59. Jahrgang, Nr. 2 – März 2008 – NADOAW 59 (2) 73-148 – ISSN 1434-4653 – D 11721

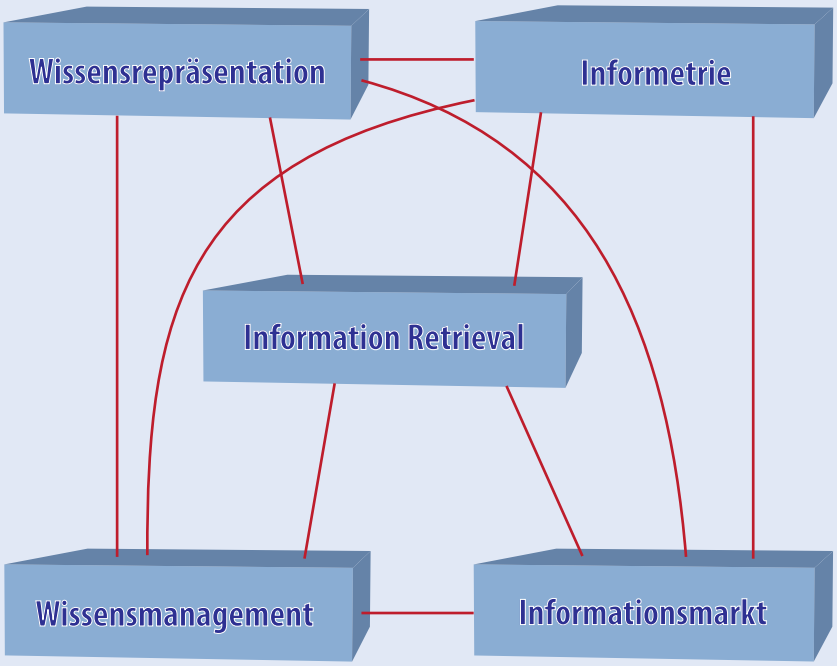
Information Wissenschaft & Praxis

2/2008

Wissensrepräsentation
Information Retrieval
Wissensmanagement

Das Forschungsprogramm der Düsseldorfer Informationswissenschaft

herausgegeben von
Wolfgang G. Stock und Sonja Gust von Loh



Düsseldorfer Informationswissenschaft

Folksonomies in Wissensrepräsentation und Information Retrieval

Ontoverse: Kollaborativer Ontologie-Aufbau

Paradigmatische und syntagmatische Relationen in der Wissensrepräsentation

Erforschung des Dark Web

Wissensmanagement in kleinen und mittelständischen Unternehmen

Klickbetrug bei der Werbung im World Wide Web

Digitale Kluft in Bosnien-Herzegowina

iwp Herausgeber: Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e. V. DGI

Sind Sie bereits mit dem neuen **Gesicht der Forschung** vertraut?

„Ich verwende Benachrichtigungen in Web of Knowledge und werde so informiert, sobald eine meiner Fakultäten genannt wird. Ist ein neues Zitat in Web of Knowledge verfügbar, sende ich der betreffenden Person eine E-Mail, um sie darüber zu informieren. Das ist äußerst praktisch. ... So besteht die Möglichkeit, einen persönlichen Service bereitzustellen und auf unsere Ressourcen, vor allem auf Web of Knowledge, hinzuweisen."

— Ann Jacobson
Bibliothekarin

AKTUELLE WIEDERGABE

isiwebofknowledge.com
newfaceofresearch.com

ISI Web of Knowledge℠

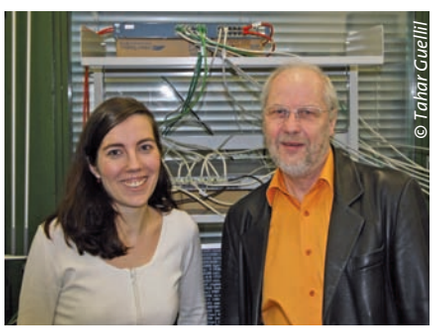
THOMSON™

Wissensrepräsentation – Information Retrieval – Wissensmanagement

Das Forschungsprogramm der Düsseldorfer Informationswissenschaft

Die Informationswissenschaft im Sinne des Düsseldorfer Forschungsprogramms untersucht das Auswerten, Bereitstellen, Suchen und Finden von relevantem (vorwiegend digital vorliegendem) Wissen. Sie umfasst ein Spektrum von fünf Teildisziplinen. Eine zentrale Rolle spielt das *Information Retrieval*, die Wissenschaft der Suchmaschinen unter Berücksichtigung technischer Systeme und Informationsbedürfnisse der Nutzer (Stock, 2007). *Wissensrepräsentation* thematisiert Metadaten sowie Methoden und Werkzeuge für den Einsatz beim Indexieren und Referieren (Stock & Stock, 2008). Anwendungsnahe Teilgebiete sind das *Wissensmanagement* mit der Konzentration auf das (Ver-)Teilen unternehmensinternen Wissens und der Integration externen Wissens in die betriebliche Informationswirtschaft sowie Forschungen zum *Informationsmarkt* mit einem weiten Gegenstandsbereich, der sich von der Informationsinfrastruktur eines Landes über die Branche elektronischer Informationsdienste und dem Informationsrecht bis hin zur Netzökonomie erstreckt. Informationswissenschaft geht – soweit möglich – empirisch vor und erfasst ihre Gegenstände mittels quantitativer Methoden. *Informetrische Forschungen* umfassen die deskriptive Informetrie, die Webometrie, die Szientometrie genauso wie die Evaluation von Informationssystemen und Nutzer- sowie Informationsbedarfsanalysen (Stock & Weber, 2006).

Die Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf verfolgt ein Forschungsprogramm, das sich allen Teildisziplinen widmet. Unser Begriff von „Forschungsprogramm“ folgt den wissenschaftstheoretischen Überlegungen von Imre Lakatos (1982). Ein wissenschaftliches Forschungsprogramm ist eine Reihe von Theorien, wobei die konzipierten Theorien zu Problemverschiebungen führen. Zu unterscheiden ist die progressive Problemverschiebung



Sonja Gust von Loh und Wolfgang G. Stock, Düsseldorf

von der degenerativen. Erstere führt uns zur Entdeckung bislang nicht bekannter Tatsachen; die Theorien haben einen „empirischen Gehaltsüberschuss“ (Lakatos, 1982, 33), während bei einer degenerativen Problemverschiebung Theorien nur gebastelt werden, „um mit den bekannten Tatsachen zurechtzukommen“ (Lakatos, 1982, 5). Ein Forschungsprogramm ist nach Lakatos nur dann erfolgreich, wenn es zu einer progressiven Problemverschiebung führt. Die Artikel in diesem Themenheft stellen theoretische Ansätze vor, deren empirischer Gehalt zweifellos in gewissem Ausmaß bereits gegeben ist, sie zeigen aber auch – besonders bei den (durchaus umfangreichen) Listen offener Probleme (als Indikator für den Gehaltsüberschuss) –, dass es gilt, die Theorien und die Tatsachen noch weitaus besser aufeinander abzustimmen.

Ein Forschungsprogramm wird von negativer wie positiver Heuristik getragen. Die negative Heuristik umfasst Ansätze, die man meiden soll, weil sie so klar sind, dass sie nicht zur Diskussion stehen. Hier liegt der harte Kern des Forschungsprogramms, der nicht infrage gestellt wird. Die positive Heuristik dagegen deutet auf „Wege, denen man folgen soll“ (Lakatos, 1982, 47). Die hier gesammelten Artikel sind Ausdruck der positiven Heuristik; sie gehen vom (derzeit nicht hinterfragten) harten Kern aus und verfolgen theoretische oder praxisorientierte Forschungswege.

Was macht den harten Kern des Düsseldorfer Forschungsprogramms aus? Wir sehen Informationswissenschaft als Grundlagenforschung und auch als angewandte Forschung Sie ist Grundlagendisziplin für Teilbereiche der Informatik, der Bibliothekswissenschaft, der Szientometrie sowie der Wirtschaftswissenschaften. Andererseits baut sie auf Ergebnissen dieser Fächer, sowie beispielsweise der Linguistik, auch auf. Informationswissenschaft blickt – selbst bei Fragen der Grundlagenforschung – auch auf die technische Realisierung und bezieht grundsätzlich den Nutzer bzw. die Nutzung mit ein. Sie findet ihren Gegenstand in Auseinandersetzung mit bestehenden Systemen (wie Suchmaschinen, Web 2.0-Services und elektronischen Informationsdiensten), indem sie diese analysiert, aber auch für sie wissenschaftlichen Vorlauf schafft. Methoden der Wissensrepräsentation werden ungeachtet ihrer historischen Herkunft betrachtet und fließen zu einem einzigen Methodenrepertoire zusammen: Wissensrepräsentation umfasst sowohl typisch informatische Aspekte (Ontologien), ursprünglich bibliothekarische oder dokumentarische Bemühungen (Nomenklaturen, Klassifikationen, Thesauri) als auch Ad-hoc-Entwicklungen im kollaborativen Web (Folksonomies).

Die positive Heuristik des Forschungsprogramms findet ihren Niederschlag in den Artikeln dieses Heftes. Die ersten drei Artikel sind Resultate von Bemühungen zur Wissensrepräsentation. Es geht mit den Folksonomies um eine gänzlich neue Methode (Peters & Stock). Ontologien führen zu einer detaillierten und formalisierten Abbildung fachlicher Wissensdomänen, die optimal kollaborativ zu erar-

beiten sind (Mainz et al.). Einen Einblick in die Werkstatt der theoretischen wie empirischen Erforschung semantischer Relationen gibt die Arbeit von Peters & Weller. Information Retrieval profitiert von elaborierten Ansätzen der Wissensrepräsentation, wie dies Heesemann & Nellißen anhand facettierter Wissensordnungen und dem dynamischen Klassieren am Beispiel eines experimentellen (mit Convera RetrievalWare realisierten) Retrievalsystems zum „Dark Web" aufzeigen. In den beiden Artikeln von Gust von Loh geht es um die Implementierung von Wissensmanagement in kleinen und mittleren Unternehmen (KMU), wobei schnell klar wird, dass gerade bei KMU Informationsbedarfsanalysen notwendige Vorarbeiten darstellen. Auf dem Informationsmarkt werden derzeit Milliardenumsätze mit „Sponsored Links" erzielt. Soubusta zeigt, dass bei dort üblichen Pay-per-Click-Abrechnungen betrügerisch generierte Klicks möglich (und realisiert) sind, die solche Geschäftsmodelle nachhaltig infrage stellen. Ergebnisse aus Informationswissenschaft und -praxis können nur dann genutzt werden, wenn in einem Land überhaupt eine hinreichende Informationsinfrastruktur gegeben ist. Am Beispiel von Bosnien-Herzegowina analysiert Trkulja die digitale Kluft zwischen fortgeschrittenen Informationsgesellschaften und einem Land, das gerade am Anfang einer Entwicklung hin zu einer Wissensgesellschaft steht. (Nicht vertreten sind in diesem Heft Beiträge zur Informetrie, da dieser Disziplin ein eigenes Themenheft der IWP (57(2006)8) gewidmet war.)

In Düsseldorf sind Forschung und Lehre eng miteinander verbunden. Leitidee ist der „(mit-)forschende Student". Da bereits Mitte 2007 die ersten Absolventen des neuen Bachelor- und Masterstudiengangs „Informationswissenschaft und Sprachtechnologie" (Stock, 2004) ihr Studium erfolgreich beendet haben, können wir hier mit zwei überarbeiteten akademischen Abschlussarbeiten aufwarten: Der Artikel von Soubusta baut auf ihrer Bachelorarbeit auf, die Beiträge von Gust von Loh auf ihrer Masterarbeit.

Das Forschungsprogramm der Düsseldorfer Informationswissenschaft ist bei weitem noch nicht abgeschlossen. Insbesondere gilt es, die Grundlagenforschung in der theoretischen Informationswissenschaft noch besser mit der empirischen informetrischen Basis und der angewandten Informationswissenschaft, also dem Wissensmanagement und dem Informationsmarkt, zu verbinden. Die konzeptionellen Forschungen dürfen nicht im Theoretischen verbleiben; die als praxistauglich vermuteten Ansätze sollen via (Informations-)Technik auch in (erfolgreich evaluierte) Anwendungen münden. Ein Beispiel für solch einen Übergang ist im Artikel von Mainz et al. über Ontoverse skizziert worden: Grundlagenforschung in Ontologien und kollaborativem Wissensmanagement führten inzwischen zu einem Prototypen eines Ontologie-Wikis. Ähnlich erfolgreich konnte bei Heesemann & Nellißen die theoretische Konzeption facettierter Wissensordnungen zu gänzlich neuen Optionen im Information Retrieval technisch umgesetzt werden. Allerdings scheint Geduld angesagt. „Man muss aufkeimende Forschungsprogramme gewähren lassen; sie können Jahrzehnte brauchen, bis sie richtig in Fahrt kommen und empirischen Fortschritt zeitigen" (Lakatos, 1982, 6).

Die Düsseldorfer Informationswissenschaft wünscht sich eine Kooperation mit anderen informationswissenschaftlichen Institutionen im Sinne der nordamerikanischen *I-Schools*, um den wissenschaftlichen Austausch zu fördern bzw. anzuregen und dadurch Kompetenzen besser lokalisieren zu können und Kenntnisse über andere Forschungsvorhaben zu erlangen. Wir sehen dieses Themenheft als einen ersten Schritt in diese Richtung und hoffen, dass vergleichbare Darstellungen der Forschungsprogramme anderer Institute folgen werden.

Literatur

Lakatos, I. (1982). Die Methodologie der wissenschaftlichen Forschungsprogramme. Braunschweig, Wiesbaden: Vieweg.

Stock, W.G. (2004). Die Düsseldorfer Informationswissenschaft. Information. Wissenschaft und Praxis, 55, 21-22.

Stock, W.G. (2007). Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

Stock, W.G., & Stock, M. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München, Wien: Oldenbourg.

Stock, W.G., & Weber, S. (2006). Facets of informetrics. Information. Wissenschaft und Praxis, 57, 385-389.

Gegründet von H.-K. Soeken †
unter dem Titel Nachrichten für Dokumentation (NfD)
Herausgegeben von der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI)
Präsidentin: Prof. Dr. Gabriele Beger
Hanauer Landstraße 151-153
D-60314 Frankfurt am Main
Telefon: (0 69) 43 03 13
Telefax: (0 69) 4 90 90 96
mail@dgi-info.de
www.dgi-info.de
Mitteilungsblatt des Normenausschusses Bibliotheks- und Dokumentationswesen im DIN Deutsches Institut für Normung e.V., der Fachgruppe Dokumentation im Deutschen Museumsbund und der Arbeitsgemeinschaft der Spezialbibliotheken (ASpB)

Redaktionsbeirat
Dr. Sabine Graumann, München (Informationswirtschaft)
Prof. Dr. Hans-Christoph Hobohm, Potsdam (Management von Informationseinrichtungen)
Prof. Dr. Rainer Kuhlen, Konstanz (Informationswissenschaft)
Prof. Dr. Dirk Lewandowski, Hamburg (Suchmaschinen, Internet)
Prof. Dr. Wolfgang Ratzek, Stuttgart (Informationspraxis)
Prof. Dr. Ralph Schmidt, Hamburg (Newcomer Report, Medien)
Prof. Dr. Wolf G. Stock, Düsseldorf (Wissensrepräsentation und Information Retrieval)

Redaktion
Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V.
Marlies Ockenfeld (verantwortlich)
Viktoriaplatz 8, 64293 Darmstadt
Telefon: (0 61 51) 86 98 12
Telefax: (0 61 51) 86 97 85
ockenfeld@dgi-info.de
Daniel Ockenfeld (Redaktionsassistenz)

Gastherausgeber
Prof. Dr. Wolfgang G. Stock und Sonja Gust von Loh

Verlag
Dinges & Frick GmbH
Greifstraße 4
65199 Wiesbaden
Postfach 1564
65005 Wiesbaden
Telefon: (06 11) 9 31 09 41
Telefax: (06 11) 9 31 09 43
Bankverbindung:
Wiesbadener Volksbank
BLZ 510 900 00, Kto-Nr. 714 22 26
Postbank Frankfurt
BLZ 500 100 60, Kto.-Nr. 267 204-606

Objektleitung
Erwin König,
e.koenig@dinges-frick.de

Anzeigenservice
Ursula Hensel Anzeigenservice
Hermann-Schuster-Straße 39
65510 Hünstetten-Wallbach
Telefon: (0 61 26) 57 08 82
Telefax: (0 61 26) 58 16 47
ursula.hensel@t-online.de
Rocco Mischok
Verlag Dinges & Frick GmbH
Greifstraße 4
65199Wiesbaden
Telefon: (06 11) 3 96 99-60
Telefax: (06 11) 3 96 99-30
r.mischok@dinges-frick.de

Gestaltung
Meinhard Zielke, Wiesbaden

Druck
Dinges & Frick GmbH
Greifstraße 4
65199 Wiesbaden
Postfach 2009
65010 Wiesbaden
Telefon: (06 11) 3 96 99-0
Telefax: (06 11) 3 96 99-30
Leonardo: (06 11) 93 20 79
Twist: (06 11) 9 10 23 78
df@dinges-frick.de

Hinweis
Die Aufsätze stellen ausschließlich die Meinung der Autoren dar. Der Inhalt wurde sorgfältig und nach bestem Wissen erarbeitet. Dennoch kann von Verlag und Redaktion eine Gewährleistung auf Richtigkeit und Vollständigkeit nicht übernommen werden. Die Beiträge und die grafischen Darstellungen unterliegen dem Urheberrecht. Nachdruck und Vervielfältigung jeglicher Art bedürfen der Genehmigung des Verlages und der Autoren.

Erscheinungsweise/ Bezugspreise
Sieben Hefte jährlich (Doppelausgabe September/Oktober)
Jahresabonnement EUR 179,–
Schüler/Studenten EUR 130,–
Einzelheft EUR 30,–
inkl. Versandkosten/Porto.
Das Abonnement gilt für mindestens ein Jahr und kann danach bis sechs Wochen zum Ende des Bezugszeitraums gekündigt werden.

Redaktionsschluss für

Heft 3/2008	10. März 2008
Heft 4/2008	21. April 2008
Heft 5/2008	16. Juni 2008

IMPRESSUM

Inhalt

2/2008

SCHWERPUNKT: DAS FORSCHUNGSPROGRAMM DER DÜSSELDORFER INFORMATIONSWISSENSCHAFT

Gastherausgeber:
Prof. Dr. Wolfgang G. *Stock* und Sonja *Gust von Loh*

73 **EDITORIAL**

Wolfgang G. *Stock* und Sonja *Gust von Loh*
Wissensrepräsentation – Information Retrieval – Wissensmanagement. Das Forschungsprogramm der Düsseldorfer Informationswissenschaft

WISSENSREPRÄSENTATION

77 Isabella *Peters* und Wolfgang G. *Stock*
Folksonomies in Wissensrepräsentation und Information Retrieval

91 Indra *Mainz*, Katrin *Weller*, Ingo *Paulsen*, Dominic *Mainz*, Jochen *Kohl* und Arndt *von Haeseler*
Ontoverse: Collaborative Ontology Engineering for the Life Sciences

100 Isabella *Peters* und Katrin *Weller*
Paradigmatic and Syntagmatic Relations in Knowledge Organization Systems

INFORMATION RETRIEVAL

108 Silke *Heesemann* und Hans-Dieter *Nellißen*
Facettierte Wissensordnungen und dynamisches Klassieren als Hilfsmittel der Erforschung des Dark Web

WISSENSMANAGEMENT

118 Sonja *Gust von Loh*
Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen. Teil 1: Grundlagen des Wissensmanagements

127 Sonja *Gust von Loh*
Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen. Teil 2: Wissensmanagement in KMU

INFORMATIONSMARKT

136 Simone *Soubusta*
On Click Fraud

142 Violeta *Trkulja*
Die Digitale Kluft: eine Analyse der Informationsinfrastruktur in Bosnien-Herzegowina

74 **IMPRESSUM**

Beilagenhinweis:
Dieser Ausgabe liegt eine Einladung der Technischen Universität Ilmenau zur PATINFO 2008 bei. Wir bitten um freundliche Beachtung

Missing Link
Westerstrasse 114-116 | 28199 Bremen
Tel.: [0421] 504348 | Fax: [0421] 504316

Erwerbungspartner, mit denen Sie rechnen können

www.missing-link.de
info@missing-link.de

Internationale Versandbuchhandlung

Folksonomies in Wissensrepräsentation und Information Retrieval

Isabella Peters und Wolfgang G. Stock, Düsseldorf

Die populären Web 2.0-Dienste werden von Prosumern – Produzenten und gleichsam Konsumenten – nicht nur dazu genutzt, Inhalte zu produzieren, sondern auch, um sie inhaltlich zu erschließen. Folksonomies erlauben es dem Nutzer, Dokumente mit eigenen Schlagworten, sog. Tags, zu beschreiben, ohne dabei auf gewisse Regeln oder Vorgaben achten zu müssen. Neben einigen Vorteilen zeigen Folksonomies aber auch zahlreiche Schwächen (u. a. einen Mangel an Präzision). Um diesen Nachteilen größtenteils entgegenzuwirken, schlagen wir eine Interpretation der Tags als natürlichsprachige Wörter vor. Dadurch ist es uns möglich, Methoden des Natural Language Processing (NLP) auf die Tags anzuwenden und so linguistische Probleme der Tags zu beseitigen. Darüber hinaus diskutieren wir Ansätze und weitere Vorschläge (Tagverteilungen, Kollaboration und akteurspezifische Aspekte) hinsichtlich eines Relevance Rankings von getaggten Dokumenten. Neben Vorschlägen auf ähnliche Dokumente („more like this!“) erlauben Folksonomies auch Hinweise auf verwandte Nutzer und damit auf Communities („more like me!“).

Folksonomies in Knowledge Representation and Information Retrieval

In Web 2.0 services “prosumers” – producers and consumers – collaborate not only for the purpose of creating content, but to index these pieces of information as well. Folksonomies permit actors to describe documents with subject headings, “tags“, without regarding any rules. Apart from a lot of benefits folksonomies have many shortcomings (e.g., lack of precision). In order to solve some of the problems we propose interpreting tags as natural language terms. Accordingly, we can introduce methods of NLP to solve the tags’ linguistic problems. Additionally, we present criteria for tagged documents to create a ranking by relevance (tag distribution, collaboration and actor-based aspects). Besides recommending similar documents („more like this!“) folksonomies can be used for the recommendation of similar users and communities („more like me!“).

Inhaltserschließung durch die kollektive Intelligenz in Web 2.0-Diensten

In den frühen Jahren des World Wide Web sind einige wenige Experten in der Lage, Wissen über dieses Medium zu verteilen. Die Mehrzahl all derjenigen, die mit dem WWW zu tun haben, beschränkt ihren Umgang mit dem Web ausschließlich auf die passive Nutzung des Mediums. Mit dem Beginn des 21. Jahrhunderts kommen Dienste auf, die sehr leicht zu bedienen sind und die damit Nutzern gestatten, Content problemlos zu publizieren. Aus dem (passiven) Nutzer wird nunmehr zusätzlich ein (aktiver) Webautor. Der Konsument von Wissen ist gleichsam auch sein Produzent geworden, ein „Prosumer“ im Sinne Tofflers (1980). Da die Autoren (zumindest manchmal) ihre Dokumente wechselseitig korrigieren und fortschreiben, kann man in diesem Zusammenhang durchaus von „kollektiver Intelligenz“ sprechen: „With content derived primarily by community contribution, popular and influential services like Flickr and Wikipedia represent the emergence of ‚collective intelligence‘ as the new driving force behind the evolution of the Internet“ (Weiss, 2005, 16). „Kollektive Intelligenz“ entsteht dank der Zusammenarbeit von Autoren und Nutzern in „kollaborativen Diensten“, die zusammengefasst als „Web 2.0“ (O’Reilly, 2005) etikettiert werden können. Solche Dienste widmen sich beispielsweise dem Führen von „Tagebüchern“ (Weblogs) und deren Suche (z.B. Technorati), dem Aufbau einer Enzyklopädie (z.B. Wikipedia), der Ordnung von Lesezeichen zu Webseiten (Del.icio.us), von Bildern (Flickr) oder von Videos (YouTube). Die Inhalte von Diensten, insofern diese sich komplementär ergänzen, werden gelegentlich als „mash-ups“ (Weiss, 2005, 23) zusammengeführt (beispielsweise Housingmaps.com als „mash-up“ aus Immobilieninformationen aus Craigslist mit Landkarten und Satellitenbildern aus Google Maps). Die Kooperation endet nicht beim Bereitstellen von Content, sondern schließt bei einigen Web-2.0-Services auch die inhaltliche Erschließung des bereitgestellten Wissens mit ein. In Abbildung 1 sehen wir eine „tag cloud“ (Sinclair & Cardew-Hall, 2008) aus den Dingen, die die Nutzer in ihrem Leben noch tun oder erreichen wollen. Die Häu-

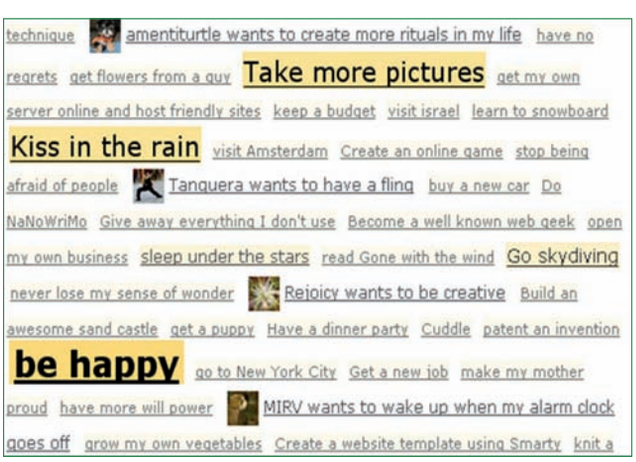


Abbildung 1: Tag Cloud der Homepage „43 Things“. Quelle: http://www.43things.com.

Der Artikel ist eine übersetzte und überarbeitete Fassung eines Vortrags der Verfasser bei der 70. Jahrestagung der American Society for Information Science and Technology in Milwaukee, Wisconsin, am 24. Oktober 2007. Isabella Peters trug eine Version des Vortrags am 29.Oktober 2007 beim Center for Intelligent Information Retrieval (University of Massachusetts, Amherst) vor. Der Studienaufenthalt von Isabella Peters wurde durch eine anteilige Finanzierung der Deutschen Forschungsgemeinschaft (DFG) ermöglicht. Wir danken der DFG für das gewährte Reisestipendium.

figkeit der Worte, der Tags, wird durch die Schriftgröße angedeutet. Am häufigsten wurde *be happy*, am zweithäufigsten *take more pictures* und weniger oft *get a new job* als Lebensziel genannt.

Diese Art der freien Schlagwortvergabe durch jedermann bezeichnen wir als „Folksonomy", wobei die freien Schlagworte hier „Tags" genannt werden. Die Indexierung mithilfe von Folksonomies ist demnach „Tagging". Peter Merholz (2004) bezeichnet das Verfahren als „metadata for the masses", der Schriftsteller James Surowiecki (2004) nennt es „the wisdom of crowds" (McFedries, 2006, 80). Folksonomies sind eine Form der Inhaltserschließung im Sinne einer freien Verschlagwortung von Dokumenten im Internet (Furnas et al., 2006). Die Produzenten oder auch die Nutzer der Dokumente werden selbst zu Indexern. Es gibt keine zentrale Kontrolle, die die Tags und die Arbeit der Prosumer prüft.

Der Begriff „Folksonomy" als Kunstwort aus „Folk" und „Taxonomy" geht auf einen Beitrag in einem Blog zur Informationsarchitektur zurück, in dem Gene Smith (2004) Thomas Vander Wal zitiert:

> „Last week I asked the AIfIA (i.e. the „Asilomar Institute for Information Architecture", d. A.) member's list what they thought about the social classification happening at Furl, Flickr and Del.icio.us each of these systems people classify their pictures/bookmarks/web pages with tags …, and then the most popular tags float on the top …
> Thomas Vander Wal, in his reply, coined the great name for these informal social categories: *a folksonomy*.
> Still, the idea of socially constructed classification schemes (with no input from an information architect) is interesting. Maybe one of these services will manage to build a social thesaurus".

Smith gebraucht das Wort „classification" zur Umschreibung von Folksonomies. Dies weist – genau wie Taxonomie – in eine falsche Richtung. Folksonomies sind gerade *keine* Klassifikationen, denn sie arbeiten weder mit Notationen noch mit Relationen. Wichtig erscheint uns der Hinweis Smith', dass man, aufbauend auf Folksonomies, Thesauri kooperativ erstellen kann.

Weite Verbreitung finden Folksonomies bei den Web 2.0-Diensten, sie können aber genauso in firmeneigenen Intranets angewendet werden (Fichter, 2006), z.B. um Corporate Blogs, Podcasts oder Vodcasts (Peters, 2006a; Peters, 2006b; Peters & Stock, 2006), Corporate Bookmarking-Dienste (Millen, Feinberg, & Kerr, 2006) oder Schwarze Bretter (Murison, 2005) zu erschließen. In Spiteri (2006 und 2007) wird der Gebrauch von Folksonomies bei Katalogen öffentlicher Bibliotheken diskutiert. Smith (2006) und Trant (2006a, 2006b) berichten über den Einsatz dieser Indexierungsmethode in Museen. Heller (2007) stellt die Nutzung von Folksonomies beim Aufbau gemeinschaftlich erstellter Bibliographien dar, ähnlich Hänger und Krätzsch (2007) bei der Erschließung ansonsten unindexierter Dokumente. Stock (2007b) ist ein Plädoyer für ein Mash-up von bewährten „traditionellen" Indexierungsmethoden und Folksonomies bei kommerziellen Informationsdiensten (siehe auch Peters, 2007)

Broad und Narrow Folksonomies

Mit Thomas Vander Wal (2005) lassen sich zwei Formen von Folksonomies unterscheiden: broad, weite, und narrow, enge, Folksonomies (siehe auch Dye, 2006). In der Broad Folksonomy (Abbildung 2) können viele verschiedene Nutzer (A bis F in der Abbildung) ein Dokument mit Tags versehen. So wird der Dokumentinhalt aus zahlreichen Ansichten mit gleichen, ähnlichen oder völlig verschiedenen Schlagworten beschrieben (1 bis 5). Das bekannteste Web 2.0-Angebot, das diese Art der multiplen Verschlagwortung nutzt, ist der Social Bookmarking-Dienst Del.icio.us. Auch für den Wissenschaftsbereich existieren mit CiteULike, Connotea (Hammond, Hannay, Lund, & Scott, 2005; Lund, Hammond, Flack, & Hannay, 2005) und BibSonomy (Hotho, Jäschke, Schmitz, & Stumme, 2006c) entsprechende Systeme.

In Narrow Folksonomies werden Tags nur einmal pro Dokument vergeben und auch nur einmal erfasst. Deswegen können nur neue Tags zugeordnet werden. Daher besteht keine Möglichkeit, Taghäufigkeiten zu zählen und Verteilungen zu beobachten. Häufig vergibt der Autor (oder der Content Creator) die ersten Tags; manchmal ist es auch anderen Usern erlaubt, weitere Tags hinzuzufügen. Dieses Vorgehen erinnert an die Inhaltserschließung mit den kontrollierten Vokabularien aus Nomenklaturen, Thesauri oder Klassifikationssystemen (Stock & Stock, 2008, Kap. 11 bis 13) durch professionelle Indexer; abgesehen davon, dass in Folksonomies nur unkontrollierte Terme zum Einsatz kommen. Web 2.0-Dienste, die mit Narrow Folksonomies arbeiten, sind u. a. Flickr (Fotos; Beaudoin, 2007), Technorati (Blogeinträge) und YouTube (Videos).

Narrow Folksonomies können keine bestimmten Häufigkeitsverteilungen der Indexierungs-Tags zeigen, da bei ihnen alle Tags gleichwertig sind (alle kommen genau einmal vor). Allerdings ist es möglich, Tagverteilungen aus den Such-Tags zu gewinnen. Die Frage ist hier: Welches Dokument wird mit welchen Tags gefunden? Jedem (angezeigten und angeklick-

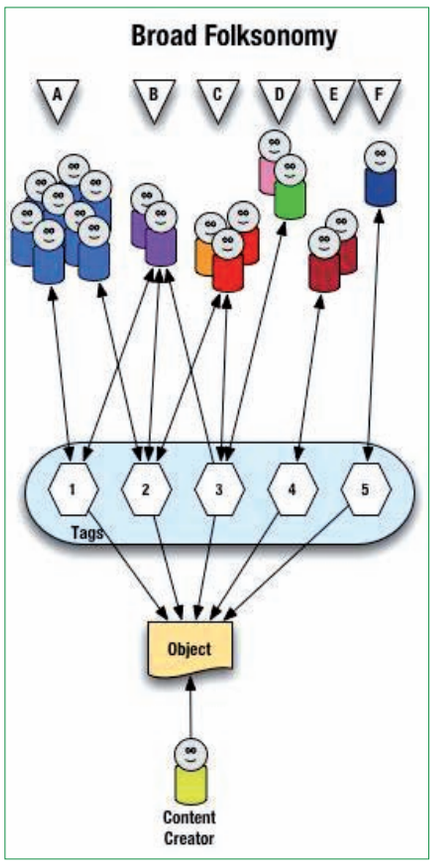


Abbildung 2: Folksonomy mit Mehrfachvergabe gleicher Tags („Broad Folksonomy").
Quelle: Vander Wal (2005).

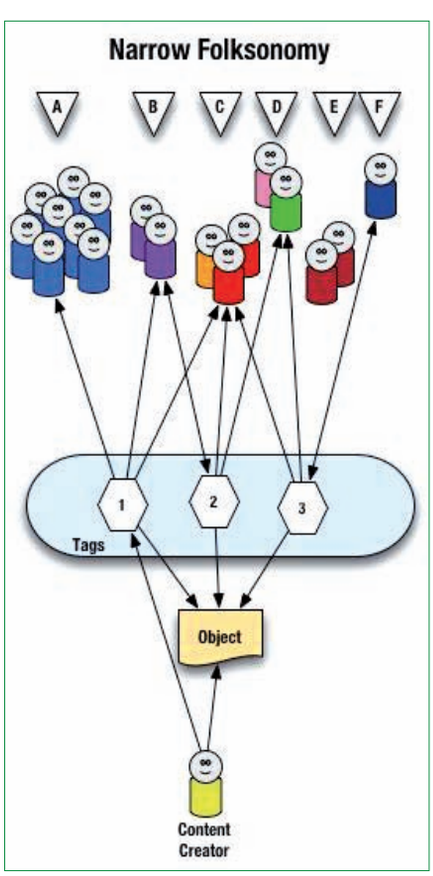


Abbildung 3: Folksonomy mit singulärer Vergabe von Tags („Narrow Folksonomy").
Quelle: Vander Wal (2005).

ten) Dokument werden diejenigen Terme als Such-Tags zugeordnet, mit denen der jeweiligen User erfolgreich das Dokument recherchiert hat. Solche Untersuchungsoptionen bestanden eigentlich schon immer; u. W. fanden sie jedoch noch nie Einsatz.
Im Vergleich zu den Broad Folksonomies werden Narrow Folksonomies i. d. R. mit weniger Tags indexiert. Das ist ein Grund, warum hier kein besonders langer Long Tail wird entstehen können. Die Häufigkeitsverteilungen der Such-Tags können aber durchaus gewisse Regelmäßigkeiten aufweisen.

Häufigkeitsverteilungen der Tags

Vander Wal (2005), Shirky (2005), Munk & Mork (2007) u. a. stellen bei Broad Folksonomies fest, dass die Verteilung der indexierten Tags einer Power Law-Kurve nach Lotka's Gesetz (Egghe & Rousseau, 1990; Egghe, 2005) gleicht. Diese Verteilung folgt der Gleichung

$$f(x)= C / x^a ,$$

bei der C eine Konstante, x der Rang des gegebenen Tags und a ein konstanter Wert (normalerweise zwischen 1 und 2) ist. Sollte diese Annahme richtig sein, wären am linken Anfang der Kurve nur wenige Tags mit hohen Werten platziert, das rechte Ende der Kurve würde allerdings aus zahlreichen nahezu gleichhäufigen Tags bestehen und den sog. „Langen Schwanz", den „Long Tail", bilden. Wie sieht dies konkret aus?
In Abbildung 4 wird deutlich, dass sich bestimmte Tags durch ihre Häufigkeit von anderen Tags deutlich abgrenzen. Die Tags *politik* und *bundestag* auf der linken Seite der Kurve vereinen fast 80 Prozent der gesamten Taghäufigkeit auf sich, während die rechte Seite der Kurve mit Tags wie *culture*, *institutionen* und *news* zwar eine größere Vielfalt der Content-Beschreibung aufweist, die aber jeweils nicht so häufig genannt werden. Diese Menge an niedrigfrequenten nahezu gleichhäufigen Tags bildet den Long Tail. Nach dem Aufrufen der Webseite kann man erkennen, dass die Tags am Beginn der Kurve (*politik*, *bundestag*) den Website-Content in einem hohen Maße adäquat und allgemein, die Long Tail-Tags ihn jedoch sehr speziell beschreiben. Die Kurve zur Webseite www.bundestag.de und somit die Tagverteilung folgt hier dem Power Law mit einem Exponenten von ungefähr a = 1. Eine idealtypische Power Law-Verteilung zeigt Abbildung 5.
In einer Untersuchung gemeinsamen Auftretens von Tags („co-tags" in Analogie zu den „co-citations" der Zitationsanalyse) können Lux, Granitzer und Kern (2007) zeigen, dass rund 80% der Tags derart mit anderen Tags vorkommen, dass diese anderen Terme einem Power Law folgen. Wie lässt sich dies erklären? Einen sehr interessanten theoretischen Ansatz legen Cattuto (2006) sowie Cattuto, Loreta und Pietronero (2007) vor. Auch hier geht es – wie bei Lux et al. – um das Auftreten der Tags auf einer Makroebene (der gesamten Datenbank oder experimentell abgetrennter Teile davon) und nicht – wie in unserem Beispiel (Abbildung 4) – um die Mikroebene, der Zuordnung gewisser Tags zu konkreten Dokumenten. Cattuto et al. wenden den Ansatz der „Semiotics Dynamics" auf die Kreation neuer bzw. dem Verwenden alter Tags in einer Folksonomy an. "Folksonomies … do exhibit dynamical aspects also observed in human language, such as the emergence of naming conventions, competition between terms, takeovers by neologisms, and more" (Cattuto, 2006, 33). Das Erklärungsmodell hat seinen Ursprung im Yule-Simon-Ansatz, der Wahrscheinlichkeiten für das Auftreten von Worten in Texten thematisiert. An jeder Stelle eines Textes gilt: Ein bestimmtes Wort hat die Wahrscheinlichkeit p, ein neues Wort zu sein (d.h., noch nicht im bisherigen Text vorzukommen), oder es hat die Wahrscheinlichkeit 1 – p, eine Kopie eines bereits vorhandenen Wortes zu sein. Die Größe von 1 – p hängt davon ab, wie häufig das gegebene Wort schon im Text vorgekommen ist, wobei ein einfacher Zusammenhang hergestellt wird: Je häufiger das Wort bereits vorkommt, desto wahrscheinlicher ist es, dass es erneut Anwendung findet. Der Yule-Simon-Ansatz ist damit eine auf Texte angewandte Variante des altbekannten „Erfolg gebiert Erfolg" oder des „Wer hat, dem wird gegeben". Um dieses Modell auf Folksonomies anzuwenden, muss man nur statt von „Worten in Texten" von „Tags in Datenbanken" sprechen. Je häufiger ein Tag bereits in einem Informationsdienst auftritt, desto höher wird seine Wahrscheinlichkeit, dass er auch erneut zur Indexierung herangezogen wird. Cattuto et al. verfeinern den Yule-Simon-Ansatz um eine Zeitkomponente derart, dass neuere Tags eine größere Indexierungswahrscheinlichkeit haben als alte. „It seems more realistic to assume that users tend to apply recently added tags more frequently than old ones, according to a skewed memory kernel" (Cattuto, 2006, 35). Abbildung 6 zeigt das Erklärungsmodell im Überblick. Die Wahrscheinlichkeit p drückt aus, dass ein völlig neuer Tag zur Indexierung herangezogen wird, 1 – p entsprechend, dass ein alter gewählt wird. Letztere Wahrscheinlichkeit hängt von der Zeit ab, zu der der (alte) Tag genutzt worden ist. Die Zeitangaben sind im Gedächtnis Q gespeichert und steuern eine Gewichtung nach der Regel „je neuer, desto wahrscheinlicher".

