visit of the site the new messages will be bundled and displayed anew.

Another decisive factor for the (further) development of mash-ups and software are legal aspects in Web 2.0. Licences such as 'Copy Left' (Stallman, 2004) or 'Open Source' (Grassmuck, 2004) as well as 'Creative Commons' (Lessig, 2003; Dobusch & Forsterleitner, 2007) provide for the collaborative editing and subsequent use of intellectual property which would otherwise be protected by copyright law (Federal Ministry of Justice, 2008). In this context, open-source licences apply to software with a disclosed source code, creative commons may however be used

[4] RSS is short for ‚Really Simple Syndication.'
[5] http://www.bloglines.com.
[6] http://www.newsgator.com.

on any resources. The modular structure of creative commons licences in particular allows authors to accurately determine the further usage of their work, e.g. to allow it to be edited but prohibit any commercial activity.

The social aspect is much more pronounced in the world of Web 2.0 as well, since the exchange of information between users and the accumulation of extensive contacts is heavily valued. It must, however, be stressed that for the first time since the creation of the internet, the user ego emphatically takes center stage: blogging software allows users to create and publish diary-like commentaries, they can post updates of their current status around the clock in social networks (e.g. "Kate is… on her way to the gym!"), online merchants count on users' readiness to actively improve and complete product information, and homemade videos, published on video platforms, attain cult status and propel their makers to great fame. The consumer press sees in this the change from average media consumers to self-determined 'prosumers,' to cite Toffler's (1980) fusion of 'producer' und 'consumer':

> Viewed most optimistically, the new sharing driving today's Internet evolution could lead the way to a truly democratic network, where producers and consumers are one and the same (Weiss, 2005, 23).
>
> The new 'I-media' have visibly changed the face of the World Wide Web in a few short years. Everybody is a potential media producer now: as columnist, diarist, critic, expert, like-minded individual, product tester (Wegner, 2005, 94).
>
> The internet has become a chaotic and colorful marketplace that's open to everyone, where you can sit in the stands or play on the stage according to your whims […] Their concept is wholly different from that of earlier internet pioneers. They do not regard their audience as passive 'users,' but as creative, communicative authors and creators who constantly want to swap ideas. They produce a previously rare and expensive product for free: content (Hornig, 2006, 62f.).

This is the reason the term 'social software' (Bächle, 2006; Stegbauer & Jäckel, 2008; Tepper, 2003; Alby, 2007) forms a certain competition to 'Web 2.0.' Bächle (2006) defines 'social software' as follows: "Social Software is the name given to software systems which support human communication and collaboration" (Bächle, 2006, 121). Lange (2006) differentiates the terms 'Web 2.0' and 'social software' thus:

> Social Software describes programs and applications that allow users to create social networks. A dating site is social software, and weblogs and Wikipedia are. […] They are all programmed for one purpose: they encourage communication, interaction and collaboration between users. That is why they should be easy to get to grips with; users should be able to act as intuitively as possible (Lange, 2006, 16).

This book supports these definitions of 'Social Software,' yet does not see the term as synonymous to Web 2.0; rather, it defines social software as a component part of Web 2.0. Alby (2007) suggests a dichotomy of the term:

> Social Software, in which communication is prized above all (and which generally does not keep records). Social Software, in which there is still commu-

> nication, but which also focuses on content, user-made or user-enriched in some way – the idea of the community is crucial (Alby, 2007, 90f.)

This recommendation will not be investigated further at this point, since it seems insufficiently differentiated to paint a picture of the diversity of applications in the area of social software. Instead, three of social software's main functions will be defined in order to adopt another course in the conceptualization of social software:

- communication and socializing,
- the building of a knowledge base and
- resource management

The first of these aims for the creation of social networks and the communicative exchange with other users of the same software. An important aspect of this is that such communication is not private, as in a one-on-one conversation, but either public (e.g. in message boards) or directed simultaneously to multiple interlocutors (as in microblogging[7] or instant messaging). Social networks provide a public platform where users can make new contacts or deepen pre-existing ones (Alby, 2007; Schmidt, 2007). The second function is a vital part of social software and Web 2.0, since it is a direct reflection of 'architecture of participation.' Users create a knowledge base by publishing so-called 'user-generated content' and making it accessible to other users. This can be text resources (e.g. blogs, wikis or rating services), but also audio or video resources (e.g. podcasts). Here users make their knowledge available to other users, provide help and advice, give explicit recommendations (as in rating services such as Ciao[8]) or speak their mind (Reinmann, 2008; Efimova, 2004; Efimova & de Moor, 2005; Doctorow, 2002). The aggregation and condensation of content from other media (as in news portals such as digg[9]), and of users' own knowledge, as well as the collaboration in the recording of encyclopaedic knowledge (as in wikis such as Wikipedia[10]), can be located within this area. The third function concerns resource management, where users may manage, structure and edit their own resources.

Here it is important to emphasize that the three functions do not have clearly outlined boundaries, and that several functions, or at least their attributes, may be found in a specific application. Thus blogs also abet communication with other users and personal resource management due to their commentary function, which may be enriched by links or photos – nevertheless, their main function remains the archiving of text posts and thus the creation of a knowledge base. The allocations applied in Figure 1.2 only take into account the respective main function of each single application and should not be seen as too rigid.

Social software with an emphasis on resource management may again be divided into two sections: personal resource management and collaborative information services. Personal resource management occurs when the structuring and management of a user's own personal resources takes place in private (e.g. Google's webmail program Gmail, which allows for the tagging of e-mails but can only be maintained privately). **Collaborative Information Services** mainly serve the management of personal resources, but also allow for the collaborative creation of a public database, accessible to each (where necessary, registered) user. Furthermore, collaborative in-

[7] Using twitter, for example: http://www.twitter.com.
[8] http://www.ciao.de.
[9] http://www.digg.com.
[10] http://de.wikipedia.org.

formation services' resources are indexed by the users, that is represented on the platform via tags[11]. That is why collaborative information services may also be termed 'tagging systems' (Smith, 2008b, 6). Smith (2008b) includes personal resource management in his usage of the term, which is not the case for the definition used here. Users collaborate to create an information service which provides access to different resources and is organized by themselves. It is these two aspects which make up the difference to social software applications, the main functions of which are in the service of the creation of a knowledge base. Wikis are created collaboratively, but they are not indexed by users via tags; blogs are represented via tags, but they are not created collaboratively. Common to all collaborative information services is the fact that users can profit from one another's activities (e.g. the publishing of photos), without having to contribute anything towards the information service themselves. However, the information service's usefulness for each individual user increases if he or she will participate in its collaborative construction (McFedries, 2006):

> Each contributor gains more from the system than she puts into it (Ankolekar, Krötzsch, & Vrandecic, 2007).
>
> The commitment of many individuals guarantees the services' success (Elbert et al., 2005).

Millen, Feinberg and Kerr (2006) mention another property shared by all collaborative information services and which mainly comes to bear on information gathering using such platforms – 'pivot browsing': "We call this ability to reorient the view by clicking on tags or user names, 'pivot browsing'; it provides a lightweight mechanism to navigate the aggregated bookmark collection" (Millen, Feinberg, & Kerr, 2006, 112). The term 'pivot' stems from spreadsheet processing, where it describes the function that collects homogenous data sets and displays them in new charts. Here the user may freely choose the data sets as well as the output fields. In collaborative information services, this functionality is achieved via links – all elements in collaborative information services (user – tag – resource) can generally be clicked on by the user. Selecting an element, e.g. the user, leads to a hit list, which can be changed by clicking on another element, e.g. a tag. In this way the user browses through the collaborative information service's database and discovers resources relevant to him or her, more or less by accident. Sinha (2006) uses a metaphor:

> The metaphor that comes to mind for pivot browsing is walking in the forests or some other open space, stopping to smell the pine, taking a break. You get the lay of the land as you walk around. The point is not just the destination, the point is the journey itself (anyone who has wasted time looking around on one of the tagging system will know what I mean) (Sinha, 2006).

Generally, two sorts of collaborative information services may be distinguished:

1. social bookmarking services and
2. sharing services.

This dichotomy will be adhered to throughout this book.

Smith (2008b) differentiates between five different tagging systems:

1. Managing Personal Information (e.g. Email),

[11] Tags are descriptors which are freely chosen by the users. For the detailed description, see chapter three.

2. Social Bookmarking,
3. Collecting and Sharing Objects,
4. Improving the E-Commerce Experience (e.g. Amazon)
5. Other Uses (e.g. ESP Game[12]) (Smith, 2008b, 7ff.).

Since the definition of collaborative information services at hand precludes the purely personal use of tags as a means for resource management, item 1 is untenable for the further elucidations on collaborative information services over the course of this book. The aspects 'Social Bookmarking' and 'Improving the E-Commerce Experience' are summarized under 'social bookmarking services.' The aspect 'Other Uses' will also be illustrated, but not allocated to collaborative information services, since the constituent property of the creation of a knowledge base accessible to all users is not in evidence here.

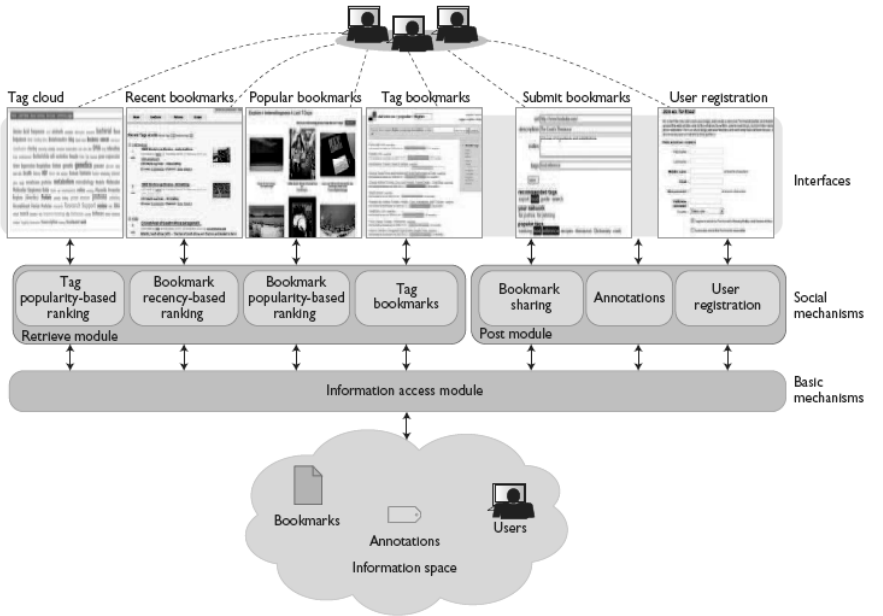

Figure 1.4: The Structure of a Collaborative Information Service, Using the Example of a Social Bookmarking System. Source: Heymann, Koutrika, & Garcia-Molina (2007, 37, Fig. 1).

Social bookmarking services allow users to store (web) resources in their personal user profile, put tags on them and thus render the resources retrievable. The resources were either found online by the users (e.g. links on del.icio.us) or provided by the social bookmarking service. Example for the latter variant are the electronic marketplace Amazon, where users may mark the products with tags, or library catalogs, where users may tag books. This procedure allows social bookmarking services to accumulate the tags thus attached to the resources, and they can now display all tags for a book as well as their frequency of allocation for the user's benefit.

[12] http.://www.gwap.com.

Heymann, Koutrika and Garcia-Molina (2007) exemplarily describe the structure of a social bookmarking service (see Figure 1.4). The creation of an account is the first step a user must take in order to join the information service and to be able to manage resources (see Figure 1.4 'user registration,' upper right-hand corner). Then the user may transfer the desired resource into the system, 'submit bookmarks' and perhaps index it by describing the resource via tags or with further information such as title, content, location ('Annotations') etc. Initially, these are the only actions a user may or must take, respectively, as 'content creator.' Further user activities mainly concern the retrieval of resources (see Figure 1.4, upper left-hand corner), since users mostly gain access to the resources via the interfaces 'Tag cloud,' 'Recent bookmarks,' 'Popular bookmarks' or 'Tag bookmarks.' This means that the user – located in the information platform's information space (see Figure 1.4 bottom center) – can reach the following resources via the interfaces:

> Tag cloud: a list of the most commonly occurring tags in the system;
> Recent bookmarks: a list of the most recent URLs posted to the system by any user;
> Popular bookmarks: a list of the most popular URLs ordered as a function of number of postings and the amount of recent activity for each URL;
> Tag bookmarks: for a given tag *t*, a list of the most recently posted URLs annotated with tag *t*;
> User bookmarks [not shown in Figure 1.4, A/N]: for a given user *u*, a list of the most recent URLs that user posted (Heymann, Koutrika, & Garcia-Molina, 2007, 38).

Sharing services also enable users to create their own profile, but also allow them to upload their own various resources (e.g. photos and videos) to the servers. These resources are then made accessible to the other users, who may view and edit them. The user can provide his or her own resources with tags and thus render them searchable.

Other developments in the area of social software that will not go unmentioned are varieties of collaborative information services that make heavy use of tags and folksonomies and incorporate another main aspect of social software besides resource management. Tagging-based social networks (such as for example 43things[13]) combine the features of collaborative information services and social software and may be termed 'goal-sharing services.' These platforms are meant to bring like-minded people together and to encourage their communication. This is implemented via tags which the users publish on the platform itself. However, they do not upload any resources but merely describe their goals in life (hence the term) via the tags in their profiles.

Social software has not gone without problems or without facing criticism. With regard to social networks, there are serious misgivings about data and user privacy (Gross, Acquisti, & Heinz, 2005; Krishnamurthy & Wills, 2008; Gissing & Tochtermann, 2007 a.o.), wikis and especially Wikipedia must answer questions about suitable criteria of quality (Stvilia et al., 2005; Hammwöhner, 2007 a.o.) and collaborative information services all too often slip into the legal mire due to alleged violations of intellectual property (Leistner & Spindler, 2005; Lohrmann, 2008 a.o.). Even the collaborative creation and publication of great quantities of data in the

[13] http://www.43things.com.