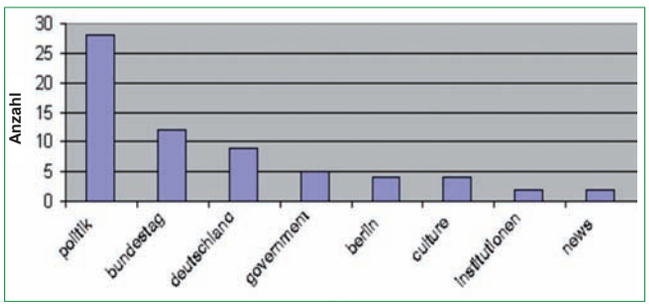


Abbildung 4: Tagverteilung zur Webseite www.bundestag.de. Quelle: http://del.icio.us (Stand: November 2007; Anzahl indexierender User: 60).

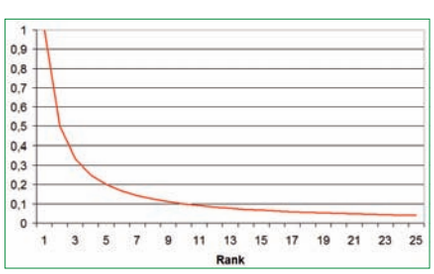


Abbildung 5: Power Law.

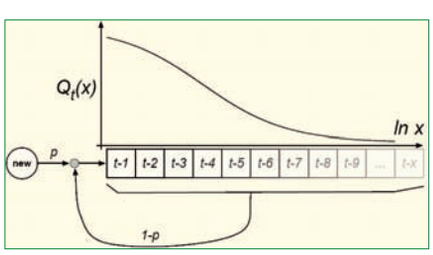


Abbildung 6: Ein Yule-Simon-Prozess der Auswahl von Indexierungs-Tags mit Berücksichtigung des Alters der Tags. Quelle: Cattuto, Loreto, & Pietronero (2007, Fig. 5).

Für die Makroebene erscheint der Ansatz von Cattuto und des Yule-Simon-Modells vielversprechend. Offen ist, ob sich diese Theorie auch auf die Mikroebene übertragen lässt. Wenn ja, müssten sich die Tags auf der Dokumentebene auch nach dem Power Law verteilen. Dies scheint aber nicht – oder zumindest nicht immer – zuzutreffen. Lux et al. (2007) konnten (auf der Makroebene) nur 80% der Co-Tag-Verteilungen als Power Law auszeichnen. 20% folgen offenbar anderen Regeln. Betrachten wir ein zweites Beispiel einer dokumentspezifischen Tag-Verteilung (Abbildung 7)!
Wie man sieht, folgt die Tagverteilung hier nicht dem Power Law. Es scheint, als würden sich hier zwei „Lange

Schwänze“ bilden: rechts der bekannte „Long Tail“ und links ein „Long Trunk“, der „Lange Rüssel“ Die Tags grenzen sich in ihren Häufigkeitsverteilungen nicht stark genug voneinander ab; es werden also viele Tags ähnlich häufig genannt. Der lange Rüssel nimmt bei www.asis.org die Tags *associations*, *library*, *information*, *ia*, *technology* und *professional* auf. Hiernach hat die Verteilung einen Wendepunkt, an den sich der lange Schwanz anschließt. Die Tags im langen Rüssel beschreiben den Inhalt der Webseite – anderes als im Fall der ersten zwei, drei Tags bei einer Power Law-Verteilung – nur wenig treffend. Die Tags *associations* und *technology* sind äußerst allgemein, und auch die restlichen Terme geben nur vage Hinweise auf den Content. Hier hat die kollektive Intelligenz offenbar Schwierigkeiten, ein „wahres“ Bild des Dokuments zu zeichnen. Es scheint jedoch auch invers-logistische Verteilungen zu geben, bei denen die Tags links vom Wendepunkt eine Webseite adäquat beschreiben. Margaret Kipp (zitiert in Peters, 2008) kann sogar demonstrieren, dass bei ihrem Beispiel die Tags im langen Rüssel im Zeitverlauf konstant bleiben. Für Kipp haben diese Power Tags den Charakter von kontrolliertem Vokabular – lediglich durch die Community der Prosumer „geprüft“.

Stock (2006) konnte in einer theoretischen Studie zeigen, dass es mindestens zwei verschiedene Verteilungen von Relevanz gibt: die bekannte Power Law-Verteilung und die invers-logistische Verteilung (siehe Abbildung 8). Die invers-logistische Verteilung hat viele mehr oder minder relevante Tags im „Long Trunk“ und zusätzlich ebenfalls einen Long Tail. Dabei folgt sie der Formel

$$f(x) = e^{-C'(x-1)^b}$$

bei der e die Euler'sche Zahl und x der Rang des Tags ist. C' (im Beispiel von Abbildung 8 ist es 0,1) ist eine Konstante und der Exponent b ist stets ungefähr 3. In vielen Fällen ist der „Long Trunk“ kürzer als der „Long Tail“.

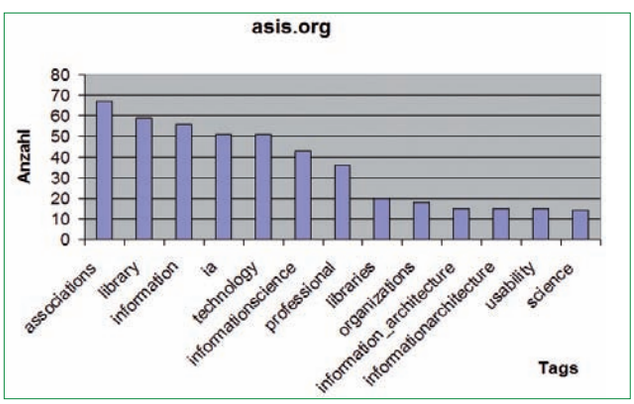


Abbildung 7: Tagverteilung zur Webseite www.asis.org. Quelle: http://del.icio.us (Stand: November 2007; Anzahl indexierender User: 293).

Einige Untersuchungen scheinen von der stillschweigenden Voraussetzung auszugehen, es gäbe ausschließlich die Power Law-Verteilung. Da bei beiden Verteilungskurven ein langer Schwanz auftritt, ist in der Tat eine Verwechslungsmöglichkeit gegeben. Man muss demnach besonderes Augenmerk auf den Anfang der Verteilung legen. Das entscheidende Kriterium, das ein Power Law ausschließt, ist das Vorliegen des langen Rüssels. Approximiert man eine Kurve unter Auslassen der ersten Rangplätze (besonders eklatant beispielsweise bei Capocci und Caldarelli (2007, Fig. 3) oder bei Cattuto (2006, Fig. 5)), so sind Fehlzuschreibungen nicht auszuschließen.

Bei der Beobachtung von Tags sollten beide Verteilungen im Auge behalten werden, da diese bei der Entwicklung und Erstellung von Information Retrieval-Werkzeugen für getaggte Dokumente wieder zur Sprache kommen müssen. Es haben nämlich die Tags im langen Rüssel bei der invers-logistischen sowie die ersten n Tags beim Power Law eine Sonderstellung, die wir sowohl für eine spezifische Rechercheoption als auch für das Relevance Ranking ausnutzen wollen.

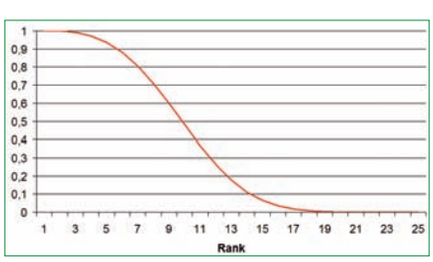


Abbildung 8: Invers-logistische Verteilung. Quelle: Stock (2006), S. 1127 (modifiziert).

Nicht unerwähnt bleiben sollen Dienste, die informetrische Analysen von Tags anbieten. So zeigt beispielsweise Technorati Zeitreihen von Blogbeiträgen in Bezug auf Themen (für ein Beispiel siehe Stock & Weber, 2006), während Cloudalicious Zeitreihen von Tags zu bestimmten URLs in Del.icio.us generiert (Russell, 2006). Hiermit öffnet sich ein Forschungsfeld für die Analyse von Regelmäßigkeiten zeitlicher Tagverteilungen.

Abbildung 9: Getaggtes Foto in Flickr. Quelle: Flickr.com.

Ofness – Aboutness – Ikonologie

Wir kommen zu einem nicht unerheblichen Problem bestimmter Anwendungen von Folksonomies, nämlich zur kollaborativen Erschließung nicht-textueller Dokumente. Manche Web 2.0-Dienste verzeichnen und sammeln Fotos (z.B. Flickr) oder Videos (z.B. YouTube). Nicht-textuelle Dokumente (und auch einige fiktionale Texte) haben nicht nur eine Ebene des jeweils Thematisierten, im Gegensatz etwa zu wissenschaftlichen Zeitschriftenartikeln. Fachliche textuelle Dokumente tragen auf einer einzigen semantischen Ebene Content, gewöhnlich mit „Aboutness“ beschrieben (Stock & Stock, 2008, 31-36). Nach Panofsky (2006) gibt es für Kunstwerke drei verschiedene semantische Stufen der Interpretationsmöglichkeiten (Smith, 2006). Wir wollen Panofskys Theorie an einem Beispiel aus Flickr erklären (siehe Abbildung 9). Das Foto zeigt eine Gruppe von Skulpturen („Mönch – Arzt – Händler“), die von der deutschen Künstlerin Katharina Fritsch erstellt wurden und im Düsseldorfer Museum K21 ausgestellt werden. Auch die zu dem Foto annotierten Tags sind aufgeführt.

Panofskys grundlegendes semantisches Level wird prä-ikonographisch genannt und beschreibt die Welt der primären bzw. natürlichen Objekte. Auf dem Foto sind drei Figuren zu sehen: eine rote Figur mit einem Pferdefuss, eine weiße mit einem Totenkopf und eine schwarze bekleidet mit einer Kutte. Für dieses Level der Interpretation benötigt der Interpret keine weiteren Kenntnisse als die aus alltäglicher Erfahrung. In der Informationswissenschaft nennt man das Ergebnis der prä-ikonographischen Interpretation die Ofness eines Dokuments (Lancaster, 2003; Layne, 2002; Markey, 1986; Shatfort, 1986; Turner, 1995). Die Ofness-Tags in diesem Beispiel sind (interessanterweise nur in englischen Termen ausgedrückt) *black*, *white* und *red*.

Das ikonographische Level umfasst die sekundären Objekte; dies ist die Welt

der Bilder, Anekdoten und Allegorien (Panofsky, 1975, 50). Der Interpret muss hier einiges an Erfahrung und Wissen im soziokulturellen sowie thematischen Umfeld des Kunstwerkes für die Interpretation mitbringen. Dieses Level bezeichnet man als die Aboutness des Dokuments. Aboutness-Tags sind *monk*, *doctor* und *dealer* sowie die deutschen Übersetzungen davon. (Nunmehr arbeitet unser Tagger mit englischen und mit deutschsprachigen Worten.)

Das dritte Level bezeichnet Panofsky als ikonologisch; es besteht aus dem „wahren" Inhalt, aus der Welt der symbolischen Werte (Panofsky, 1975, 50). Umfassende Kenntnisse des thematischen Gebiets sind hier für eine adäquate Interpretation notwendig. Informationswissenschaftler berücksichtigen normalerweise dieses Level nicht in der Inhaltserschließung; User der Web 2.0-Dienste können aber den semantischen Gehalt ihrer Tags selbst bestimmen. In dem Beispiel ist nur ein, allerdings sehr schwacher, ikonologischer Tag zu finden: *art* (bzw. *Kunst*). Eine stärkere ikonologische Beschreibung könnte z.B. *unheilige Dreifaltigkeit* sein.

Darüber hinaus werden auch nicht-inhaltliche Informationen vom Nutzer zum Dokument hinzugefügt, wie z.B. der Name des Museums, *K21*, oder der Name der Künstlerin, *KatharinaFritsch*. Diese Arten von Tags beschreiben in der Diktion von Peter Ingwersen (2002) die Isness des Dokuments. Katharin Fritsch *ist* (is) die Künstlerin des Objekts; K21 *ist* (is) das ausstellende Museum. Folksonomies machen keinen Unterschied zwischen den verschiedenen Beschreibungs- und Interpretationsebenen; sie fließen zu einem einzigen semantischen Level zusammen.

Nach Ingwersen (2002) werden vier Aspekte von Aboutness unterschieden: a) die Autor-Aboutness beschreibt den Content, wie er ist, b) die Indexer-Aboutness die zweckgebundene Interpretation des Contents, c) die Anfrage-Aboutness den Content als Ausdruck der Suchanfrage und letztlich d) die User-Aboutness, die Interpretation des Contents durch den Nutzer. Das Taggen von Fotos, Videos usw. ist als nutzerzentriert zu klassifizieren, d.h. beim Level 1 handelt es sich um die User-Ofness, Level 2 umfasst die User-Aboutness und Level 3 die User-Ikonologien; hinzu tritt als vierte Ebene die Isness aus Nutzersicht. All diese verschiedenen Stufen von Interpretationen werden in einer Tag Cloud zusammengefasst und nicht unterschieden. Trotz dieser Problematik stellt Smith in Bezug auf das Taggen von Museumsobjekten durch die Nutzer fest: "Tagging has the potential of increasing access to artwork images and records dramatically for searchers of all levels of expertise" (Smith, 2006, 13). Studien über den Einsatz neuer nutzerzentrierter Dokumentationsmethoden in Museen zeigen, dass die meisten der vom Nutzer kreierten Schlagworte nicht in der bisher verwendeten Wissensordnung vorkommen, sondern neu sind (Trant, 2006b, 21) und dass der Einsatz von Folksonomies neue Möglichkeiten schafft, ein Kunstwerk zu finden (Trant, 2006a).

In zahlreichen informationswissenschaftlichen Veröffentlichungen wird das Thema Inhaltserschließung von Fotos und Bildern behandelt (z.B. Rasmussen, 1997), allerdings nicht mit Bezug auf Folksonomies und Web-2.0-Dienste, sondern beispielsweise auf Pressebilder.

Ornager (1995, 214) hat dazu herausgefunden, dass sich die Inhaltsanalyse von Fotos und Bildern auf hauptsächlich fünf Kategorien stützt: „named person (who), background information about the photo (when, where), specific events (what), moods and emotions (shown or expressed), size of photo." Markkula und Sormunen (2000, 270 f.) stellen fest, dass "(t)he most often used index terms referred to specifics, i.e. to individual objects, places, events and linear time, and to the theme of the photo". Zur Inhaltserschließung und zur Beschreibung der abgebildeten Objekte verwenden die Indexer häufig Nomen, jedoch werden Handlungen und Ereignisse auch mit Verben beschrieben („Czechs *celebrate* the victory..."; Markkula & Sornumen, 2000, 271). Eine wichtige Beobachtung ist: "(T)he output of the indexing process seemed to be quite inconsistent" (Markkula & Sornumen, 2000, 273). Diese Indexierungsinkonsistenz dürfte in der doch stark subjektiven Färbung der Bildbetrachtung und -auswertung begründet liegen (Neal, 2007).

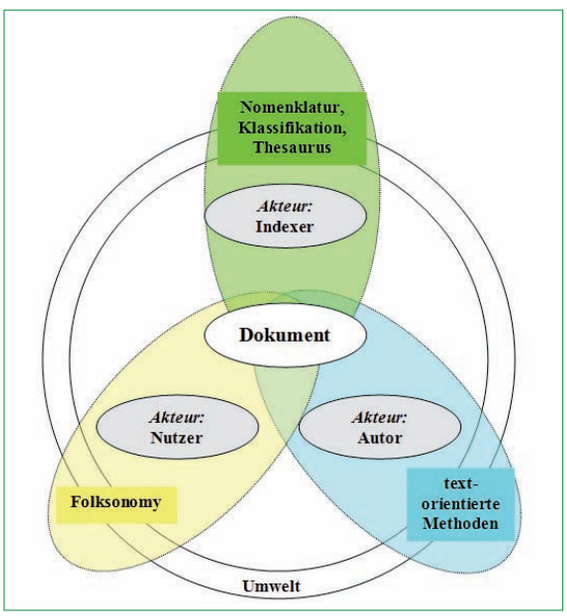


Abbildung 10: Methoden der Wissensrepräsentation und deren Akteure. Quelle: Stock & Stock (2008, S. 41).

Die Vorteile von Folksonomies

Folksonomies sind eine neue Methode, um den Inhalt von Dokumenten zu erfassen. Die Prosumer – oder die „Akteure" im Sinne der Analyse kognitiver Arbeit (Fidel, 2006) – indexieren selbst das Dokument. Dabei spiegelt die Folksonomy die Sprache und das Wissen der Prosumer authentisch wider (Quintarelli, 2005). Diese Art der Inhalterschließung führt zu einer Vielzahl an „multiple interpretations", unterschiedlichen (und manchmal unvereinbaren) Meinungen und „multicultural views" von ein und derselben Information (Peterson, 2006). Diese "shared intersubjectivities" ermöglichen es dem Nutzer "to benefit, not just from their own discoveries, but from those of others" (Campbell, 2006, 10).

Folksonomies komplettieren die Werkzeuge und Methoden der Wissensrepräsentation, wie dies in Abbildung 10 skizziert ist. Wir unterscheiden drei Gruppen von Akteuren, die Dokumente indexieren: Autoren, professionelle Indexer und Nutzer (Kipp, 2006b). Hierbei werden unterschiedliche Wege der Indexierung beschritten, die wahrscheinlich auch unterschiedliche Merkmale der Dokumente jeweils in den Vordergrund rücken. Textorientierte Methoden machen in einem großen Maße Gebrauch von der Sprache des Autors, z.B. bei der Textwortmethode oder bei der Zitationsindexierung. Im Gegensatz zu diesen Methoden berücksichtigen Folksonomies nicht nur die Sprache des Autors, sondern auch die Sprache der Nutzer. Wissensordnungen wie Nomenklaturen, Thesauri oder Klassifikationssysteme benötigen für ihren Gebrauch Interpreten oder Übersetzer. Zum einen sind dies die Experten, die das Vokabular erstellen, und zum anderen die, die das Vokabular zur Inhaltserschließung benutzen. Die Entwickler von kontrollierten Vokabularien müssen zunächst die „literature, needs, actors, tasks, domains, activities, etc." (Mai, 2006, 17) analysieren, um daraus das kontrollierte Vokabular zu erstellen. Dies ist ohne Zweifel eine zeitaufwändige und teure Prozedur. Dagegen sind Folksonomies günstig im Unterhalt, denn die Pflege und Indexierung geschieht durch die freiwillige und kostenlose Zusammenarbeit vieler Nutzer. "Tagging has dramatically lower costs because there is

no complicated, hierarchically organized nomenclature to learn. User simply create and apply tags on the fly" (Wu, Zubair, & Maly, 2006, 111). Darüber hinaus gibt es eine hohe Bereitschaft seitens der User, selbst zu taggen und so den Taggingprozess in der Community anzuregen und zu verstärken (Fichter, 2006).

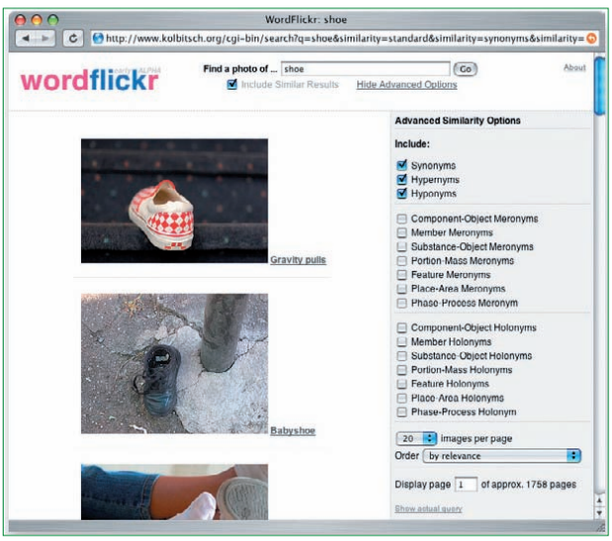


Abbildung 11: WordFlickr: Synthese aus Web-2.0-Dienst und Thesaurus. Quelle: Kolbitsch (2007a, S. 85).

Aus praktischen Gründen ist es wohl unmöglich, alle URLs, Blogeinträge, Fotos und Videos im World Wide Web intellektuell mit kontrollierten Vokabularien zu erschließen. Algorithmische Suchmaschinen wie Google oder Yahoo! können textuelle Dokumente automatisch indexieren. Für Dokumente wie Fotos oder Videos gestaltet sich eine rein automatische, content-basierte Erschließung jedoch schwierig, da sie keine Metadaten enthalten.
Welchen Nutzen bewirken Folksonomies? Wir wollen an dieser Stelle einige kurze Überlegungen anstellen.

1) Die Entwicklung und die Pflege von bestehenden kontrollierten Vokabularien kann von Folksonomies profitieren (Aurnhammer, Hanappe, & Steels, 2006; Christiaens, 2006; Gendarmi & Lanubile, 2006; Macgregor & McCulloch, 2006; Mika, 2005; Zhang, Wu, & Yu, 2006). Die Tags, ihre Häufigkeit und ihre Verteilungen können als Quelle für neue kontrollierte Terme, für Term-Modifikationen und Term-Löschungen dienen. Diese „bottom-up categorization" (Vander Wal, 2004) garantiert schnelle Reaktionszeiten bezüglich neuer Themen und Innovationen in der Wissensdomäne und ist damit ein wichtiges Instrument sowohl zum Neuaufbau als auch zur Pflege von Nomenklaturen, Klassifikationssystemen und Thesauri. Doch inwieweit können Folksonomies praktisch zur Entwicklung von kontrollierten Vokabularien, von Ontologien bis hin zum Semantischen Web beitragen (Al-Khalifa, 2007; Specia & Motta, 2007)? Die bis jetzt entwickelten Modelle arbeiten hauptsächlich mit Co-Occurences (Schmitz, 2006), einfachen Cluster-Algorithmen (Grahl, Hotho, & Stumme, 2007) oder dem Vektorraum-Modell (Dimensionen: Dokumente, Vektoren: Tags, Tagähnlichkeit: Cosinus) (Heymann & Garcia-Molina, 2006). Die entstandenen Ähnlichkeitswerte können als Grundlage für einen Ähnlichkeitsgraphen dienen, bei dem die Position des Tags Auskunft über die hierarchische Lokalisierung in einer „latent hierarchical taxonomy" (Heymann & Garcia-Molina, 2006, 4) geben kann.

2) Folksonomies ergänzen kontrollierte Vokabularien in der Recherchepraxis. Kann man bei Folksonomies und Ontologien überhaupt von Rivalen sprechen? Für Tom Gruber (2006, 994) ist dies "nonsense, and it is time to embrace a unified view". Die Grundidee hierbei ist es, Tags in ihrem semantischen Umfeld darzustellen, aus dem der Nutzer zusätzliche Suchargumente gewinnen kann. Da Folksonomies nie explizit paradigmatische Umfeldterme nennen, müssen diese aus anderen Wissensordnungen gewonnen werden. Infrage kommen alle fachlichen wie linguistischen Begriffsordnungen. WordNet (Miller, 1998) konnte bereits erfolgreich bei Daten aus Del.icio.us (Laniado, Eynard, & Colombetti, 2007) sowie bei Flickr (Kolbitsch, 2007b) zum Einsatz kommen. Abbildung 11 zeigt die experimentelle Oberfläche von WordFlickr, einem Mash-up aus WordNet und Flickr.

3) Journalisten oder Wissenschaftler müssen an der Hürde der Gatekeeper und ihrer Qualitätskontrolle vorbei, wollen sie ihre Werke veröffentlichen. Es stellt sich die Frage, ob es solche Gatekeeper auch im Internet gibt. Clay Shirky (2005) sagt dazu: "The Web has an editor, it's everybody". Eine Qualitätskontrolle des Contents findet statt – jedoch erst nach seiner Veröffentlichung. Je mehr Nutzer ein Dokument taggen, desto mehr Relevanz scheint dieses Dokument für sie zu haben. Ist dies aber eine ernstzunehmende Qualitätskontrolle? Wird etwas zu „geprüfter" Qualität, nur weil viele Leute dies so sehen? (Wenn viele Studenten bei einer Mathematikklausur die gleiche – falsche – Lösung bringen, wird diese nicht dadurch qualitativ wertvoll, sondern bleibt falsch.[1]) Quantität bedeutet nicht Qualität. Andererseits weist es in eine bestimmte Richtung, wenn viele Nutzer ein Stück Information mit *stupid* und ein anderes mit *cool* taggen. Dieser Content könnte für das Relevance Ranking verwertet werden.

4) Das Retrieval mit Folksonomies kann auf zweifache Weise erfolgen: a) der Nutzer gibt einfach mögliche Tags in die Suche ein oder b) er nutzt die Tags, um sich zum gewünschten Dokument „durchzuhangeln". Dieser letzte Punkt ist für Adam Mathes (2004) die zentrale Eigenschaft und Stärke von Recherchen mit Folksonomies – Serendipity. "The long tail paradigm is also about discovery of information, not just about finding it" (Quintarelli, 2005). Das Suchen mit Tags ist für Laien viel einfacher und schneller als die Recherche mit elaborierten Information Retrieval-Werkzeugen, wie z.B. der Internationalen Patentklassifikation (IPC). Einige Nutzer verzichten sogar völlig auf die Eingabe von Suchargumenten und klicken sich durch die Tag Clouds zur gewünschten Information durch (Sinclair & Cardew-Hall, 2008).

5) Es gibt Tags, die in Folksonomies zum ersten Mal auftauchen, sog. Neologismen. In Mathes (2004) werden die Beispiele *sometaithurts* (für „so meta it hurts") und *flicktion* diskutiert. „Although small, there is a quick formation of new terms to describe what is going on, and others adopting that term and the activity it describes" (Mathes, 2004). Dieser unerwartete und unvorhersehbare Gebrauch von Sprache und Tags reflektiert deutlich die „communication and ad-hoc group formation facilitated through metadata" (Mathes, 2004).

6) Betrachtet man alle Dokumente, Tags und Nutzer als Knoten in einem Netzwerk, so kann man leicht die Communities identifizieren, die ein Thema oder ein Interessensgebiet teilen: „Collaborative tagging systems have the potential of becoming a technological infrastructure for harvesting social knowledge" (Wu, Zubair, & Maly, 2006, 114).

7) Auf der Basis gemeinsam verwendeter Tags oder gemeinsam indexierter Dokumente wird es möglich, Folksonomies für Zwecke von impliziten kollaborativen Recommendersystemen zu nutzen (Diederich & Iofciu, 2006; Niwa, Doi, & Honiden, 2006).

8) Ein weiterer Vorteil von Folksonomies ist (auch wenn es dafür noch keine wissenschaftliche Evidenz gibt), dass sie dem Nutzer die Schwierigkeiten, aber

1 Dieses illustrative Beispiel verdanken wir unserem Studenten Peter Bücker.

auch Nützlichkeit von Inhaltserschließung und Indexierungsmethoden näher bringen und ihn dafür sensibilisieren.

Wir wollen die Vorteile der Folksonomies noch mal zusammenfassen. Folksonomies:

- spiegeln die Sprache der Nutzer authentisch wider,
- erlauben verschiedene Interpretationen,
- sind eine günstige Form der Inhaltserschließung,
- sind die einzige Möglichkeit, Massen-Information im Web zu erschließen,
- sind Termquellen für die Entwicklung und Pflege von Ontologien und kontrollierten Vokabularien,
- geben die Qualitätskontrolle an die Nutzer weiter,
- erlauben konkretes Suchen *und* Browsing,
- berücksichtigen Neologismen,
- tragen dazu bei, Communities zu identifizieren,
- geben eine Basis für Recommendersysteme,
- sensibilisieren Nutzer für die Inhaltserschließung.

Die Nachteile der Folksonomies

Es resultieren zahlreiche Probleme aus dem Fehlen eines kontrollierten Vokabulars. „Lack of precision is a function of user behaviour, not the tags themselves", wiegelt Shirky (2004) ab. Nichtsdestotrotz findet man in Folksonomies oftmals Uneindeutigkeiten, wie verschiedene Wortformen, Nomen in Plural, Nomen in Singular oder Abkürzungen, die das Suchen und Finden von Dokumenten erschweren. Weitere Beispiele präsentiert der Web 2.0-Dienst Del.icio.us (siehe oben Abbildung 7). Da hier jeweils nur ein Wort als Tag eingegeben werden kann, werden Phrasen (*informationarchitecture*) aus zwei oder mehr Wörtern kreiert oder einzelne Wörter mit einen Unterstrich (*information_architecture*) verbunden. Abkürzungen (*ia*) finden Verwendung; Singular- (*library*) wie Pluralformen (*libraries*) kommen nebeneinander vor. Es fehlt die Zusammenfassung von Synonymen, auch über Sprachgrenzen hinweg, und die Trennung von Homonymen. Daneben finden sich viele Rechtschreib- oder Tippfehler und unterschiedliche Schreibweisen. Daher lassen sich Folksonomies auch als unstrukturierte Liste von Schlagwörtern bezeichnen (Reamy, 2006). „Currently most users don't give much thought to the way they tag resources, and bad or ‚sloppy' tags are ten-a-penny in Folksonomies" (Guy & Tonkin, 2006). Das Tagging-Verhalten der User ist also unberechenbar und nicht vorhersehbar, was das Information Retrieval drastisch erschwert. Ein paar Zahlen sollen dies verdeutlichen: Nach Guy und Tonkin (2006) sind fast 40 Prozent der Flickr-Tags und 28 Prozent der Del.icio.us-Tags „either misspelt, from a language not available via the software used, encoded in a manner that was not understood by the dictionary software, or compound words consisting of more than two words or a mixture of languages".

Die Prosumer, die ein Dokument erschließen, handeln in unterschiedlichen Kontexten und aus verschiedenen Motivationen: Ein Nutzer annotiert Tags zu einem Dokument aus beruflichen Hintergründen, ein anderer hat sein Freizeitvergnügen im Hinterkopf. Es fehlt daher ein einheitliches, grundlegendes Level der Inhaltserschließung. "There is […] systematic variation across individuals in what constitutes a basic level" (Golder & Huberman, 2006, 200). Erfahrungen, sprachliche Ausdrucksstärke, kognitive Fähigkeiten und Motivationen sind alles Faktoren, die den Grad der Inhaltserschließung bestimmen und beeinflussen. Darüber hinaus muss berücksichtigt werden, dass "a significant amount of tagging, if not all, is done for personal use rather than public benefit" (Golder & Huberman, 2006, 207).

Web 2.0-Services finden weite Verbreitung und erfreuen sich wachsender Popularität – und das auf der ganzen Welt. Die Problematik, die hier auftaucht, betrifft die Vielsprachigkeit unter den Usern. Das „merging of languages" (Gordon-Murnane, 2006) führt zu Problemen in der intersprachlichen Synonymie oder Homonymie (z.B. deutsch: *Gift* und englisch: *gift*) und schwächt die Präzision der Suchresultate bzw. zeigt u. U. falsche Ergebnisse. Darüber hinaus gibt jeder Nutzer seine Tags vornehmlich (auch) in seiner Muttersprache ein und kann wahrscheinlich nur einen begrenzten Teil anderssprachiger Tags verstehen – und somit auch nur nach einem begrenzten Teil in der gesamten Datenbank suchen und darauf zugreifen.

Folksonomies machen im Gegensatz zu Thesauri und Klassifikationssystemen keinen Gebrauch von Relationen zwischen den Begriffen, sog. paradigmatischen Relationen, wie z.B. Äquivalenzrelation, Hierarchierelation oder Assoziationsrelation. Jedoch bestehen in der aktuellen Co-Occurrence von mehreren Tags zu einem Dokument zwischen diesen syntagmatische Relationen (Stock, 2007a, 451). Wir wollen dies an einem Beispiel verdeutlichen (Abbildung 12).

Die Tags *UK* und *england* sind durch eine geographische Meronymie verbunden, während die Tags *seat* und *stadium* durch eine Komplex-Komponenten-Relation gekennzeichnet sind. Diese syntagmatischen Relationen sind eigentlich paradigmatisch und könnten zur Suchfrageerweiterung oder -einschränkung genutzt werden (Weller & Peters, 2007; Peters & Weller, 2008). Jedoch sind diese nützlichen Relationen in der Tagliste verborgen und werden nicht von Folksonomies eingesetzt.

Professionell erstellte Metadaten werden in verschiedene Felder segmentiert, z.B. in *Dokumenttyp*, in *Autor* oder in die *Notationen eines Klassifikationssystems*. Die Inhaltserschließung berücksichtigt hier die formalen Aspekte eines Dokuments und die Aspekte der Aboutness und unterscheidet sie voneinander. Folksonomies vernachlässigen diese strikte Trennung von unterschiedlichen Metadaten, machen aber sehr wohl Gebrauch von Tags, die nicht nur den Inhalt des Dokuments beschreiben, wie z.B. einen Namen für den Besitzer des Dokuments oder den Hinweis, dass es sich um ein Buch handelt. Daneben werden auch Wertungen vorgenommen (*stupid*). Diese user-spezifischen Tags beschreiben oder evaluieren ein Dokument nur aus der Sicht des einen Users, sodass diese Tags „are virtually meaningless to anybody except their creators" (Pluzhenskaia, 2006, 23). Zahlreiche andere Tags können performativ genannt werden, d.h. dass sie als Ausdruck einer geplanten oder erledigten Aktivität annotiert werden, z.B. *toread* oder *todo* (Kipp, 2006). Auch synkategorematische Tags kommen vor, d.h. sie können nur im Kontext verstanden werden. Ein typisches Beispiel für einen synkategorematischen Tag ist *me* bzw. *ich*, wenn er zur Beschreibung der Eigenaufnahme des Fotografen annotiert wird. Es ist auch nicht auszuschließen, dass einige der Tags absichtlich als Spam-Tags von Nutzern verwendet werden. "(I)t involves an unethical user who propagates […] tags in order to corrupt a system" (Kroski, 2005). Beim Tagging von Fotos und Videos kommt die Schwierigkeit hinzu, dass sich unterschiedliche Levels der Indexierung bezüglich Ofness, Aboutness, Ikonologie und Isness miteinander vermischen, sich nicht unterschei-

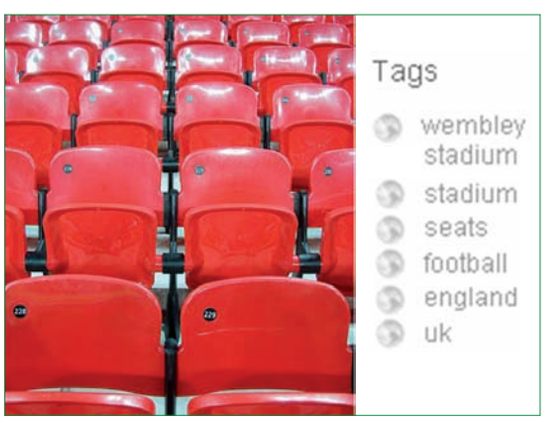


Abbildung 12: Syntagmatische und paradigmatische Relationen zwischen Tags. Quelle: Flickr.com.

den lassen und in ihrer Inter- und Intraindexerkonsistenz erheblich schwanken. Wir wollen die Nachteile der Folksonomies noch einmal kurz festhalten:

- fehlendes kontrolliertes Vokabular,
- verschiedene Levels der Indexierung,
- Vermischung von Sprachen,
- versteckte paradigmatische Relationen bleiben ungenutzt,
- fehlende Trennung von formalen bzw. bibliographischen Tags und Aboutness-Tags,
- Spam-Tags, nutzer-spezifische Tags und andere uneindeutige Schlagworte,
- Verschmelzung von Ofness, Aboutness, Ikonologie und Isness.

NLP – eine Problemlösung für Folksonomies?

Um die oben genannten Probleme zu lösen, könnten sich zwei Vorgehensweisen etablieren, die sich gegenseitig ergänzen. Zum einen kann man sich auf die Indexer bzw. Prosumer konzentrieren und versuchen, diese zu einer besseren „tag literacy" (Guy & Tonkin, 2006) zu erziehen. Dazu muss man zunächst das Nutzerverhalten bezüglich Tagging beobachten (siehe z.B. Bar Ilan et al.; 2006; Lin et al., 2006; Morrison, 2007; Veres , 2006a oder Winget, 2006). Die Nutzerforschung steht hinsichtlich der Folksonomies allerdings erst am Anfang ihrer Bemühungen.

Um die Nutzer bei der Auswahl „richtiger" Tags zu unterstützen und sie in der Tagvergabe zu trainieren, bietet es sich an, Tags durch das System vorzuschlagen (MacLauri, 2005; Xu, Fu, Mao, & Su, 2006). Dieses Vorschlagssystem kann auf zwei Ebenen arbeiten. Wir unterscheiden zwischen syntaktischen Vorschlägen (z.B. den Tag *Bild* durch den Tag *Bilder* ergänzen) und relationalen Hinweisen (z.B. der Nutzer annotiert den Tag *Bild*, und das System schlägt *Grafik* vor, weil unser Nutzer bereits den Tag *Grafik* verwendet hat). Der Einsatz von weitergehender Tag-Recommendation bei der Indexierung ist nicht frei von Problemen – im Gegensatz zu Vorschlägen von Such-Tags im Retrieval (s. u.!). Schlägt ein System nämlich dem Indexer die jeweils bereits am häufigsten vergebenen Tags eines Dokuments vor und orientieren sich die indexierenden Nutzer tatsächlich daran, so entsteht – in einer Art self-fulfilling prophecy – stets eine Tag-Verteilung nach dem Power Law.

Zum anderen kann man die Tags als Elemente der natürlichen Sprache ansehen und sie mit den etablierten Methoden des Natural Language Processing (NLP) bearbeiten (Stock, 2007a, Kap. 14 bis 18). Studien zeigen, dass circa 90 Prozent der Tags Nomen sind (Guy & Tonkin, 2006). Ob auch andere Wortarten, vor allem Verben, einmal häufig genutzt werden, bleibt abzuwarten. Da Adjektive an Nomen gebunden sind, müssen wir diese stets mitbeachten. Aus Gründen der Einfachheit nehmen wir an, dass der größte Teil der Tags Nomen (einschließlich Adjektiven) sind. Deswegen können wir während der Tag-Bearbeitung mittels NLP-Verfahren alle anderen Wortformen vernachlässigen.

Für die NLP-Verarbeitung der Tags bevorzugen wir einen wort-basierten Ansatz, da dieser bessere Möglichkeiten zur Bearbeitung bietet (siehe Abbildung 13 die drei untersten Ebenen). Eine (hier nicht weiter verfolgte) alternative Möglichkeit bietet die Implementierung von *n*-Gramm-Verfahren (Stock, 2007a, Kap. 13).

Problematisch ist, dass wir nicht davon ausgehen können, dass alle Dokumente getaggt sind. Für textuelle Dokumente können wir jedoch einen Ersatz generieren: Bei Blogs haben wir den gesamten Text als Basis, bei Fotos und Videos zumindest die Titel, Beschreibungen und Kommentare. Die wichtigsten Terme bestimmen wir automatisch (Brooks & Montanez, 2006a; 2006b). Mit dem Ergebnis aus WDF*IDF lässt sich ein Ranking erstellen; die ersten n (etwa n = 3) Terme werden als Tags bestimmt. Dieses Verfahren darf nicht zu häufig Anwendung finden, da die automatische Indexierung dem Grundgedanken des intellektuellen Taggings zuwiderläuft. Eine Studie von Al-Khalifa und Davis (2007) zeigt im Trend, dass die von den Prosumern indexierten Tags nur wenig mit Termen korrelieren, die auf der Basis automatischer Verfahren gewonnen werden.

Nach der Spracherkennung und dem Parsing müssen wir die kontextspezifischen synkategorematischen Tags, wie *me* und *ich*, bearbeiten. Für diese Problemfälle schlagen wir vor, sie automatisch durch den Nutzernamen des angemeldeten Users zu ersetzen. Sucht dieser spezielle Nutzer nach sich selbst, indem er den Tag *me* nutzt, erhält er die passenden Resultate. Alle nicht angemeldeten User erhalten für eine Suchanfrage mit *me* keine Ergebnisse. Nur die Suche nach dem Benutzernamen zeigt ein Resultat.

Der Bearbeitungsalgorithmus folgt den typischen NLP-Aufgaben (Fehlererkennung, Wortformzusammenfassung, Identifikation von Eigennamen, Phrasenerkennung und Dekomposition), wie in Abbildung 13 dargestellt. „Höhere" Bearbeitungsschritte verlangen nach einem Mash-up aus Folksonomy und kontrolliertem Vokabular. Während der Erkennung von Homonymen und Synonymen muss man Wissensordnungen wie beispielsweise WordNet (Miller, 1998) zur Hilfe nehmen. Außerdem könnte es hilfreich sein, bei der Homonymunterscheidung auch Co-Occurrence-Statistiken der Tags einzubeziehen (Butterfield et al., 2006).

Man stelle sich ein Tag Cluster vor, dass die Tags *Java*, *Perl* und *Programmierung* enthält, und ein weiteres Cluster, welches *Java*, *Sumatra* und *Indonesien* als Tags aufweist. Das System könnte beide Tag-Cluster anzeigen und den User, der nach *Java* gesucht hat, selbst auswählen und entscheiden lassen, welche Bedeutung des Tags er tatsächlich gesucht hat. Ein Merging von Thesauri, Klassifikationssystemen oder Ontologien mit Folksonomies kann die Nutzung von Relationen (wie Hierarchie oder Assoziationsrelation) ermöglichen (Gruber, 2007) und so die Suchanfrage erweitern bzw. eingrenzen. Im Sinne des multilingualen Retrievals erleichtert der Einsatz von mehrsprachigen maschinenlesbaren Wörterbüchern den Zugriff auf fremdsprachige Dokumente.

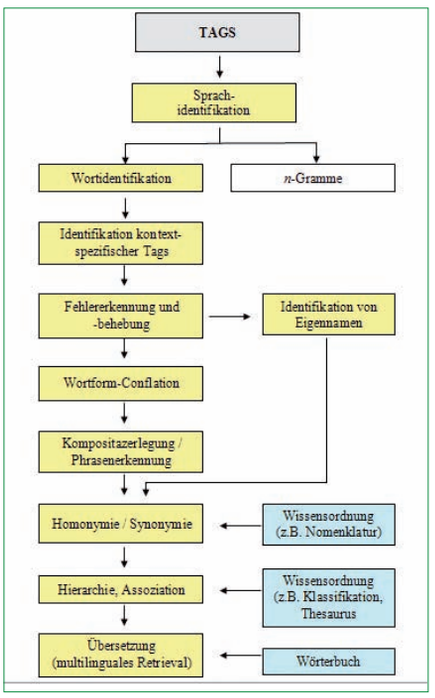


Abbildung 13: Der Aufgaben-Algorithmus der Tag-NLP.

Recherchen nach Power Tags

Wir wollen kurz einen Blick zurück auf die Abbildungen 5 und 8 werfen. Hier gibt es zwei Möglichkeiten, den Einfluss der Power Tags im Retrieval zu verstärken: a) Im Fall einer Power Law-Verteilung werden die ersten n Tags genutzt. n wird hier – in Abhängigkeit von dem Exponenten a – eher klein gehalten; bei a = 1 bietet sich beispielsweise n = 4, bei a = 2 etwa n = 2 an. b) Im Fall der invers-logistischen Verteilung wählen wir alle Tags des linken „Long Trunk" bis zum Wendepunkt der Kurve aus. Es wäre wünschenswert, eine zusätzliche, optionale Retrievalfunktionalität zu installieren, die nur Ergebnisse aus Übereinstimmungen von Suchargument

und Power Tag erzielt („Power Tags only"). Demnach würde das Beispiel in Abbildung 4 nur mit den Tags *politik* und *bundestag* gefunden werden. Alle anderen Terme und Tags blieben unberücksichtigt. Wir erhoffen uns durch das Angebot von Power Tags als Suchoption eine Verbesserung der Präzision der Suchergebnisse. Begründen können wir diese Vermutung durch das gegenläufige Verhältnis von Recall und Precision. Steigt das eine, fällt das andere. Im Falle der Recherche ausschließlich nach Power Tags wird der Recall – da der gesamte dokumentspezifische „Lange Schwanz" abgeschnitten wird – drastisch verkleinert.

Relevance Ranking der getaggten Dokumente

Ein bislang noch weitgehend ungelöstes Problem ist das Ranking von getaggten Dokumenten, die man als Antwort auf seine Suchanfrage erhält. Die Menge der Dokumente, die man nach der NLP-Bearbeitung bekommt, ist nämlich noch unsortiert.

Es gibt eine Patentanmeldung von Yahoo! für Flickr, die sich mit diesem Problem auseinandersetzt (Butterfield et al., 2006). Für ein Ranking nach „Interestingness" in Narrow Folksonomies wurden fünf generelle Kriterien definiert: 1) die Zahl der Tags zu einem Dokument, 2) die Zahl der Nutzer, die ein Dokument taggen, 3) die Zahl der Nutzer, die das Dokument nach einer Suche erhalten, 4) die Zeit (je älter desto weniger relevant) und 5) die Relevanz der Tags. Zusätzlich gibt es zwei weitere Rankingkriterien f,,ür den „personalized interestingness rank", 6) Nutzerpräferenzen (z.B. Favoritenlisten) und 7) der Wohnort des Nutzers. Beim Faktor der Zeit melden wir Bedenken an. Es ist nicht notwendig, dass ein Dokument im Laufe der Zeit stets an Relevanz verliert. Dann wäre die Relevanz von, sagen wir, der „Mona Lisa" heutzutage nahe Null, was Kunsthistoriker sehr wahrscheinlich bestreiten würden. Butterfield et al. berücksichtigen u. E. zu Recht die kollaborativen Aspekte, die wir bei Web 2.0-Diensten vorfinden; sie vernachlässigen jedoch insbesondere altbekannte Retrievalmodelle wie beispielsweise das Vektorraummodell oder die Linktopologie. Auch sollte man konkrete Aktionen der Prosumer nicht außer Acht lassen.

Für das Relevance Ranking der getaggten Dokumente gibt es nämlich drei Mengen von Faktoren:

- die (informationslinguistisch „gereinigten") Tags selbst,
- die Kollaboration in Web 2.0-Diensten,
- nutzer-spezifische Rankingkriterien (siehe Abbildung 14).

Jede Faktorenmenge und auch jeder einzelne Faktor kann zudem gewichtet werden. Jeder einzelne der Faktoren führt entweder zu einem Ansteigen (bei positiven Faktoren) oder zu einem Abfallen (bei negativen Faktoren) des Retrievalstatuswertes der Dokumente.

Set eins berücksichtigt die konkreten Tags und bearbeitet sie für das Relevance Ranking. Das Vektorraummodel kann genutzt werden, um die Ähnlichkeit zwischen Dokumenten zu bestimmen (Stock, 2007a, Kap. 20). Die Dimensionen sind dabei die verschiedenen Tags in der Datenbank, der Wert der Dimension wird durch WDF*IDF bestimmt (1a). Der WDF-Wert errechnet sich aus der dokumentspezifischen Häufigkeit (freq) eines bestimmten Tags t in Relation zur Gesamthäufigkeit L aller Tags, die dem jeweiligen Dokument d (auch u. U. mehrfach) zugeordnet worden sind:

$$WDF(t,d) = [ld (freq(d,t) + 1] / ld L.$$

Der IDF-Wert bezieht die Gewichtung eines Tags t auf sein Vorkommen in der gesamten Datenbank. Sei N die Mächtigkeit der Menge aller Datensätze in der Datenbank und n die Anzahl derjenigen Dokumente, die mittels t indexiert worden sind, so gilt:

$$IDF(t) = ld (N/n).$$

Die Dokumente werden durch Vektoren abgebildet, und die Ähnlichkeit zwischen Dokument und Suchfrage wird letztlich durch den Cosinus (1b) bestimmt. Der Wert des WDF hängt in Broad Folksonomies von der Anzahl der Index-Tags ab; bei Narrow Folksonomies wird der WDF-Wert dagegen durch die Zahl der Such-Tags determiniert.

Es gibt Ansätze, die für das Ranking einen modifizierten PageRank einsetzen. Das Motto heißt hier: "The basic notion is that a resource which is tagged with important tags by important users becomes important itself" (Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G., 2006a, 417). Jedoch ist diese Schlussfolgerung problematisch und kann zu unbefriedigenden Ergebnissen führen. Die gleiche Studie stellt den (bei BibSonomy eingesetzten) FolkRank-Algorithmus vor, mit dem Ziel „to focus the ranking around the topics defined in the preference vector" (419). Der FolkRank verfolgt die Idee eines „Super-Posters" oder „Super-Autors", der eine große Menge an Content veröffentlicht und damit ein Experte (in einem bestimmten Bereich) zu sein scheint (1c). Daher sollten Inhalte dieser User beim Ranking von Suchergebnissen höher gewertet werden (Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G., 2006a, 2006b). Abhängig von ihrer Häufigkeitsverteilung werden manche Tags als Power Tags markiert (1d). Der WDF-IDF-Gewichtungswert von Power Tags kann mittels eines festzulegenden Faktors f (f > 1) justiert werden.

Set zwei bezieht die aktive Zusammenarbeit der User als positive Faktoren mit in das Ranking ein. So können die Klick-Raten (2a) bezogen auf einzelne Dokumente zum Ranking herangezogen werden (Culliss, 1997). Für Jung, Herlocker und Webster (2007) sind die Klick-Raten ein implizites Relevance Feedback in der Internetsuche; sie sind im hohen Maße ein kollaborationsorientiertes Ranking-Kriterium im Sinne des Web 2.0. Möglicherweise ist auch die Anzahl der unterschiedlichen Nutzer, die ein Dokument indexieren, ein sinnvoller Ranking-Faktor (2b). Hoch frequentierte Diskussionen auf der Basis gewisser Dokumente verweisen auf Dokumente von hoher Bedeutung für die Community. Die Zahl der Kommentare zu einem Dokument (2c) kann die Stellung im Ranking bestimmen. Insbesondere Blogpostings, aber auch andere Dokumente wie Bilder und Videos, insofern auf diese Links verweisen, bieten Potentiale für einen quantitativen Ausdruck ihrer linktopologischen Stellung entweder mittels des Kleinberg-Algorithmus' (Kleinberg, 1999) oder des PageRank (Brin & Page, 1998). Authorities (Dokumente mit vielen In-Links) befinden sich im Zentrum des Interesses von bestimmten Themen und sollten daher höher gerankt werden.

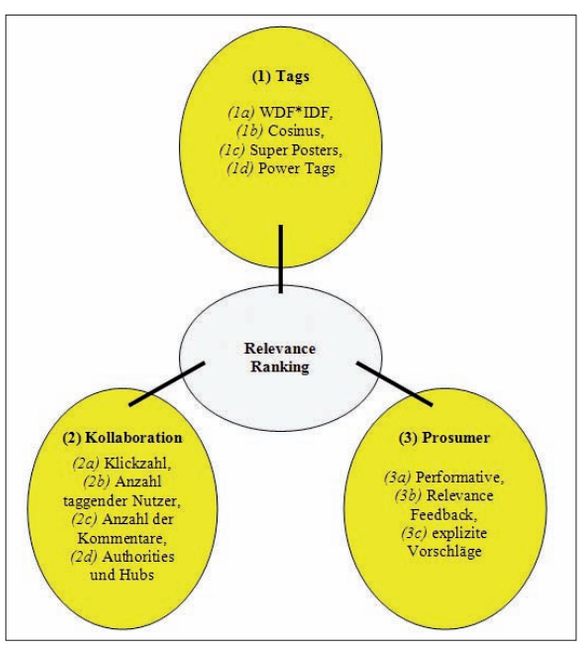


Abbildung 14: Kriterien für das Relevance Ranking von getaggten Dokumenten.

Das dritte Set der Rankingkriterien bezieht sich auf den Nutzer selbst. Dokumente, die mit *todo* oder *toread* getaggt sind, scheinen sehr wichtig für den Nutzer zu sein und sollten daher ihre Stellung im Ranking erhöhen (3a). Negative Werturteile (wie z.B. *stupid*) führen – abhängig von der Häufigkeit ihres dokumentspezifischen Vorkommens – zu einer Senkung des Retrievalstatuswertes. Man kann auch den Nutzer aktiv am Ranking beteiligen und ihm die Gelegenheit zum Feedback geben. Das kann auf zweifache Weise geschehen: (3b) das Relevance Feedback geschieht nach der Anzeige und Bewertung der Suchergebnisse (z.B. Rocchio, 1971) oder (3c) mittels eines formalen Recommendersystems, wie es z.B. in einer Sternchen-Bewertung oder mittels eines Frage-Dialogs nach dem Stil „Hat Ihnen dieses Dokument geholfen?“ gegeben ist. Der User bekommt die Gelegenheit, das Dokument zu bewerten und so aktiv an der Bestimmung des jeweiligen Retrievalstatuswertes mitzuwirken – ein demokratischer Weg des Relevance Ranking. Beide Faktoren (3b und 3c) können in das Ranking einfließen; einmal implizit als Gewichtungsfaktor für den Retrievalstatuswert und einmal explizit („Andere Nutzer fanden dieses Dokument sehr hilfreich“ oder „Das Dokument erhält 5 von 6 Relevanzpunkten“). Beide Möglichkeiten haben u. W. bis jetzt noch keinen Eingang in die Web 2.0-Dienste gefunden.

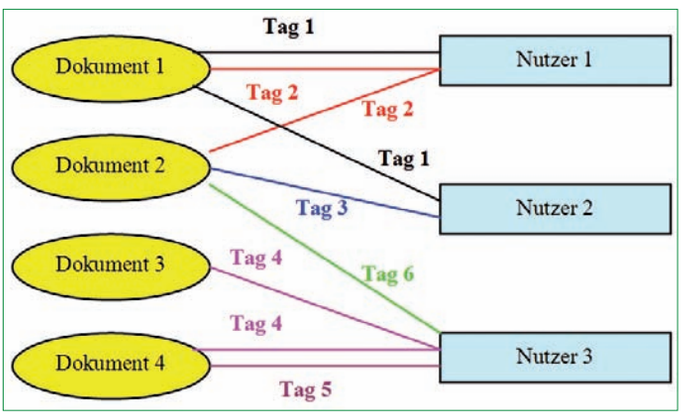


Abbildung 15: Dokumente, Tags und Nutzer in einer Folksonomy.

Aus Nutzersicht wichtig ist bei solch ausgefeilten Rankingalgorithmen, dass er sie ganz oder teilweise abschalten kann. Ein Drop-Down-Menü scheint eine optimale Vorgehensweise, um die Nutzer in die Einstellung der Rankingfaktoren der Suchresultate einzubeziehen. Er selbst kann entscheiden, nach welchen Kriterien er die Ergebnisse sortieren möchte. Vielleicht möchte ein User lieber nach formalen Kriterien (Veröffentlichungszeitpunkt, Autor usw.) sortieren oder er bevorzugt aus unserem Werkzeugkasten für Relevance Ranking nur ganz bestimmte Kriterien.

Vorschlagssysteme auf der Basis von Folksonomies: Recommendation von Dokumenten und Nutzern

Innerhalb einer Folksonomy sind wir mit drei unterschiedlichen Aspekten konfrontiert (Marlow, Naaman, Boyd, & Davis, 2006; Yeung, Gibbins, & Shadbolt, 2008):

- die zu beschreibenden Dokumente,
- die Tags (Worte), die zur Beschreibung gewählt werden,
- die Nutzer (Prosumer), die die Indexierung ausführen.

Nutzer untereinander sowie Dokumente untereinander sind in einem sozialen Netzwerk miteinander verknüpft, wobei als Pfade jeweils die Tags fungieren. Dokumente sind erstens *thematisch* miteinander verbunden, wenn sie durch dieselben Tags indexiert worden sind. Die Dokumente 1 und 2 sowie 3 und 4 aus Abbildung 15 sind jeweils thematisch verknüpft (Dokumente 1 und 2 über Tag 2; Dokumente 3 und 4 über Tag 4). Zusätzlich sind Dokumente zweitens *über gemeinsame Nutzer* verknüpft. So sind die Dokumente 1 und 2, 3 und 4, aber auch 2 und 4 über ihre Nutzer verknüpft.

Nutzer sind miteinander verknüpft, wenn sie entweder dieselben Tags verwenden oder dieselben Dokumente indexieren. Nutzer sind *thematisch* verbunden, wenn sie dieselben Tags verwenden (im Beispiel die Nutzer 1 und 2 über Tags 1 und 2); sie sind *über gemeinsame Dokumente* gekoppelt, wenn sie diese jeweils inhaltlich beschreiben (Nutzer 1, 2 und 3 über Dokument 2).

Wir wollen diese Zusammenhänge ausnutzen, um auf der Basis der Stellung eines Tags, eines Nutzers und eines Dokuments im Netzwerk automatisch Vorschläge zu generieren (Diederich & Iofciu, 2006; Höhfeld & Kwiatkowski, 2007, 272 f.). Unsere Ausführungen konzentrieren sich auf Empfehlungen ähnlicher Dokumente (ausgehend von einem Musterdokument: „More like this!“) sowie von ähnlichen Nutzern (ausgehend von einem bestimmten Nutzer, beispielweise sich selbst: „More like me!“), da beide Vorschläge über recht einfache Algorithmen auf der Basis von Co-Occurences arbeiten.

Möchte man durch die Empfehlungen weiterer Tags dem Nutzer Synonyme und Quasi-Synonyme an die Hand geben, so eignen sich Co-Occurences nicht, denn Synonyme kommen selten gemeinsam syntagmatisch verbunden als Tags vor (ein User, der sein Bild mit *Hochzeit* indexiert hat, wird dieses kaum zusätzlich mit *Eheschließung* und *Heirat* beschreiben). Hier sind andere Algorithmen gefordert. Der Fall wird genau dann leicht, wenn wir auf eine Begriffsordnung (wie beispielsweise WordNet) zurückgreifen können, da dann die Empfehlungen aus den dort eingetragenen Synonymen gebildet werden.

Der erste Schritt bei Empfehlungen von Dokumenten und Nutzern besteht darin, Ähnlichkeiten zwischen dem Ausgangsdokument bzw. -nutzer und allen anderen Dokumenten (Nutzern) der Datenbank zu berechnen. Dazu kommen einschlägige Algorithmen wie Jaccard-Sneath, Dice oder Cosinus infrage (Stock & Stock, 2008, Kap. 20). Möchte man eine Community (also ähnliche Nutzer), ausgehend von Nutzer 1, identifizieren, so errechnet man nach Jaccard-Sneath

$$\text{Ähnlichkeit (Nutzer 1, Nutzer i)} = g / (a + b - g).$$

Die Formel findet zweifach Anwendung: 1) a ist die Anzahl der Tags, die Nutzer 1 vergeben hat, b die Anzahl der Tags, die Nutzer i zum Indexieren heranzog und g ist die Anzahl derjenigen Tags, die die beiden Nutzer gemeinsam verwenden; 2) a ist nunmehr die Anzahl der Dokumente, die Nutzer 1 ausgewertet hat, b die Anzahl der von Nutzer i indexierten Dokumente und g die Zahl der von beiden beschriebenen Dokumente.

Der zweite Schritt besteht entweder in einer einfachen oder in einer elaborierteren Vorgehensweise. Der einfache Weg fasst – in der Reihenfolge der beiden errechneten Ähnlichkeiten – die jeweils ersten k Dokumente bzw. Nutzer (k kann auf etwa 20 eingestellt werden) zu einer Quasiklasse zusammen (Methode der k-nearest neighbors). Ein aufwändigerer zweiter Weg arbeitet mit der Clusteranalyse (z.B. Single Linkage oder Complete Linkage) oder vergleichbaren Verfahren. Auch hier erhalten wir Quasiklassen – nun jedoch möglicherweise mehrere und diese ggf. in einer hierarchischen Ordnung. Dem Nutzer werden die Quasiklassen als ähnliche Dokumente bzw. als ähnliche Nutzer (oder als „seine“ Community) vorgeschlagen.

Es gibt bisher nur sehr wenige empirische Studien zum Einsatz von Recommentation bei Folksonomies. Erwähnenswert ist die Untersuchung von Jäschke, Marinho, Hotho, Schmidt-Thieme und Stumme (2007), die die Wirkung von vorgeschlagenen Suchtags auf die Retrievalleistung (anhand von BibSonomy und dem Mu-

sikdienst Last.fm) untersuchten. Der Recall-Precision-Graph zeigt durchgehend Steigerungen der Retrievalqualität beim Einsatz von Recommendation-Methoden.

Fazit und offene Forschungsprobleme

Folksonomies haben in informationswissenschaftlicher Sicht sowohl Vorteile (authentische Sprache der Nutzer, Vereinigung unterschiedlicher Sichtweisen auf ein Dokument, Quelle für die Entwicklung kontrollierter Vokabularien, Ermöglichung von Suchen *und* Browsen, Identifikation von Communities usw.) als auch Nachteile (Mangel an Präzision, unterschiedliche Indexierungsebenen, Vermischung von Sprachen, versteckte paradigmatische Relationen, Spam-Tags, nutzerspezifische Tags und andere unbrauchbare Terme, Verschmelzung der semantischen Ebenen bei Bildern und Videos). Da die Vorteile überwiegen, müssen wir dafür Sorge tragen, dass die Nachteile möglichst minimiert werden.

Ein erster Schritt der Problemlösung ist die Behandlung der Tags als natürlichsprachige Eingaben, auf die (bekannte wie neue) Methoden der Informationslinguistik (Natural Language Processing) angewandt werden. Damit sollten sich zumindest Tags unterschiedlicher Sprachen trennen, die Tags auf Grund- oder Stammformen bringen sowie Eingabefehler korrigieren lassen. Über die Kopplung der Folksonomies mit kontrollierten Vokabularien sollten – in einem gewissen Maße – Homonyme getrennt und Synonyme zusammengeführt werden; ebenso können die semantischen Relationen der eingesetzten Wissensordnungen (z.B. die Hierarchien) bei der Recherche nutzbar werden. Liegen maschinenlesbare Wörterbücher vor, wird auch multilinguales Retrieval möglich.

Eine große Herausforderung ist die Kreation passender Algorithmen des Relevance Ranking. Hier sehen wir drei Kriterienbündel, die Berücksichtigung finden müssen: die (via NLP „gereinigten“) Tags selbst (WDF*IDF, Vektorraummodell, Berücksichtigung von Super-Postern sowie von Power Tags), Merkmale der Kollaboration (Klickraten, Anzahl der indexierenden Nutzer, Anzahl der Kommentare und linktopologische Ansätze bei verlinkten Dokumenten) sowie Aspekte der Prosumer (Performative, Relevance Feedback, formale Bewertungen). Ein neues Suchfeld nach *Nur Power Tags* verringert den Ballast und steigert somit die Präzision der Suchresultate.

Über Ähnlichkeitsberechnungen und den Einsatz von Clusteranalyse oder verwandter Verfahren erhalten wir Algorithmen, um sowohl ähnliche Dokumente („More like this!“) zu finden als auch um ähnliche Nutzer („More like me!“) und damit Communities zu identifizieren.

Das Forschungsgebiet der Folksonomies ist noch sehr jung. Teilweise haben wir außer Ansätzen noch nicht viel Konkretes an brauchbaren (d.h. in Informationsdiensten praktisch anwendbaren) Resultaten vorzuweisen. Wir sehen folgende offenen Forschungsfragen:

- Bisher nicht untersucht sind die Such-Tags. Insbesondere bei Narrow Folksonomies brauchen wir diese jedoch, um Tagverteilungen herzustellen und – soweit dort auch vorhanden – Power Tags zu identifizieren.
- Wir wissen, dass viele Tagverteilungen bei Broad Folksonomies dem Power Law folgen; wir wissen aber auch, dass dies nicht auf alle dokumentspezifischen Tagverteilungen zutrifft. Wie groß sind die Anteile der Tagverteilungen auf der Mikroebene, die dem Power Law und die der invers-logistischen Kurve folgen? Gibt es darüber hinaus weitere Verteilungsformen?
- Das Entstehen einer Power-Law-Verteilung ist durch den Yule-Simon-Ansatz theoretisch zu fassen. Wie erklären wir aber die Entwicklung einer invers-logistischen Verteilung?
- In Web 2.0-Diensten werden häufig Bilder und Videos indexiert. Wir haben als Problem die Verschmelzung der semantischen Ebenen bei der Indexierung durch Tags (Ofness, Aboutness, Ikonologie, Isness) markiert. Des Weiteren kennen wir die Inter-Indexer-Inkonsistenz bei der Indexierung solcher Dokumente. Eine Lösung können wir nicht bieten.

CONTENTdm—Digital Collection Management Software

Werden Sie Teil einer aktiven Anwendergemeinschaft mit einem gemeinsamen Ziel: wichtige Bestände, Sammlungen und Kulturschätze zu digitalisieren und einem breiten Publikum zugänglich zu machen.

CONTENTdm bietet die vollständige Lösung zur Speicherung, Verwaltung, Selektion und Internet-Präsentation Ihrer digitalen Sammlungen.

Mit CONTENTdm generieren Sie aus Ihren Spezialsammlungen vollständig suchbare Online-Ressourcen, auf die von jedem Browser aus zugegriffen werden kann.

CONTENTdm—digital collection management software

OCLC™
The world's libraries. Connected.

Grünwalder Weg 28g
82041 Oberhaching, Deutschland
T +49-(0)89-613 08 300
F +49-(0)89-613 08 399
E deutschland@oclc.org

www.oclc.org

- Unsere Lösungsoptionen für die Mängel bei Tags beruhen auf informationslinguistischen Ansätzen. Können wir uns in der Tat auf die Behandlung von Substantiven und Adjektiven beschränken und Verben außer acht lassen? Wie häufig kommen Verben überhaupt als Tags vor?
- In den dokumentspezifischen Indexaten sind paradigmatische Relationen „versteckt". Wie lassen sich die versteckten semantischen Relationen aufdecken?
- Bislang setzt kein System durchgehend NLP zur Bearbeitung von Tags ein. Entsprechend fehlen Erfahrungen mit informationslinguistisch bearbeiteten Tags.
- Wir behaupteten, dass die Retrievaloption „Power Tags only" Verbesserungen der Retrievalqualität in Form einer Steigerung der Precision mit sich bringt. Hier fehlt noch jede empirische Bestätigung.
- Der vorgestellte Werkzeugkasten für Kriterien des Relevance Ranking für getaggte Dokumente ist ein erster Vorschlag, Algorithmen klassischer Retrievalmodelle mit den Gegebenheiten der Web 2.0-Dienste (Kollaboration, Prosumer) zu verbinden. Hier fehlt nicht nur die technische Umsetzung, sondern auch – im Anschluss daran – deren Evaluation.
- Nach welchen Algorithmen kann man Nutzern ähnliche Tags (vor allem Synonyme und Quasi-Synonyme ihrer ursprünglichen Suchterme) vorschlagen, wenn keine Begriffsordnung einsetzbar ist?
- Eignen sich Verfahren nach k-nearest neighbors oder nach der Clusteranalyse für die Identifikation von ähnlichen Dokumenten und von Communities? Werden sie von den Nutzern angenommen?

Literatur

Al-Khalifa, H.S. (2007). Automatic Documet-Level Semantic Metadata Annotation Using Folksonomies and Domain Ontologies. PhD-Thesis. University of Southampton / School of Electronics and Computer Science.

Al-Khalifa, H.S., & *Davis*, H.C. (2007). Exploring the value of folksonomies for creating semantic metadata. International Journal on Semantic Web and Information Systems 3(2007)1, 12-38.

Aurnhammer, M., *Hanappe*, P., & *Steels*, L. (2006). Augmenting navigation for collaborative tagging with emergent semantics. Lecture Notes in Computer Science 4273, 58-71.

Bar-Ilan, J., *Shoham*, S., *Idan*, A., *Miller*, Y, & *Shachak*, A. (2006). Structured vs. unstructured tagging – A case study. In: Proceedings of the 15th International WWW Conference. Collaborative Web Tagging Workshop.

Beaudoin, J. (2007). Flickr image tagging: Patterns made visible. Bulletin of the American Society for Information Science and Technology 34(2007)1, 26-29.

Brin, S., & *Page*, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107-117.

Brooks, C.H., & *Montanez*, N. (2006a). An analysis of the effectiveness of tagging in blogs. In: *Nicolov*, N., *Salvetti*, F., *Liberman*, M., & *Martin*, J.H. (eds.): Computation Approaches to Analyzing Weblogs. Papers from the 2006 AAAI Spring Symposium (S. 9-15). Menlo Park, Cal: AAAI Press. (Technical Report SS-06-03. American Association for Artificial Intelligence).

Brooks, C.H., & *Montanez*, N. (2006b). Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th International World Wide Web Conference (S. 625-632). New York: ACM.

Butterfield, D.S., *Costello*, E., *Fake*, C., *Henderson-Begg*, C.J., & *Mourachow*, S. (2006). Interestingness ranking of media objects. Patentanmeldung Nr. US 2006/0242139 A1. Eingereicht: 8.2.2006. Publiziert: 26.10.2006.

Butterfield, D.S., *Costello*, E., *Fake*, C., *Henderson-Begg*, C.J., *Mourachow*, S., & *Schachter*, J.E. (2006). Media object metadata association and ranking. Patentanmeldung Nr. US 2006/0242178 A1. Eingereicht: 8.2.2006. Publiziert: 26.10.2006.

Campbell, D.G. (2006). A phenomenological framework for the relationship between the semantic Web and user-centered tagging systems. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Capocci, A., & *Caldarelli* (2007). Folksonomies and clustering in the collaborative system CiteULike. (arXiv:0710.2835v2).

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

Cattuto, C (2006). Semiotic dynamics in online social communities. European Physical Journal C 46(2006)Suppl. 2, 33-37.

Cattuto, C., *Loreto*, V., & *Pietronero*, L. (2007). Semiotic dynamics and collaborative tagging. Proceedings of the National Academy of Sciences of the United States of America 104(2007)5, 1461-1464.

Christiaens, S. (2006). Metadata mechanisms: From ontology to folksonomy … and back. Lecture Notes in Computer Science 4277, 199-207.

Culliss, G.A. (1997). Method for organizing information. Patent-No. US 6.006.222. Granted: Dec. 21, 1999. Applied: Jan. 8, 1997.

Diederich, J., & *Iofciu*, T. (2006). Finding communities of practice from user profiles based on folksonomies. In Proceedings of the 1st International Workshop on Building Technology Enhanced Learning Solutions for Communities of Practice (TEL-CoPs'06).

Dye, J. (2006). Folksonomy: A game of high-tech (and high-stakes) tag. E-Content 29(2006)3, 38-43.

Egghe, L. (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Amsterdam: Elsevier Academic Press.

Egghe, L., & *Rousseau*, R. (1990). Introduction to Informetrics. Amsterdam: Elsevier.

Fichter, D. (2006). Intranet applications for tagging and folksonomies. Online 30(2006)3, 43-45.

Fidel, R. (2006). An ecological approach to the design of information systems. Bulletin of the American Society for Information Science and Technology 33(2006)1, 6-8.

Furnas, G.W., *Fake*, C., *von Ahn*, L., *Schachter*, J., *Golder*, S., *Fox*, K., *Davis*, M., *Marlow*, C., & *Naaman*, M. (2006). Why do tagging systems work? In: CHI ,06 Extended Abstracts on Human Factors in Computing Systems (S. 36-39). New York: ACM.

Gendarmi, D., & *Lanubile*, F. (2006). Community-driven ontology evolution based on folksonomies. Lecture Notes in Computer Science, 4277, 181-188.

Golder, S.A., & *Huberman*, B.A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science 32(2006)2, 198-208.

Gordon-Murnane, L. (2006). Social bookmarking, folksonomies, and Web 2.0 tools. Searcher - The Magazine for Database Professionals 14(2006)6, 26-38.

Grahl, M., *Hotho, A.*, & *Stumme*, G. (2007). Conceptual clustering of social bookmarking sites. In: Proceedings of I-KNOW ,07, Graz, Austria, September 5-7, 2007 (S. 356-264).

Gruber, T. (2006). Where the social Web meets the semantic Web. Lecture Notes in Computer Science 4273, 994.

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. International Journal on Web Semantics and Information Systems 3(2007)1, 1-11.

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

Hänger, C., & *Krätzsch*, C. (2007). Collaborative Tagging als Service von Hochschulbibliotheken. In: *Ball*, R. (Hrsg.): Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek Forschungszentrum Jülich (S. 123-134). Jülich: Forschungszentrum Jülich GmbH / Zentralbibliothek.

Hammond, T., *Hannay*, T., *Lund*, B., & Scott, J. (2005). Social bookmarking tools. A general review. Part 1. D-Lib Magazine, 12(1).

Heller, L. (2007). Bibliographie und Sacherschließung in der Hand vernetzter Informationsbenutzer. Bibliothek, 31(2), 162-172.

Herget, J., & *Hierl*, S. (2007). Top down versus bottom up: Knowledge representation in transition. From controlled vocabulary to folksonomies. In: Proceedings of the 2nd International CODATA Symposium on Generalization of Information (S. 150-157).

Heymann, P., & *Garcia-Molina*, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging system. InfoLab Technical Report; 2006-10. Stanford: Stanford University / Computer Science Dept.

Höhfeld, S., & *Kwiatkowski*, M. (2007). Empfehlungssysteme aus informationswissenschaftlicher Sicht – State of the Art. Information – Wissenschaft und Praxis 58(2007)5, 265-276.

Hotho, A., *Jäschke*, R., *Schmitz*, C., & *Stumme*, G. (2006a). Information retrieval in folksonomies: Search and ranking. Lecture Notes in Computer Science 4011, 411-426.

Hotho, A., *Jäschke*, R., *Schmitz*, C., & *Stumme*, G. (2006b). Trend detection in folksonomies. Lecture Notes in Computer Science, 4306, 56-70.

Hotho, A., *Jäschke*, R., *Schmitz*, C., & *Stumme*, G. (2006c). BibSonomy: A social bookmark and publication system. In: Proceedings of the Conceptual Structure Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures.

Huang, H. (2006). Tag distribution analysis using the power law to evaluate social tagging systems: A case study in the Flickr database. In 17th ASIS&T SIG/CR Classification Research Workshop. Abstracts of Posters (S. 14-15).

Ingwersen, P. (2002). Cognitive perspectives of document representation. In: CoLIS 4: 4th International Conference on Conceptions of Library and Information Science (S. 285-300). Greenwood Village: Libraries Unlimited.

Jäschke, R., *Marinho*, L., *Hotho*, A., *Schmidt-Thieme*, L., & *Stumme*, G. (2007). Tag recommendations in folksonomies. Lecture Notes in Artificial Intelligence 4702, 506-514.

Jung, S., *Herlocker*, J.L., & *Webster*, J. (2007). Click data as implicit relevance feedback in web

search. Information Processing & Management, 43, 791-807.

Kim, H.G., *Hwang*, S.H., *Kang*, Y.K., *Kim*, H.L., & *Yang*, H.S. (2007). An agent environment for contextualizing folksonomies in a triadic context. Lecture Notes in Computer Science 4496, 728-737.

Kim, H.L., *Hwang*, S.H., & *Kim*, H.G. (2007). FCA-based approach for mining contextualized folksonomy. In: Proceedings of the 2007 ACM Symposium on Applied Computing (S. 1340-1345). New York: ACM.

Kipp, M.E.I. (2006a). @toread and cool: Tagging for time, task and emotion. In: 17th ASIS&T SIG/CR Classification Research Workshop. Abstracts of Posters (S. 16-17).

Kipp, M.E.I. (2006b). Complementary or discrete contexts in online indexing: A comparison of user, creator and intermediary keywords. Canadian Journal of Information and Library Science 30(2006)3, in prep.

Kipp, M.E.I. (2007). Tagging for health information organisation and retrieval. In: Proceedings of the North American Symposium on Knowledge Organization, vol. 1 (S. 63-74).

Kipp, M.E.I., & *Campbell*, D.G. (2006). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In: Proceedings of the 69th Annual Meeting of the American Society for Information Science & Technology (Vol. 43), Austin, TX (CD-ROM).

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM 46(1999)5, 604-632.

Kolbitsch, J. (2007a). Aspects of Digital Libraries. Diss. TU Graz.

Kolbitsch, J. (2007b). WordFlickr: A solution to the vocabulary problem in social tagging systems. In: Proceedings of I-MEDIA ,07 and I-SEMANTICS ,07. Graz, Austria, September 5-7, 2007 (S. 77-84).

Kroski, E. (2005). The Hive Mind: Folksonomies and user-based tagging. [Blog post; 2005-12-07]. Online: http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging.

Lancaster, F.W. (2003). Indexing and Abstracting in Theory and Practice. Champaign: University of Illinois. 3rd ed.

Laniado, D., *Eynard*, D., & *Colombetti*, M. (2007). A semantic tool to support navigation in a folksonomy. In: Proceedings of the 18th Conference on Hypertext and Hypermedia (S. 153-154). New York: ACM.

Layne, S. (2002). Subject access to art images. In: *Baca*, M. (ed.): Introduction to Art Image Access (S. 1-19). Los Angeles: Getty Research Institute.

Lin, X., *Beaudoin*, J.E., *Bui*, Y., & *Desai*, K. (2006). Exploring characteristics of social classification. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Lüth, J. (2007). Inhaltserschließung durch Nutzerinnen und Nutzer. Ergebnisse eines Tests mit Internetquellen der virtuellen Fachbibliothek EconBiz. In: *Ockenfeld*, M. (Hrsg.): Information in Wissenschaft, Bildung und Wirtschaft. 29. Online-Tagung der DGI. 59. Jahrestagung der DGI (S. 113-120). Frankfurt am Main: DGI.

Lund, B., *Hammond*, T., *Flack*, M., & *Hannay*, T. (2005). Social bookmarking tools (II). A case study – Connotea. D-Lib Magazine, 11(2005)4.

Lux, M., *Granitzer*, M. & *Kern*, R. (2007). Aspects of broad folksonomies. In: Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007) (S. 283-287). Washington, DC: IEEE Computer Society.

Macgregor, G., & *McCulloch*, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review 55(2006)5, 291-300.

MacLaurin, M.B. (2005). Selection-based item tagging. Patentanmeldung Nr. US 2007/0028171 A1. Eingereicht: 29.7.2005. Publiziert: 1.2.2007.

Mai, J.M. (2006). Contextual analysis for the design of controlled vocabularies. Bulletin of the American Society for Information Science and Technology 33(2006)1, 17-19.

Markey, K. (1986). Subject Access to Visual Resource Collections. Westport: Greenwood.

Markkula, M., & *Sormunen*, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. Information Retrieval 1(2000), 259-285.

Marlow, C., *Naaman*, M., *Boyd*, D., & *Davis*, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the 17th Conference on Hypertext and Hypermedia (S. 31-40). New York: ACM.

Mathes, A. (2004). Folksonomies – Cooperative Classification and Communication Through Shared Metadata. Urbana, Ill.: University of Illinois Urbana-Campaign / Graduate School of Library and Information Science.

McFedries, P. (2006). Folk wisdom. IEEE Spectrum, 43(2006)Febr., 80.

Merholz, P. (2004). Metadata for the masses. [Blog post, 2004-10-19]. Online: www.adaptivepath.com/publications/essays/archives/000361.php.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. Lecture Notes in Computer Science 3729, 522-536.

Millen, D.R., Feinberg, J., & Kerr, B. (2006). DOGEAR. Social bookmarking in the enterprise. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (S. 111-120). New York: ACM.

Miller, G.A. (1998). Nouns in WordNet. In Fellbaum, C. (ed.), WordNet. An Electronic Lexical Database (S. 23-46). Cambridge, Mass., London: MIT Press.

Müller-Prove, M. (2007). Taxonomien und Folksonomien – Tagging als neues HCI-Element. In i-com, 6(1), 14-18.

Morrison, P.J. (2007). Why are they tagging, and why do we want them to? Bulletin of the American Society for Information Science and Technology 34(2007)1, 12-15.

Munk, T.B., & *Mork*, K. (2007). Folksonomy, the power law and the significance of the least effort. Knowledge Organization 34(2007)1, 16-33.

Murison, J. (2005). Messageboard topic tagging: User tagging of collectively owned community content. In: Proceedings of the 2005 Conference on Designing for User eXperience (art. no 5). New York: American Institute of Graphic Art.

Neal, D. (2007). Folksonomies and image tagging: Seeing the future? Bulletin of the American Society for Information Science and Technology 34(2007)1, 7-11.