World Wide Web can lead to problems that mainly come to bear on the user's information gathering: "But simply adding more information isn't virtually free. It comes with a cost: noise" (Kalbach, 2008, 36). To find the relevant information in the mass of available information becomes ever more difficult. Schütt (2006) summarizes this problem using the example of blogs:

> But it's hardly begun and already there's talk of a 'blog overflow' because every Tom, Dick and Harry has his own blog, and so the quality of the content seems to be drowning once again in the ocean of information. This cries for artistic limitations or specialized blog search engines that help find the diamond (Schütt, 2006, 33).

This is why tools have been outlined in Figure 1.4 that provide access to the information resources in social software, and which will be explained over the course of this chapter. Technorati[14] would be one such search engine that searches blogs and thus makes heavy use of tags.

The flood of information may also be met by improved access paths to the resources. In information science, this access is generally warranted through the coverage and indexing of the resources; for collaborative information services, this is done via tags and folksonomies, respectively. Figure 1.2 mentions tagging games, which help during the indexing of mainly non-textual resources. Principally, however, the indexing via tags is implemented by the users of collaborative information services themselves, so that once again the amalgamation of producing and consuming users in Web 2.0 becomes apparent: "[Folksonomy, A/N] represents another example of the fuzziness separating consumers and creators on the Web today" (Godwin-Jones, 2006, 10). After all, users use tags (their own as well as others') to gain access to the information resources:

> It may be more accurate, therefore, to say that folksonomies are created in an environment where, although people may not actively collaborate in their creation and assignation of tags, they may certainly access and use tags assigned by other (Spiteri, 2007).

Below, I will introduce various collaborative information services, explain their functionalities and discuss their applicability in the World Wide Web. Filtered from the mass of all information services available online will be the most popular and thus basically the prototypical representatives, where provisions will be made for all distinct resource types, such as photos, videos, music, links etc. In the presentation of collaborative information services, particular attention will be paid to the question of to what degree they use folksonomies for indexing and research purposes. The goal of this analysis is to provide an overview of the applicabilities of folksonomies in collaborative information services.

Social Bookmarking Services

Social bookmarking services have upgraded a function of conventional desktop applications for Web 2.0 by enabling the user to save the web browser's list of favorites online, independently of the desktop, and thus make resources already found more easily retrievable (Gordon-Murnane, 2006, 28): "One of the greatest chal-

[14] http://www.technorati.com.

lenges facing people who use large information spaces is to remember and retrieve items that they have previously found and thought to be interesting“ (Millen, Feinberg, & Kerr, 2006, 111). The user does not even have to use his own browser anymore in order to access his bookmarks or favorites, but can instead use any browser to connect to his social bookmarking service of choice and save, manage and retrieve his links (Noruzi, 2006): “Social Bookmarking systems are a class of collaborative applications that allow users to save, access, share and describe shortcuts to web resources“ (Braly & Froh, 2006). Thus bookmark collections are booming once more after having been nearly replaced by search engines and their fast search algorithms (Hammond et al., 2005). The use of corporate social bookmarking services in-house by companies is also wide-spread (Pan & Millen, 2008; Hayman, 2007; Hayman & Lothian, 2007; John & Seligmann, 2006; Damianos, Griffith, & Cuomo, 2006; Millen, Feinberg, & Kerr, 2006; Farrell & Lau, 2006). Spiteri (2006b) tells an anecdote about the problems caused by desktop-dependent lists of favorites and their limited organization possibilities:

> Our bookmarks or favorite lists are mushrooming out of control. Many of us have folders within folders within folders. We find ourselves bookmarking the same site a dozen times because we can’t remember where we filed it. Alternatively, we simply ‘Google it’ to save time. […] The problem is exacerbated when people use different computers (e.g., one at work, one at home, a laptop, etc.); they do not keep the same information across the different computers they use (Spiteri, 2006b, 78).

Feinberg (2006) defines the following main functions for social bookmarking services: a) “keeping frequently used resources handy,“ b) “grouping resources needed for a particular project,” c) “identifying potentially interesting but non-critical resources” and d) “locating occasionally used resources that are difficult to recall” (Feinberg, 2006).

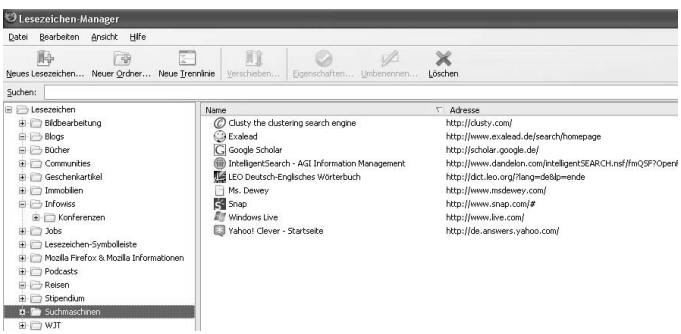

Figure 1.5: Bookmark Manager of the Browser ‘Firefox’ with its Folder Structure.

Apart from their decentralized bookmark storage facilities, social bookmarking services distinguish themselves mainly through another aspect: collaboration. They enable users to not only manage their own links privately, but also to make their collections of favorites publicly accessible to other users of the same service (Heller, 2007b). Thus the entire online community can share in and profit from a single

user's activity – only then does the actual meaning of the term 'social' manifest itself:

> Just as long as those hyperlinks (or let's call them plain old links) are managed, tagged, commented upon, and published onto the Web, they represent a user's own personal library placed on public record, which - when aggregated with other personal libraries - allows for rich, social networking opportunities (Hammond et al., 2005).

Regulski (2007) reports on a magazine that uses social bookmarking services specifically in order to raise the profile of their own articles and thus increase the number of their citations and of access paths leading to these articles.

The activities that users may perform on their lists of links mainly encompass the organization and categorization of URLs. In contrast to traditional browser-oriented classification systems, the URL no longer has to be saved in a folder (see Figure 1.5), but can be described and at the same time rendered retrievable via metadata, in this case via tags. To achieve this, the user attaches as many tags as needed to the URL. The community's other users may then use the tags or the link address to search for resources. If they do so using tags, their search results will be all URLs that were indexed with the requested tags, as well as any available information as to what other tags the displayed resources were allocated, and which users have stored the URL in their profile. A search via link address will also lead to these users and furthermore, to all tags the resource was indexed with by the community. Thus the social bookmarking systems changes from a personal URL management program to an information service that uses the main features of social networks and thus enables the community to search more effectively. It now serves not only as a storage space (as described above by Feinberg (2006)), but also as a provider of interesting or relevant web resources – after all, these searches lead the user to URLs that he might never have found browsing through the net, and which have been indexed intellectually by the community using tags. This is where social bookmarking services have the upper hand over search engines: "Social bookmarking is already used by many as an alternative to search engines, or at least as an addition – an addition which has the advantage that the information has been pre-filtered by like-minded people" (Reinmann, 2008, 54). At this point Skusa and Maaß (2008) and Graefe, Maaß and Heß (2007) argue that social bookmarking services' search results can never be as up to date as algorithm-based search engines, and that services such as news websites are seldom included in social bookmarking services and thus rendered researchable besides.