Niwa, S., *Doi*, T., & *Honiden*, S. (2006). Web page recommender systems based on folksonomy mining for ITNG ,06 submissions. In: Third Conference on Information Technology: New Generation (ITNG'06) (S. 388-393). Washington, DC: IEEE Computer Society.

Noruzi, A. (2006). Folksonomies: (un)controlled vocabulary? Knowledge Organization 33(2006)4, 199-203.

Noruzi, A. (2007). Folksonomies: Why do we need controlled vocabulary? Webology 4(2), editorial 12.

O'Reilly, T. (2005). What is Web 2.0. Design patterns and business models for the next generation of software. Online: www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Ornager, S. (1995). The newspaper image database. Empirical supported analysis of users' typology and word association clusters. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (S. 212-218). New York: ACM.

Ott, E.S. (2006). Picture tagging. Patentanmeldung Nr. US 2007/0079321 A1. Eingereicht: 17.2.2006. Publiziert: 5.4.2007.

Panofsky, E. (1975) Sinn und Deutung in der bildenden Kunst. Köln: DuMont.

Panofsky, E. (2006). Ikonographie und Ikonologie. Köln: DuMont.

Paolillo, J.C., & *Penumarty*, S. (2007). The social structure of tagging internet videos on del.icio.us. In: Proceedings of the 40th Hawaii International Conference on System Sciences.

Peters, I. (2006a). Inhaltserschließung von Blogs und Podcasts im betrieblichen Wissensmanagement. In *Ockenfeld*, M. (Hrsg.): Content. 28. Online-Tagung der DGI, 58. Jahrestagung der DGI. Proceedings (S. 143-151). Frankfurt am Main: DGI.

Peters, I. (2006b). Against folksonomies: Indexing blogs and podcasts for corporate knowledge management. In: *Jezzard*, H. (ed.): Preparing for Information 2.0. Online Information 2006. Proceedings (S. 93-97). London: Learned Information Europe.

Peters, I. (2007). Web 2.0 und Fachinformation. Folksonomies im Web 2.0. Password 22(2007)5, 10-11.

Peters, I. (2008). ASIS&T Annual Meeting 2007: Milwaukee, Wisconsin. Joining Research and Practice: Social Computing and Information Science. Password 23(2008)1, 16-19.

Peters, I., & *Stock*, W.G. (2006). Corporate Blogs im Wissensmanagement. Wissensmanagement 8(2006)6, 40-41.

Peters, I., & *Stock*, W.G. (2007a): Web 2.0 im Unternehmen. Wissensmanagement 9(2007)4 22-25.

Peters, I., & *Stock*, W.G. (2007b): Folksonomy and information retrieval. In: Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology Vol. 45 (S. 1510-1542). (CD-ROM).

Peters, I., & *Weller*, K. (2008). Paradigmatic and syntagmatic relations in knowledge organization systems. Information – Wissenschaft und Praxis, 59(2008)2, 100-107.

Peterson, E. (2006). Beneath the metadata. Some philosophical problems with folksonomies. D-Lib Magazine 12(2006)11.

Pluzhenskaia, M. (2006). Folksonomies or fauxsonomies: How social is social bookmarking? In: 17th ASIS&T SIG/CR Classification Research Workshop. Abstracts of Posters (S. 23-24).

Quintarelli, E. (2005). Folksonomies: Power to the people. Paper presented at the ISKO Italy UniMIB meeting, Milan, June 24, 2005.

Rasmussen, E.M. (1997): Indexing images. Annual Review of Information Science and Technology 32, 169-196.

Reamy, T. (2006). Folksonomies and complexity theory: Evolved information structures. In: *Jezzard*, H. (ed.): Preparing for Information 2.0. Online Information 2006. Proceedings (S. 111-113). London: Learned Information Europe.

Rocchio, J.J. (1971). Relevance feedback in information retrieval. In: *Salton*, G. (ed.): The SMART Retrieval System – Experiments in Automatic Document Processing (S. 313-323). Englewood Ciffs, N.J.: Prentice Hall.

Russell, T. (2006). Cloudalicious: Folksonomy over time. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (S. 364). New York: ACM.

Schmitz, C., *Hotho*, A., *Jäschke*, R., & *Stumme*, G. (2006). Mining association rules in folksonomies. In: *Batagelj*, V., *Bock*, H.H., *Ferligoj*, & *Žiberna*, A. (eds.): Data Science and Classification (S. 261-270). Berlin, Heidelberg: Springer.

Schmitz, P. (2006). Inducing ontology from Flickr tags. In: Proceedings of the 15th International WWW Conference. Collaborative Web Tagging Workshop.

Shatfort, S. (1986). Analyzing the subject of a picture. A theoretical approach. Cataloguing and Classification Quarterly 6(1986)3, 39-62.

Shirky, C. (2004). Folksonomy. [Blog post; 2004-08-24]. Online: http://many.corante.com/archives/2004/08/25/folksonomy.php

Shirky, C. (2005). Ontology is overrated: Categories, Links, and Tags. www.shirky.com/writings/ontology_overrated.html.

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

Smith, G. (2004). Folksonomy: Social classification. [Blog post; 2004-08-03]. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.

Smith, M.K. (2006). Viewer tagging in art museum: Comparisons to concepts and vocabularies of art museum visitors. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Specia, L., & *Motta*, E. (2007). Integrating folksonomies with the semantic Web. Lecture Notes in Computer Science, 4519, 624-639.

Spiteri, L.F. (2006). The use of folksonomies in public library catalogues. The Serials Librarian 51(2006)2, 75-89.

Spiteri, L.F. (2007). Structure and form of folksonomy tags: The road to the public library catalogue. Webology 4(2007)2, art. 41.

Spyns, P., *de Moor*, A., *Vandenbussche*, J., & *Meersman*, R. (2006). From folksonomies to ontologies: How the twain meet. Lecture Notes in Computer Science 4275, 738-755.

Stock, W.G. (2006). On relevance distributions. Journal of the American Society for Information Science and Technology 57(2006)8, 1126-1129.

Stock, W.G. (2007a). Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

Stock, W.G. (2007b). Folksonomies and science communication. A mash-up of professional science databases and web 2.0 services. Information Services & Use, 27(2007)3, 97-103.

Stock, W.G., & *Stock*, M. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München, Wien: Oldenbourg.

Stock, W.G., & *Weber*, S. (2006). Facets of informetrics. Information – Wissenschaft und Praxis 57(2006)8, 385-389.

Surowiecki, J. (2004). The Wisdom of Crowds. New York: Doubleday (dt. Die Weisheit der Vielen. München: Bertelsmann, 2005).

Toffler, A. (1980). The Third Wave. New York: Morrow.

Tonkin, E. (2006a). Folksonomies: The fall and rise of plain-text tagging. Ariadne, 47.

Tonkin, E. (2006b). Searching the long tail: Hidden structure in social tagging. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Trant, J. (2006a). Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. New Review of Hypermedia and Multimedia 12(2006)1, 83-105.

Trant, J. (2006b). Social classification and folksonomy in art museums: Early data from the steve.museum tagger prototype. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Vander Wal, T. (2004). Feed on this. [Blog post; 2004-10-03]. www.vanderwal.net/random/category.php?cat=153.

Vander Wal, T. (2005). Explaining and showing broad and narrow folksonomies. [Blog post; 2005-02-21]. www.vanderwal.net/random/category.php?cat=153.

Veres, C. (2006a). The language of folksonomies: What tags reveal about user classification. Lecture Notes in Computer Science 3999, 58-69.

Veres, C. (2006b). Concept modeling by the masses: Folksonomy structure and interoperability. Lecture Notes in Computer Science 4215, 325-338.

Voß, J. (2007). Tagging, folksonomy & co. – Renaissance of manual indexing? In: *Oßwald*, A., *Stempfhuber*, M., & *Wolff*, C. (Hrsg.): Open Innovation. Neue Perspektiven im Kontext von Information und Wissen. ISI 2007 (S. 243-254). Konstanz: Universitätsverlag.

Wang, X., *Bai*, R., & *Liao*, J. (2007). Chinese weblog pages classification based on folksonomy and support vector machines. Lecture Notes in Computer Science 4476, 309-321.

Weiss, A. (2005). The power of collective intelligence. netWorker, 9(2005)3, 16-23.

Winget, M. (2006). User-defined classification on the online photo sharing site Flickr … Or, how I learned to stop worrying and love the million typing monkeys. In: 17th ASIS&T SIG/CR Classification Research Workshop.

Weller, K., & *Peters*, I. (2007). Reconsidering relationships for knowledge representation. In: Proceedings of I-KNOW ,07. Graz, Austria, September 5-7, 2007 (S. 493-496).

Wu, H., *Zubair*, M., & *Maly*, K. (2006). Harvesting social knowledge from folksonomies. In: Proceedings of the 17th Conference on Hypertext and Hypermedia (S. 111-114). New York: ACM.

Xu, Z., *Fu*, Y., *Mao*, J., & *Su*, D. (2006). Towards the semantic Web: Collaborative tag suggestions. In: Proceedings of the 15th International WWW Conference. Collaborative Web Tagging Workshop.

Yeung, C.M.A., *Gibbins*, N., & *Shadbolt*, N. (2008). Mutual contextualization in tripartite graphs of folksonomies. Lecture Notes in Computer Science 4825, 966-970.

Zhang, L., *Wu*, X., & *Yu*, Y. (2006). Emergent semantics from folksonomies: A quantitative study. Lecture Notes in Computer Science 4090, 168-186.

Inhaltliche Erschließung, Zusammenarbeit, Folksonomy, Wissensrepräsentation, Information Retrieval, Prosumer, Broad Folksonomy, Narrow Folksonomy, Index-Tag, Such-Tag, Tag-Verteilung, Long Tail, Long Trunk, Power Law, invers-logistische Verteilung, Yule-Simon-Prozess, Ofness, Aboutness, Isness, Flickr, Del.icio.us, YouTube, Natural Language Processing, Relevance Ranking, Interestingness Ranking, ähnliche Dokumente, ähnliche Nutzer, Community

Die Autoren

Isabella Peters, M.A.

ist wissenschaftliche Mitarbeiterin der Abteilung für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf. Ihre Forschungsschwerpunkte liegen bei Web-2.0-Diensten, Folksonomies, der Blogo- und Podcastosphäre sowie im betrieblichen Wissensmanagement.
isabella.peters@uni-düsseldorf.de

Univ.-Prof. Dr. Wolfgang G. Stock

ist Leiter der Abteilung für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf. Seine Forschungen liegen vor allem in Gebieten des Information Retrieval, der Wissensrepräsentation und der Informetrie.
stock@phil-fak.uni-duesseldorf.de

Effektive Wissensverteilung mit DOMESTIC

Entdeckungs- und Erinnerungs-Recherche!

Alles dazu unter ...

www.domestic.de

Ontoverse:

Collaborative Ontology Engineering for the Life Sciences

Indra Mainz, Katrin Weller, Ingo Paulsen, Dominic Mainz, Jochen Kohl, Düsseldorf
& Arndt von Haeseler, Wien (Österreich)

We introduce ontologies as a new method for detailed and formalized knowledge representation, with particular focus on their use for the life sciences. Furthermore, we demonstrate the need for collaborative approaches in constructing elaborated scientific ontologies. The Ontoverse ontology wiki is presented as one tool to support all phases of collaborative ontology engineering.

Ontoverse: Gemeinschaftlicher Ontologieaufbau für die Lebenswissenschaften

Ontologien werden als eine neue Methode zur detaillierten und formalisierten Wissensrepräsentation vorgestellt, dabei liegt der Schwerpunkt auf dem Einsatz von Ontologien in den Lebenswissenschaften. Wir zeigen, dass für ausgereifte wissenschaftliche Ontologien Ansätze zur gemeinschaftlichen Erarbeitung notwendig sind. Das Ontoverse Ontologie-Wiki wird vorgestellt; es ist ein Hilfsmittel, das alle Phasen des gemeinschaftlichen Ontologieaufbaus unterstützt.

Introduction

Recently, two trends within the World Wide Web have emerged: the Semantic Web and the Social Web (or Web 2.0). The Semantic Web aims at sophisticated information integration (Antoniou & van Harmelen, 2004; Alexiev et al., 2005; Breitman et al., 2007, Davies et al., 2006), whereas the Social Web combines the power of user communities to produce and to organize different types of content (O'Reilly, 2005; Gordon-Murnane, 2006). The increasing merging of both social and semantic approaches has crucial effects on economics as well as on scientific communication and research strategies.

Within science management and funding policies, the term of e-Science has come up as the new perspective for global scientific collaboration (Hey & Trefethen, 2005; Goble et al., 2006). The overall goal of e-Science, meaning "enhanced Science", is to provide more sophisticated infrastructures for a global scientific network. This comprises distributed computing (grid-computing), the efficient management of resources, and interlinking of information, knowledge and people (Brünger-Weilandt, 2007; Weller et al., 2007). One of the key technologies in e-Science as well as for the Semantic Web are data structures called ontologies. Ontologies are formalized knowledge models and are currently viewed as the fundamental component to integrate, share and exchange knowledge and semantically annotated content, respectively.
Among the different scientific disciplines, the life sciences have proved themselves as particularly interested in these current developments; especially in the use of ontologies as means for formal knowledge representation and effective information integration. This is due to the vast amount of data produced in this field, but also to inconsistent terminologies and definitions within this large research area, which make scientific exchange eminently difficult. To fulfil the promises of e-Science the sustained construction of high-quality ontologies is one of the key challenges.
This article reports our work in progress from the Ontoverse research project. This project develops a web-based platform for collaborative ontology engineering and management in life sciences. We will explain the importance of collaborative ontology engineering in general and exemplified for the design and implementation of a domain ontology for bioinformatics called BIO2Me (BioInformatics Ontology for Tools and Methods) (Mainz, 2006). We discuss conceptual and technical problems in providing a platform for collaborative ontology management and present our approach for a technical solution.

1 Ontologies for Scientific Knowledge Representation

1.1 Knowledge Representation for the Semantic Web

Ontologies are knowledge representation systems which are defined as formal conceptualizations of a knowledge domain (Gruber, 1993), expressed in systems of concepts (classes), individuals (individuals are called instances if they are associated member of at least one class) and the relations between them. Thus they may depict a domain-specific vocabulary and its term-interconnections, collect synonyms (like in a thesaurus), capture hierarchies (like classification systems) and establish various self-defined interrelations between terms (figure 1). They also allow adding constraints and axioms, e.g. to control class memberships.
Ontologies are a key component in research efforts to establish a Semantic Web (which in return is often viewed as one part of e-Science programs). They should be used for making meaning of documents (or other information units) explicit and unambiguous. They are intended to support interpersonal communication as well as human-computer and inter-computer interactions. To capture these complex ontologies, mark-up languages can be used (e.g. XML or RDF) (Lassila & Swick, 1999), and more specific ontology languages have been developed (like OWL) (McGuinness & van Harmelen, 2004). To create ontologies on the basis of these formal languages, so-called ontology editors are available. These tools are often still under improve due to some deficiencies and/or new user requirements. They enable creating, storing, managing and exchanging of ontologies. A well-known editor currently used is Protégé[1]. It is supported by different additional plug-ins, e.g. for different visualization models, reasoning or search functionalities.

1 Protégé: http://protege.stanford.edu/.

With all this, ontologies can be more formal and complex than traditional knowledge representation systems and can thus be used for detailed depiction of knowledge domains and contexts. But they also often resemble traditional thesauri, classification systems or topic maps. Ontologies can be created for different purposes and with different aims. They may be regarded as a tool for document indexing and for enhancing of information retrieval processes. But they are also often intended to be a shared conceptualization of a domain of interest in terms of a knowledge base for a community (Gruber, 1993; Gruber, 2005). They may contain detailed, explicit information about (real-world) objects and contexts and may thus in some cases strongly resemble databases. Retrieving information within ontologies might then require additional knowledge in specific query languages. Reasoning mechanisms or inference engines are used to make information explicit which is implicitly included in the ontology and to check the logical consistency of the knowledge model (e.g. Gómez-Pérez et al., 2004).

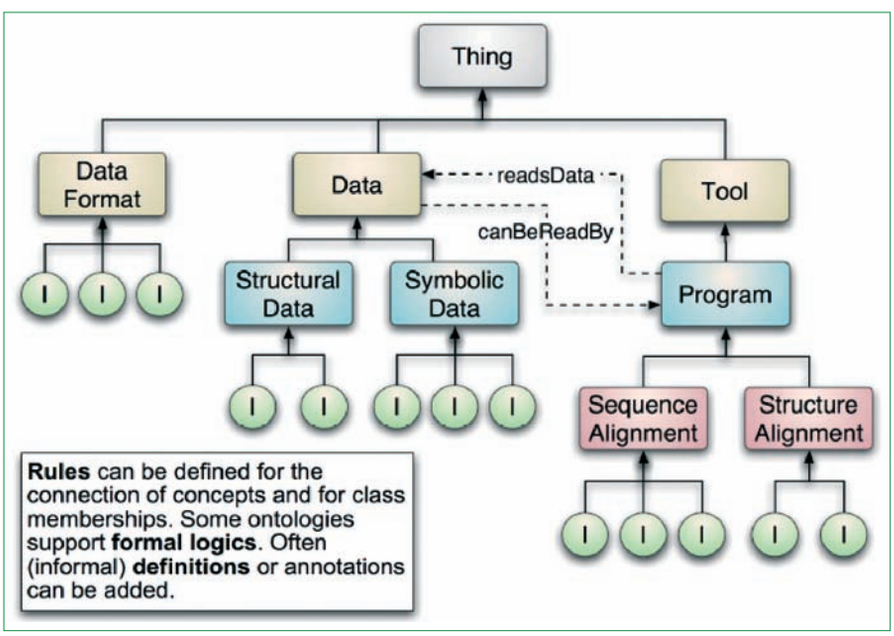


Figure 1: Schematic depiction of a cut-out of an ontology on bioinformatics, with a basic hierarchical structure and additional interrelations. Instances (I) can be added to the concepts.

This is realized by the use of attributes or restrictions added to the knowledge relations and by incorporating instances (individuals) as representations of real-life entities. Relations within an ontology can for example be defined as transitive, functional and symmetric, inverse relations can be bound together – all to describe contexts more precisely. In OWL DL[2] cardinalities may be applied to relations, e.g. for stating that an article must have at least one author or a dog must have exactly four legs (Weller, 2007). Instances that refer to real-life entities (not to abstract concepts) usually represent the lowest hierarchical level within an ontology. The inclusion of instances is often regarded as the crucial step from an ontology to a knowledge base. Including instances, complex interrelations and axioms, ontologies can be the foundation for knowledge-based information systems or expert level answering and advice systems (Jackson, 1998; Giarratano & Riley, 2005), that help to answer users' questions on a certain domain.

1.2 Ontologies in the Life Sciences

A great interest in ontologies as knowledge representation systems is found in the life sciences, mainly biology and medicine, and also bioinformatics (e.g. Baclawski & Niu, 2005; Bodenreider & Stevens, 2006). Within these domains we find some of the currently most regarded ontologies, first of all the Gene Ontology (GO) (Gene Ontology Consortium, 2000), but also the TAMBIS Ontology, the RiboWeb Ontology, The Ontology for Molecular Biology, EcoCyc etc. (Stevens et al., 2000; Bodenreider & Stevens, 2006). Some of them, among them Gene Ontology, are listed in the OBO (Open Biomedical Ontologies) collection[3]. The OBO collection is an umbrella community for a range of ontologies being designed for biomedical domains. A prominent characteristic of OBO is, that the collected ontologies should be orthogonal and interoperable to form an exclusive reference set.

The Gene Ontology consists of three controlled vocabularies (sub-ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions. They are used to precisely and consistently annotate gene products across different databases. GO was created jointly, the main work is done by a team of about ten experts. Additionally, anyone may post requests for new terms and suggest changes via the GO Website, a community has formed up to use and to discuss GO (Bodenreider & Stevens, 2006). The Gene Ontology currently comprises about 24,700 terms. It could not have been built and maintained by a single person. We see from this example, that collaborative efforts are needed to handle scientific ontologies. The actual implementation of collaboration in this context may be optimized; because discussions and planning tools are not adjusted to the needs of the ontology engineering process (see chapter 2.2).

We can also learn from this example that collaboration occurs on two different levels: The approach of the Gene Ontology team lets a core designer team be responsible for almost every decision, they are the ones who decide on the structure of the ontology. Thus, on this level we have a relatively small but very competent and active group, which requires high-quality in cooperation support. The community of interest, which can suggest changes and additions, constitutes a second level. These people only have limited influence on the actual structure of the Gene Ontology. This very strict two level approach may not be appropriate for every context. It does not entirely make use of the intellectual power of a large community. Yet, it points out a recurrent problem: the (apparent) reciprocity of quality control and openness of a collaborative system.

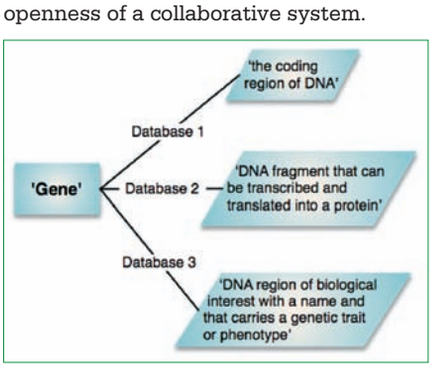


Figure 2: Inconsistent interpretations of the concept "Gene" within different settings. (Stevens et al., 2000).

A variety of reasons can explain the growing interest in ontologies within the life sciences: (a) Within these scientific disciplines large amounts of data are produced. This means, that actual data (not just text) is produced as a result of scientific research, which requires some

2 OWL Description Logics. One of the three subtypes of OWL.

3 OBO: http://obo.sourceforge.net.

DB Entry

SCIENTIFIC NAME : Shewanella putrefaciens
SYNONYM : Pseudomonas putrefaciens
SYNONYM : Alteromonas putrefaciens
SYNONYM : Shewanella putrefaciens (Lee et al. 1981) MacDonell and Colwell 1986
SYNONYM : Alteromonas putrefaciens (ex Derby and Hammer) Lee et al. 1981
SYNONYM : "Pseudomonas putrefaciens" (Derby and Hammer) Long and Hammer 1941
MISSPELLING : Shewanella putrifaciens
MISSPELLING : Alteromonas putrifaciens

Figure 3: Different names for the same species may exist and have to be bound together. Source: NCBI Taxonomy Database (Wheeler et al., 2000).

kind of annotation and structuring. Furthermore, one can currently observe an enormous, exponential growth in the size of data sets, which requires sophisticated structuring, accessibility, context information and information integration. (b) Some domains seem to contain taxonomies per se. Fields such as zoology particularly require global structuring and have long been a subject to structuring approaches (Bodenreider and Stevens, 2006). Ontologies can now be viewed as a new dimension in these ordering approaches. (c) Within the life sciences, researchers from a diversity of sub-disciplines have to collaborate, but (because of historical reasons) often do not use exactly the same vocabulary. Definitions may vary even for key concepts of the domain, a prominent example is the inconsistent definition of "gene" used in different scientific databases (figure 2) (Stevens et al., 2000). Sometimes also different names for the same thing (e.g. species) emerge within different scientific communities (figure 3). Both, the different views on the same concept and the unification of synonyms, have to be captured and formalized to enable more efficient collaboration.

1.3 An Ontology for Bioinformatics

Within the Ontoverse project, we have created an ontology for Bioinformatics Tools and Methods (BIO2Me) to provide a usage scenario for a detailed domain ontology and to gain first-hand experiences into collaborative ontology engineering processes (Mainz, 2006).
The need for such an ontology is based in the complexity and dynamics of this domain. Bioinformatics provides tools for the efficient processing of experimental data, of sequence analyses, and structure prediction and visualization (amongst others). There is a variety of programs available, which handle similar biological questions with sometimes completely different computational methods. This plethora of programs can currently not be overviewed with reasonable effort. Even for experts in bioinformatics it is hard to decide which tool fits their individual requirements best. Respectively, the BIO2Me ontology classifies bioinformatics tools and describes their characteristics as detailed as possible. Tools are categorized according to their application ranges (with bioinformatics perspective). Further properties are specified, like supported biological tasks, utilized computational methods, processed data formats etc. The ontology can thus help to enable searching for bioinformatics tools that meet a user's needs and can provide information about certain tools and computational methods. This is necessary almost every time, a researcher is confronted with new phenomena within his scientific field or with new working tasks.
Let us imagine the following scenario:
A new harmful virus type X has appeared; an outbreak of this virus should of course be prevented. A research group will be in charge of investigating the nature of this virus and suggesting possible counter-actions. In serious cases as this, a fast and comprehensive overview on all supportive tools, that can be used to analyze and process the problem at hand, is needed. The team will have to identify the properties of the virus and thereupon estimate its hazardous nature and identify possible molecular targets for a therapy. Different tools for such purposes exist and some of them will be known to every researcher within the responsible team. But there currently is no comprehensive overview on all existing tools, on technical requirements and interoperability.
A supposable expert system on bioinformatics tools (using an ontology like BIO2Me as its knowledge base) could provide answers to the following problems:

- Which tools and methods exist, that can handle the given tasks?
- Which of them are most suitable to do the research or most powerful?
- How can these tools be combined into a process framework?
- Which data output do they provide?

Of course, the underlying ontology will have to be exhaustive and kept up-to-date, i.e. it should constantly be enriched with newly derived information (e.g. new tools, algorithms, data formats). This requires a large contributing community, e.g. scientists that directly integrate new tools, which they have developed. On the other hand it requires constantly checking of the quality of the ontology: on consistency, accuracy and correctness.
If the ontology-based expert system envisioned so far would be furthermore integrated into a comprehensive e-Science platform, new options for handling the problem from our scenario become feasible (figure 4):

- We may directly couple it with a publication database to perform some basic research.
- Other scientific ontologies may be provided, e.g. some that clearly describe biological foundations.
- Ontologies can be used to enhance information retrieval processes.
- Expert profiles should be collected and social networks supported to enable searching for other experts for current problems.

The vision of such a comprehensive e-Science platform has highly influenced the design of the Ontoverse platform, which is described below (see Figure 4).

So far, BIO2Me has been created within the Ontoverse research team and with some help of associated researchers. It has been integrated to the Ontoverse ontology editing platform (see section 3) and will be subject to future collaborative refinements, before a stable version will be released.
Lessons learned from the construction of this ontology are as follows:

- The domain of BIO2Me eminently points out the need for collaborative ontology engineering. To represent

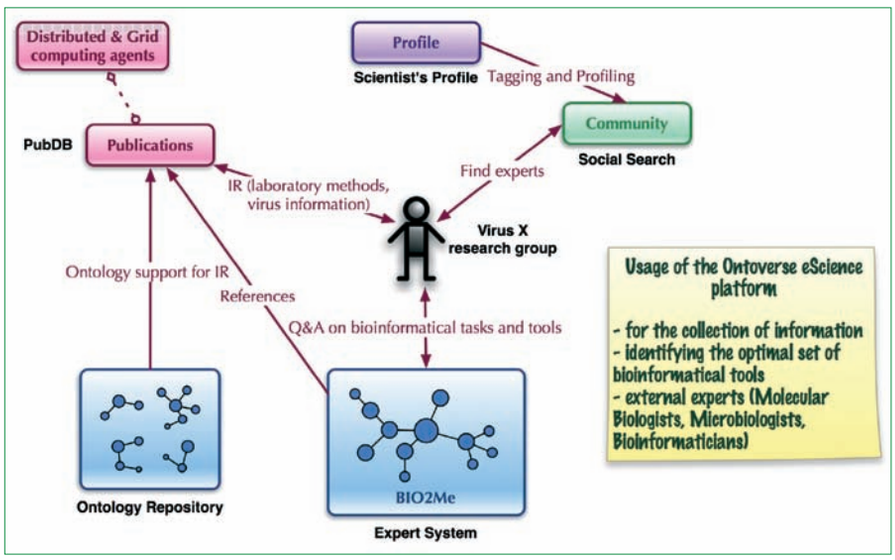


Figure 4: Scenario for using a scientific ontology as basis for an expert system and e-Science research platform.

bioinformatics tools with their applications, the whole bioinformatics research field and biology itself have to be displayed adequately in a structured way. Different expert knowledge is needed, to characterize different functions and application areas of bioinformatics tools. Furthermore, a vital community is needed to add information on the different tools which should be represented.

- In long term, the major challenge with this ontology will be, to keep it up-to-date. It will be necessary to constantly keep track of new developments in bioinformatics tools, e.g. as new tools or new versions of existing tools may be published. Ideally, the developers would directly report this to the community. The questions arise, how permanent changes in the domain of interest can be handled. Is it possible to work with different version releases?
- The basic challenge of this particular ontology was to define its basic structure. This is where the highest quality control is needed, because it is most difficult to change fundamental structures at a later point in time. It is also the part of the work, which requires the most discussion and planning. Less fundamental aspects, like adding new instances to existing concepts, can however easily and freely be handled by a large community.

2 Collaborative Ontology Engineering

As we have seen from Gene Ontology and BIO2Me development, a major problem of building ontologies currently lies in the enormous need for manpower to cover especially broad or deep domains of interest. Thus, collaborative ontology building is the key to more effective and comprehensive ontology engineering and maintenance. This opens different new fields of research: (a) Creating an ontology is laborious and requires careful considerations on how to represent a domain of interest adequately (Michlmayr, 2005). There are currently only few guidelines on how to model an ontology appropriately. Deeper understanding of problems in ontology planning and conceptualization is required. (b) A comprehensive understanding of social dynamics within collaborating communities is needed. (c) Based on the previous aspects, adjusted tools that support communities in building and maintaining ontologies have to be designed.

2.1 User Communities

As we can see with the current trends of Web 2.0 and Social Software, web users are principally willing to share their knowledge and to collaborate in creating new content (most prominently within the online encyclopaedia Wikipedia[4]). With folksonomies large communities have also started to take over some kind of indexing process for web documents (Guy & Tonkin, 2006; Mathes, 2004, Peters & Stock, 2007). A basic understanding of the functions of content-descriptive keywords and document indexing emerges, which will also be a useful basis for the understanding of ontologies. Yet, ontology engineering requires more specific skills, and collaborative ontology engineering poses new challenges to Social Software design. One crucial point is that for the construction of a domain ontology for scientific purposes, both ontology engineers or information architects as well as domain experts are needed to formalize precise representations (Paulsen et al., 2007). On the one hand ontology languages like OWL require specific technical skills and expertise, the process of ontology engineering requires experiences with knowledge representation methods. On the other hand, as we have also seen in the examples above, scientific ontologies often focus on very specialized domains of interest and demand for very detailed expert knowledge in these domains, respectively.

The ideal community for building scientific ontologies would therefore comprise both ontology designers (OD) and domain experts (DE). Only if they collaborate and combine their knowledge, ontologies can be created both well-formed and based on solid expert knowledge (of course it may also be the case, that one person possesses both kinds of expertise and can thus act as OD as well as DE). All this requires a special environment, which not only provides the technical means to plan, discuss and create ontologies, but also encourages the users to actually participate and share their knowledge with others. Although user-involvement does not seem to be a problem with Social Web applications of rather private nature, knowledge sharing and web collaboration in scientific contexts will only be accepted with reservation. It raises open questions regarding intellectual property rights, scientific reputation and motivation dynamics. Collaborative ontology engineering can be envisioned in two different settings: within closed and rather small groups, e.g. within a specific scientific research group, or with open worldwide communities. These two different scenarios will pose different requirements to a collaborative work environment particularly in regard of security aspects, access rights to data, user identification and authentication or authorship labelling. Yet, the basic aspects of the workflow in ontology engineering will be the same in both cases. Furthermore, the user community should comprise ODs as well as DEs.

2.2 Ontology Engineering Processes and Knowledge Elicitation

Ontology engineering is a process of several phases; mainly a planning and conceptualization phase, a phase of actual editing the structural data, a maintenance phase that includes updates, corrections and ontology enrichment, and as an optional last phase the reuse of the ontology (e.g. as starting point for a new ontology) (Weller, 2006). Other segmentation approaches are also possible, different methodologies for ontology engineering are discussed, e.g. in (Corcho et al., 2003; Gómez-Pérez et al., 2004; Stevens et al., 2000).

Every ontology project should begin with a planning phase to define the aim and scope of the ontology as well as design guidelines. It is useful to provide an Ontology Requirement Specification Document (ORSD) (Sure et al., 2002) for that purpose, e.g. as a standard form which lists the basic planning points and which can be filled with the respective decisions. Particularly in collaborative ontology engineering this phase is crucial to receive a common view on the domain of the ontology. During this planning phase, thematic discussions of the user community are extremely helpful to capture the characteristics of the domain. Thus, the planning phase passes directly into a more content-centred conceptualization phase, to design a first sketch of the structure. Figure 5 shows a first conceptualization for the BIO2Me ontology.

The construction of a formal ontology always depends heavily on an iterative process of knowledge elicitation and knowledge formalization. This also reflects the interactions of domain experts, who bring in their knowledge, and ontology designers, who formalize it. This interaction is the key to successful collaborative ontology engineering. Some requirements for collaborative ontology editors can be derived: DEs need specific tools to enter informal or semi-formalized knowledge and to discuss it with each other. Specific communication channels have to be established for fundamental discussions on the purpose and on the structure of the ontology.

The iterative character of ontology conceptualization and formalization also implies that the ontology will not be consistent (and of course not complete) during the entire process. Discrepancies and contradictions are necessary to a certain degree, but may cause problems in handling and using the ontology. The same holds true for changing variations of the ontology over time. One approach to this

4 Wikipedia: www.wikipedia.org.

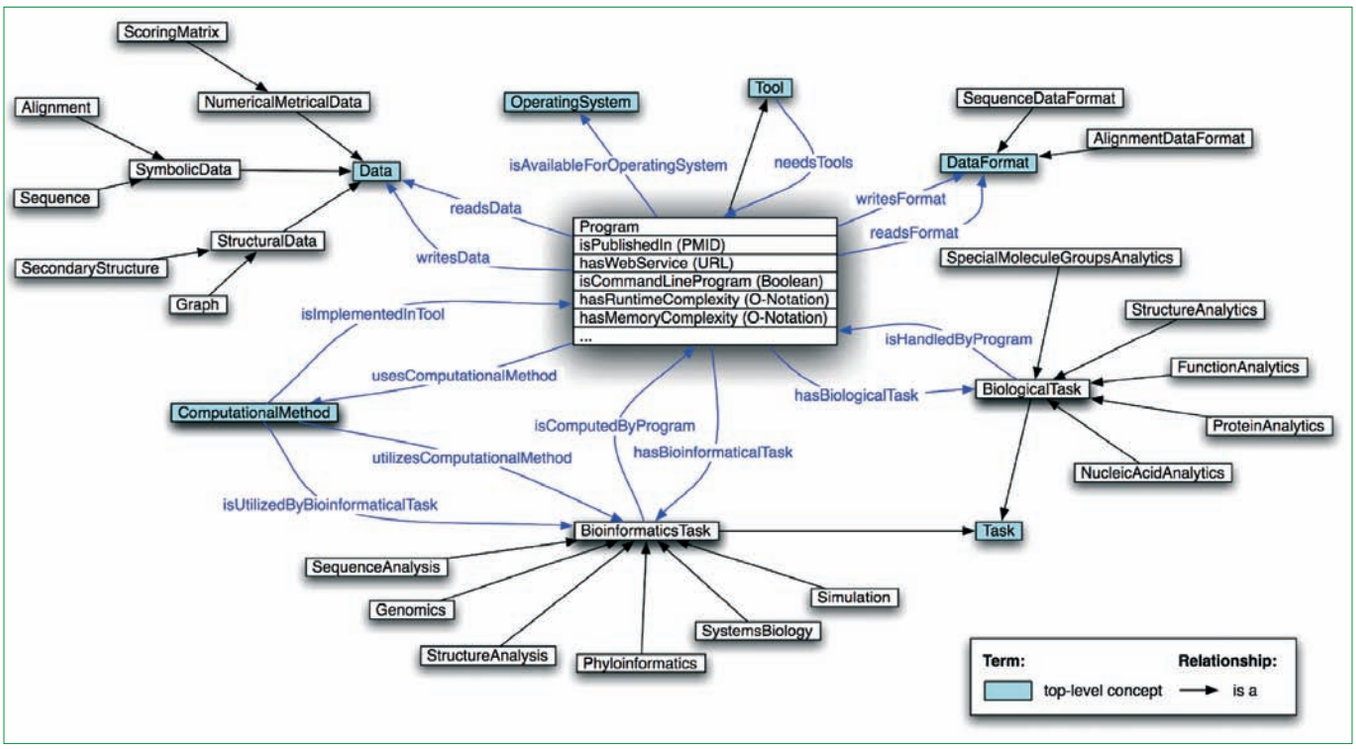


Figure 5: First conceptualization to structure the domain of bioinformatics for the BIO2Me development.

may be to establish a basic foundation of the ontology first, which provides a fixed frame and which cannot be altered after a certain stage. It will then have to be constantly filled with detailed information.

2.3 Tools for Collaborative Ontology Engineering

It requires a lot of coordination, communication and support by adjusted tools to enable such collaborative development of ontologies, especially in early phases when the basic structure of the ontology is planned. Current ontology editors do not explicitly support planning and conceptualization of ontologies, but concentrate on the formal editing process. This support will be vital, especially for broad and heterogeneous communities (but also for small collaboration groups). The idea of collaborative ontology editing and supportive tools is not entirely new. First examples date back to the early days of ontology engineering research (e.g. Farquhar et al., 1996). Some support single features to enable multi-user projects, like change resolution or labelling of authorship information. Others aim at open web-based solutions. Yet, no tool could so far establish itself with a community, which jointly creates formal domain models on large scale. Lately the huge success of Social Software applications and Web 2.0 may have encouraged new works in this regard, several new projects regarding some kind of collaborative ontology engineering come up, principles and problems are discussed (Noy et al., 2006). Some of the more recent efforts to develop collaborative tools for creation of knowledge models or knowledge representations are for example Wiki@nt (Bao & Honavar, 2004), OntoWiki (Hepp et al., 2005), Collaborative Protégé[5] or Knoodl[6].
Other approaches combine databases and content production with some kind of vocabulary structuring or semantic annotation, e.g. Soboleo (Zacharias & Braun, 2007), DBin (Tummarello et al., 2006) or Semantic Wikipedia (Völkel et al., 2006). They do not focus on the construction of explicit formal knowledge models; they follow a more practice-oriented approach which then along the way leads to a domain representation model.
The field should be monitored, as some more tools are expected to become published soon.

3 The Ontoverse Approach to Collaborative Ontology Engineering

3.1 The Ontoverse Ontology Wiki

The fundamental idea for the Ontoverse project was to provide a place, were users are free to create, edit and modify ontologies and to do this as a community of users with the same interests. A formal ontology editor is combined with (modified) basic wiki characteristics (easy to handle, web-based, multi-user support, change tracking and notifications, undoing changes, options for discussing the current state). The Ontoverse platform will handle multiple ontologies and allow users to start new ontology projects. It is thus also a repository for different scientific ontologies, so that another major role for an interested community will be to interlink and classify current ontology projects to make them accessible and traceable. This will enable effective ontology reuse and prevent redundant projects.
A web-based collaboration platform is combined with formal ontology engineering functionalities, social networking aspects, social and semantic tagging and information extraction mechanisms. Major innovations are:

- An open collaborative approach is used, which takes into account all the people who have expert knowledge in a certain knowledge area. A major challenge thus lies in constructing a system that brings together ontology designers and domain experts and regards their different types of expertise, so that ontology quality profits from their different abilities. This approach requires further discussions on collaboration models for DEs and ODs. Different models for access rights and division of work can be envisioned. Yet, each of them requires vital community and social networking support. The Ontoverse platform supports the search for users with certain qualities, e.g. fields of interest.
- All different phases of ontology development are supported. The wiki should also encourage thematic dis-

5 Collaborative Protégé: http://protege.stanford.edu/doc/collab-protege/.
6 Knoodle : http://knoodl.com/ui/home.html.

cussions and the adding of preliminarily unstructured data in the forefront of the formal ontology editing process. Current ontology editors do hardly include supporting mechanisms for the planning and conceptualization phase in ontology engineering. This step however is very important for the development of an ontology. Ontoverse introduces a distinction of proto-ontological and ontological data: Ontological data refers to formalizations of ontological elements (mainly concepts, relations/properties, instances). In contrast, proto-ontological data can be every piece of information not yet being brought into formal structure. This ranges from the ontology requirement specification, over semi-formal concept graphs to the rating and annotation of concepts and relations. The Ontoverse platform facilitates the organization of these data by establishing a proto-ontological area, in which domain experts can offer their knowledge and discuss it with other experts. Therefore different forums are available as well as a wiki, which can be used to enter and organize data, and to plan and discuss an ontology (including a prefabricated form for the ORSD and options to include graphical schemes of a domain, figures 6 and 7). This wiki is directly interlinked with a web-based ontology editor (figure 8).