A detailed description of the functionalities of social bookmarking services will be provided by the example of two popular services: del.icio.us[15] and BibSonomy[16]. Del.icio.us represents social bookmarking services that are directed at the broad mass of users and explicitly do not specialize in any subject area (as do, for example, Blinklist[17], Furl[18], Ma.gnolia[19] oder Simpy[20]) – even though six of the ten most popular bookmarks in del.icio.us are about web programming (Wetzker,

[15] http://del.icio.us.
[16] http://www.bibsonomy.org.
[17] http://www.blinklist.com.
[18] http://www.furl.net.
[19] http://www.ma.gnolia.com.
[20] http://www.simpy.com.

Zimmermann, & Bauckhage, 2008) and thus directed at a technology-savvy usership, while BibSonomy mainly caters to the scientific community and manages scientific links as well as references on scientific publications (as do CiteULike[21] and Connotea[22], a.o.; Lund et al., 2005). All platforms cited are free of charge.

Del.icio.us was created in 2003 by Joshua Schachter as a social bookmarking service (Alby, 2007), and as of September, 2007, could boast more than 5m registered users and in excess of 150m saved bookmarks (Arrington, 2007; Heymann, Koutrika, & Molina, 2008). Del.icio.us was taken over by Yahoo! (Schachter, 2005).

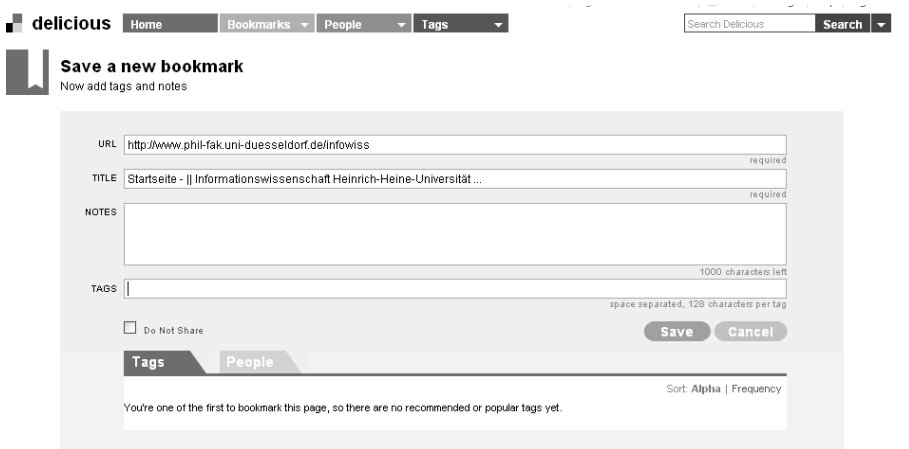

Figure 1.6: User Interface for Manually Adding Bookmarks in del.icio.us.

After creating an account on del.icio.us' homepage, users can directly begin collecting and managing their bookmarks (Orchard, 2006). To do so, they either enter the URLs manually (see Figure 1.6) or use a browser add-on which enables them to add bookmarks with the click of a button. Apart from the URL, they can also add a title for and notes on the bookmark and attach tags to it. The maximum number of tags attachable is 1,000 characters, the maximum tag length is 128 characters, and two tags are separated from one another by a blank. This way of separating multiple tags is described by Smith (2008b, 124ff.) as 'character-delimited,' since the user can pick any system-prescribed symbol (e.g. blank, comma or semicolon) as a separating symbol and thus index several tags for a single resource at once. Choosing 'Do Not Share' keeps the bookmark private, so that other members of the community cannot access them. Figure 1.6 displays the user interface for saving the first bookmark, which is why none of the user's previously added tags are being recommended for indexing at the bottom (blue tab 'Tags'). In this case, the URL http://www.phil-fak.uni-duesseldorf.de/infowiss has not even been saved in del.icio.us' database, so that none of its other users' tags are available for recommendation.

[21] http://www.citeulike.org.
[22] http://www.connotea.org.

The user interface is constructed differently when the user and the about-to-be-saved bookmark both already have several tags (see Figure 1.7). In this case, the user is shown his or her tags for indexing, so that he or she can compile a personal and fairly consistent indexing vocabulary. The most popular tags for this bookmark are equally displayed. If the user enters a tag that has been used for this bookmark before (in this case 'search'), a selection of tags as well as information as to their frequency of usage for this bookmark will be displayed via a 'Type Ahead' functionality.

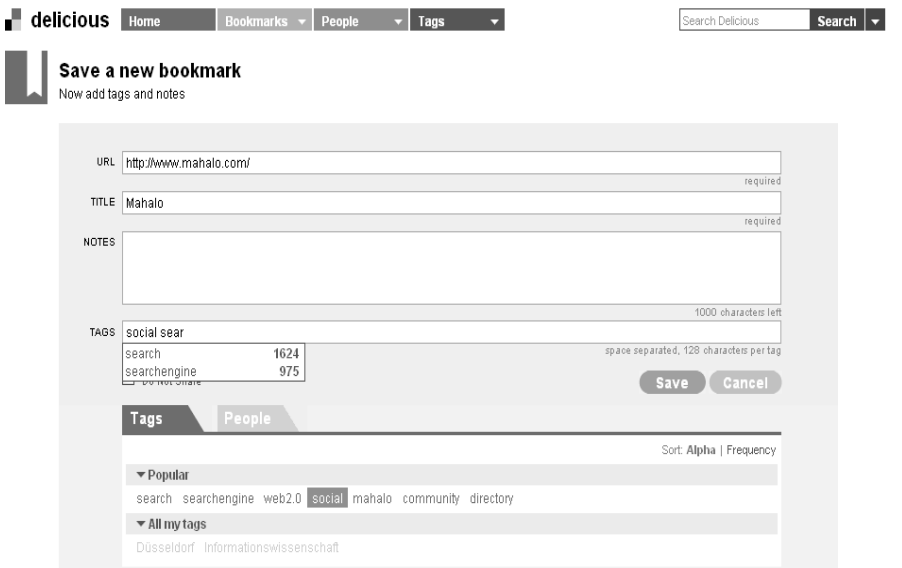

Figure 1.7: Adding an Already Indexed Bookmark in del.icio.us.

Figure 1.8 shows a user account. Here the user is shown information regarding his saved URLs (date of entry, name, number of users who have also saved the URL, tags attached to the URL) and his used tags in a sort of list. Clicking on one of the tags in the list or adding a tag in 'Username → Type a Tag' displays all bookmarks that have been indexed with this tag. Entering in the search field has the advantage that tags can be linked with an AND. This interface also enables the user to apply different editing functions to the bookmarks and tags. Thus information regarding the URL can be changed or the URL deleted altogether, the way the tags are displayed can be adjusted via 'tag options' (see Figure 1.8).

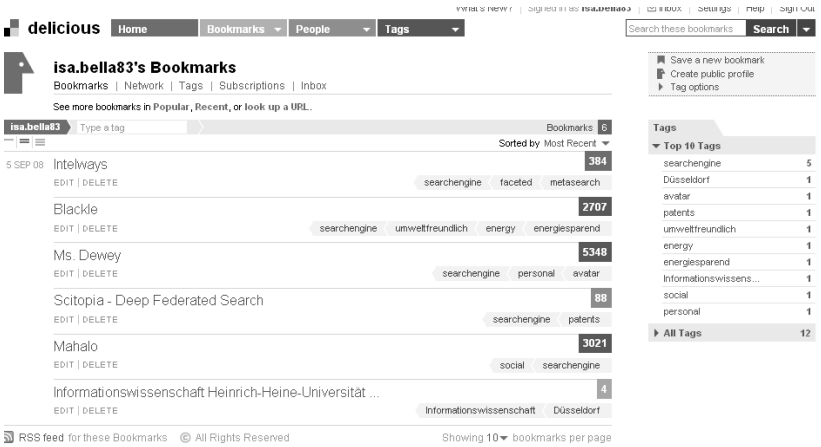

Figure 1.8: Interface of a User Account in del.icio.us.

Additionally, the user can personally change the tags: they are renameable and deletable, and they can also be compiled into so-called 'tag bundles' in order to create a hierarchical tag order. This functionality's greatest use is in helping users manage their own tags.

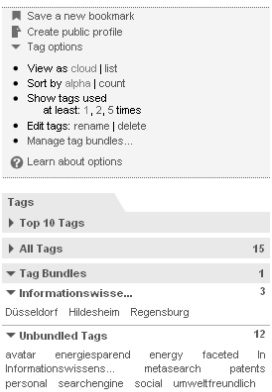

Figure 1.9: Options for the Display of Tags via 'Tag Options' in del.icio.us.

Helping users index their own bookmarks is not the tags' sole purpose, however – they are equally employed during searches for information or new URLs. They can be used to order 'subscriptions' (see Figure 1.10), in which the user is automatically sent all URLs indexed with the searched-for tag (in this case 'folksonomy') directly into their account. The subscription may also be limited to one particular user. Thus the user is kept up to date on the latest postings or developments within his area of interest, or of their favored community member, without having to actively search

for the information. The automatic reception of new URLs can also be implemented by subscribing to RSS feeds, which then react to either tags or users.

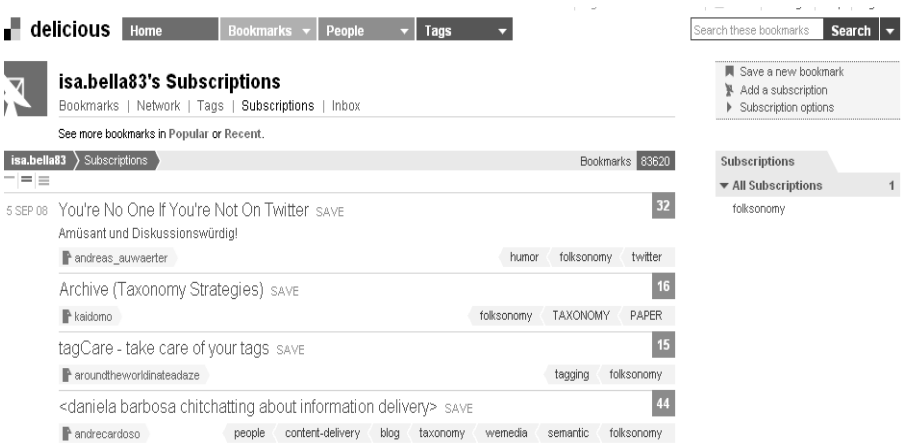

Figure 1.10: Subscriptions for the Tag 'Folksonomy' in del.icio.us.

Also, the user can use tags to search for a URL, either by entering a tag into the search mask on the del.icio.us homepage or by browsing and clicking through tag clouds (see Figure 1.11). After a search, both variants yield all tags within del.icio.us that were indexed with the searched-for tag. The tag search field again has the advantage that the search terms can be linked with an AND. However, the search tags must be entered into the different fields one by one (see Figure 1.12).
This form of user guidance enables users to not only restrict search results with further tags, and thus limit the number of results, but also to remove already entered tags with a click of the mouse and so re-adjust the search results. In Figure 1.12, an initial search for the tags 'web,' 'design,' 'graphic,' 'journals' and 'css' was executed, which yielded one result. If the user removes the tag 'css' from the AND link, the number of results will increase to three.

The exclusion of search terms is not an option; an OR link can be activated via tag bundles. The use of upper- and lower-case letters for search terms is ignored, only compounds must be spelled correctly. Search requests via search field as well as by clicking on tags in the tag cloud are supplemented by del.icio.us with 'related tags,' that is tags similar to the search term, which can serve to refine the request (see Figure 1.13). The search result is re-adjusted by clicking on a related tag. These similar tags are calculated through co-occurrence analyses of tags in del.icio.us.

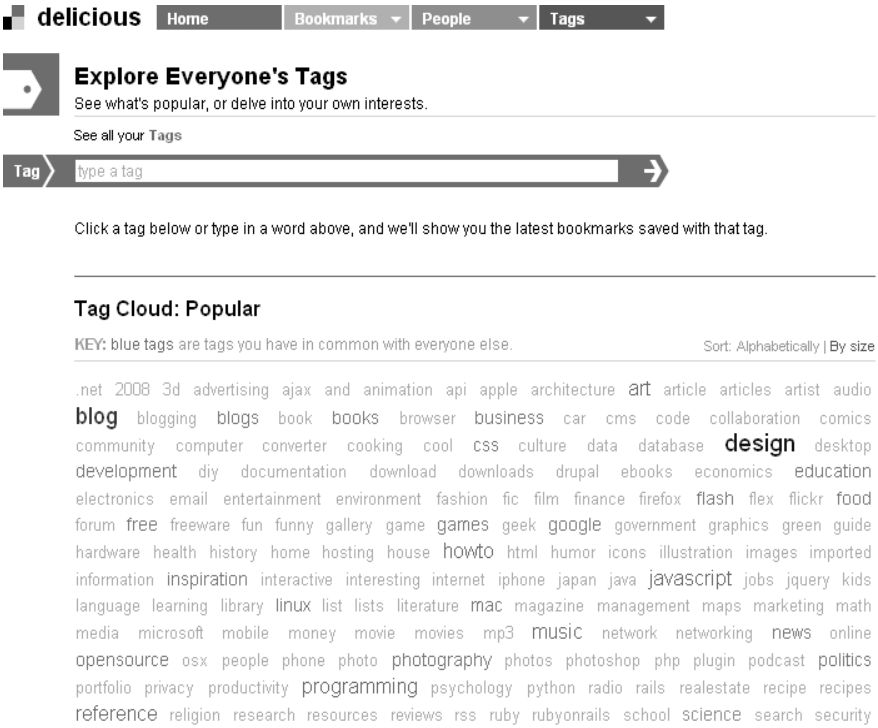

Figure 1.11: Search Field and Tag Cloud for Information Retrieval in del.icio.us.