- For the formal ontology editor, new technical solutions are being developed to facilitate data storing and collaborative editing of data. Interactive ontology visualizations are included as well as options for integrative ontology mapping and merging (Malzahn et al., 2007; El Jerroudi & Ziegler, 2007).

- Integration of a publication database to store and manage thematic documents and adding information extraction (IE) functionalities (Pazienza, 2003): Publications are one kind of knowledge source to gather information about a domain. Within the Ontoverse project a publication database is used as knowledge source for an information extraction application that supports the community in identifying relevant concepts, instances and relations from texts, which can then be incorporated into the ontologies (figure 9). In return, the newly developed ontologies themselves will help to retrieve relevant documents from the database. The publications can be indexed with concepts from the ontology. This will be an effective way to enhance current metadata and keywords as semantic browsing of documents becomes possible.

- Ontoverse combines different types of content indexing: (a) Original keywords assigned to the documents (by their authors or in professional indexing processes) are recorded. (b) Ontology terms from any ontology kept in the Ontoverse platform can be added to publications automatically. (c) Registered users of Ontoverse can also tag publications with their own keywords. For future, refinements are planned to incorporate the human actor as evaluator of the content. A seeking user benefits from the Human tagged content in combination with information about the tagger (like scientific reputation).

- Ontoverse focuses particularly on scientific ontologies and will become a repository for a multitude of collaboratively designed ontologies for the life sciences. The tools for developing ontologies and the actual ontologies are kept within one web-accessible platform.

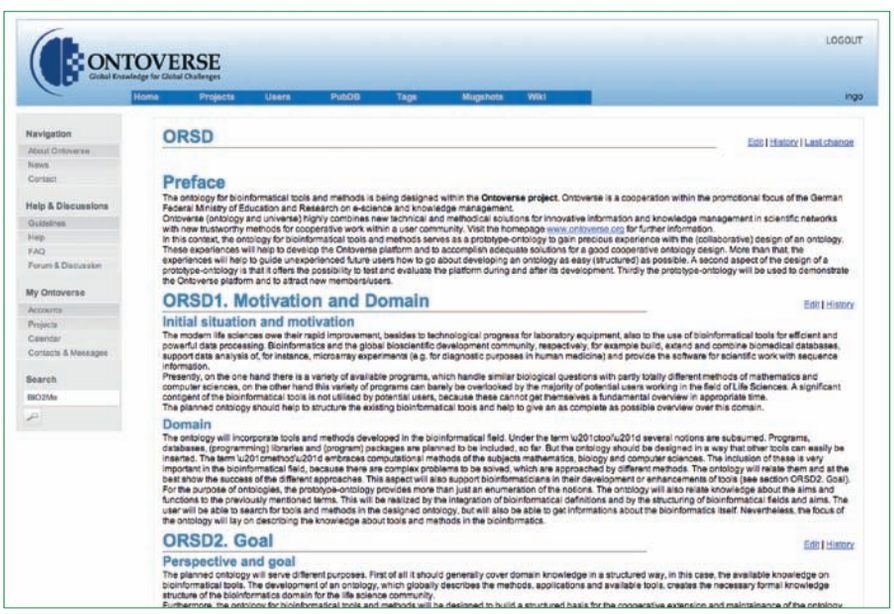


Figure 6: Ontoverse wiki for informal knowledge input, planning and discussion: Planning of a new ontology project with the ORSD.

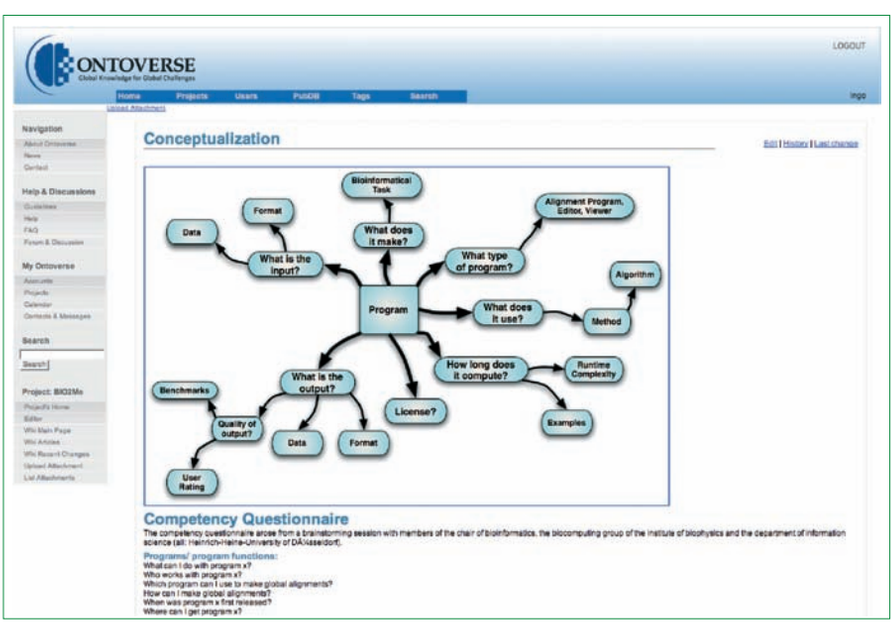


Figure 7: Ontoverse wiki for informal knowledge input, planning and discussion: Conceptualization of a domain of interest.

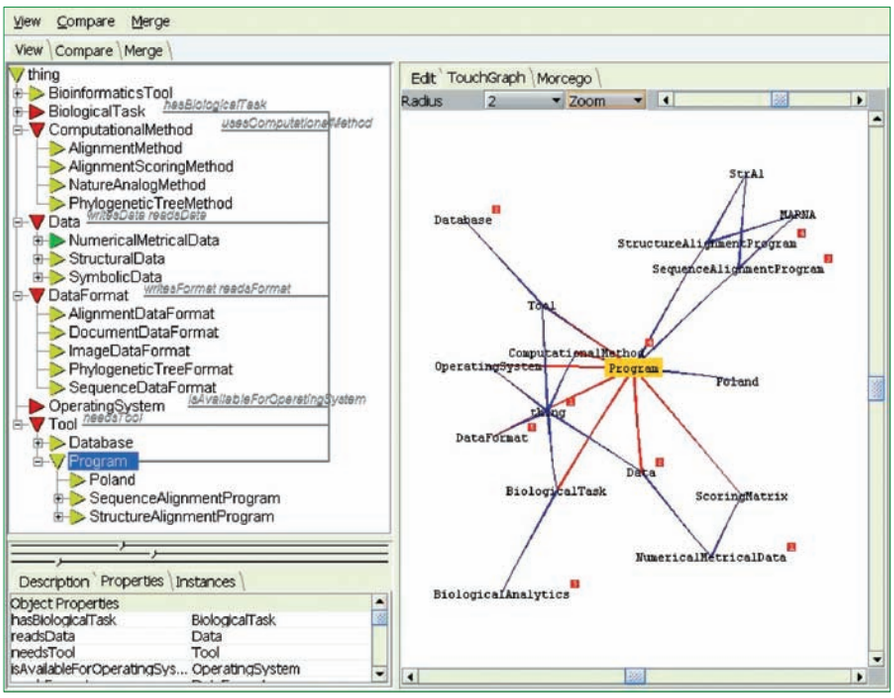


Figure 8: Ontoverse's collaborative editor for formal ontology engineering with additional ontology visualization. (El Jerroudi & Ziegler, 2007).

3.2 Working with the Ontoverse Platform

The combination of all the aspects above add up to an environment for knowledge exchange and collaborative knowledge formalization: the Ontoverse platform[7].
Ontology editing is organized in projects: for each ontology, one project is generated, which is the collaborative workspace for this ontology. Users can then join these projects to work on the ontology, respectively. For every project, there is a wiki space to support planning and conceptualization (figure 6 and 7). This wiki is interrelated with the formal ontology editor (figure 8), which will load the project's ontology if it is opened.

Community features. Every member of the platform has a user profile, which can be filled by the user himself (the approach of Ontoverse is to allow only real names to be used for members to support quality control, authorship notifications and to better respond to scientific communication structures). Users can browse existing profiles to find new research contacts.
General and project-specific forums can be used for discussions and to get in contact with other experts. Personal contact lists can be used to build up social scientific networks. Event calendars help to organize work schedules and to announce events relevant for the whole community or single projects.

Project Organization. The typical workflow can be described as follows: A scientist needs an ontology of a specific domain (or some specific information out of it). S/he checks the Ontoverse platform for existing ontology projects. If one of them already fits her/his needs, s/he may join the project and either actively contribute to the project or passively view and use the collected data. If none of the projects was suitable, s/he may initiate her/his own ontology project, hoping for more users to join it and collaborate. The founder of a new project will become its project administrator (PA) and is responsible for final decisions regarding the ontology (in this regard, more detailed structures for usage rights and internal workflows are under development).

Ontology planning and engineering. Within the project-specific wiki for each ontology project, a form for an ORSD is provided to capture basic requirements and conventions. It is then useful, to begin with a loose collection of relevant proto-ontological data. These proto-ontological data may be textual, e.g. glossaries, single concepts enriched with explanations or definitions, unstructured collections of concepts or other notes and references to external sources. Like in conventional wikis, documents can be uploaded and integrated; that will mainly be images, e.g. schematic depictions of a subject. Throughout the whole wiki, changes can be traced and assigned to their authors. All this collected domain knowledge is needed as a basis to set up a formal ontology. This should be done by experienced ontology designers. Optimal results will be achieved if ODs pose concrete questions regarding the domain directly to the DEs, ideally even express their information needs concretely within the ontology wiki. ODs then build up the actual ontology via another channel: the web-based, multi-user ontology editor (Malzahn et al, 2007; El Jerroudi & Ziegler, 2007; Paulsen et al., 2007). It produces an ontology in OWL format, stored in a web repository. Within the editor, changes during the engineering process will be highlighted.

Information extraction (IE) and ontology population. The internal publication database (PubDB, basically covering scientific papers from PubMed) can be used as the basis for semi-automatic information extraction processes. This will help to identify new concepts for an ontology based on existing literature. The IE tool suggests new concepts, based on the concepts already existing in the ontology and kind of hierarchical structures identified in the publications' texts. The results of this process for enriching the BIO2Me ontology with concepts from bioinformatics literature may look like this: 'StructSorter' *is_a* 'Alignment Program' (figure 9). AlignmentProgram is a concept in BIO2Me, StructSorter has been identified by the IE tool. A user may now decide, whether he wants to add StructSorter as a sub-concept of Alignment Program in the ontology.
Within the projects, the working groups can collect relevant publications. Publications may be tagged with free keywords, but one can also index them with terms from an existing ontology.

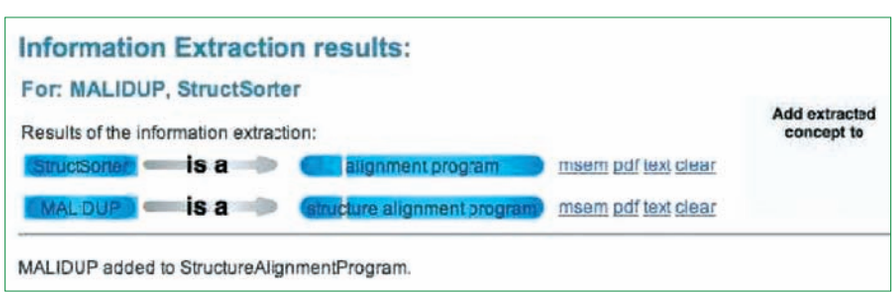


Figure 9: Results from Ontoverse's information extraction tool, which finds new concepts or instances for an existing ontology.

7 Ontoverse Demonstration: http://www.ontoverse.org.

3.3 Future Work

While the workflow described above operates on our internal version of the system, additional work has to be done before the Ontoverse platform can be opened for an interested community. So far we have established basic communication and networking structures. We provided the shared workspaces for ontology planning, conceptualization and formalization as well as options for retrieving literature from a publication database. Some information extraction functionalities can be used.

Topics and problems which will now have to be extended are:

- Community management, usage roles and rights
- Authorship notifications, copyright issues, intellectual property
- Prevent of misuse and vandalism, community dynamics, incentives

Some features are under development, e.g. mapping and merging of ontologies, versioning support, more specific change notification systems. We may also discuss the types of data formats supported by Ontoverse. Currently, ontologies are created in OWL, other languages might be needed.

The BIO2Me ontology will be refined, "completed" and updated constantly. New ontologies will be started as projects within the platform.

Conclusion

Ontologies are the core element for an upcoming Semantic Web and a vital focus of current research activities in different areas. Besides computer science, information science, computational linguistics and related fields, a growing interest in ontologies is observed in the life sciences. They need ontologies to formalize given information in a uniform and explicit way, to generate a shared knowledge model and to annotate and interrelate research results.

Ontologies for such scientific purposes have to be maintained over long time spans and have to be of high quality. This is achieved best, if a community of experts with different abilities jointly develops and maintains an ontology. Expert knowledge for the different aspects of the domain to be captured is needed as well as understanding of principles and methods of ontology engineering.

Different tools are currently under development, which enable shared development of domain models and ontologies. Ontoverse provides a solution, which exclusively focuses on the needs of shared ontology engineering for a heterogeneous scientific community. This solution will serve as a basis for cooperative knowledge management and knowledge structuring in form of domain ontologies and will at the same time become a platform for scientific dialog to support institutions and professional users in confidential knowledge exchange.

Acknowledgments

Financial support was provided by the German Bundesministerium für Bildung und Forschung (BMBF). Arndt von Haeseler is also supported by the Wiener Wissenschafts-, Forschungs und Technologiefonds (WWTF).

We thank all our colleagues within the Ontoverse project. Further thanks go to all participants of the Ontoverse Workshop 2007 for fruitful discussions and suggestions.

References

Antoniou, G. & *van Harmelen*, F. (2004). A Semantic Web Primer. Cambridge, Mass.: MIT Press.

Alexiev, V. et al. (2005). Information Integration with Ontologies: Experiences from an Industrial Showcase. Chichester: Wiley & Sons.

Baclawski, K. & *Niu*, T. (2005). Ontologies for Bioinformatics. Cambridge, Mass: MIT Press.

Bao, J., & *Honavar*, V. (2004). Collaborative Ontology Building with Wiki@nt: A Multi-agent Based Ontology Building Environment. In: Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools (EON), Hiroshima 2004. Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools. http://km.aifb.uni-karlsruhe.de/ws/eon2004/EON2004_Proceedings.pdf [10.01.2008].

Bodenreider, O. & *Steves*, R. (2006). Bio-Ontologies: Current Trends and Future Directions. Briefings in Bioinformatics 7(2006)3, 256-274.

Breitman, K., *Casanova*, M.A., & *Truszkowski*, W. (2007). Semantic Web. Concepts, Technologies and Applications. London: Springer.

Brünger-Weilandt, S. (2007). E-Science: Advancing New Ways of Scientific Communication. Information Services & Use 27(2007)4, 161-166.

Corcho, O., *Fernández-López*, M. & *Gómez-Pérez*, A. (2003). Methodologies, Tools and Languages for Building Ontologies. Where is their Meeting Point? Data & Knowledge Engineering 46(2003)1, 41-64.

Davies, J., *Fensel*, D. & *van Harmelen*, F. (Eds.) (2003). Towards the Semantic Web: Ontology-Driven Knowledge Management. Chichester: Wiley & Sons.

El Jerroudi, Z. & *Ziegler*, J. (2007). Interaktives Vergleichen und Zusammenführen von Ontologien. I-com, Zeitschrift für interaktive und kooperative Medien 6(2007)3, 44-49.

Farquhar, A., *Fikes*, R. & *Rice*, J. (1996). The Ontolingua Server. A Tool for Collaborative Ontology Construction. Technical report, Stanford KSL96-26.

Gene Ontology Consortium (2000). Gene Ontology: Tool for the Unification of Biology. Nature Genetics 25(2000), 25-29.

Giarratano, J. & *Riley*, G.D. (2005). Expert Systems: Principles and Programming. 4th Edition. Boston, MA: Course Technology.

Goble, C. et al. (2006). E-Science and the Semantic Web: A Symbiotic Relationship. In: N. *Lavrac*, L. *Todorovski*, & K.P. *Jantke* (Eds.): Proceedings of Discovery Science, 9th International Conference, DS 2006 (pp. 1-12). Berlin, Heidelberg: Springer.

Gómez-Pérez, A., *Fernández-López*, M., & *Corcho*, O. (2004). Ontological Engineering. London: Springer.

Gordon-Murnane, L. (2006). Social Bookmarking, Folksonomies, and Web 2.0 Tools. Searcher. The Magazine for Database Professionals 14(2006)6, 26-38.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specification. Knowledge Acquisition 5(1993)2, 199-220.

Gruber, T. (2005). Ontology of Folksonomy. A Mash-Up of Apples and Oranges, In 1st On-Line Conference on Metadata and Semantics Research (MTSR '05). http://tomgruber.org/writing/ontology-of-folksonomy.htm [10.01.2008].

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

Hepp, M., *Bachlehner*, D. & *Siorpaes*, K. (2005). OntoWiki. Community-driven Ontology Engineering and Ontology Usage Based on Wikis. In Proceedings of the 2005 International Symposium on Wikis (pp.143-144). New York: ACM Press.

Hey, T. & *Trefethen*, A. (2005). Cyberinfrastructures for e-Science. Science, 308, 817-821.

Jackson, P. (1998). An Introduction to Expert Systems. 3rd Edition. Boston, MA: Addison-Wesley.

Lassila, O. & *Swick*, R.. (Eds.) (1999). Resource Description Framework (RDF): Model and Syntax Specification. W3C Candidate Recommendation. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ [10.01.2008].

Mainz, I. (2006). Entwicklung einer Prototypontologie für bioinformatische Werkzeuge. Bachelor Thesis at the Heinrich-Heine-Universität Düsseldorf, October 2006.

Malzahn, N. et al. (2007). Collaborative Ontology Development: Distributed Architecture and Visualization. In: Proceedings of the German eScience Conference 2007 (GES), Max Planck Digital Library, ID: 315470.0.

Mathes, A. (2004). Folksonomies: Cooperative Classification and Communication Through Shared Metadata. http://adammathes.com/academic/computer-mediated-communication/folksonomies.html [10.01.2008].

McGuinness, D. & *van Harmelen*, F. (Eds.) (2004). OWL Web Ontology Language Overview. W3C Recommendation. http://www.w3.org/TR/owl-features/ (retrieved January 10, 2008).

Michlmayr, E. (2005). A Case Study on Emergent Semantics in Communities. In G. *Stumme* et al. (Eds.), Proceedings ot the Workshop on Semantic Network Analysis, International Semantic Web Conference (pp. 1-14). www.kde.cs.uni-kassel.de/ws/sna2005/SNA05-proc.pdf [10.01.2008].

Noy, N.F. et al. (2006). A Framework for Ontology Evolution in Collaborative Environments. In: I. *Cruz* et al. (Eds.): Proceedings of the 5th International Semantic Web Conference (pp. 544-558). Berlin, Heidelberg: Springer.

O'Reilly, T. (2005). What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [10.01.2008].

Pazienza, M.T. (Ed.) (2003). Information Extraction in the Web Era. New York: Springer.

Paulsen, I. et al. (2007). Ontoverse. Collaborative Knowledge Management in the Life Science Network. In: Proceedings of the German eScience Conference 2007 (GES), Max Planck Digital Library, ID 316588.0.

Peters, I. & *Stock*, W.G. (2007). Folksonomy and Information Retrieval. In: Proceedings of the 70th Annual Meeting of the American Society for Information Science and Technology. Vol. 45 (pp. 1510-1542) (CD-ROM).

Stevens, R., *Goble*, C. & *Bechhofer*, S. (2000). Ontology-Based Knowledge Representation for Bioinformatics. Briefings in Bioinformatics 1(2000)4, 398-414.

Sure, Y., *Staab*, S. & *Studer*, R. (2002). Methodology for Development and Employment of Ontology Based Knowledge Management Applications. ACM SIGMOD Record, Special Section on Semantic Web and Data Management 31(2002)4, 18-23.

Tummarello, G., *Morbidoni*, C. & *Nucci*, M. (2006). Enabling Semantic Web Communities with DBin: an Overview. In: I. *Cruz* et al. (Eds): Proceedings of the 5th International Semantic Web Conference (pp. 943-950). Berlin, Heidelberg: Springer.

Völkel, M. et al. (2006). Semantic Wikipedia. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, May 2006 (WWW06) (pp. 585-594). New York: ACM Press.

Weller, K. (2006). Kooperativer Ontologieaufbau. In: M. *Ockenfeld* (Hrsg.): Content, 28. Online-Tagung der DGI, 58. Jahrestagung der DGI, Proceedings (pp. 227-234). Frankfurt am Main: DGI.

Weller, K. (2007). Folksonomies and Ontologies. Two New Players in Indexing and Knowledge Representation. In: H. *Jezzard* (Ed.): Online Information Conference Proceedings (pp. 108-115). London: Learned Information Europe.

Weller, K. et al. (2007). Semantisches und vernetztes Wissensmanagement für Forschung und Wissenschaft. In: R. *Ball* (Hrsg.): Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek Forschungszentrum Jülich (WissKom 2007) (pp. 33-46). Jülich: Forschungszentrum Jülich Zentralbibliothek GmbH Verlag.

Wheeler, D.L. et al. (2000). Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 28(2000)1, 10-14.

Zacharias, V, & Braun, S. (2007). SOBOLEO. Social Bookmarking and Lightweight Ontology Engineering. In: Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), 16th International World Wide Web Conference (WWW 2007), Banff, Alberta, Canada, May 2007. www2007.org/workshops/paper_41.pdf [10.01.2007].

Ontology, ontology engineering, collaboration, social software, ontology editor, life sciences

Zusammenarbeit, Wissensrepräsentation, Ontologie, Software, Projekt, Lebenswissenschaften

The Authors

From left to right:
Jochen Kohl, Arndt von Haeseler, Dominic Mainz, Katrin Weller, Indra Mainz and Ingo Paulsen

Dipl.-Biol. Indra Mainz, B. Sc.

is researcher in the Bioinformatics Group at the Institute for Biophysics, Heinrich-Heine-University Düsseldorf. She is member of the Ontoverse research project where she is mainly engaged with collaborative ontology engineering processes and with the development of a wiki for proto-ontological knowledge management. She furthermore is concerned with the development of the BIO2Me ontology and respective practical applications.
mainzi@biophys.uni-duesseldorf.de

Katrin Weller, M.A.

is researcher at the Department of Information Science, Heinrich-Heine-University Düsseldorf. She is member of the Ontoverse research project. Her main research foci are ontology development and knowledge engineering, social networking and collaborative ontology management.
weller@uni-duesseldorf.de

Dr. Ingo Paulsen

is co-initiator of the research project Ontoverse and works at the Bioinformatics Department, Heinrich-Heine-University Düsseldorf. His researches focus on Semantic Web technologies, computational linguistics and biological databases.
paulsen@cs.uni-duesseldorf.de

Dipl.-Biol. Dominic Mainz

is co-initiator of the research project Ontoverse and works at the Bioinformatics Department, Heinrich-Heine-University Düsseldorf. His research focus is on semantic technologies to process ontological data, ontology editors and logic programming.
mainz@cs.uni-duesseldorf.de

Dipl.-Biol. Jochen Kohl

is researcher at the Bioinformatics Department, Heinrich-Heine-University Düsseldorf. He is member of the Ontoverse research project and is particularly engaged in the database-supported management of ontologies.
kohlj@uni-duesseldorf.de

Prof. Dr. Arndt von Haeseler

acts as the scientific coordinator of the Ontoverse project. Since September 2005 he is the scientific director of the Center for Integrative Bioinformatics Vienna (CIBIV) and has a full professorship for Bioinformatics.
arndt.von.haeseler@univie.ac.at

Paradigmatic and Syntagmatic Relations in Knowledge Organization Systems

Isabella Peters und Katrin Weller, Düsseldorf

Classical knowledge representation methods have been successfully working for years with established – but in a way restricted and vague – relations such as synonymy, hierarchy (meronymy, hyponymy) and unspecified associations. Recent developments like ontologies and folksonomies show new forms of collaboration, indexing and knowledge representation and encourage the reconsideration of standard knowledge relationships for practical use. In a summarizing overview we show which relations are currently used in knowledge organization systems (controlled vocabularies, ontologies and folksonomies) and which relations are expressed explicitly or which may be inherently hidden in them.

Paradigmatische und syntagmatische Relationen in der Wissensrepräsentation

Klassische Methoden der Wissensrepräsentation arbeiten seit Jahren erfolgreich mit etablierten – aber eingeschränkten und vagen – Begriffsrelationen, wie Synonymie, Hierarchien (Meronymie, Hyponymie) und unspezifizierten Assoziationen. Jüngste Entwicklungen wie Ontologien und Folksonomies stellen neue Möglichkeiten für Zusammenarbeit, Indexierung und Wissensrepräsentation dar und regen gleichzeitig dazu an, die bestehenden Standard-Relationen zu überdenken. Wir geben einen zusammenfassenden Überblick über Relationen, die derzeit in der Wissensrepräsentation eingesetzt werden, und erkunden explizite und versteckte Relationen innerhalb von Ontologien und Folksonomies.

Introduction

Most methods used in knowledge organization systems (KOS) organize knowledge in form of concepts and the relations between them, though the level of differentiating relations varies. Major knowledge representation methods make some of the relations explicit, but they use only a small number of paradigmatic semantic relations so far. This approach is somewhat contrary to the detailed and highly-professional theoretical reflections on relationships that are currently available from fields such as information science, knowledge engineering, linguistics, artificial intelligence studies and related disciplines, e.g. (Green et al. 2002), (Löbner 2002), (Khoo & Na 2006), (Storey, 1993), (Sun & Zheng 2005).

Our aim is to point out some recent developments in knowledge representation and indexing, which motivate to reconsider the classification of relations for practical use. First of all, the current efforts to establish a *Semantic Web* demand for more sophisticated depiction of knowledge relationships. Ontologies, the core technique for indexing documents with semantic representations, can make use of a wider range of relations than classical KOS. A closer look at the expressiveness of logic-based ontology languages and an examination of current ontologies will help to redefine a classification of relationship types. The second and less obvious stimulus lies in upcoming *folksonomies* and the relations implicitly existent between user-created tags in social tagging systems (Peters & Stock 2007; 2008).

Both new methods of knowledge representation may be a valuable resource for reconsidering existing relations and establishing new generalizable ones. Finally, we point out several aspects in establishing knowledge-relationships, which are of importance for information structuring and retrieval and should become subjects of future discussions.

Relation Types in Knowledge Organization Systems

In knowledge organization systems (Zeng 2000) two fundamental semantic relations can be distinguished, based on their strength in bonding terms together: *Paradigmatic relations* are fixed, rigidly coupled concept relations applied to controlled vocabularies. An example would be the hierarchical relation between "vehicle" and "bicycle" formalized within a classification scheme. *Syntagmatic relations* are not attached to concepts but originate merely in the actual co-occurrence of terms within a certain setting. Thus, syntagmatic relations may for example exist between two expressions within a single document or between two keywords assigned to a document. These relations can be captured and expressed in form of networks (Stock 2007, p. 451). In context of information retrieval applications, both syntagmatic and paradigmatic relations help the user to browse concept systems for appropriate search terms and enable query expansion.

In knowledge representation *generalizable* relations play a major role; these are paradigmatic relations that can be meaningfully used in all (or in most of all) general or domain-specific knowledge representation and organization models. The relations currently considered to be generally applicable and implemented in practice are (Bean & Green 2001), (Bertram 2005, p. 34):

- *Relations of equivalence*. Synonyms and quasi-synonyms have (almost) the same meaning and are therefore exchangeable in a given context, e. g. "car" and "automobile" are synonyms (Löbner 2002, p. 117). Control over synonyms is particularly important for indexing and documentation contexts, to enable a consistent use of a vocabulary and enhance recall in information retrieval.
- *Hierarchical relations*. Two concepts are hierarchically related if one concept includes the extension of the other concept. This comprises meronymy (mereology, part-of relation, part-whole relation, partonomy) (Gerstl & Pribbenow 1996), (Pribbenow 2002) and hyponymy (kind-of-rela-

This article is an extended and revised version of the poster presentation "Reconsidering Relationships for Knowledge Representation" presented at the Triple-I Conference 2007 in Graz, Austria. In a report format we would like to provide insights to our current work-in-progress and like to encourage discussion and suggestions on our ideas, approaches and preliminary findings.

tion, taxonomic relation, taxonomy) (Cruse 2002), i. e. "lips" is a meronym of "mouth", "duck" is a hyponym of "bird". The different sub-concepts for a shared broader concept are siblings (e.g. "duck" and "lark" are both sub-concepts of "bird", so they are siblings).

- *Associative relations*. They are unspecified connections of concepts that can have any kind of relation (except synonyms and hierarchical relations). Associative relations are undirected and reversible.

Classical Methods of Knowledge Representation and Supported Relations

In information science and documentation, the recently most important methods of knowledge representation comprise nomenclatures (*Schlagwortmethode)*, classifications and thesauri (Bertram 2005), (Lancaster 2003). These indexing languages are built, maintained and applied by information professionals as well as by trained domain experts.

A *nomenclature* (Stock & Stock 2008) is a rather simple form of knowledge representation which works with controlled keywords extracted from natural language. The annotation or representation of knowledge follows specific rules, as for example the *Regeln für den Schlagwortkatalog* (RSWK) (Umlauf 1999) for use in German libraries. This method clearly focuses on capturing synonyms and quasi-synonyms for controlling the vocabulary. Hierarchical constructions are not indicated in this method (otherwise the emerging set of keywords would form a thesaurus).

Classifications consist of non-verbal notations which represent concepts and relations between them. The aim is to organize the whole knowledge of a database in a uniform and language-independent way and to display it in a proper structure. Classifications are structured hierarchically, without further distinguishing different types of hierarchies (DIN 32705 1987).

Thesauri are built of controlled terms extracted from natural language, sometimes called descriptors. They pay much attention to the relation of equivalence, which results in a collection of synonyms as non-descriptors. Furthermore, thesauri split up the hierarchical relation into hyponymy and meronymy and make use of (entirely undifferentiated) associative relations (Aitchison et al. 2000).

Paradigmatic relations of classifications and thesauri are even bindingly regulated by national as well as international norms (amongst others (DIN 1463/1 1987)).

Figure 1 showes the mentioned knowledge representation methods and the relations in use. Moreover, it explains their classification due to their complexity in structuring relationships between concepts.

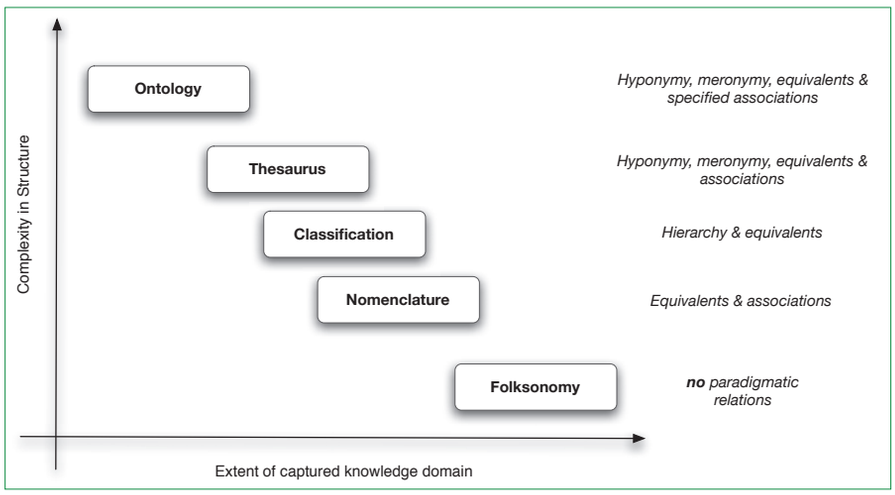


Figure 1: Classification of popular knowledge representation systems according to complexity in relational constructions. Source: (Weller & Peters 2007).

Additional Relations in Ontologies

During the last 15 years, ontologies have been a research focus in computer science and information science as a new knowledge representation method (although some confusion arises as a couple of authors use the term "ontology" as a synonym for KOS in general). They are designed for semantic indexing and information integration (Stuckenschmid & van Harmelen 2005). Elaborated ontology languages such as OWL (W3C 2004) support the use of as many as needed types of relations between concepts. Most ontologies make use of hyponymy and meronymy, as in a thesaurus. These relationships are often labelled *is_a* (or ISA, isa) and *part_of*, but may also be named differently, e.g. *subclass_of*. There are no naming conventions or other common construction guidelines yet; discussions have just begun (Schulz et al. 2006). In addition, associative relations can be split into a diversity of domain-dependent, specified paradigmatic relations. They may be very narrowly defined, like *has_ingredient* and reversely *is_ingredient_of* in a food ontology, or *develops_from* for gene developments in the Gene Ontology (Gene Ontology Consortium 2000).

In some regard, ontologies transcend the traditional approach of KOS as means for document indexing. Ontologies may intend to capture a domain of interest with all its aspects, which is more than providing a controlled vocabulary for indexing and retrieval purposes. They themselves become an independent knowledge base. Fur this purpose, the use of detailed self-defined relations is necessary. Besides relations to interlink different concepts within the ontology, other types of relations may be defined to add fixed values as properties for a certain concept.

Some hundred (narrow) domain ontologies are currently used experimentally; some are also already practically applied (e.g. the Open Biomedical Ontologies, OBO[1]). A challenging research task will now be to investigate them regarding the relationships in use. This may lead to new views on how relationships can be distinguished, which might also be of use for other knowledge representation methods. On the other hand the community of ontology designers needs further evaluation and studies on advantages and disadvantages of certain relational constructions. Often knowledge engineers have to decide on whether they should establish hierarchical structures or rather describe concepts by self-defined relations; a question which is difficult to handle without guidance and instructions. We think that some guidelines can be derived that provide help for this modeling conflict, after analyzing semantic relations in more detail, as also proposed e.g. by (Hovy 2002).

Syntagmatic Relations in Folksonomies

During the last two years we have been experiencing a boom of *collaborative services* (or *social software*) within the so-called *Web 2.0* (Gordon-Murnane 2006), (O'Reilly 2005). Web-users become authors and publish their own content (e.g. bookmarks, photos, videos) via special platforms on the WWW (e.g. Del.icio.us, Flickr, or YouTube). Furthermore, users are also taking over the indexing of online content. This is referred to as *social*

1 http://obo.sourceforge.net.

tagging; the users index documents with freely chosen tags. The totality of applied tags within one system is called *folksonomy*. There are no explicit rules or guidelines on how to annotate documents in social tagging applications; the basic assumption is that collective intelligence will ensure satisfying knowledge representation (Weiss 2005), (Quintarelli 2005), (Mathes 2004), (Hammond et al. 2005). It is known that about 90% of applied tags are nouns (Guy & Tonkin 2006), which resembles the preference for nouns to be used in controlled vocabularies.

As there are no structures and formats provided in folksonomies, all relations between the tags are by definition syntagmatic ones, i.e. they implicitly exist in the co-occurrence of tags. Relationships between tags are not labeled at all; they are not even discussed by the users. And still we can assume that a user does keep the idea of certain relationships in mind during the process of tagging content. For example, a certain picture on the platform www.flickr.com might be tagged with "Düsseldorf, Germany, Altstadt" another with "Brasil, Brazil" or "countryside, landscape", which implicitly includes a part-of relation in the first case and kinds of synonymy in the other cases.

Social tagging provides empirical data on the user's (not the expert's) language use in terms of knowledge representation. This may also reflect the users' perception of relations in practical indexing contexts. Thus, folksonomies implicitly include data on how non-experts compose content-descriptive tags, which should be evaluated to gain a better understanding of the underlying ideas. A lot of research in this direction has been made, especially regarding the automatic derivation of ontologies (Schmitz 2006) or of hierarchies (Fan et al. 2007), (Heymann & Garcia-Molina 2006), (Wang et al. 2006) from tags, or the extraction of specific semantic information like "place" or "event" (Rattenbury et al. 2007). The underlying approach of this automatic creation of taxonomies or controlled vocabularies is based on statistical evaluations of the co-occurrence of tags. In contrast, we are particularly interested in the kinds of relations that can be intellectually identified within folksonomies and in how they can be transferred into general paradigmatic relations. Stock & Stock (2008) emphasize the need for intellectual interpretation of syntagmatic relationships as well: „Diese Interpretation des jeweiligen Zusammenhangs und damit der Übergang von der syntagmatischen zur paradigmatischen Relation bedarf der intellektuellen Arbeit der Experten. Sie entscheiden, ob Äquivalenz-, Hierarchie- oder Assoziationsrelation vorliegen" (Stock & Stock 2008, p. 251).

Research Goal: Subdividing Associative and Hierarchical Relations

As already mentioned, ontologists have started to decompose what was previously summed up as associative relations in thesauri. By now, there are few structured attempts to collect and classify the new, specific relations in use. Some suggestions on that have already been made but were not adapted yet. Dextre Clarke, (2001, p. 48) proposes following specifications of associative relations:

- terms with overlapping meanings: *ships – boats*,
- discipline – phenomena: *seismology – earthquakes*,
- process – instrument: *velocity measurement – speedometers*,
- occupation – person in that occupation: *accountancy – accountant*,
- action – product of the action: *road-making – roads*,
- action – its patient: *teaching – students*,
- concept – its origins: *water – water wells*,
- causal dependence: *erosion – wear*,
- thing / action – counter-agent: *pests – pesticides*,
- raw material – product: *hides – leather*,
- action – property associated with it: *precision measurement – accuracy*,
- concept – opposite: *tolerance – prejudice.*

Schmitz-Esser (2000, p. 79) differentiates usefulness (*job creation – economic development*) and harmfulness (*over-fertilisation – diversity of species*) within the associative relations whereas the Standard Thesaurus Wirtschaft (Stock 1999) makes use of relations of contextual connection (*body care – soap*) and of related branch (*coal – mining*) in economic contexts.
But not only associative relations can be specified, hierarchical relations have also been a subject to similar investigations (Winston et al. 1987), (Cruse 2002). Especially the subdivisions of meronymic relations have been explored and have been considered regarding problems with transitivity and possible applications in information retrieval, e.g. query expansion (for details see (Weller & Stock 2008)).
Further refinements and differentiations of associative relations should now be investigated by using existing ontologies and folksonomies as resources. Our aim is to find new paradigmatic generalizable relations – at least for certain domains – which could be a basis for ontology engineering or for consistent construction guidelines. Furthermore, generalizable paradigmatic relations can enhance traditional knowledge representation methods regarding complexity and consistency and they can provide a fine-grained field structure for information indexing and information retrieval.

Analysis of Relations within Ontologies and Folksonomies

In this section we will discuss single examples from a case study with chosen ontologies and folksonomies.

Figure 2 to figure 4 show how relations of different types are used in ontologies of travel, tools and virtual humans.
The Travel Ontology (fig. 2) links concepts by using several relations. It aims at structuring the domain of "travel" for a practical application, e.g. as basis for a travel planning system. It makes use of traditional relations like hyponymy, e.g. in "Hotel" *is_a* "Accomodation", and meronymy, e.g. in "Destination" *has_part* "Beach". Indirect part-of-relations can be found in "Destination" *has_accomodation* "Accomodation". An "Accommodation" obviously is a certain part of a "Destination" – alternatively we may deduce that an accommodation is placed in a certain location, but that is not stated directly in this case. Quite similarly is the relation in "Destination" *has_activity* "Activity",

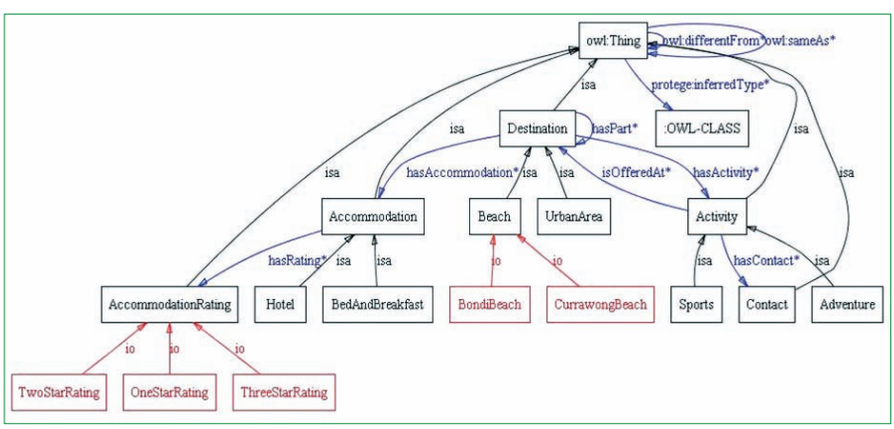


Figure 2: Travel Ontology, an excerpt displayed with Protégé Ontoviz plug-in.
Source: http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library by Holger Knublauch.