Tag-based search is not del.icio.us' default setting, but it can be activated via the tab 'Explore Tags' on the homepage or, after a search, via the tab 'Tags.' Searching for and browsing through URLs and usernames is also an option. Likewise, the search can be manually limited to one's own bookmarks, one's own network (provided the user is linked to others via the 'friendship' function) or one's own tags. The list of search results is sorted after the latest save or indexing date – the latest entry will be displayed at the top of the page.

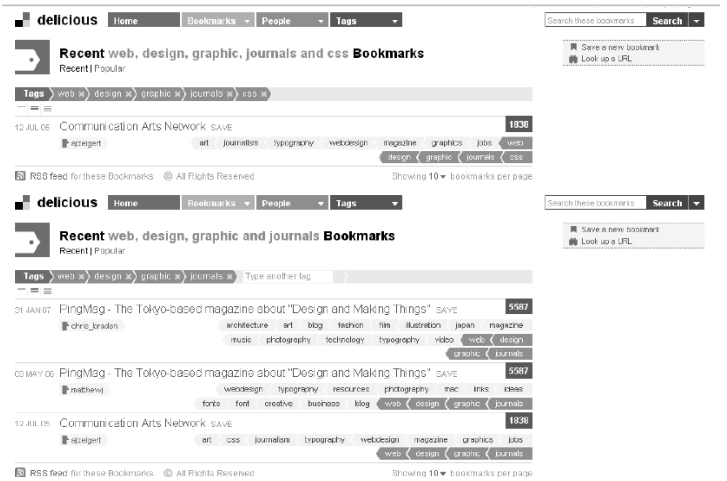

Figure 1.12: Linking Tags in a Search on del.icio.us.

Figure 1.13: Related Tags for the Refinement of Search Requests in del.icio.us.

BibSonomy (Hotho et al., 2006a; Hotho et al., 2006b; Jäschke et al., 2007; Schmitz et al., 2006; Regulski, 2007) has committed itself to the storage of scientifically oriented URLs and scientific references, and in 2007 reported more than 5,000 registered users (Jäschke et al., 2007). This social bookmarking service thus offers many further functionalities which cannot be found in del.icio.us and which are mainly directed at scientists, while considering their ways of working. BibSonomy serves as a decentralized URL storage facility on the one hand, and on the other hand as a storage facility for references and publications in the BibTeX format (Patashnik, 1988). Furthermore, BibSonomy facilitates the automatic compilation of bibliographies from saved references.

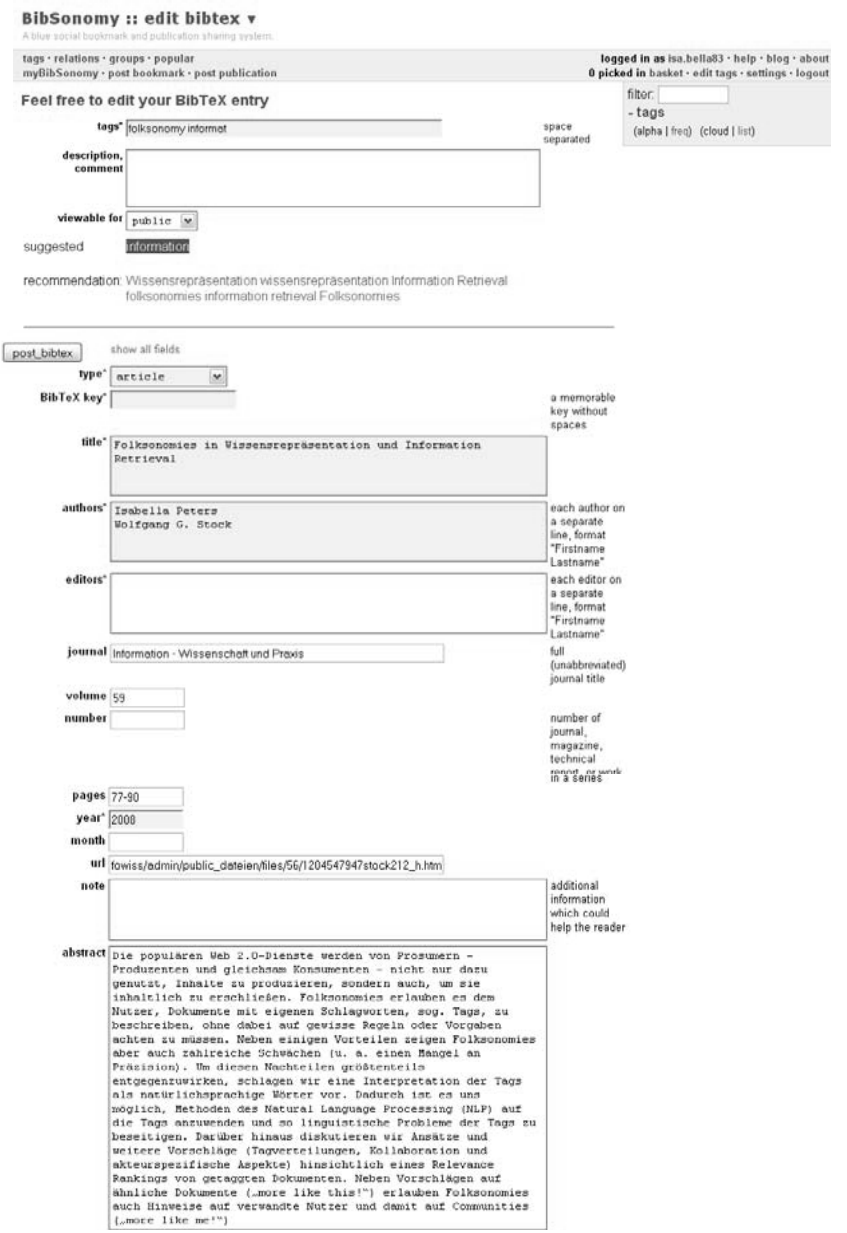

Figure 1.14: Manually Entering a Publication into the Designated Fields Equipped with Tagging Functionality in BibSonomy.

After creating an account, users can directly begin managing their bookmarks and publications. The information (such as author name, title etc.) can be either entered directly into the designated fields or imported from a BibTeX program or added 'post publication' via a browser button. Here the user can choose manually what sort

of publication the resource to be saved is to be classified as. In a second step, an enhanced user interface appears (see Figure 1.14) which enables the user to add even more information to the resource, such as personal remarks, abstracts or tags. The fields are heavily influenced by the parameters set by the two literature management programs BibTex and EndNote, in order to facilitate the import and export of the saved publications. To enter a URL (see Figure 1.15), the user is only required to fill out the fields 'URL,' 'title' and 'tags.' While entering the tags, the user receives tag suggestions in two different ways: a) 'suggested' tags are displayed during text input via a 'Type Ahead' functionality and calculated from the user's available tags, b) 'recommendation' tags are calculated from the user's available tags or, as in Figure 1.15, extracted from the title of a resource. The separate tags are then separated via a blank. Using a drop-down menu, the user can also decide whether the resource (bookmark or publication) should be made accessible to the entire BibSonomy community, just to his friends or only to him- or herself:

> These three accessibility options for an entry facilitate the platform's usage as a private, portable information basis, as an instrument for collaborative research, or as a forum for the exchange of recommendations and information regarding research-relevant literature (Regulski, 2007, 180).

The user can access their account via a permanent URL. As displayed in Figure 1.15, the user is here provided an overview of his saved URLs and publications as well as of the tags used for indexing these resources. The way the tags are represented can be adjusted and displayed as a tag cloud (as in the picture) or as a list, for example, sorted after frequency of occurrence or alphabetically or with a minimum of indexing frequency. The user can also use the field 'Filter' to browse through their tags.

Figure 1.15: User Account with Information on Saved URLs and Publications and Used Tags in BibSonomy.

For the bookmarks, the user can edit or delete details on all URLs either at the same time or one by one. Furthermore, he can export the bookmarks via RSS or XML feeds. For the publications, additional functions are at the user's disposal. Thus publications can be exported using different formats (RSS, BibTex, RDF, a.o.) or deposited in a 'basket' using the 'pick' button. This last function is available to the user

not only in the user account, but on the entire BibSonomy platform, so that he can compile a bibliography while browsing and export it using various formats.

BibSonomy, like del.icio.us, also offers a tag editing function (see Figure 1.16). This enables the user to either delete their tags or replace them with others, and to link them via unspecific relations and thus create a tag hierarchy (Hotho et al., 2006b) – an example would be a link between the tags 'knowledge representation' and 'folksonomy.' While doing so, it is important to enter the superordinate tag into the right field and the subordinate tag into the left field.

Figure 1.16: Editing Functions for the User's Tags and Relations in BibSonomy.

The greatest effect of these relations is on searches. If the user clicks on the superordinate tag in the 'relations' menu, he is not only shown all publications and bookmarks indexed with this 'supertag' but also all resources indexed with its 'subtag' (see Figure 1.17). However, this functionality is limited to the user's own personal resources and tags and cannot be applied to the platform at large.

Figure 1.17: Subtags and Supertags in BibSonomy.

Research within BibSonomy is possible in various different ways: either by clicking on the tags in the tag cloud (see Figure 1.18) or by entering terms into the search field. The search field's advantage is that the user can directly limit his search on a

tag, a user, a group, an author, a concept, a BibTex Key[23] or his own account. Additionally, he can link multiple search terms with an AND or an OR (the same goes for tags, users etc.).

Figure 1.18: Search Functions in BibSonomy.

A tag cloud search and a traditional search via search field (set to 'tags') both yield a list of resources indexed with the tag in question (see Figure 1.19).

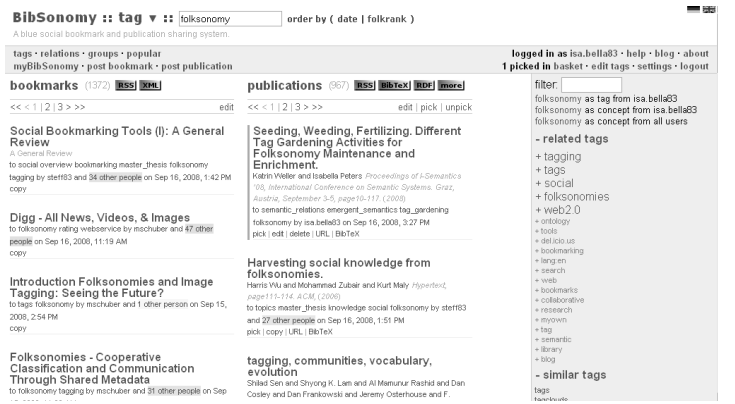

Figure 1.19: List of Search Results for the Tag 'Folksonomy' in BibSonomy.

Furthermore, on the right-hand side the user is offered the possibility of limiting the search to his own account or of changing the tag to a concept [24] and then repeating the search, either on the whole platform or in his account. Also displayed on the right-hand side are 'related tags' and 'similar tags' which can be activated as an alternative to the search term by clicking on them and do not serve to refine the search result via an AND link:

> Related tags are those tags which were assigned together to a post. If e.g. a user has tagged a post with *java* and *programming*, then those two tags are re-

[23] The BibTex-Key is a resource-specific key that serves the precise definition of the data set in a literature management program.

[24] "In order to distinguish between simple tag queries and those involving subtags, we call the latter one a query for java as a concept [occurs when 'Java' was designated as the hyponym of another term 'X,' A/N]" (Jäschke et al., 2007).

lated. Similar tags on the other side are computed by a more complex similarity measure coming from the research on information retrieval, namely cosine similarity in the vector of the popular tags. Similar tags are in many cases synonym tags[25].

The tags' font size reflects the degree to which they resemble the search term. The user can structure the list of results according to the date the resources were recorded or according to their FolkRank value. FolkRank is an adaptation of the idea of PageRank and aims to find the most relevant resources for each tag. To achieve this, it examines the link structure between users, tags and resources. BibSonomy has also implemented several shortcuts for search purposes, so that the required information can be entered directly into the browser address field and searched. Attach the following to www.bibsonomy.org:

a) *tag/SEARCHTERM* in order to retrieve all resources indexed with the search term.
b) *user/USERNAME* in order to retrieve all of the user's resources.
c) *group/NAME* in order to retrieve all of the goup's resources.
d) *relations/USERNAME* in order to retrieve all of the user's relations.

These search options can also be accessed via the drop-down menu 'myBibSonomy.' Here the user can also gain access to the PDF documents he uploaded to the BibSonomy server, to any duplicates and to the advanced search functions, called 'mySearch.' Advanced search offers the user more elaborate search functionalities for the publications in his account (see figure 1.20). Here he can link several tags or authors with ANDs or Ors for search purposes, and use the operators to link tags and authors with each other. Additionally, he can enter search terms that do not correspond with the tags into the free text search field. The search is run over the publications' title fields. Advanced search is not available for URLs.

Figure 1.20: Advanced Search 'mySearch' for a User's Publications in BibSonomy.

[25] See http://www.bibsonomy.org/faq.

E-Commerce

Tagging functionalities are growing ever more popular in e-commerce (Tschetschonig et al., 2008; Hayman, 2007; Hayman & Lothian, 2007), where they are used to siphon users' knowledge into the presentation, classification[26] and marketing of products.

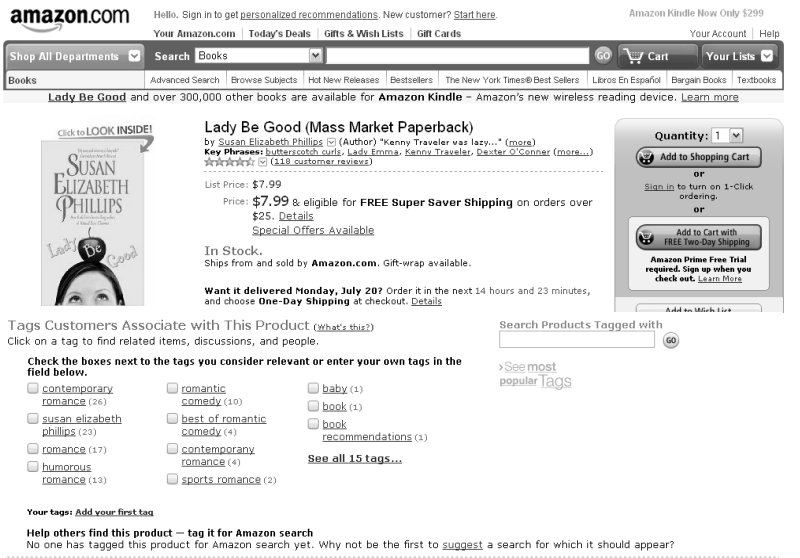

Figure 1.21: Product with Tags, Tag Search Option and the Option for Adding Tags in Amazon.