59(2008)2, 100-107 Information Wissenschaft & Praxis

which states that in some way an activity is part of a certain destination, and the reverse "Activity" *is_offered_at* "Destination" captures the information that "Activity" is located at a certain destination. We may thus regard these two relations as specific occurrences of meronymy with a spatial reference. As soon as individual examples for destinations are added and interlinked with individual activities or accommodations, the ontology will become an independent knowledge base.

A somehow new relation type of responsibility is provided in "Activity" *has_contact* "Contact". Yet, the included aspect of responsibility (some contact person is responsible for providing a certain activity) can only be deduced intellectually. On the formal level the ontology only indicates, that some kind of relationship exists between the concepts "Activity" and "Contact".

The Tools Ontology in figure 3 reflects information on software tools and therefore aims at capturing the basic metadata to describe such tools. We can see that most of the relations are bound to the central concept "Software Tool". The domain of software tools is not represented in a hierarchy of such tools (which would also be possible, if we would start with a concept "Software Tools" and then refine them by more specific subclasses like "Visualization Tools" and "Statistical Tools"). An alternative approach was chosen. All relevant properties applying to a software tool are added directly to this concept, which will fulfill the function of describing and structuring existing tools according to different characteristics. This is particularly useful if different dimensions would be needed to describe and classify subclasses of software tools (e.g. formats and functions) at the same time. What is not shown in the figure is that a hierarchical structure is used on another level to represent different possible types of tool functions and different types of data types ("Shape Types") which can be processed by a tool.

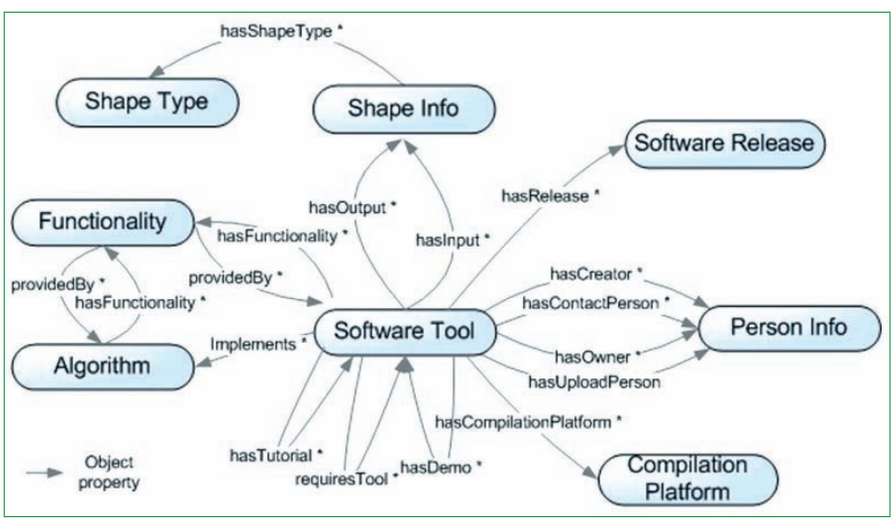


Figure 3: Common Tools Ontology. Source: AIM@Shape ontologies, http://dsw.aimatshape.net/tutorials/cto.jsp.

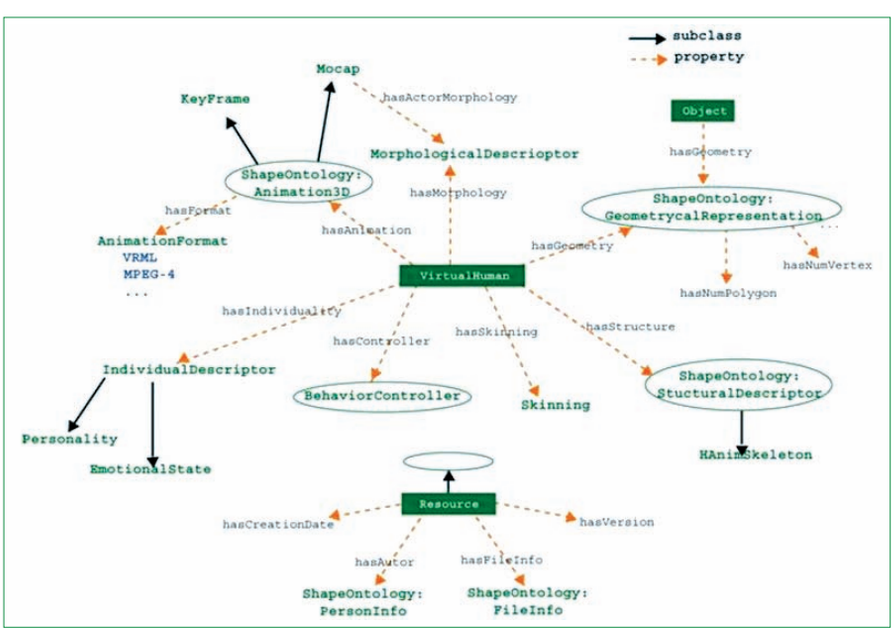


Figure 4: Virtual Human Ontology. Source: AIM@Shape ontologies, http://vrlab.epfl.ch/~alegarcia/VHOntology/long.html.

Figure 4 presents the relations used in the Virtual Human Ontology. It makes use of relations of form or format, e.g. "Animation3D" *has_format* "AnimationFormat", "VirtualHuman" *has_animation* "Animation3D", or "VirtualHuman" *has_structure* "StructuralDescriptor", and relations of shape as in "VirtualHuman" *has_skinning* "Skinning" or "VirtualHuman" *has_geometry* "GeometricalRepresentation". Relations with temporal regards are given in "Resource" *has_creation* date and "Resource" *has_version* while the connection between "Resource" *has_Autor* "PersonInfo" describes a possessive relation of author and work. The link of "VirtualHuman" and "IndividualDescriptor" is expressed with *has_individuality* and shows a relation of attributes; "Resource" and "FileInfo" are connected with *has_FileInfo* and attach additional information in this way. A relation of impact is given in "VirtualHuman" *has_controller* "BehaviorController".

As we have now seen in the three examples, very specific relations can be made explicit within ontologies. Folksonomies provide only syntagmatic relations between tags, specified semantic relations have to be deduced completely intellectually. This means, one should use ac-

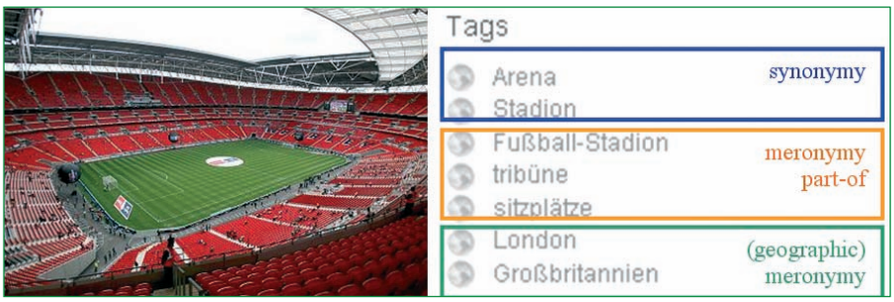


Figure 5: Hidden paradigmatic relations in folksonomies. Source: www.flickr.com.

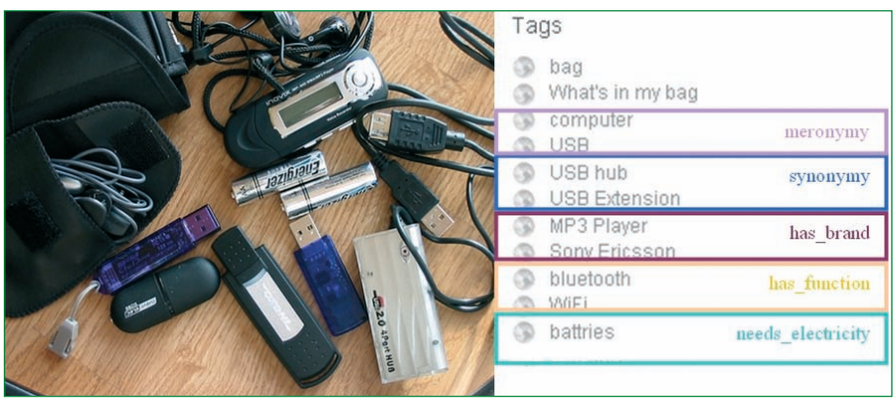


Figure 6: Hidden paradigmatic relations in folksonomies. Source: www.flickr.com.

Figure 7: Hidden paradigmatic relations in folksonomies. Source: www.flickr.com.

tual co-occurrences of tags as indicators for existing relations. Figure 5 to figure 7 present several examples of co-occurrences of tags – hence their several syntagmatic relations. We have used them exemplarily to demonstrate, which kinds of relations the folksonomy users may have intended to represent.

In figure 5 we identify known paradigmatic relations such as meronymy and synonymy: e.g. "sitzplätze" *is_part_of* "tribüne" and "tribüne" *is_part_of* "Fußball-Stadion"; "London" *is_part_of* "Großbritannien" indicates a part-of relation which can be described as geographic meronymy; "Arena" is a *(quasi-) synonym* for "Stadion". These relations are not made explicit by the content creator during the tagging process, but afterwards they are inherently hidden in the syntagmatic relations of tags.

Relations of equivalence and hierarchy can be found in figure 6 as well: "USB" *is_part_of* "computer". "USB hub" can be viewed as a sibling of "USB Extension", both are subcategories for a class "USB devices" (which is not included here). The content creator annotated some more tags which are linked with each other in different kinds of relations. We can find a relation of ownership in "MP3 Player" *has_brand* "Sony Ericsson" and relations of function in "computer" *has_function* "bluetooth" and "computer" *has_function* "WiFi". A relation of requirement is given in "MP3 Player" *needs_electricity* "batt(e)ries".

In figure 7 the co-occurrence of tags represent relations of location, e.g. in "cloud gate" *is_located_in* "Millennium Park" or "Millennium Park" *is_located_in* "chicago". A relation of production/ fabrication can be found in "sculpture" *is_made_of* "metal". There is an attributive relation regarding the sculpture's format, e.g. "sculpture" *has_shape* "bean". Standard relations like synonymy ("vacation", "trip" and "travel") and like both relations of hierarchy, hyponymy ("cloud gate" *is_a* "sculpture") and meronymy ("chicago" *is_part_of* "america") can be found as well.

Findings

So far we have analyzed random samples of both folksonomies and ontologies, more comprehensive studies will follow. Yet, these first examinations already showed some interesting tendencies in relation-modelling within ontologies and in tagging behaviour in folksonomies.
We have seen that both ontologies and folksonomies include different types of semantic relations, among them standard relations as well as new relationship types. Within folksonomies, we did not expect users' awareness on the principles of semantic relations, but hoped that the choice of tags would represent a certain relational perspective of real-world structures. This was confirmed. On the other hand, we expected ontology designers to create sophisticated and well-founded relations between concepts. This could not be completely confirmed. Although we find different explicit relations, these relations in use are often inconsistent, of varying specificity and without theoretical foundations. They are not expressions of detailed analysis of real-world correlations, but are rather displaying ad-hoc approaches or spontaneous ideas to picture a domain of interest. At least for the majority of small or application-specific ontologies the analysis of relationship types resembles to some degree the analysis of folksonomies: the creators of these ontologies are free to do what they want, no guidelines are provided, no knowledge on principles of semantic relations or background in knowledge representation is required. The resulting concept interrelations will therefore have to be subject to intellectual analysis before new insights to relationship types can be gained. Still we expect large and renowned ontology projects (e.g. the Cyc Ontology[2]) to provide relationship types of different (higher) quality.

In the beginning of our studies, we particularly assumed that ontologies would perform the specification of so far undifferentiated associative relations. We now encountered that associations are to some degree specified, but not as explicitly as expected.

2 www.cyc.com.

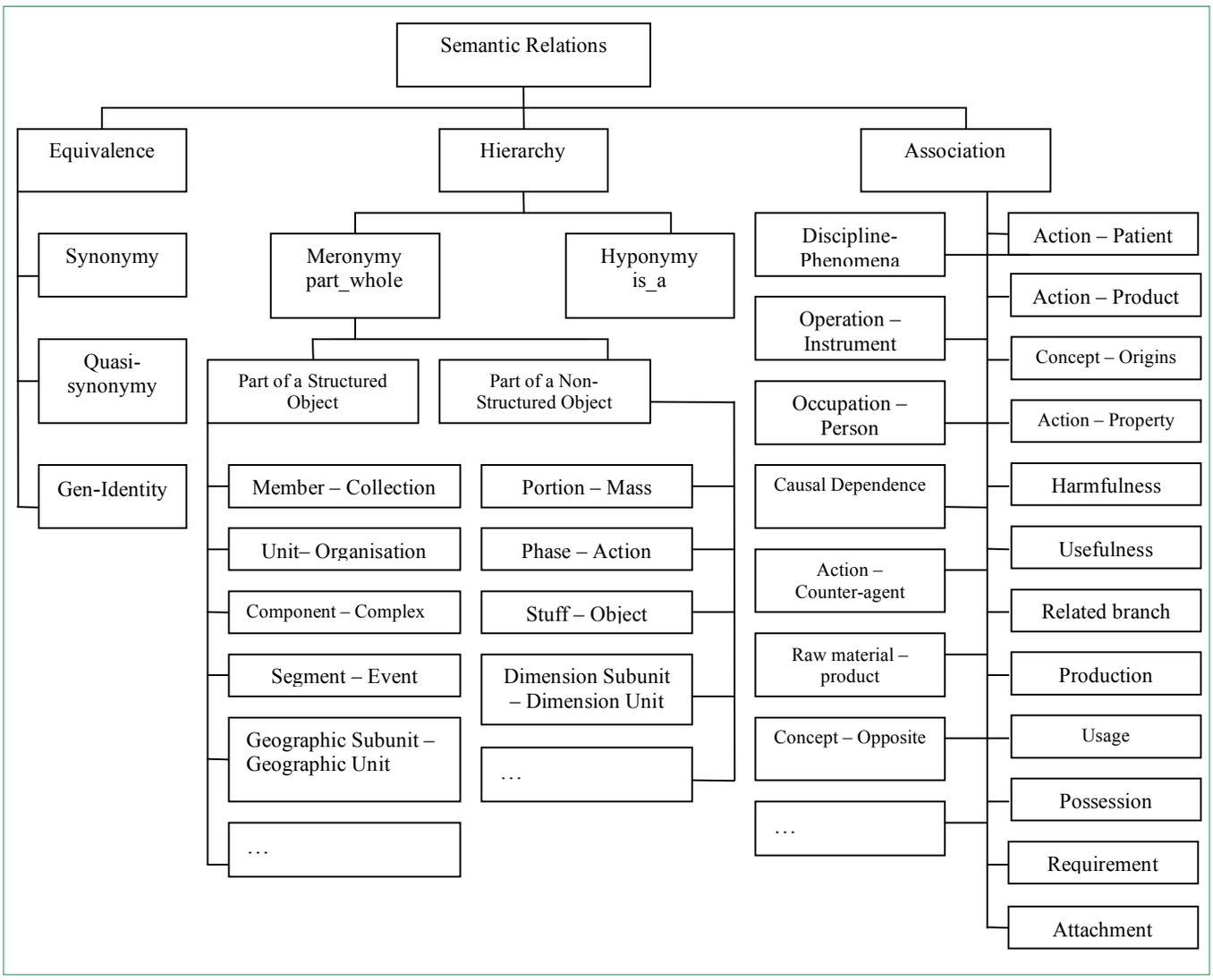


Figure 8: Semantic relations and their specifications.

One major problem of current ontologies is, that they include relations which seem to be specified associations on first sight, but which do not provide any additional information if compared to traditional unspecified associations. They interrelate concepts without describing the nature of the relationship. Within the Travel Ontology we find the statement "Destination" *has_accomodation* "Accomodation". This does only indicate that there is some kind of relation between destination and accommodation, but does not explain at all the nature of this relation. Relations of this kind can be found frequently, probably due to lack of consciousness in the community. They are not more specific then general associations. They are unsubstantial, and may thus even dilute the expressiveness of ontologies. The underlying intention behind these relations will have to be captured intellectually to refine the spectrum of useful semantic relations (still it may not always be possible to capture the underlying ideas).

Other considered relations were used to describe the entitative characteristics or features of an object, e.g. *is_located_in*, *is_made_of* or *has_format*. Such aspects of physical characteristics are particularly common, but we also encounter other characteristics, for example with temporal references (e.g. *has_release*). We further may regard certain forms of meronymy as relations that describe the nature of an object, in this case by naming its characteristic and constituent parts. We have seen the case of an "Algorithm" which is implemented in a "Software Tool" and thus becomes an elementary part of this tool (fig. 3).

Although there is no strict line, we suggest to distinguish these types of attribute-specifying relations from concept-interrelating ones. The former are mainly needed for precise descriptions of individuals within KOS, e.g. to state that an individual "car" has a certain "colour" (without paying much attention to the nature of the concept "colour" itself). Relations of the latter kind connect different concepts and should specify the relationship between them, e.g. to state that the nature of the relationship between "automotive manufacturer" and "car" is that of producer – product. We have seen some examples for this, e.g. "Software Tool" *has_creator* "Person". Others regard for example aspects of ownership (*has_owner*) or conditions (*needs_electricity*).

Based on existing studies (Storey 1993), (Cruse 2002), (Schmitz-Esser 2000), (Dextre Clarke 2001), (Winston et al. 1987), the examples above and further examples on specifications of semantic relations we have created a systematization of relations, which should be the basis for future discussions and refinements (fig. 8), with the final aim of a general theory of relations.

Conclusions and Future Work

The advent of ontologies and folksonomies asks for reconsidering the types of relations in knowledge representation from a practical point of view. Differentiated relations that can be explicitly formalized in ontologies or that may be

inherently hidden in folksonomies seem to be a key factor on the way towards a Semantic Web.

We believe in the necessity of reconsidering relationships regarding their classification and generalizability for the use in knowledge representation and information retrieval. New methods will have to be developed for structuring relationship types and identifying generalizable ones. Theoretical studies on semantics of relations are available and may be used as starting points. We have shown that the investigation of knowledge relationships in two complementing empirical approaches a) by collecting and classifying existing relations used in ontologies and b) by empirically analyzing user-created tags in folksonomies gives valuable results in terms of regarding real-used language and concept linking. What is more, the analysis showed that guidelines for ontology (relation) engineering and tagging are of great importance to enable more precise and suitable knowledge representation.

The currently suggested system of relationships will be refined. One critical point therein will be to also examine the properties of relations in more detail. Features such as transitivity (Johansson 2004), reflexivity and symmetry can be used to describe and to dissociate different types of relations.

Acknowledgements

We thank our colleagues at the Heinrich-Heine-Universität for inspirations and support, particularly Wolfgang G. Stock for his fruitful suggestions. Katrin Weller is financially supported by the Bundesministerium für Bildung und Forschung (BMBF), Bonn, of the Federal Republic of Germany within the research project *Ontoverse*.

References

Aitchison, J., *Gilchrist*, A. & *Bowden*, D. (2000). Thesaurus Construction and Use. London: Fitzroy Dearborn.

Bean, C.A. & *Green*, R. (Eds.) (2001). Relationships in the Organization of Knowledge. Dordrecht: Kluwer.

Bertram, J. (2005). Einführung in die inhaltliche Erschließung. Grundlagen, Methoden, Instrumente. Würzburg: Ergon.

Cruse, D.A. (2002). Hyponymy and its varieties. In: *Green*, R., *Bean*, C.A., & *Myaeng*, S.H., (Eds.), The Semantics of Relationships (pp. 3-21). Dordrecht: Kluwer.

Dextre Clarke, S.G. (2001). Thesaural relationships. In: *Bean*, C.A. & *Green*, R. (Eds.). Relationships in the Organization of Knowledge (pp. 37-52). Boston, Kluwer.

DIN 1463/1 (1987). Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri. Berlin: Beuth.

DIN 32705 (1987). Klassifikationssysteme. Erstellung und Weiterentwicklung von Klassifikationssystemen. Berlin: Beuth.

Fan, J., Gao, Y. & Luo, H. (2007). Hierarchical classification for automatic image annotation. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 111-118). New York: ACM.

Gerstl, P. & *Pribbenow*, S. (1996). A conceptual theory of part-whole relations and its applications. Data & Knowledge Engineering, 20, 305-322.

Gene Ontology Consortium (2000). Gene Ontology. Tool for the Unification of Biology. Nature Genetics, 25, 25-29.

Green, R., *Bean*, C.A. & *Myaeng*, S.H. (Eds.) (2002). The Semantics of Relationships. Dordrecht: Kluwer.

Gordon-Murnane, L. (2006). Social bookmarking, folksonomies, and Web 2.0 tools. Searcher - The Magazine for Database Professionals, 14(6), 26-38.

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

Hammond, T. et al. (2005). Social Bookmarking Tools. A General Review. Part 1. D-Lib Magazine, 11(2005)4.

Heymann, P. & *Garcia-Molina*, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. In: InfoLab Technical Report 2006-10,Department of Computer Science, Stanford University, Stanford, CA, USA.

Hovy, E. (2002). Comparing Sets of Semantic Relations in Ontologies. In: *Green*, R., *Bean*, C.A. & *Myaeng*, S.H., (Eds.). The Semantics of Relationships (pp. 91-110). Dordrecht: Kluwer.

Lancaster, F. (2003). Indexing and Abstracting in Theory and Practice. 3rd ed, Champaign: University of Illinois.

Löbner, S. (2002). Understanding Semantics. London: Edward Arnold Publishers.

Johansson, I. (2004). On the transitivity of the parthood relations. In: *Hochberg*, H., & *Mulligan*, K., (Eds.). Relations and Predicates (pp. 161-181). Frankfurt am Main: Ontos.

Khoo, C.S.G. & *Na*, J.C. (2006). Semantic relations in information science. Annual Review of Information Science and Technology 40(2006), 157-228.

Mathes, A. (2004). Folksonomies – Cooperative Classification and Communication Through Shared Metadata. Online: http://adammathes.com/academic/computer-mediated-communication/folksonomies.html [10.01.2008].

O'Reilly, T. (2005). What is Web 2.0. Design patterns and business models for the next generation of software. www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [10.01.2008].

Peters, I. & *Stock*, W.G. (2008). Folksonomies in Wissensrepräsentation und Information Retrieval. Information – Wissenschaft & Praxis, 59(2008)2, 77-90.

Peters, I. & *Stock*, W.G. (2007). Folksonomies and information retrieval. In: Joining Research and Practice: Social Computing and Information Science. Proceedings of the 70th ASIS&T Annual Meeting, 44 (pp. 1510-1542).

Pribbenow, S. (2002). From classical mereology to complex part-whole-relations. In: *Green*, R., *Bean*, C.A., & *Myaeng*, S.H. (Eds.). The Semantics of Relationships (pp. 35-50). Dordrecht: Kluwer.

Quintarelli, E. (2005). Folksonomies: power to the people. Paper presented at the ISKO Italy-UniMIB meeting. www.iskoi.org/doc/folksonomies.htm [10.01.2008].

Rattenbury, T., *Good*, N. & *Naaman*, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 103-110). New York: ACM.

Schmitz-Esser, W. (2000). EXPO-INFO 2000. Visuelles Besucherinformationssystem für Weltausstellungen. Berlin [u.a.]: Springer.

Schmitz, P. (2006). Inducing ontology from flickr tags. In: Proceedings of the Workshop on Collaborative Web Tagging at WWW2006 Edinburgh, Scottland, May 22-26.

Schulz, S., *Kumar*, A. & *Bittner*, T. (2006). Biomedical ontologies: What part-of is and isn't. Journal of Biomedical Informatics 39(2006)3, 350-361.

Stock W.G. & *Stock*, M. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München. Wien: Oldenbourg.

Stock W.G. (2007). Information Retrieval. Informationen suchen und finden. München; Wien: Oldenbourg.

Stock, M. (1999). Standard-Thesaurus Wirtschaft. Ein neuer Standard der Wirtschaftsinformation? Password 14(1999)1, 22-29.

Storey, V.C. (1993). Understanding semantic relationships. VLDB Journal 2, 455-488.

Stuckenschmidt, H. & *van Harmelen*, F. (2005). Information Sharing on the Semantic Web. Berlin, Heidelberg: Springer.

Sun, X. & *Zheng*, Q. (2005). An approach to acquire semantic relationships between terms. In: Proceedings of the 2005 ACM Symposium on Applied Computing (pp. 1630-1633). New York: ACM.

Umlauf, K. (1999). Regeln für den Schlagwortkatalog. Die Grundregeln der RSWK. Berlin: Berliner Handreichungen zur Bibliothekswissenschaft; 66.

W3C Recommendation (2004). OWL Web Ontology Language Overview. www.w3.org/2004/OWL/ [10.01.2008].

Wang, T. et al. (2006). Automatic extraction of hierarchical relations from text. Lecture Notes of Computer Science, 4011, 215-229.

Weiss, A. (2005). The Power of Collective Intelligence. netWorker, 9 (3), 16-23.

Weller, K. & *Stock*, W.G. (2008). Transitive meronymy. Automatic concept-based query expansion using weighted transitive part-whole relations. Information – Wissenschaft & Praxis, 59(2008)3 (submitted).

Weller, K. & *Peters*, I. (2007). Reconsidering relationships for knowledge representation. In: Proceedings of I-Know ,07, Graz, September 5-7 (pp. 501-504).

Winston, M.E., *Chaffin*, R. & *Herrmann*, D. (1987). A taxonomy of part-whole-relations. Cognitive Science 11(1987), 417-444.

Zeng, M.L. (2000). Taxonomy of knowledge organization sources / systems. http://nkos.slis.kent.edu/KOS_taxonomy.htm [10.01.2008].

Relation, paradigmatic relation, syntagmatic relation, knowledge organization system, equivalence, hierarchy, generalizable relation, association, ontology, folksonomy

Relation, Hierarchie, Ontologie, Wissensrepräsentation

The Authors

Isabella Peters, M.A.

is researcher at the Heinrich-Heine-University, Düsseldorf, and works in the Department of Information Science. Her main research focuses the topic Web 2.0, especially the use and characteristics of folksonomies in terms of indexing and information retrieval and the application of blogs and podcasts in corporate knowledge management.
isabella.peters@uni-duesseldorf.de

Katrin Weller, M.A.

is researcher at the Heinrich-Heine-University, Düsseldorf, and works in the Department of Information Science. Currently she is involved in the research project Ontoverse, where she is concerned with Ontologies as new forms of knowledge organization systems, knowledge engineering processes, and collaborative ontology management.
weller@uni-duesseldorf.de

Facettierte Wissensordnungen und dynamisches Klassieren als Hilfsmittel der Erforschung des Dark Web

Silke Heesemann und Hans-Dieter Nellißen, Düsseldorf

Die Identifizierung und Überwachung terroristischer und/oder extremistischer Aktivitäten und Planungen im Internet, speziell im Dark Web, rücken immer mehr in den Fokus rechtsstaatlicher Ermittlungen. Die Menge an relevanten Informationen erfordert eine Fokussierung auf einen bestimmten Jargon, mit dem verdächtige Entwicklungen erkannt werden. Auf dieser Basis wird ein Analysesystem aufgebaut. Der Versuch eines optimalen Retrieval-Ansatzes zu dieser Thematik wird im Folgenden dargestellt.

Faceted Knowledge Organization Systems and dynamic classification as tools for the research of the Dark Web

The identification and observation of terrorist and/or extremist activities and plans on the internet, especially on the Dark Web, are moving more and more into the centre of inquiry. The amount of relevant information requires to focus on a specific jargon, with the help of which suspicious developments can be detected and an analysis system can be built up. An attempt at an optimal information retrieval approach to this issue is depicted below.

Einleitung

Anfang Mai 2007 wurde das europäische Informationsportal „Check the Web“ bei Europol freigegeben. Im Fokus steht die „Beobachtung und Auswertung des Internets zur Bekämpfung des internationalen Terrorismus“ (BMI, 2007). Das Herausfiltern von zur Terrorismusbekämpfung relevanten Informationen aus dem Internet stellt eine Herausforderung für die Ermittlungsbehörden dar.

Many Internet pages in various languages have to be monitored and evaluated, which requires enormous technical and human resources. Due to the huge quantity of Internet pages in use, problems arise on a national and international level concerning the quantity and quality of resources, especially with a view to the language skills needed (EU Council Secretariat 2007).

Im Dezember 2007 belief sich die Anzahl von Websites im WWW auf über 155 Millionen, mit einem Zuwachs von 50 Millionen seit Dezember 2006, dies entspricht einem prozentualen Anstieg von 48 Prozent (Netcraft Web Server Survey December 2007). Das Interesse der Sicherheitsbehörden und Nachrichtendienste konzentriert sich dabei zusätzlich auf die Überwachung nicht direkt zugänglicher Informationen im Web, so wie Emails, Chatrooms und weiterer Kommunikationswege, die von einem klassischen Crawler nicht abdeckbar sind. Wir befinden uns an dieser Stelle teilweise im Deep Web (Bergman 2001), auch Hidden Web oder Invisible Web genannt (Sherman & Price 2001), das heißt bei Web-Content, der nicht durch Hyperlinks erreichbar ist, wie z.B. Datensätze in Datenbanken. Das Deep Web soll laut einer Studie aus dem Jahr 2000 400- bis 550-mal größer sein als das Oberflächenweb (Bergman 2001). Auch wenn die Zahlen dieser Studie stark angezweifelt werden (Lewandowski 2005), kann man doch von einer großen Menge von Seiten im Deep Web ausgehen. Der Begriff *Dark Web* wird im allgemeinen Sprachgebrauch entweder missverständlich als Synonym für Deep Web verwendet oder er bezeichnet Webseiten mit terroristischem, extremistischem oder anderem kriminellen Inhalt, ob nun im Oberflächen- oder im Deep Web. Unser Thema ist das Web der Terroristen. Zur Überwachung und Informationsgewinnung aus dem Deep Web sind spezielle Techniken nötig, um diese Seiten für einen Crawler überhaupt erreichbar zu machen, da eine einfache Linkverfolgung hier nicht greift. Aufgrund der Komplexität dieses Themas kann an dieser Stelle nicht weiter darauf eingegangen werden.

Die aus Oberflächenweb und Deep Web zusammengesetzte enorme Datenmenge stellt zunächst eine Herausforderung für das Crawlen nach zur Terrorismusbekämpfung relevanten Inhalten dar. Die aus dem Crawlvorgang gewonnene Datenmenge bildet dann die Basis für das Herausfiltern der dort befindlichen Informationen. Es ist daher von Vorteil, ein themenspezifisches, fokussiertes Crawling durchzuführen (Kwiatkowski & Höhfeld 2007) und den Inhalt ebenso thematisch fokussiert durchsuchbar zu machen, um Informationsballast zu vermeiden und qualitativ hochwertige Informationen zu gewinnen. Hierzu benötigt man eine Wissensordnung, die das Focused Crawling und das Retrieval steuert. Doch unabhängig davon wo man nach brisantem Inhalt oder Informationen sucht, wichtig ist die Aufarbeitung des gewonnen Materials in einem Retrievalsystem, welches das Suchen und Finden relevanter Informationen ermöglicht.

Die Untersuchung des Webs konzentriert sich bisher hauptsächlich auf die Aufdeckung und Visualisierung von sozialen Netzwerken, das heißt von Verbindungen zwischen terroristischen Organisationen und deren Mitgliedern sowie deren Identifikation.

Ein Forschungsteam um Hsinchun Chen, Professor of Management Information Systems an der Universität von Arizona, gilt in der wissenschaftlichen Forschung zu diesen Themen momentan als führend (AI Lab 2007). Unter der Leitung von Chen wurde z.B. die Software Coplink entwickelt, welche soziale Netze visualisiert und Ermittlungsbehörden bei der Fahndung nach Straftätern hilft (Chen et al. 2003). Ein weiteres Programm namens Writeprint ermöglicht die Identifizierung anonymer Autoren mit Hilfe von Mustererkennung des Schreibstils und inhaltlicher Analysen, wie die Fokussierung auf Themen und häufig verwendete Schlüsselbegriffe (Li, Zheng & Chen 2006). Writeprint ist Teil des „Dark Web Project“ (Chen et al. 2005). Dieses hat zur Aufgabe, Webseiten extremistischen Ursprungs und Inhalts zu sammeln und auszuwerten (Quin et al. 2005). Hierfür wurden automatisierte Systeme entwickelt:

59(2008)1, 108-117 Information Wissenschaft & Praxis

> To address this research gap, we explore an integrated approach for identifying and collecting terrorist/ extremist Web contents. We also propose a Dark Web Attribute System (DWAS) to enable quantitative Dark Web content analysis from three perspectives: technical sophistication, content richness, and Web interactivity (Qin et al. 2007).

Neben der Aufdeckung sozialer Netze wird also auch die technische und inhaltliche Entwicklung extremistischer Webseiten erforscht. Das Attribute System basiert auf den Untersuchungen Demchaks der ähnliche Untersuchungen für Regierungs-Webseiten unternommen hat (Demchak 2000; Zhou 2005). Demchaks System wurde unter anderem um die Attribute Propaganda und Ideologie erweitert; es werden aber nur Cluster von Webseiten identifiziert, in denen entsprechende Begriffe auftreten (Reid et al. 2005). Der Fokus dieser Untersuchungen liegt folglich nicht in der inhaltlichen Erschließung der gewonnenen Dokumente. Dieser Punkt der inhaltlichen Analyse wurde bisher selten beachtet. Die steigende Anzahl von Dokumenten, die gewonnen wird, muss von den Ermittlungsbehörden aber auch inhaltlich effizient durchsuchbar gemacht werden. Die Datensammlung des „Dark Web Project“ beinhaltet 500 Millionen Dokumente (Webseiten, Postings, Videos etc..) von 10.000 Websites; eine Datengröße von insgesamt zwei Terabyte (AI Lab, 2007). Das zugehörige „Dark Web Research Portal“ bietet zwar eine Suchoberfläche an, mit welcher nach Schlüsselwörtern im Datenbestand gesucht werden kann, allerdings liegt hier keine komplexe Ontologie zugrunde (Reid & Chen 2006). Die Möglichkeiten, die das Information Retrieval durch den Einsatz semantischer Netze in Verbindung mit facettierten Begriffsordnungen bietet, werden nicht genutzt. Gerade hier liegt ein Schwerpunkt des Instituts für Sprache und Information, Abteilung Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf.

Im Rahmen eines Projekts der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf wurde ein exemplarischer Thesaurus über den Terror in Nordirland erstellt und mit Hilfe des Retrievalsystems RetrievalWare der Firma Convera realisiert (RetrievalWare wurde mittlerweile vom Wettbewerber FAST übernommen).

Mit Hilfe facettierter Thesauri und dynamischer Klassierung, wie sie die Software RetrievalWare ermöglicht, können Wissensdomänen unter Einsatz eines semantischen Netzes abgebildet und Datenbasen unterschiedlicher Formate und Sprachen nach Informationen durchsucht werden. Mit dem in RetrievalWare integrierten Tool Knowledge Workbench kann man ein kontrolliertes Vokabular in Form einer Taxonomie erstellen, die dann durch Begriffserweiterungen in Form von Relationen ein semantisches Netz bildet. Die Software ermöglicht die Erstellung eines facettierten, multilingualen Thesaurus, auf dessen Grundlage eine dynamische Klassierung der Dokumente aus der Datenbasis erfolgt und dann dem Nutzer in der Suchoberfläche vielfältige Recherchemöglichkeiten bietet. Dreh- und Angelpunkt dieses Retrievalsystems ist der Thesaurus oder besser gesagt, sind die Thesauri.

Facettierte Wissensordnungen

Der Einsatz von Facetten zur Inhaltserschließung ist nicht neu (Stock & Stock 2008). Bereits 1930 stellt Ranganathan sein System facettierter Klassifikation vor (Ranganathan 1933). Die von ihm entwickelte Colon-Classification ordnet Themen in die Facetten Personality (Individualität), Matter (Material), Energy (Energie), Space (Raum) und Time (Zeit) ein. In einer Facettenklassifikation wird ein Gegenstand oder Thema hiermit aus unterschiedlichen Fragestellungen zu seiner Beschaffenheit oder des thematischen Inhalts klassifiziert und fällt somit mehreren Klassen zu. Das Thema oder der Gegenstand wird aus unterschiedlichen Dimensionen betrachtet:

> In short, faceted classification is a method of multidimensional description and arrangement of information resources by their concept, attributes or "aboutness". It addresses the fact that users may look for a document resource from any number of angles corresponding to its rich attributes (Uddin & Janecek 2007).

Sucht man z.B. in einer Datenbank nach einem Artikel über Architektur im 19. Jahrhundert, lässt sich die Suche in einer Vorauswahl auf den entsprechenden Zeitrahmen oder eine Stilrichtung einschränken, und es werden nur die Dokumente angezeigt, die diesen Einschränkungen entsprechen. Allerdings werden die von Ranganathan aufgestellten starren Regeln zur Facetten-Erstellung im Allgemeinen nicht streng übernommen. Die Facetten werden frei zum entsprechenden Thema passend erstellt, wobei genau diese Flexibilität ein großer Vorteil der Facetten-Klassifikation darstellt. Jede Facette kann einer eigenen Struktur in der Klassifikation folgen. So kann z.B. eine materielle Facette als Hyponomie hierarchisch aufgebaut sein, während eine räumliche Facette als Meronymie, das heißt Teil-Ganzes Beziehung aufgebaut ist (Kwasnik 1999). Die Kategorien einer facettierten Klassifikation unterliegen keinem starren enumerativen Schema, sondern sind frei untereinander kombinierbar (Vickery 1969). Der erste Schritt bei der Erstellung einer facettierten Wissensordnung ist die Facettenanalyse, an deren Ende eine Liste von Kategorien steht, welche die Facetten bilden (Broughton 2006).

Heute sind facettierte Klassifikationen im Web häufig zu finden, da sie die Navigation durch den Content erleichtern. So bieten viele Webshops die Möglichkeit, unter verschiedenen Spezifikationen nach Waren des Sortiments zu suchen. In der Suchoberfläche lässt sich die Artikelauswahl z.B. nach Produktart, Preisrahmen und/oder Hersteller durch ein Dialogfeld oder ein Verzeichnis einschränken. Die Eingabe eines Suchworts ist dabei kaum noch nötig. An der University of California in Berkeley wurde die Open-Source-Software „Flamenco“ (FLexible information Access using MEtadata in Novel COmbinations) entwickelt, mit der man eine Suchoberfläche erstellen kann, die Dokumente in Kategorien einordnet, durch die Nutzer dann durch Hyperlinks in den Verzeichnissen browsen können (Elliott 2001). Diese flexible browsing-basierte Suche in kategorisierten Verzeichnissen erwies sich in Nutzerstudien als vorteilhaft (Hearst et al. 2002; Yee et al. 2003).

Dem Nutzer muss keine Suchterminologie bekannt sein, um komplexe Anfragen zu stellen. Er kann seine Suche schrittweise nach seinen Ansprüchen einschränken oder erweitern (English et al. 2002a). Wählt man eine Kategorie aus einer Facette aus, so werden die Ergebnisse aus dieser angezeigt, sowie eine Navigationsleiste mit den anderen Kategorien, die ebenfalls mit den Ergebnissen übereinstimmen und weitere Unterkategorien der ersten Auswahl. Die Suche kann so immer wieder spezifiziert und um andere Aspekte erweitert oder wieder abstrahiert werden, ohne dass der Nutzer den Überblick verliert (English et al. 2002b). Eine oft frustrierende Versuch-und-Irrtum-Interaktion mit dem Suchsystem ist somit nahezu ausgeschlossen (Sacco 2006).