The online merchant **Amazon**[27] has been making use of the (implicit) shopping and browsing behavior of its customers and of their readiness to share their impressions of products with other users ever since its foundation in 1994 (Spalding, 2007a). Every user's transactions and viewed products are saved in Amazon's system, evaluated and used to interest him in other products (Linden, Jacobi, & Benson, 1998; Linden, Smith, & York, 2003; Gaul et al., 2002). Typical examples for this procedure are the recommendations underneath the product description: "Customers Who Bought This Item Also Bought Items B, C, D etc." or "Frequently Bought Together: Items A, B and C." While these are calculate purely from the products' sales statistics, registered users receive product recommendations based on their own shopping and browsing behavior. Furthermore, registered users can write and publish customer reviews to the purchased products on Amazon, as well as rate the items with stars ('x out of five stars'). These reviews can also be commented on by other users, thus facilitating not only the exchange of communication between Amazon and its customers but also between customers. Especially active users with a

[26] See also the online marketplace for handmade products www.etsy.com, the product category system of which is based on user-generated tags.
[27] http://www.amazon.com.

large number of reviews under their belt are rated 'Top Reviewers' by Amazon, creating an incentive for users to keep publishing intelligent comments. Since 2007, Amazon has enabled users to exchange their knowledge via a product wiki, Amapedia[28] (Gilbertson, 2007).

The success of this collaborative trading platform based on customers' and manufacturers' product information has led to the introduction of tags to describe and categorize products. Since late 2005 Amazon customers have been able to attach tags to products (Arrington, 2005), where the actual allocation of tags is the prerogative of registered users – anonymous users can only use tags to search for products (see Figure 1.21). After the registration, the user may annotate[29] each product with a tag, a comment or a review (see Figure 1.21). Underneath the formal product description, the user finds tags that have already been used by others users. On display are the six most commonly attached tags. Should more tags be available, the user can call these up and arrange them by their popularity, alphabetically or by their indexing date (see Figure 1.22). Additionally, one can see which other users have tagged the product (e.g. Mariam "meme05" was the first to add a tag 'modern romance' and that the last added tag was 'smoochbook').

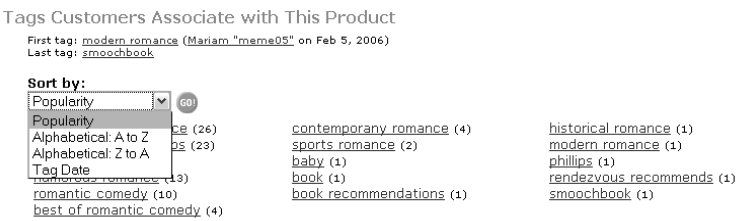

Figure 1.22: Sorting Options for Tags and Access Path to Other Users in Amazon.

To use already available tags for his own indexing purposes, the user only has to click on them (exemplified by the checked tags); to add a new tag (e.g. 'smoochbook'), he can use the input field (see Figure 1.23). Apart from the tags, the user is provided information as to how often each tag has been used for this particular resource. Each user may add as many tags to the resource as he wants. The input field for tags is interesting because it is designed for entering single tags only, which means that tags separated by a blank will be combined into one single tag if entered into the field one after the other (e.g. 'grandma_xmas smoochbook') and then clicking on the 'add' button. This may result in vast tag chains if the user is unfamiliar with this system. On the other hand, Amazon tries to avoid the forming of compounds – the linking of tags via symbols such as dashes or underscores etc. is not corrected, like the tag 'grandma_xmas' in Figure 1.23. Any capitalization, however, is immediately changed to lower-case letters by Amazon.

[28]http://amapedia.amazon.com orr
http://www.amazon.com/gp/help/customer/display.html?ie=UTF8&nodeId=43568011.

[29] First, however, he must decide whether he wants to use a pseudonym or his ‚real' name, generated from his credit card / bank account information.

Figure 1.23: Adopting Tags and Adding Personal Tags in Amazon.

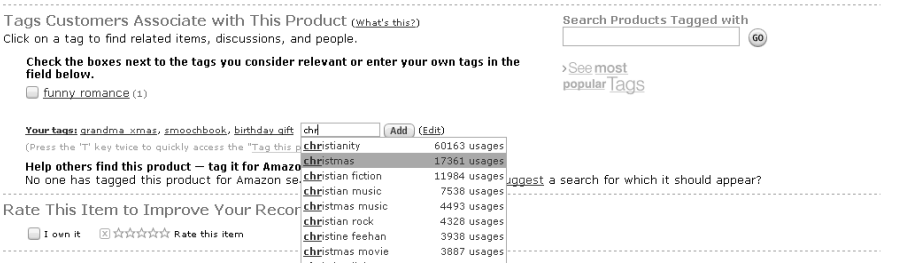

Figure 1.24: Type-Ahead Functionality and Frequency of Tag Usage Displayed on Amazon.

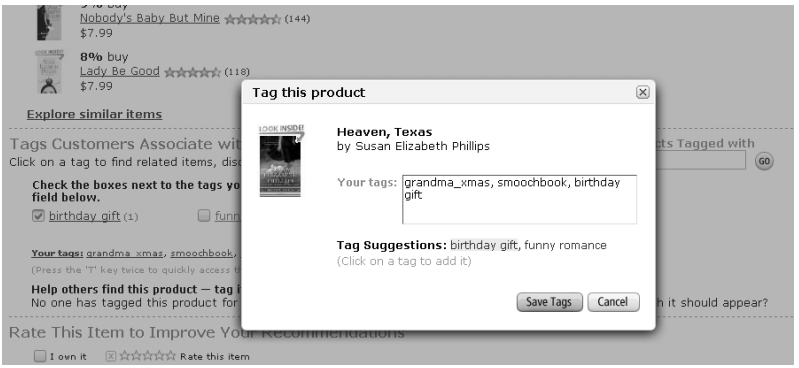

Figure 1.25: Quick Tagging Function by Double-Typing 't' in Amazon.

While manually entering tags, the user is supported by Amazon via a Type-Ahead functionality, which additionally shows how frequently the tag in question has been used (see Figure 1.24). A function unique in the manner of its implementation is quick tagging. Typing 't' twice in a row opens a browser window (see Figure 1.25) allowing the user to add or edit tags, as well as showing him tags already used for this resource.

With regard to adding tags, the user is here alerted to the fact that multiple tags in the input field must be separated by a comma. Additionally, the indexing process is