Für Produkte und Bilder müssen in den jeweiligen Dokumenten Metadaten (Tags) hinzugefügt werden, die das Objekt entsprechend den Kategorien beschreiben. Bei der Suche innerhalb von Texten bestehen diese Metadaten aus dem Text selbst. Dokumente, in denen ein Wort vorkommt, welches mit einer Kategorie übereinstimmt, werden als Ergebnis angezeigt. In komplexen Retrievalsystemen geschieht dies jedoch nicht allein durch eine thematische Volltextindexierung, sondern durch ein semantisches Netz.

Facettierter Thesaurus

Ein Dokumentationsgebiet bzw. Wissensgebiet wird durch eine Begriffsordnung dargestellt, die dieses Wissensgebiet abbildet. Es wurden bereits vielfältige Thesauri unterschiedlichster Wissensgebiete erarbeitet, wie z.B. der umfassende Standard-Thesaurus Wirtschaft.
Bei einem facettierten Thesaurus wird ein Wissensgebiet ebenso begrifflich dargestellt, allerdings aus mehreren Blickwinkeln betrachtet. Dem Nutzer steht in einer entsprechenden Suchoberfläche nicht nur eine einzige Begriffsordnung zur Verfügung, sondern gleich mehrere, die untereinander kombinierbar sind. Die facettierte Klassifikation wird durch Relationen erweitert, so dass ein facettierter Thesaurus entsteht. Ein früher facettierter Thesaurus ist der „Thesaurofacet" des Wissensgebiets Technik und verwandten Themen (Aitchison, Gomershall & Ireland 1969). Der Nachrichtendienst Factiva arbeitet mit den vier Facetten Unternehmen, Branchen, Geographica und Themen, allerdings ausschließlich mit Hierarchierelationen (Stock 2002). Es ist aber möglich, noch komplexere multilinguale facettierte Ontologien umzusetzen.

RetrievalWare arbeitet bei der Suche nach Informationen mit semantischen Netzen. Ein Thesaurus stellt so ein semantisches Netz dar. Begriffe eines Themengebiets werden systematisiert und mit einer Vorzugsbenennung, dem Deskriptor, in einer Baumstruktur dargestellt. Synonyme verweisen auf den entsprechenden Deskriptor. Weitere Relationen, wie z.B. Ober- und Unterbegriffe, bilden das Konzept um den jeweiligen Begriff ab. Diese Verbindung von Konzepten mittels Relationen bildet das semantische Netz. Es kann jedoch nicht nur ein einziger Thesaurus erstellt und in RetrievalWare integriert werden, sondern es sind gleich mehrere Thesauri zum Indexieren, Klassifizieren und Suchen miteinander kombinierbar. Es entsteht so ein facettierter Thesaurus, der dann (unter ausschließlicher Nutzung der Hierarchierelation) als facettierte Klassifikation eine facettierte Suche nach Informationen ermöglicht.
Für den Beispielthesaurus „Terror in Nordirland" musste zunächst überlegt werden, welche Facetten das zu erfassende Themengebiet umfasst. Hierzu war eine Recherche nötig, um sich mit dem Thema vertraut zu machen. Im Rahmen des zeitlich begrenzten Projekts konnte natürlich nur ein rudimentärer Einblick in das Themengebiet gewonnen werden. Es musste überlegt werden, welche Aspekte für die Begriffsordnung relevant sind. Es wurden drei Thesauri erstellt: Ein Thesaurus beinhaltet geographische Angaben, die zum Teil bis auf die Straßenebene führen oder spezielle Lokalitäten beinhalten, wie z.B. die Haftanstalt Maze Prison, in der viele Mitglieder von terroristischen Organisationen inhaftiert waren. Ein weiterer Thesaurus stellt Organisationen zusammen, sowohl Terrororganisationen als auch Regierungsorganisationen, denen zum Teil Personen zugeordnet wurden. Der dritte Thesaurus befasst sich mit allgemeinen, themenrelevanten Begriffen, wie z.B. Waffen, terroristischen Aktivitäten und bereits vorgefallenen Anschlägen, im Projekt „misc" genannt.
Es werden somit drei Fragestellungen abgedeckt: Wer? (Organisationen), Wo? (Geographie) und Was? (misc). Nach Ranganathan wären dies die Facetten Personality, Space und Matter. Dem Nutzer können in diesem Beispiel drei Thesauri zur Verfügung gestellt werden, die individuell kombinierbar sind und so eine Sicht auf verschiedene Aspekte des Terrorismus aus unterschiedlichen Perspektiven eröffnen.

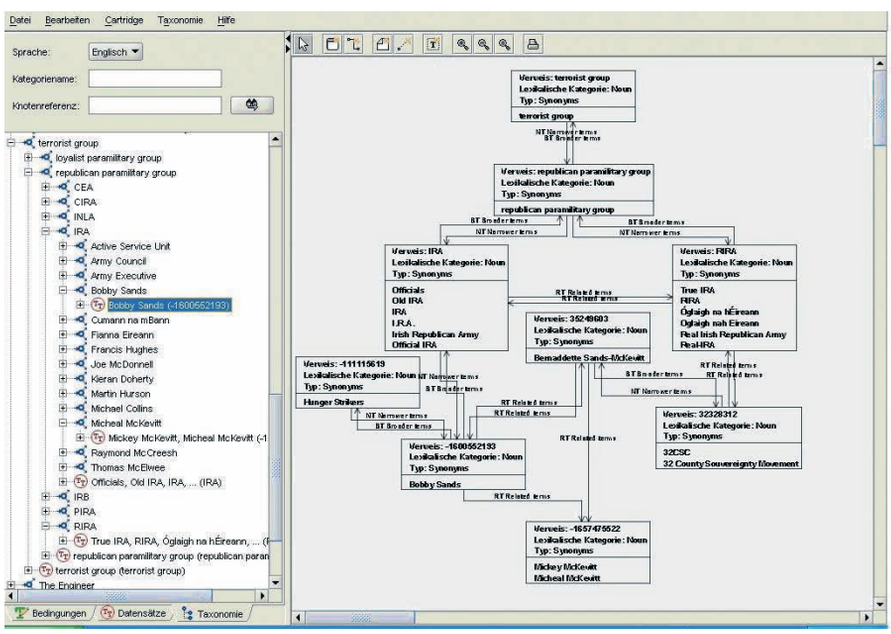


Abbildung 1: CartridgeEditor Convera.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

Ein Dokument, welches einen Anschlag der IRA in Belfast thematisiert, wird somit jeder Thesaurusfacette zugeordnet und ist aus jeder Facette erreichbar.
Der Einsatz facettierter Thesauri ermöglicht zudem eine schnelle Einsatzfähigkeit. So können die Ermittler schon mit wenigen kleinen Facetten arbeiten und das Retrievalsystem nach und nach um weitere Facetten erweitern und diese mit steigendem Hintergrundwissen und neuen Informationen modifizieren. Allerdings ist dabei zu beachten, dass ein bereits vorhandener Datenbestand bei einer Änderung bzw. Erweiterung der facettierten Thesauri erneut indexiert und klassiert werden muss.
Der oft kritisierte Mangel an Flexibilität und Skalierbarkeit von Ontologien wird durch den Einsatz facettierter Thesauri aufgehoben. Die Facetten gewähren einen leichten Überblick und können flexibel modifiziert werden. Ein einzelner Thesaurus mit 10.000 Begriffen kann so gegebenenfalls auf zwei Facetten mit je 100 Begriffen reduziert werden. Die Möglichkeit der Kombination dieser beiden Facetten, also 100·100, ergibt dann ebenfalls 10.000 Begriffe (Buchanan 1989).

Thesauruserstellung mit Convera

Begonnen wird mit der Erstellung der Taxonomie, das heißt einer monohierarchischen Klassifikation von Begriffen, welche in einer Baumstruktur dargestellt wird. Convera bietet zur Erstellung der Begriffsordnung die Software Knowledge Workbench an. Mit Hilfe des darin integrierten Programms Cartridge Editor kann ein komplexes Begriffsnetz erstellt werden. Dieses Begriffsnetz dient dann zur Indexierung, Klassierung und Gewichtung der Dokumente und zur Filterung der gesuchten Informationen innerhalb der Datenbasis. Es können aber auch bereits zur Verfügung stehende Thesauri in RetrievalWare eingebunden werden (Bayer et al. 2005).
Der Editor hat zwei Bereiche, wie in Abbildung 1 zu sehen ist. Im linken Bereich gibt es drei Registerkarten – Bedingungen, Datensätze und Taxonomie –, welche die terminologische Kontrolle unterstützen. Im rechten Bereich befindet sich eine graphische Oberfläche, Canvas genannt, in welcher die Beziehungen zwischen den Begriffen erstellt und visualisiert werden. Auf dem Register „Taxonomie" werden die Kategorien monohierarchisch in einer Baumstruktur erstellt. Zu jeder Kategorie wird automatisch ein Datensatz generiert, von Convera Synonym Set, kurz Synset genannt.

In Abbildung 2 ist ein solches Synset für den Begriff *IRA* aus der Thesaurusfacette „Organisationen“ zu sehen.
Beim Erstellen der Kategorie ist der gewählte Kategoriename in der Taxonomie

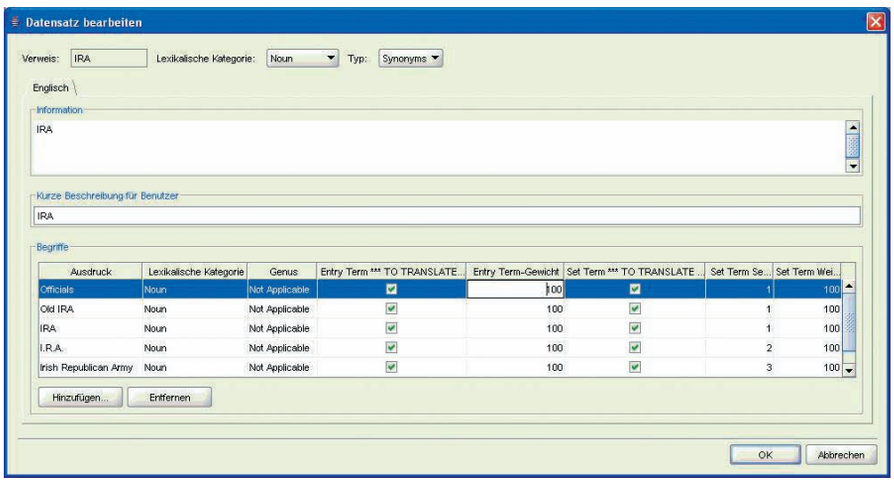


Abbildung 2: Synset für den Begriff IRA aus der Thesaurusfacette „Organisationen“. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

die Vorzugsbenennung, also der Deskriptor, welcher später in der Suchoberfläche dem Nutzer als Verzeichnis angezeigt wird. Diesem Deskriptor können Synonyme zugefügt und deren grammatikalische Kategorie und Genus angegeben werden. Des Weiteren kann für jedes Synonym ein Gewichtungswert gesetzt werden, der den Bindungsgrad der Bezeichnung an den Begriff angibt und beim Ranking der Dokumente dementsprechend berücksichtigt wird. Würde der Begriff *Old IRA* mit einem Term-Gewicht von 50 Prozent angegeben, würde dieser bei der Klassifikation und bei der Suche durch den Nutzer auch nur mit 50 Prozent in das Ranking eingehen. Bei dem gewählten Synonym-Gewicht von 100 Prozent ergibt die Suchanfrage nach *IRA* z.B. die gleichen Ergebnisse wie die Suche nach dem hinterlegten Synonym *Irish Republican Army*.
Es muss nun jede Kategorie intellektuell und manuell erstellt werden, wobei eventuell weitere Datensätze heranzuziehen sind, um das semantische Umfeld des Begriffs genau abzubilden und die Relevanz der Treffer bei der Suche zu erhöhen. Bei der Erstellung der Taxonomie muss bedacht werden, dass diese dem Nutzer später als Verzeichnis in der Suchoberfläche angezeigt wird.

> Es (...) ist dafür Sorge zu tragen, dass pro Begriff eine überschaubare Menge an Unterbegriffen (etwa maximal 20 bis 30) existiert, da ansonsten die Ausgabe unübersichtlich werden kann. Dasselbe gilt für die Anzahl der Topterme der Dokumentationssprache (Stock 2007, 457).

Bei der Erstellung der Synsets muss ein sehr gutes Verständnis des Themas vorhanden sein. So ist z.B. Robert Bates, ein Mitglied der „Shankill Butchers“, einer besonders brutalen Gruppe der „Ulster Volunteer Force“ Anfang der 70er Jahre, auch unter dem Namen „Basher“ bekannt. Es ist wichtig solche Pseudonyme zu kennen, weil sonst wichtige Dokumente nicht gefunden werden können. So erhält man für die Suchanfrage *Basher* alle Dokumente, die den Namen Robert Bates enthalten, und umgekehrt alle Dokumente, die Bates nur unter seinem Pseudonym behandeln. Durch die Eingabe des Synonyms *Basher* in den Datensatz wird dem Nutzer der Zugang zu diesen Dokumenten ermöglicht, ohne dass er alle Pseudonyme kennen muss. Im Registerblatt „Datensätze“ können alle Synsets sortiert in der Reihenfolge ihrer Eingabe angezeigt werden. Das Programm weist jedem Datensatz automatisch eine Verweisnummer als Identifikator zu; sie kann aber auch manuell eingegeben werden. Im Register „Bedingungen“ werden alle Worte, sowohl die Vorzugsbenennungen als auch die Nicht-Deskriptoren der Synsets, alphabetisch aufgelistet. Nachdem alle Begriffe eingegeben wurden, kann zu der Herstellung der Relationen im Canvas übergegangen werden. Es können beliebige Relationsformen mit unterschiedlichen Gewichtungswerten hergestellt werden. Für die Thesauri des Projekts wurden die Relationen Broader Term (BT), Narrower Term (NT) und Related Term (RT) eingesetzt (Abbildung 1).
Die manuelle Eingabe erweist sich als aufwendig und erfordert ein hohes Maß an Präzision und Kontrolle. Die erstellten Taxonomien und Thesauri können mit Hilfe von Qualitätsmetriken und eines Benchmarking auf ihre Effizienz getestet und modifiziert werden. Nach positiver Evaluation wurden die drei Thesaurusfacetten in RetrievalWare eingebunden (Abbildung 3).

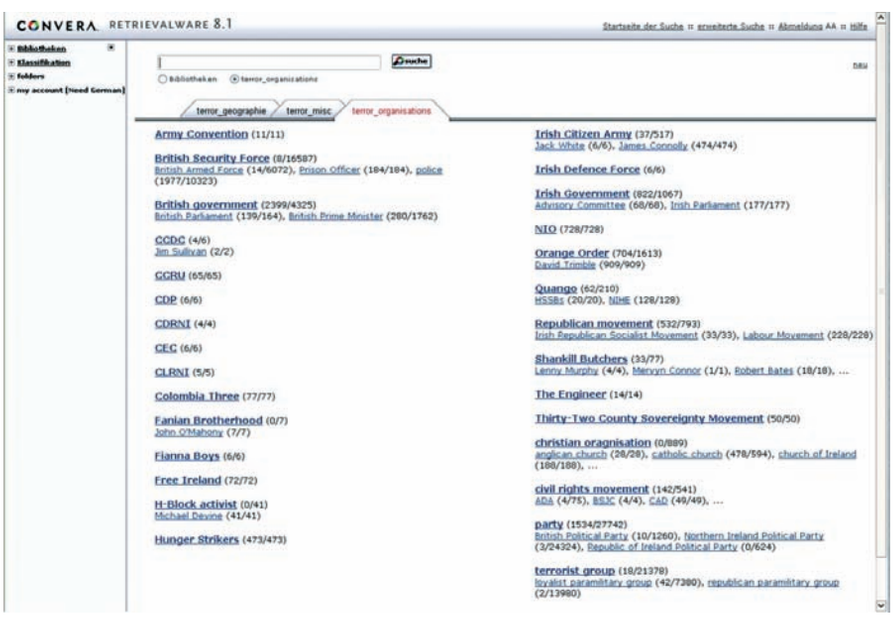


Abbildung 3: Ansicht Startseite der Suche. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

Crawling

Nachdem die drei Thesauri fertig gestellt waren, musste eine Datenbasis aus dem Web gewonnen werden. Da der Crawler von Convera nicht lizenziert war, griff die Projektgruppe auf ein Freeware-Programm unter GPL-Lizenz zurück. Ein Crawler, auch Robot oder Spider genannt, verfolgt Links, die sich auf einer Webseite befinden und speichert deren Inhalte ab (Pant, Srinivasan & Menczer 2003). Hierzu muss zunächst eine Liste mit relevanten Start-URLs, auch Seed-Liste genannt, vorhanden sein, von der aus der Crawler die dort hinterlegten Hyperlinks verfolgt. Die Seed-Liste mit den Ausgang-URL

unseres Projekts umfasste unter anderem frei zugängliche Webseiten von Organisationen (Sinn Fein), Tageszeitungen (An Phlobacht) und Datenbanken (Cain), von denen auch relevante weiterführende Links zu erwarten waren. Es ist dabei zu beachten, dass die Tiefe der Linkverfolgung nicht zu hoch eingestellt wird, weil sonst zu viele irrelevante Seiten gespiegelt werden. Der Crawler wurde so konfiguriert, dass er alle Datei-Formate übernommen hat. Insgesamt wurden für die Experimentelle IRA-Datenbank 1,5 Gigabyte an Daten zusammengestellt.

Unter realen Bedingungen, das heißt bei einem Vorhaben wie „Check the Web" wäre es in Anbetracht der vorhersehbaren enormen Datenmenge von Vorteil, einen Crawler zu verwenden, der fokussiert auf ein Thema hin crawlt (Kwiatkowski & Höhfeld 2007).

Dynamische Klassierung und facettierte Thesauri

Die Dynamische Klassierung ermöglicht das Indexieren und Kategorisieren in einem Arbeitsschritt. So kann auch eine sehr große heterogene Datenbasis, wie sie sich aus einem Crawling ergibt, schnell und präzise durchsucht werden.
Die dynamische Klassierung erfolgt auf Grundlage des erstellten facettierten Thesaurus. Die Dokumente werden vom Retrievalsystem mit den Termen der Thesauri indexiert und nach ihrer Gewichtung bewertet. In der Suchoberfläche werden die erstellten Taxonomien mit der Trefferzahl zum jeweiligen Begriff und dessen untergeordneten Begriffen als Verzeichnisse angezeigt. Die indexierten Dokumente werden dann direkt den Klassen zugeordnet (Abbildung 4).

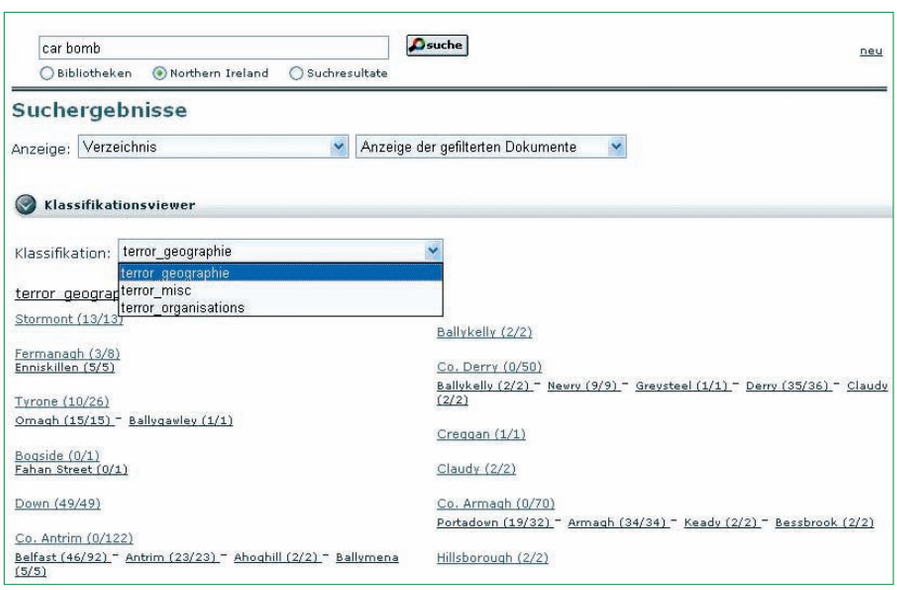
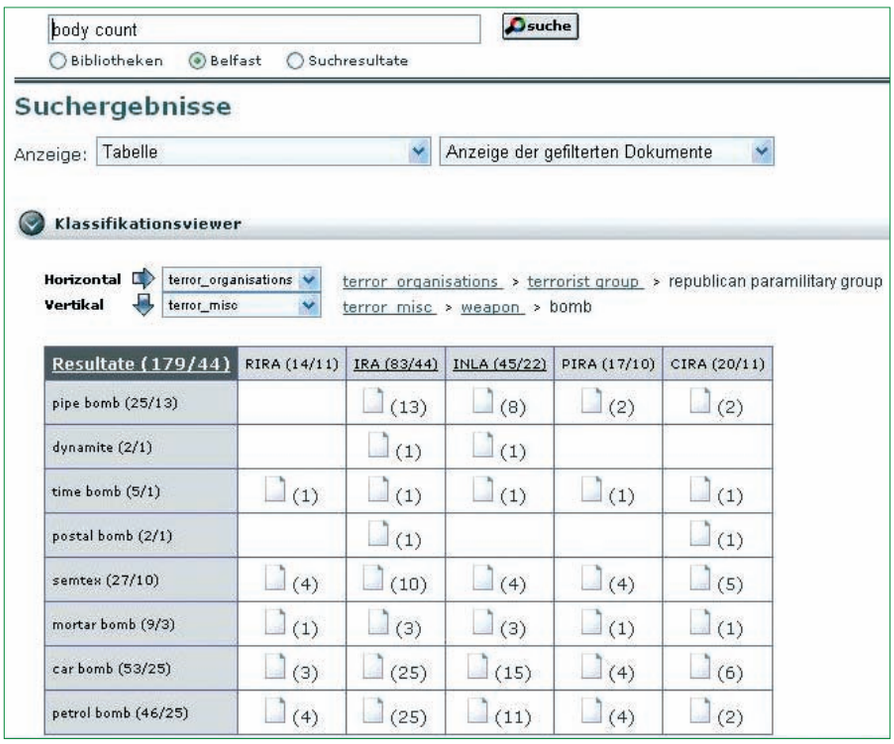


Abbildung 4: Verzeichnis-Ansicht Suchergebnisse.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft

Das Focused Crawling ermöglicht das Durchsuchen des *Dark Webs* auf der Grundlage einer Ontologie (Chakrabarti, van den Berg & Dom 1999). Ein fokussierter Crawler findet durch eine Begriffsordnung effektiv und ballastarm thematisch verwandte Seiten (Davison 2000). Das Problem einer zu hohen Link-Tiefe wird durch das fokussierte Crawling gelöst, da nur relevante Seiten verfolgt werden. Eine Beschreibung dieser sinnvollen Vorgehensweise würde den Rahmen dieses Artikels sprengen, sollte in der Praxis aber weiter verfolgt und umgesetzt werden.
Um die Aktualität des gecrawlten Datenbestandes zu gewährleisten, muss der Crawlvorgang in kurzen Abständen immer wieder neu gestartet werden, denn ein großer Teil der aktiven Web-Inhalte verändert sich täglich (Cho & Garcia-Molina 2000).

Die Suche nach *car bomb* in der Thesaurusfacette „Geographie" führt für den Begriff *Co. Antrim* zu keinem Treffer, allerdings zu 122 Treffern in den Unterbegriffen, also zu den Stadtteilen, die zu County Antrim gehören. Darunter sind wiederum 46 Treffer für den Begriff *Belfast* und 92 Treffer für dessen Unterbegriffe, wie z.B. *Maze Prison*. Es ist also möglich im Verzeichnis nach den Suchergebnissen zu browsen und so immer tiefere, das heißt speziellere Fokusse zu betrachten. Die Begriffsleiter, die der Nutzer beim Browsen durch die Begriffe hinter sich gelassen hat, wird dabei in jedem Schritt angezeigt, damit er sich besser orientieren und den Pfad seiner Suche zurückverfolgen kann. In unserem Beispiel in Abbildung 4 hat der Nutzer die Suche nach dem Begriff *car bomb* in der Vorauswahl auf die Facette „Geographie" mit Fokus auf *Northern Ireland* eingeschränkt. Dieselbe Suchanfrage lässt sich allerdings auch innerhalb dieses Fokus auf die anderen Facetten anwenden, indem im Klassifikationsviewer eine andere Facette ausgewählt wird. Die angezeigten Treffer zeigen dann z.B. bei gewählter Facette „Organisationen", die Dokumente an, die den Begriff *car bomb* sowohl im Zusammenhang mit dem Begriff *Northern Ireland*, als auch mit den Begriffen der Organisationen nennen. So ergänzen sich dynamische Klassierung und facettierte Thesauri zu einfach durchführbaren und zugleich komplexen Anfragen.
Die Anzeige der Suchergebnisse in der Tabellenansicht, wie sie Convera RetrievalWare anbietet, stellt dabei eine besonders nutzerfreundliche Visualisierung der Suchergebnisse dar und ermöglicht die dynamische Anzeige von Treffern in allen drei erstellten Facetten.

In Abbildung 5 hat der Nutzer am Beginn seiner Suche den Begriff *Belfast* im Verzeichnis „Geographie" gewählt und den Suchbegriff *body count* eingegeben. In der Tabellenansicht hat er dann die Möglichkeit, zwei weitere Facetten einzubeziehen. In diesem Beispiel steht horizontal die Begriffsleiter der Organisationen und vertikal die Begriffsleiter der allgemeinen Begriffe (misc) zur Verfügung. Auch hier wird der Pfad, den der Nutzer innerhalb der Begriffsleiter genommen hat, angezeigt. In der Facette „Organisationen" befindet er sich auf der Ebene der republikanisch paramilitärischen Gruppierungen und in der Facette der allgemeinen Begriffe auf der Ebene der verschiedenen Arten von Bomben. In der Tabelle wird die Anzahl Treffer zu dem Begriff *body count* im Kontext dieser Begriffsebenen angezeigt.
Eine einzige Suchanfrage kann hiermit auf Konzepte erweitert werden, die die Treffermenge signifikant auf den zu betrachtenden Fokus eingrenzen und Informationsballast vermeiden. Die 25 Dokumente, die im dargestellten Beispiel unter *IRA* und *car bomb* angezeigt werden, enthalten Begriffe, die im Kontext zu Todesopfern im Zusammenhang mit der Organisation IRA und Autobomben im geographischen Gebiet Belfast stehen. Dieses Beispiel zeigt eine extrem fokussierte Ansicht, die jedoch jederzeit wieder verallgemeinerbar ist, indem der Nutzer sich in der Begriffsleiter wieder nach oben bewegt. Die Tabellen-Ansicht der dynamischen Klassierung hat jedoch nicht nur den Vorteil, dass die Treffermenge kontextspezifisch noch mehr eingegrenzt werden kann als in der einfachen Verzeichnisstruktur, sie bietet auch die Möglichkeit, Muster und Verbindungen aufzudecken, die bei der initialen Suchanfrage noch gar nicht in Betracht gezogen worden sind. Die Granularität des semantischen Netzes lässt sich so an-



Resultate (179/44)	RIRA (14/11)	IRA (83/44)	INLA (45/22)	PIRA (17/10)	CIRA (20/11)
pipe bomb (25/13)		(13)	(8)	(2)	(2)
dynamite (2/1)		(1)	(1)		
time bomb (5/1)	(1)	(1)	(1)	(1)	(1)
postal bomb (2/1)		(1)			(1)
semtex (27/10)	(4)	(10)	(4)	(4)	(5)
mortar bomb (9/3)	(1)	(3)	(3)	(1)	(1)
car bomb (53/25)	(3)	(25)	(15)	(4)	(6)
petrol bomb (46/25)	(4)	(25)	(11)	(4)	(2)

Abbildung 5: Tabellen-Ansicht Suchergebnisse.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

fragespezifisch einstellen, dass entweder fokussiert oder abstrahiert wird, wobei eine Abstraktion Informationen aufdecken kann, die aus einer Fokussierung herausgefallen sind und so unerwartet Fragen beantworten kann, die noch gar nicht gestellt wurden.

Ranking und Ergebnisansicht

Die Gewichtung (und in deren Folge das Ranking) der Dokumente erfolgt bei Convera auf unterschiedlichen Ebenen. Die eingegebenen Suchterme werden mit den indexierten Dokumenten abgeglichen und unter verschiedenen Aspekten gewichtet. Der Rankingprozess beginnt mit einem grobkörnigen Ranking (coarse grain rank) und wird mit einem feinkörnigen Ranking (fine grain rank) abgeschlossen. Die aus jedem Schritt resultierenden Gewichtungswerte für jedes Dokument werden addiert. Das erste grobkörnige Ranking gewichtet die Dokumente nach der Anzahl der mit den in der Suchanfrage übereinstimmenden Terme und deren verwandten Begriffen. Danach werden die verwandten Begriffe danach gewichtet, in welchem Kontext sie stehen, das heißt nach ihrer Nähe zu anderen kontextspezifischen Begriffen. Die Begriffe werden auch nach ihrem Verwandtschaftsgrad gewichtet, das heißt stark verbundene Konzepte werden höher gewichtet als schwach verbundene Konzepte. Morphologische Variationen werden dabei ebenfalls berücksichtigt. Verfeinert wird das Ranking durch das „fine grain". Je näher die Terme der Suchanfrage im Dokument zusammen auftreten, desto höher ist der Gewichtungswert. Kommen zwei Suchterme oder deren verwandte Begriffe in einem Satz vor, so ist das Dokument relevanter als wenn diese im selben Absatz vorkommen. Diese Stellungsrestriktion macht auch eine Disambiguierung von Homonymen möglich. Des Weiteren werden die Anzahl der gefundenen Suchterme und deren Entsprechungen in Relation zur Länge des Dokuments gestellt. Das grobkörnige Ranking lässt sich bei der Erstellung der Thesauri durch einstellbare Gewichtungswerte der verschiedenen Relationen beeinflussen. Das feinkörnige Ranking kann vom Administrator vor dem Indexlauf eingestellt werden.
Die Ergebnisse aus der Suchanfrage werden dem Nutzer mit der Angabe der Dokumentrelevanz in einer Liste angezeigt. Neben einer Ansicht der Treffer mit einer kurzen Schlagzeile der Dokumente steht auch die Ansicht einer Dokumentzusammenfassung zur Verfügung, in welcher der thematische Schwerpunkt des jeweiligen Dokuments ersichtlich ist (Abbildungen 6 und 7).

Wählt man ein Dokument aus der Ergebnisliste aus, so wird dieses mit den markierten Treffern angezeigt (Abbildung 8). Dabei werden die gefundenen Begriffe aus der Suchanfrage gelb und die verwandten Begriffe blau markiert. Bewegt man den Cursor auf einen markierten Begriff, so wird der errechnete Relevanzwert für diesen angezeigt.

Darüber hinaus können die Suchterme vom Nutzer manuell gewichtet werden, indem die jeweiligen Wörter mit einem Doppelpunkt (Multiplikator) und einer Zahl von eins bis zehn versehen werden. So würde bei einer Suchanfrage mit dem Ausdruck *bomb Engineer:10* der Term *Engineer* zehnmal höher gewichtet als *bomb*. Es lassen sich zusätzlich auch ähnliche Dokumente zu einem Treffer anzeigen, wobei die am höchsten gewichteten Begriffe, die im Dokument durch die

Abbildung 6: Schlagzeilen-Ansicht Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

Abbildung 7: Ansicht Dokumentzusammenfassung Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

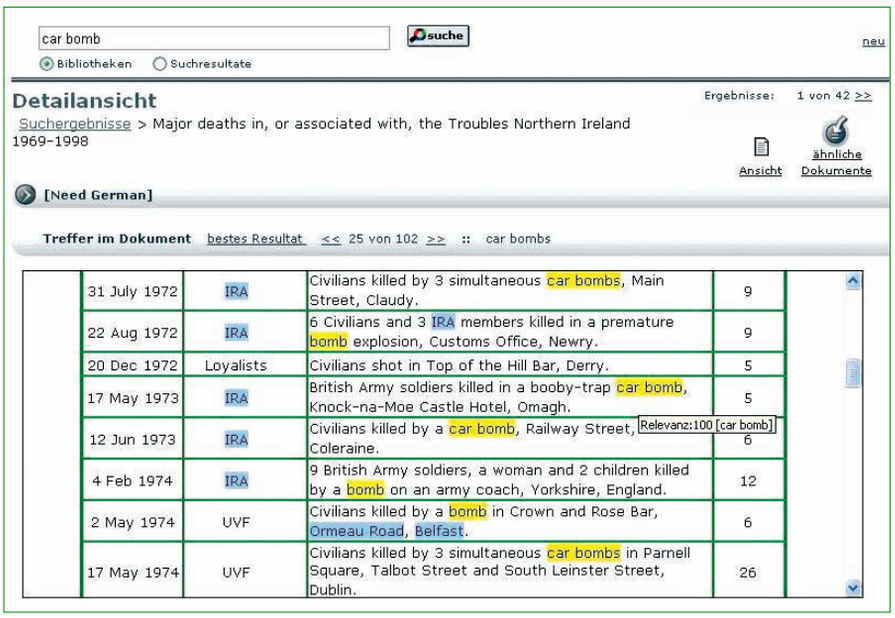


Abbildung 8: Detailansicht eines Dokuments aus der Ergebnisliste.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

initiale Suche gefunden worden sind, verwendet werden.

Sprachübergreifendes Retrieval

Der hier behandelte Thesaurus „Terror in Nordirland" wurde monolingual in Englisch erstellt. Convera bietet aber als Standard die Übersetzungen in die Sprachen Deutsch, Englisch, Französisch, Italienisch und Niederländisch an. So führt auch eine Suche nach dem Begriff *Hungerstreik*, wenn die Abfragesprache „Deutsch" gewählt wird, zu Dokumenten mit dem Begriff *hunger strike*, wie auch zu Dokumenten mit der Übersetzung in den erwähnten anderen Sprachen (Abbildung 9). Um eine sprachübergreifende Konzeptsuche zu ermöglichen, müsste der Thesaurus allerdings multilingual erstellt und gepflegt werden.

In der Suchoberfläche von RetrievalWare besteht die Möglichkeit, die Suchanfrage nicht nur um Terme der Konzepte in der Abfragesprache zu erweitern, sondern auch in anderen Sprachen. Dabei handelt es sich nicht um eine einfache Übersetzung der Suchterme, sondern um die Übersetzung in ein anderssprachiges Konzept. Markiert man die Option „Übereinstimmung mit der Abfragesprache", verhindert man dabei, dass Homonyme in anderen Sinnzusammenhängen gefunden werden, als in der Suchanfrage intendiert. Wörter mit gleicher Schreibweise in unterschiedlichen Sprachen, aber mit verschiedenen Bedeutungen werden dann in Bezug auf das Konzept der Abfragesprache erweitert. Da ein Wort in verschiedenen Sprachen in einem unterschiedlichen Sinnzusammenhang stehen kann, ist hier mit großer Sorgfalt vorzugehen.

Der Übersetzer muss über detaillierte Kenntnisse der Zielsprache verfügen, um das Konzept eines Begriffes in seinem nationalen, kulturellen Zusammenhang zu übertragen. Dem Aspekt des Sprachwandels kommt dabei eine besondere Bedeutung zu, weil sich die Bedeutung eines Wortes verändern kann. Nicht zu vergessen sind dialektale Besonderheiten oder unterschiedliche Wortbedeutungen innerhalb sozialer Gruppen. Im Zusammenhang mit Themen, die den Terrorismus oder die organisierte Kriminalität betreffen, müsste zusätzlich Insiderwissen über den Sprachgebrauch unter den Mitgliedern der entsprechenden Gruppierungen vorhanden sein, bis hin zu steganografischen Verfahren.

Erweiterte Suche

In der erweiterten Suchoberfläche sind vielfältige Einstellungen möglich, um die Suche zu optimieren und so den Recall, also die Vollständigkeit der Suchergebnisse, und die Precision, das heißt die Genauigkeit der Suchergebnisse, zu erhöhen. Es werden vier Suchtypen angeboten, die Boolesche Suche, eine feldspezifische Suche, Muster-Suche und Konzept-Suche (Abbildung 10).

(1.) Die Boolesche Suche erfolgt über die üblichen booleschen Operatoren ohne Relevance Ranking. (2.) Die Suchfelder der feldspezifischen Suche können in der administrativen Oberfläche passend zum Datenbestand erstellt werden. Bei Webseiten bieten sich z.B. Tags wie Description, Head und Body an. Das Feld „Dokumentinhalt muss enthalten" dient als Filter für eine Vorauswahl der Dokumente, auf welche dann die Suchanfragen aus dem primären Suchfeld und den spezifischen Feldern ausgeführt wird. Wird dort zum Beispiel *Belfast OR London* eingegeben und im primären Feld *car bomb*, so werden nur die Dokumente nach *car bomb* durchsucht, die entweder die Wörter *Belfast* oder *London* enthalten. (3.) In der Konzeptsuche wird je nach Einstellung das gesamte semantische Netz, das heißt das Konzept um den gesuchten Begriff, zum jeweiligen Anfrageterm hinzugezogen. Dabei kann der Konzepter-

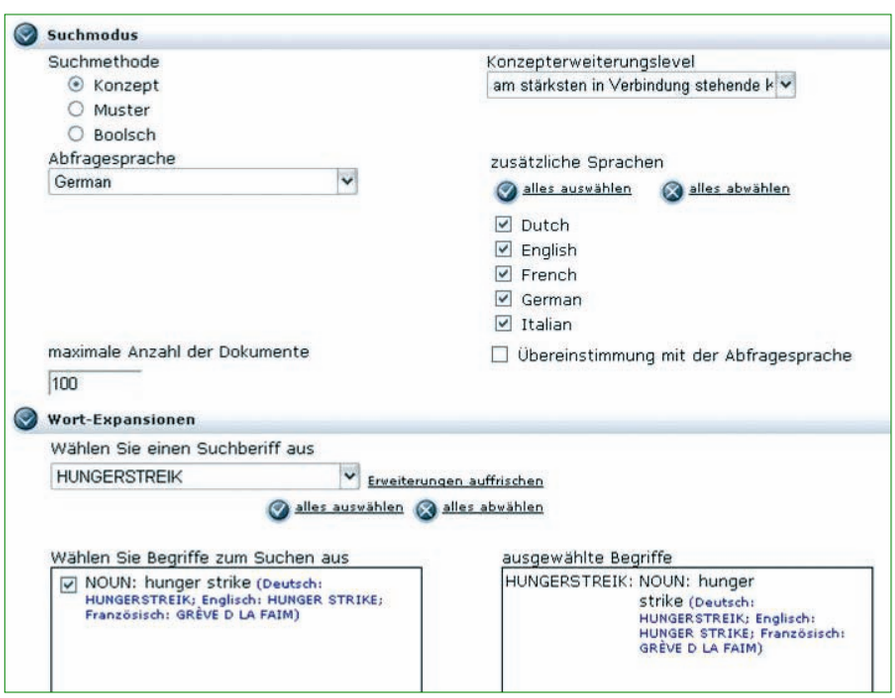


Abbildung 9: Cross-lingual Search in der erweiterten Suche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

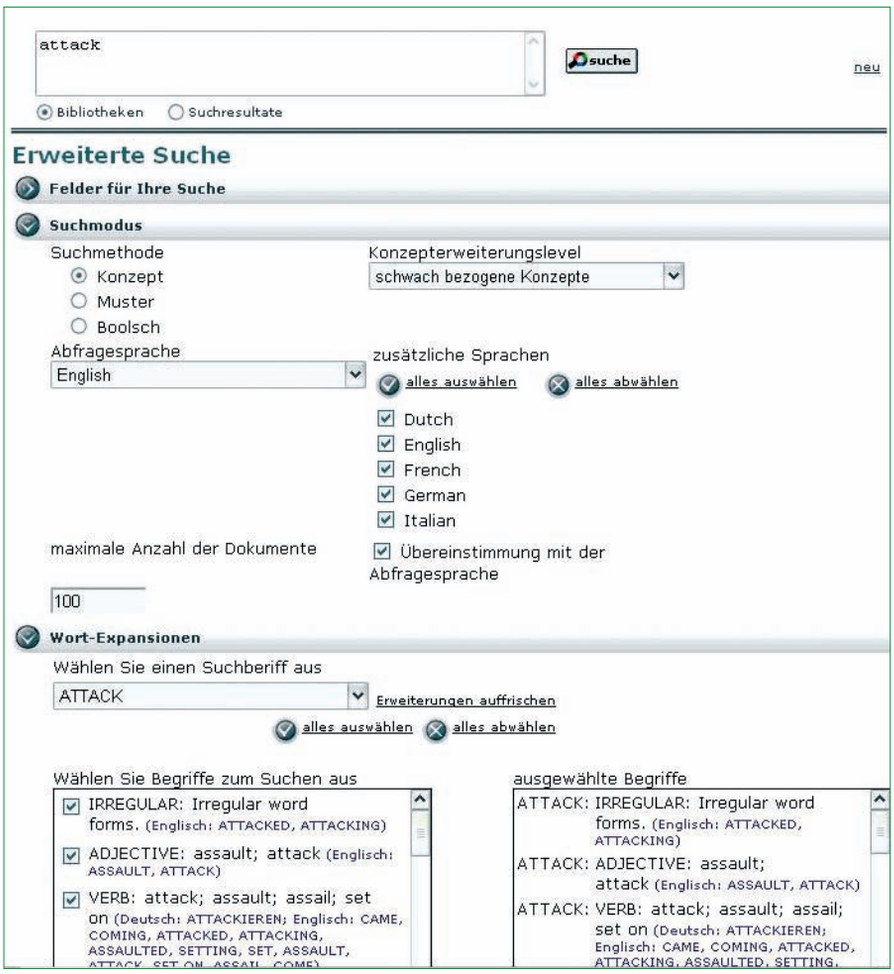


Abbildung 10: Erweiterte Suche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

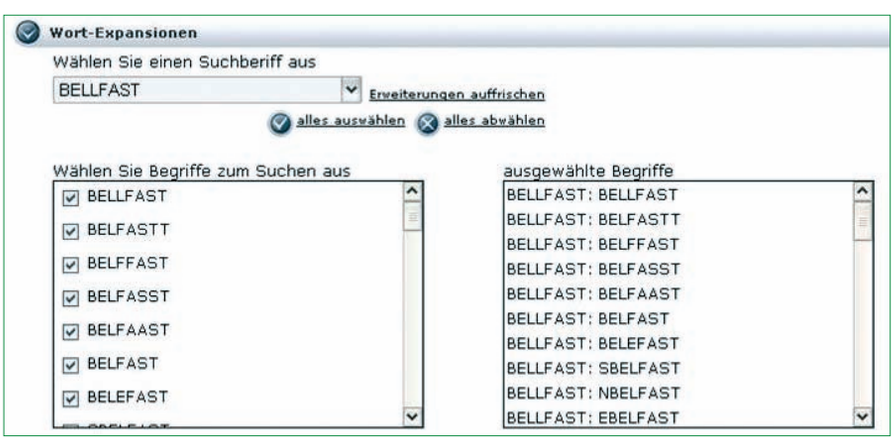


Abbildung 11: Wortexpansion der Mustersuche.
Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

weiterungsrahmen ebenfalls eingestellt und so die Granularität des semantischen Netzes justiert werden, um das der Anfrageterm erweitert wird. Zur Wahl stehen: exakte Suche, einfache Variationen, am stärksten in Verbindung stehende Konzepte, stark in Verbindung stehende Konzepte und schwach bezogene Konzepte. Je weiter der Expansionsrahmen gefasst wird, desto mehr Suchergebnisse werden angezeigt, da mehr Begriffe in die Suche und Gewichtung einbezogen werden. In der Wortexpansion werden zu den Termen konzeptuelle Erweiterungen angeboten, die man markieren und in die Suche einbeziehen kann. Die ursprüngliche Suchanfrage wird mit diesen Erweiterungen aufgefrischt. Dabei werden alle Konzepte der gewählten Sprache angezeigt. Eine Steigerung des Recalls geht bei schwach bezogenen Konzepten einher mit einer niedrigeren Precision. Hier muss der Nutzer entscheiden oder ausprobieren, ob seine Anfrage konkret oder weiter gefasst beantwortet werden soll. (4.) In der Muster-Suche werden unterschiedliche Schreibweisen, Tippfehler oder Ungenauigkeiten in der Darstellung z.B. gescannter Dokumente berücksichtigt. Convera nutzt die adaptive Mustererkennung APRP (adaptive pattern recognition process), um diese Unregelmäßigkeiten zu kompensieren. Somit würde z.B. der libysche Staatschef Muammer al-Gaddafi auch unter den Schreibweisen Mu`ammar al-Qadhafi, Muammar Qaddafi, Mo'ammar Gadhafi, Muammar Kaddafi, Muammar Qadhafi, Moammar El Kadhafi, Mu'ammar Al Qathafi, Mohammer Q'udafi oder Mu'ammar al-Qaddafi sowohl als Suchbegriff, als auch in den Dokumenten erkannt und in den Suchergebnissen enthalten sein (Bachmann 2004). Sucht man in der IRA-Datenbank nach Dokumenten, die den Bezirk Belfast behandeln, und gibt stattdessen in der Konzeptsuche *Bellfast* ein, so wird als Ergebnis ein einziges Dokument angezeigt, welches zufällig den gleichen Schreibfehler enthält. Die Mustersuche ermöglicht eine Auswahl möglicher Schreibweisen in der Wortexpansion (Abbildung 11). Die Ergebnisliste enthält daraufhin Dokumente, in denen der Begriff *Bellfast* in der korrekten Schreibweise „Belfast" vorkommt, sowie dessen verwandte Begriffe (Abbildung 12).

Es ist möglich, mit Hilfe von Operatoren die Suchtypen zu kombinieren. So kann man in der Muster- und Booleschen Suche auch die Konzeptsuche integrieren, indem man dem Term ein Ausrufungszeichen anfügt. Dieser Term wird dann semantisch um seine verwandten Begriffe erweitert, so dass z.B. Tippfehler des Nutzers oder unterschiedliche Benennungen von Begriffen zugleich mit einer Konzepterweiterung des vom System korrigierten Terms kombiniert werden. Die feldspezifische Suche ist kombinierbar mit allen drei Suchtypen. Die Dokumente müssen dann alle Terme der Felder und, optional des primären Suchfeldes, aufweisen, um in der Ergebnisliste angezeigt zu werden. Die Mustersuche kann ebenso in der Konzept- und Booleschen Suche integriert werden, indem man dem Suchterm eine Tilde ~ voranstellt. Anführungszeichen, das heißt Phrasen, verhindern jeweils die semantische Begriffserweiterung.

Fazit

Die zielgenaue Recherche in sehr großen Datenmengen – wie beim *Dark Web* – erfordert den Einsatz von hocheffizienten Mechanismen, die zum einen die relevanten Webseiten identifizieren und es zum anderen ermöglichen, die so ge-

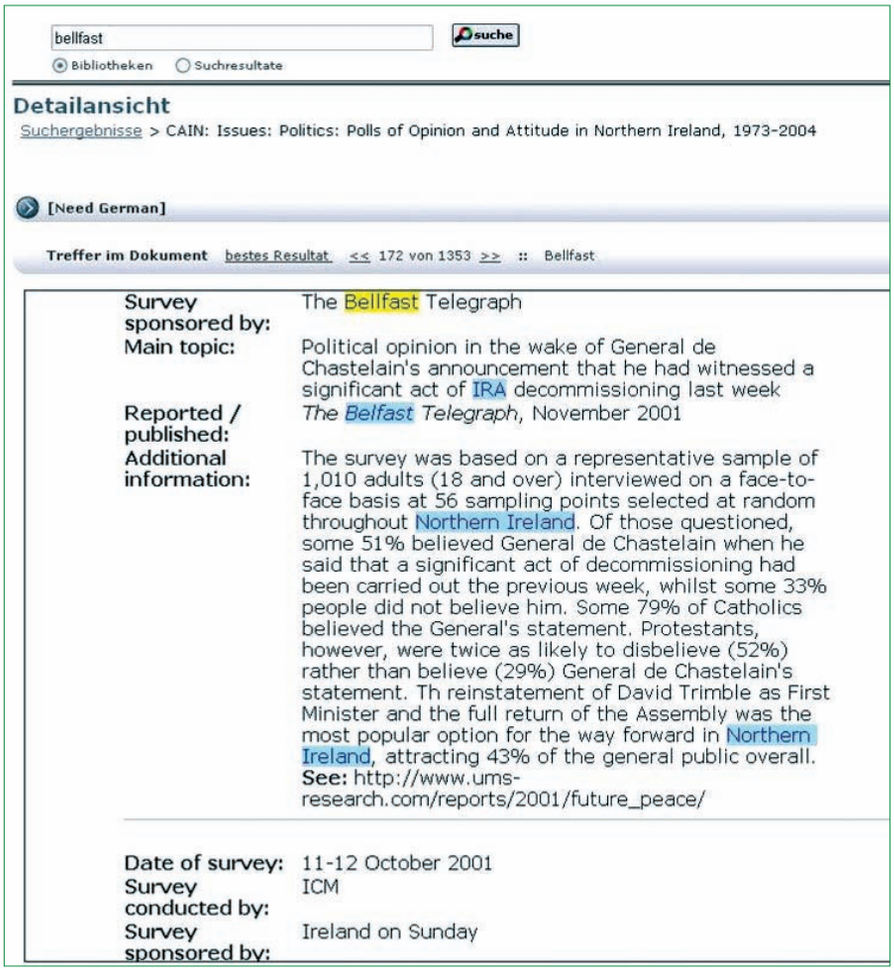


Abbildung 12: Detailansicht eines Ergebnisses aus der Mustersuche mit dem Term „Bellfast“. Quelle: Experimentelle IRA-Datenbank der Düsseldorfer Informationswissenschaft.

wonnenen Dokumente durchsuchbar und darin enthaltene Informationen auffindbar zu machen.

Am Beispiel des Thesaurus Terror Nordirland konnte exemplarisch veranschaulicht werden, wie die Integration facettierter Thesauri und dynamischer Klassierung zu einer optimalen Auswertung auch sehr großer Datenmengen führt.

Durch die dynamische Klassierung wird die Datenmenge, auf der eine Suchanfrage ausgeführt wird, automatisch klein gehalten. Die vielen irrelevanten Dokumente, die auch bei einem fokussierten Crawling gesammelt werden, treten gar nicht in Erscheinung. Bei jedem Suchvorgang, ob beim Browsen durch die Verzeichnisse oder durch Eingabe eines Suchbegriffs, werden die passenden Ergebnisse in die Kategorien eingeordnet. Die Suche basiert dadurch auf einer bereits automatisch ermittelten Ergebnismenge. Durch den Einsatz facettierter Thesauri wird diese Ergebnismenge wiederum auf bestimmte Aspekte eingegrenzt. Facettierte Thesauri sind schnell einsetzbar und können nach und nach modifiziert werden. Die intellektuelle Thesaurus-Erstellung und -Pflege ermöglicht es, die einzelnen Thesauri bei Veränderungen im Sprachgebrauch oder Wissenszuwachs innerhalb kurzer Zeit anzupassen. Automatische Verfahren können ergänzend hinzugezogen werden. Der Aufbau eines semantischen Netzes erfordert aber auch bei automatischer Wortgewinnung eine intellektuelle Bearbeitung der Wissensordnung.

Mit Hilfe der dynamischen Klassierung und dem Einsatz facettierter Thesauri ist ein multidimensionales Retrieval möglich, das ein Höchstmaß an Flexibilität während der Suche bietet. Es werden nicht nur schnell präzise Ergebnisse gefunden, es ist auch eine ungezielte Recherche möglich, durch die assoziativ neue Erkenntnisse gewonnen werden. Der Nutzer kann zwischen mehreren thematischen Perspektiven wählen oder diese kombinieren und davon ausgehend auf Spezifikationen fokussieren, die ebenfalls miteinander kombinierbar sind. Insbesondere die Tabellenansicht ermöglicht eine assoziative Suche, die Schwerpunkte, Muster und Abweichungen sichtbar macht. Der Einsatz von dynamischer Klassierung und facettierter Thesauri bildet die optimale Grundlage für präzise, ballastarme Informationsgewinnung, insbesondere bei sehr großen Datenmengen.

Besonders im Fall polizeilicher Ermittlungen im Umfeld des *Dark Web* sollte jede Möglichkeit ausgeschöpft werden, um das Ergebnis so präzise wie möglich zu gestalten und zusätzlich eine nutzerfreundliche, intuitive Suche zu ermöglichen. Die Analyse von *Dark Web*-Inhalten kann nicht bei der Identifizierung von Verbindungen und technischen Möglichkeiten verharren. Ermittlungsmaßnahmen erfordern eine genauere Auseinandersetzung mit dem zusammenzustellenden Material und den darin enthaltenen Informationen. Der Zeit- und Kostenfaktor für den intellektuellen Ontologieaufbau wird durch den Zeitgewinn und die Effizienz bei der Informationsgewinnung wieder aufgefangen.

Danksagung

Wir danken Sonja Gust von Loh für die Bereitstellung von Informationen über „Check the Web“ und für Anregungen zum derzeitigen Stand der Dark-Web-Forschungen. Ebenfalls danken wir Jasmin Schmitz für die Überprüfung der Übersetzung des Abstracts ins Englische.

Literatur

Aitchison, J., *Gomershall*, A., & *Ireland*, R. (1969). Thesaurofacet: A Thesaurus and Faceted Classification for Engineering and Related Subjects. Whetstone, Leicester: English Electric Co. Ltd.

Artificial Intelligence Lab (2007). University of Arizona. Management Information Systems (MIS) Department. http://ai.bpa.arizona.edu/research/terror/index.htm [15.01.2008].

Bachmann, H. (2004). Convera Overview. Workshop on Managing Nuclear Knowledge. International Atomic Energy Agency. Trieste. http://www.iaea.org/inisnkm/nkm/documents/trieste2004/ L12(Bachmann).pdf [15.01.2008].

Bayer, O. et al.(2005). Evaluation of an ontology-based knowledge management system. A case study of Convera RetrievalWare 8.0. Information Services & Use 25(2005), 181-195.

Bergman, M.K. (2001). The Deep Web: Surfacing hidden value. The Journal of Electronic Publishing, 7(1). www.press.umich.edu/jep/07-01/bergman.html [15.01.2008].

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings 58(2006)1/2, 49-72.

Buchanan, B. (1989). Bibliothekarische Klassifikationstheorie. München: Saur.

Bundesministerium des Innern (2007). Europa sicher leben. www.eu2007.bmi.bund.de/ nn_1035938/EU2007/DE/Bilanz/GesamtBilanz__Ratspraesidentschaft__de.html [15.01.2008].

Chakrabarti, S., *van den Berg*, M., & *Dom*, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks 31(1999), 1623-1640.

Chen, H. et al. (2003). COPLINK* Managing Law Enforcement Data and Knowledge. Communications of the ACM 46(2003)1, 28-34.

Chen, H. et al. (2005). The dark web portal: Collecting and analyzing the presence of domestic and international terrorists groups on the web. Lecture Notes in Computer Science 3495, 623-624.

Cho, J. & *Garcia-Molina*, H. (2000). The evolution of the web and implications for an incremental crawler. In: Proceedings of the 26th International

Conference on Very Large Databases (S. 200-209).

Davison, B. (2000). Topical locality in the Web. In: Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (S. 272–279). New York: ACM.

Demchak, C. C., *Friis*, C. & *La Porte*, T. M. (2000). Webbing Governance: National Differences in Constructing the Face of Public Organizations. In: *Garson*, G. D. (Ed.), Handbook of Public Information Systems (S. 71-84). New York: Marcel Dekker Publishers.

Elliott, A. (2001). Flamenco Image Browser: Using Metadata to Improve Image Search During Architectural Design. In: Proceedings of CHI 2001 (S. 69-70). New York: ACM.

English, J., *Hearst*, M., *Sinha*, R., *Swearingen* K. & *Yee*, P. (2002a). Hierarchical Faceted Metadata in Site Search Interfaces. CHI 2002 Conference Companion. http://flamenco.berkeley.edu/papers/chi02_short_paper.pdf [15.01.2008].

English, J., *Hearst*, M., *Sinha*, R., *Swearingen*, K. & *Yee*, P. (2002b). Flexible search and browsing using faceted metadata. Unpublished Manuscript. http://flamenco.berkeley.edu/papers/flamenco02.pdf [15.01.2008].

EU Council Secretariat (2007). The European Union and the Fight Against Terrorism. Brussels: Press Office of the Council of the European Union.

Hearst, M., *Elliott*, A., *English*, J., *Sinha*, R., *Swearingen*, K. & *Yee*, K.P. (2002). Finding the flow in web site search. Communications of the ACM 45(2002)9, 42-49.

Lewandowski, D. (2005). Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt am Main: DGI.

Kwasnik, B.H. (1999). The role of classification in knowledge representation and discovery. Library Trends 48(1999)1, 22-47.

Kwiatkowski, M., & *Höhfeld*, S. (2007). Thematisches Aufspüren von Web-Dokumenten – Eine kritische Betrachtung von Focused Crawling-Strategien. Information - Wissenschaft und Praxis, 58(2007)2, 69-82.

Li, J., *Zheng* R. & *Chen*, H. (2006). From fingerprint to writeprint. Communications of the ACM 49(2006), 76-82.

Netcraft Web Server Survey (2007). http://news.netcraft.com/archives/web_server_survey.html [15.01.2008].

Pant, G., *Srinivasan*, P. & *Menczer*, F. (2004). Crawling the Web. In: *Levene*, M., & *Poulovassilis*, A. (Eds.), Web Dynamics (S. 153-178). Heidelberg: Springer.

Qin, J. et al. (2005). The Dark Web Portal project: Collecting and analyzing the presence of terrorist groups on the Web. Lecture Notes in Computer Science 3495, 623-624.

Qin, J. et al. (2007). Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity. International Journal of Human-Computer Studies 65(2007), 71-84.

Ranganathan, S.R. (1933): Colon Classification. Madras: Madras Library Association.

Reid, E. & *Chen*, H. (2006). Extremist Social Movements Groups and Their Online Digital Libraries. Information Outlook 10(2006)6, 57-65.

Reid, E. et al. (2005). Collecting and analyzing the presence of terrorists on the Web: A case study of Jihad Websites. Lecture Notes in Computer Science 3495, 402-411.

Sacco, G. M. (2006). Dynamic taxonomies and guided searches. Journal of the American Society for Information Science and Technology 57(2006)6, 777-780.

Sherman, C. & *Price*, G. (2001). The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today

Stock, M. (2002). Factiva.com: Neuigkeiten auf der Spur. Searches, Tracks und News Pages bei Factiva. Password 17(2002)5, 31-40.

Stock, W.G. (2007). Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

Stock, W.G. & *Stock*, M. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München, Wien: Oldenbourg.

Uddin, M.N. & *Janecek*, P. (2007). The implementation of faceted classification in web site searching and browsing. Online Information Review 31(2007)2, 218-233.

Vickery, B.C. (1969). Facettenklassifikation. München-Pullach, Berlin: Verlag Dokumentation.

Yee, K.P. et al. (2003). Faceted metadata for image searching and browsing. In: Proceedings of the SIGCHI conference on Human factors in computing systems (S. 401-408). New York: ACM.

Zhou Y. et al. (2005). U.S. domestic extremist groups on the Web: Link and content analysis. IEEE Intelligent Systems 20(2005)5, 44-51.

Wissensordnung, facettierte Wissensordnung, dynamisches Klassieren, Information Retrieval, Dark Web, IRA-Datenbank, Convera RetrievalWare, facettierte Klassifikation, facettierter Thesaurus, Thesauruserstellung, Crawling, sprachübergreifendes Retrieval, semantisches Netz

Die Autoren

Silke Heesemann

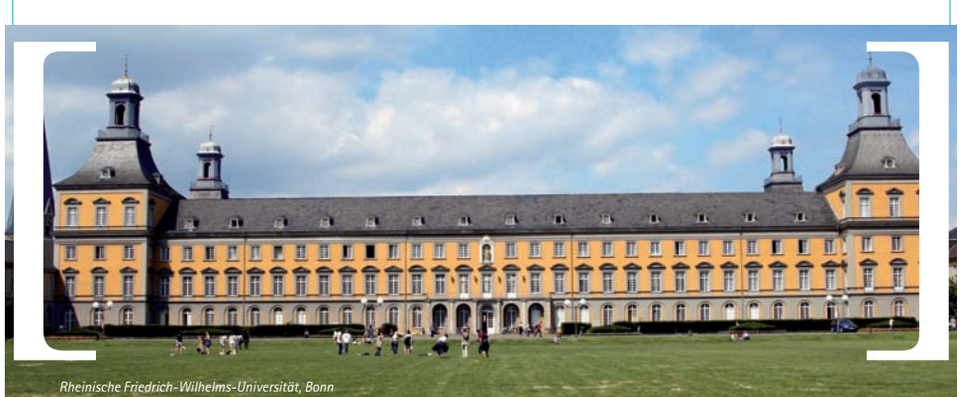
ist Mitarbeiterin der Arbeitsgruppe Datenbanken und Softwaretechnik der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf.

silke.heesemann@uni-duesseldorf.de

Hans-Dieter Nellißen

ist technischer Mitarbeiter der Arbeitsgruppe Datenbanken und Softwaretechnik der Abteilung Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf.

hans.nellissen@uni-duesseldorf.de



ImageWare inside

Installationen in deutschen Exzellenzuniversitäten: RWTH Aachen | FU Berlin | Albert-Ludwigs-Universität Freiburg | Georg-August-Universität Göttingen | Universität Konstanz | TU München | Ludwig-Maximilian-Universität München | Ruprecht-Karl-Universität Heidelberg

CeBIT Halle 3 Stand C30 HANNOVER 4.–9.3.2008

InetBib 09.–11.04.2008 Würzburg

Buchscanner Bookeye | Scansoftware BCS-2 | Dokumentenliefersystem MyBib eDoc

ImageWare Components GmbH | Am Hofgarten 20 | 53113 Bonn | Germany | www.imageware.de

Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen

Teil 1: Grundlagen des Wissensmanagements

Sonja Gust von Loh, Düsseldorf

Zentrale These des Artikels ist, dass Informationsbedarfsanalysen vor Einführung von Wissensmanagement insbesondere in kleinen und mittleren Unternehmen (KMU) zwingend durchzuführen sind, um zielgenau einzelne Bausteine implementieren zu können. Dazu werden Begrifflichkeiten und Theorien erläutert, die zu einer praktischen Anwendung notwendig sind und diese erleichtern können. Begrifflichkeiten, wie Daten, Information und Wissen müssen zwingend in ihren unterschiedlichen Meinungen erfasst werden. Auch der Unterschied zwischen Informations- und Wissensmanagement ist zu klären. Zentrale Bedeutung für Unternehmen hat das implizite Wissen. Dieses zeichnet sich dadurch aus, dass es nur schwer oder gar nicht in Sprache gefasst werden kann. Gerade das implizite Wissen macht die Umsetzung von Wissensmanagementansätzen in der Praxis so schwierig. Wie mit diesem eventuell umgegangen werden kann, zeigen Nonaka und Takeuchi mit ihrer Wissensspirale. Dieser asiatische Ansatz ist neben dem praktisch orientierten Ansatz von Probst et al., ebenfalls theoretische Grundlage des Wissensmanagements.

Knowledge Management and Information Needs Analysis in Small and Medium Sized Enterprises. Part 1: Foundations of Knowledge Management

The article's central thesis is that information needs analysis must always be done prior to the implementation of knowledge management in small and medium sized enterprises (SME). So the company can choose the best knowledge module as a starting point. We have to explain basic concepts and theories which are important for practical use. Concepts like Data, Information *and* Knowledge *as well as* Information Management *and* Knowledge Management *have different meanings. Tacit knowledge plays a key role in enterprises. This is a specific form of knowledge which can only difficultly or even not be put into words. Especially this implicit knowledge complicates Knowledge Management in the practice. How to handle this type of knowledge is explained by Nonaka and Takeuchi by means of the knowledge spiral. The practically oriented approach by Probst et al. is founded by a loop of knowledge management building blocks.*

Dieser Beitrag besteht aus zwei Teilen. Beide Teilartikel basieren auf praktischen Erfahrungen in einem kleinen und mittelständischen Unternehmen (KMU). Die Autorin hat dort von September 2006 bis Dezember 2006 eine Informationsbedarfsanalyse durchgeführt, um festzustellen, wo Schwachstellen in der Unternehmenskommunikation liegen. Diese Schwachstellen sollten durch Wissensmanagementtools im Idealfall ausgebessert bzw. behoben werden. Vorliegend handelt es sich um den theoretischen Teil der Forschungsarbeit und um Ergebnisse, die sich pauschalisieren lassen. Eingeteilt sind die beiden Artikel in eine allgemeine Studie über das Wissensmanagement (Teil 1) und einen über spezielle Thematiken bei KMU sowie darüber hinaus, wie Informationsbedarfsanalysen in das Wissensmanagement einzuordnen sind (Teil 2).

Einleitung

Im ersten Teil werden allgemeine Eigenschaften des Wissensmanagements erläutert. Zur Hinführung zum Thema dient die Begriffsklärung von Daten, Information und Wissen. Diese Begrifflichkeiten hängen unmittelbar mit dem Thema zusammen und sind deshalb von hoher Bedeutung. Nur so kann geklärt werden, was Wissensmanagement leistet.
Wissensmanagement sollte nicht mit Informationsmanagement verwechselt werden. Letzteres bezieht sich insbesondere auf Informationstechnologien und lässt die inhaltliche Komponente fast gänzlich außer Acht. Wissensmanagement hingegen legt den Fokus auch auf den Content, den Menschen und die Organisation, um sicherzustellen, dass die vorgeschlagenen Möglichkeiten auch angenommen werden.
Von besonderem Interesse für den humanen Aspekt ist das implizite Wissen, welches von Gilbert Ryle erstmals beschrieben und später von Polanyi weiter ausgeführt wurde. Das implizite Wissen ist genau der Punkt, der für das Wissensmanagement eine große Herausforderung darstellt. Implizite Wissensvorräte können gar nicht oder nur sehr schwer vermittelt werden, da sie teilweise nicht bewusst gewusst werden.
Auch verschiedene bereits bekannte Ansätze werden erläutert. Insbesondere der praktisch orientierte Ansatz von Gilbert Probst et al. ist interessant. Dieser unterteilt das Wissensmanagement in acht Bausteine und ermöglicht somit Firmen an bestimmten Punkten anzusetzen und eben nicht alles auf einmal einzuführen. Ein weiterer sehr bekannter Ansatz, der jedoch praktisch eher seltener in westlichen Ländern vorzufinden ist, ist der von Ikujiro Nonaka und Hirotaka Takeuchi. Diese beiden Asiaten führten die Wissensspirale ein, mit deren Hilfe man implizites Wissen, soweit es geht, explizieren kann.

1 Begriffsklärung: Daten, Information, Wissen

Zunächst sollen einige Grundbegriffe geklärt werden, die mit dem Wissensmanagement einhergehen. Sowohl Mitarbeiter, wie auch diejenigen, die Wissensmanagement einführen, müssen mit den Begriffen und ihren Bedeutungen vertraut sein. Leider gibt es nicht nur jeweils eine Definition von Daten, Information und Wissen. Vor allem im Bezug auf Information und Wissen gibt es eine Fülle von Auslegungen. An dieser Stelle sollen nur die aufgeführt werden, die für das Wissensmanagement am wichtigsten sind.

59(2008)2, 118-126 Information Wissenschaft & Praxis

In der Betriebswirtschaft liegt die Unterscheidung von Information und Wissen meist darin, dass letzteres zwingend an eine Person gebunden ist. Man unterscheidet in diesem Kontext zwischen individuellem und organisationalem Wissen (Brücher 2004, 11). Auch die hierarchische Anordnung von Daten, Information und Wissen – wie auf Abbildung 1 zu sehen – ist im Zusammenhang mit diesem Ansatz verbreitet.

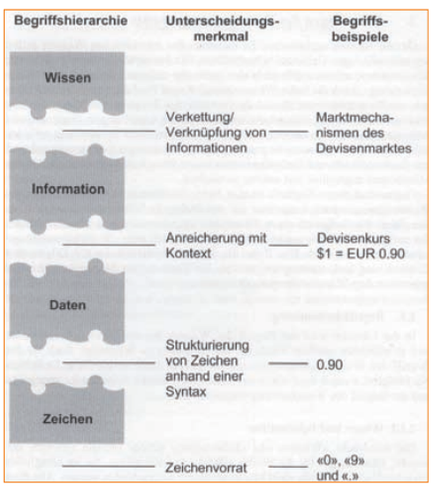


Abbildung 1: Zeichen-Daten-Information-Wissen (Quelle: Brücher 2004, 10).

Diese Trennung basiert auf einer Hierarchie und ist nicht immer haltbar, da die Grenzen zwischen den Begriffen fließend sind. Laut dieser Begriffsdefinition stehen *Zeichen* für den Zeichenvorrat, der beliebig angeordnet sein kann. Bei *Daten* handelt es sich um bereits strukturierte Zeichen. Das können alle Zeichen jeglicher Alphabete, Zahlen, Satzzeichen sowie Piktogramme sein (Wagner 1995, 15). Sie befinden sich immer auf einem Datenträger, wie z.B. auf einem Blatt Papier. Wenn diese Daten in einen Kontext gestellt werden, erhält man *Information*. Kommt es letztlich zur Verknüpfung von Informationen und zur intellektuellen Einbettung, so spricht man von Wissen. Geht man von dieser hierarchischen Strukturierung aus, so ist das Wissen tatsächlich an einen Menschen gebunden. Nur dieser verfügt über die Fähigkeit, Informationen intellektuell einzuordnen.
Informationen sind für den Empfänger meist zweckorientiertes Wissen, das im Idealfall zum Handeln motiviert. Es kann sich um neues Wissen handeln, kann aber auch bereits bekannt sein. Geht man davon aus, dass Informationen bereits bekannt sind, so handelt es sich meist um Bestätigungen von bereits Gewusstem. Eine Information ist immer eine Nachricht, die in der Regel schriftlich dokumentiert, akustisch oder visuell von einem Sender an einen Empfänger übertragen wird (Davenport 1998a, 29).
Für das Wissensmanagement sind die Informationen wichtig, die für das Unternehmen relevant sind. Mit Hilfe von Informationen sollen Mitarbeiter bestimmte Handlungen durchführen oder unterlassen; so wird deutlich, dass Informationen zweckorientiert sind. Um das zu erreichen, müssen immer wieder neue Informationen bereitgestellt werden. So wird Information zur betrieblichen Ressource und dient der Planung, Organisation und Kontrolle bestimmter Sachverhalte. In Unternehmen gewinnt diese Ressource zusehends an Bedeutung. Sie ist die Kenntnis über Sachverhalte, die ein Handelnder benötigt, um eine Entscheidung darüber zu fällen, wie er ein Ziel am günstigsten erreicht (Krcmar & Rehäuser 1996, 4) und bezieht sich damit direkt auf den Entscheider. Jedoch gibt es auch Ansätze, die Information für ein Wirtschaftsgut halten. So bezeichnet Frank Linde Information als Ware.

> Mit Hilfe von Software werden Informationsinhalte erstellt oder verarbeitet, die dann verkauft bzw. anderweitig – z. B. für unternehmensinterne Zwecke – genutzt werden können (Linde 2005, 11).

Information ist ein ökonomisches Gut, wenn sie verkauft oder gekauft wird, so z.B. der Kauf von Fachinformationen aus kommerziellen Datenbanken. Die Information ist hier nicht nur Ware, sondern wird auch zum Verbrauchsgut. Dieses ist durch Veränderungsraten gekennzeichnet, d.h. es ist schnell veraltet und muss ständig erneuert werden (ebd., 10). Eine weitere Besonderheit im Hinblick auf „Information als Ware" ist sicherlich, dass verkaufte Informationen im Gegensatz zu „greifbarer" Ware auch nach dem Verkauf noch zum Besitz des Verkäufers gehören (Stock 2000, 36f).
Unabhängig davon, ob Information als Ressource oder Ware gesehen wird, müssen die Mitarbeiter einer Firma die Information aufnehmen. Die Information wird dann in Wissen verdichtet. Dieser betriebswirtschaftliche Ansatz versucht – wie oben bereits erwähnt – Information und Wissen zu trennen bzw. Information in Wissen umzuwandeln.
Aus informationswissenschaftlicher Sicht lassen sich Wissen und Information jedoch nicht klar trennen. Rainer Kuhlen (1995, 34) stellt folgende These auf:

> Information ist Wissen in Aktion.

Er erläutert des Weiteren:

> Entsprechend interessieren aus informationswissenschaftlicher Perspektive Methoden, Verfahren, Systeme, Organisationsformen und deren jeweilige Rahmenbedingungen, mit deren Hilfe aus gesellschaftlich produziertem Wissen Information für aktuelle Problemlösungen erarbeitet bzw. aus Informationen neues Wissen produziert werden kann. Der Prozess der Erarbeitung von Information belässt Wissen nicht in seinem *Rohzustand*, vielmehr ist er als Transformations- oder, mit einer gewissen Bewertung, als Veredelungsprozess anzusehen... Diese Umwandlung von Wissen in Information nennen wir die Erzeugung informationeller Mehrwerte.

In diesem Kontext ist Wissen etwas Statisches, wohingegen Information etwas Dynamisches ist (Stock 2007, 23). Es beinhaltet Erfahrungswerte, Erkenntnisse und Schlussfolgerungen, die vor diesem Hintergrund gefällt werden. Es handelt sich um unbewusste, wie auch bewusste Schlüsse, die gezogen werden. Auch Davenport und Prusak vertreten den Standpunkt, dass Wissen eine Struktur aufweist, die es ermöglicht Informationen und neue Erfahrung in den bereits bestehenden Erfahrungshorizont einzuordnen.

> Knowledge is a fluid mix of framed experience, values, contextual information and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the mind of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices and norms (Davenport & Prusak 1998b, 5).

Auch anhand dieser Definition sieht man, dass Wissen zwar von einem einzelnen Individuum abhängig, jedoch auch in Arbeitsabläufen und Dokumenten wieder zu finden ist. Diese Definition ähnelt dem Ansatz von Stock.
Nach Stock kann Wissen in subjektives und objektives Wissen eingeteilt werden. Subjektives Wissen ist immer an eine Person gebunden und befindet sich in den Köpfen der Mitarbeiter. Das Wissen ist in diesem Zusammenhang die Fähigkeit, einen Gegenstand korrekt zu erfassen bzw. mit diesem korrekt umzugehen. Bei objektivem Wissen hingegen geht es um Wissen, das in Büchern oder Datenbanken „schlummert". Es wird dauerhaft auf einem Datenträger fixiert und handelt sich um das Gewusste an sich, den Inhalt und den sprachlichen Ausdruck. Auch wenn es sich um zwei Arten des Wissens handelt, so können das subjektive und das objektive Wissen nicht voneinander getrennt werden. Die beiden Wissensarten stehen in ständiger Wechselwirkung zueinander. Das subjektive orientiert sich am objektiven Wissen und umgekehrt (Stock 2007, 20). Dies soll im Folgenden an einem Beispiel verdeutlicht werden:
Ein Mitarbeiter benötigt eine wichtige Information, die er schließlich durch Recherche erhält. Sein subjektives Wissen

wird also durch objektives Wissen, nämlich das Wissen, welches sich z.B. in einem von ihm gelesenen Artikel befindet, beeinflusst. Das subjektive Wissen wird dann zum objektiven, wenn er sein Wissen wieder schriftlich für andere Mitarbeiter verfügbar macht.
Auch Probst liefert eine geeignete Definition von Wissen im Zusammenhang mit Wissensmanagement:

> Wissen bezeichnet die Gesamtheit der Kenntnisse und Fähigkeiten, die Individuen zur Lösung von Problemen einsetzen. Dies umfasst sowohl theoretische Erkenntnis als auch praktische Alltagsregeln und Handlungsanweisungen. Wissen stützt sich auf Daten und Information, ist im Gegensatz zu diesen jedoch immer an Personen gebunden. Es wird von Individuen konstruiert und repräsentiert deren Erwartungen über Ursache-Wirkungs-Zusammenhänge (Probst et al. 2006, 22).

Probst beschreibt hier das von Stock definierte subjektive Wissen.
Die Herausforderung für das Wissensmanagement ist es, das Wissen aus den Köpfen der Mitarbeiter für einen bestimmten Personenkreis zur Verfügung zu stellen. Individuelles Wissen soll also in kollektives Wissen umgewandelt werden. Kollektives oder auch organisationales Wissen ist immer auf die Mitglieder der Organisation zurückzuführen. Es entsteht durch sinnvolle Kombination einzelner Wissensbestände (Brücher 2004, 11). Es ist nicht nur die Summe des Wissens der einzelnen Mitarbeiter, sondern es geht um die effektive Kombination dieses Wissens (Probst et al. 2006, 21).
Abschließend ist festzuhalten, dass für Wissensmanagement die oben genannte Definition von Davenport und Prusak am vollständigsten ist, denn sie bezieht das subjektive Wissen mit ein, das an den Menschen gekoppelt ist. Trotzdem vernachlässigt sie nicht das objektive Wissen, das sich in Dokumenten wieder findet, aber auch durch Arbeitsabläufe und -prozesse widergespiegelt wird.

2 Wissensmanagement und Informationsmanagement – eine Einführung

Wissensmanagement oder englisch *Knowledge management* ist eine Disziplin, die Elemente der Betriebswirtschaftslehre, Wirtschaftsinformatik, und der Informationswissenschaft beinhaltet.
Schon immer waren Wissen und Informationen wichtige Ressourcen in der Ökonomie:

> Knowledge has always been implicated in the process of economic development; since anything we do, how we transform resources into products and services, crucially depends on the knowledge we have at our disposal for effecting such transformation. An ancient artisan, a medieval craftsman and his apprentices, and a modern manufacturing system all make use of knowledge: certain skills, techniques, and procedures are employed for getting things done (Tsoukas 2002, 2).

Jedoch wird Wissen immer wichtiger und somit findet auch Wissensmanagement mehr Zuspruch in Firmen. Mittlerweile ist auch das Interesse von kleinen und mittleren Unternehmen geweckt worden. Heute besitzt man durch die zunehmende Entstehung neuer Informationstechnologien die Möglichkeit, wissensbezogene Prozesse besser greifbar zu machen (Schneider 2001, 22). In der Informations- bzw. Wissensgesellschaft besteht ein wachsender Bedarf, die Ressource *Information* bzw. *Wissen* richtig einzuordnen. So beziehen sich Nonaka und Takeuchi auf das Werk von Drucker:

> In einer auf Wissen beruhenden Gesellschaft, so Drucker, ist der „Wissensarbeiter" die treibende Kraft. Zu seiner Definition des Wissensarbeiters gehört auch der Wissensmanager, der das Wissen im Sinne eines produktiven Nutzens zuteilen kann, so wie der Kapitalist das Kapital im Sinne eines produktiven Nutzen zuteilen konnte (Nonaka & Takeuchi 1997, 17).

Ebenfalls wichtig ist, dass die Mitarbeiter zwischen wichtiger Information und unwichtiger unterscheiden können. Auch Produkte und Dienstleistungen werden immer wissensintensiver und die Wissenskomplexität immer größer (Brücher 2004, 9). Durch die Globalisierung wird es immer wichtiger, Wissen schnell aufbereiten zu können. Vor allem für Großunternehmen, die international tätig sind, ist dieser Aspekt sehr bedeutsam. Sie müssen dafür sorgen, dass dieselben Informationen allen Mitarbeitern – überall auf der Welt – zur Verfügung stehen.

Eine einheitliche Informationsversorgung ermöglicht den Mitarbeitern den Arbeitsablauf nahezu identisch zu vollziehen (Schneider 2001, 23). Ziel des Wissensmanagements ist es, den Informationsfluss intern sowie extern zu optimieren, Wissen sinnvoll zu steuern und weiterzuentwickeln, aber auch neues Wissen zu generieren. *Good practices*, die bereits vorhanden sind, müssen kommuniziert und vor allem verständlich vermittelt werden (Dani et al. 2006, 1725). Krcmar beschreibt die Ziele wie folgt:

- Ausgleich von Informationsnachfrage und Informationsangebot
- Versorgung der Entscheidungsträger mit relevanten Informationen
- Gewährleistung einer hohen Informationsqualität. Informationen müssen gepflegt werden, so dass sie immer aktuell sind, und Informationsballast sollte vermieden werden
- Einsatz von Informations- und Kommunikationstechnik
- Zeitliche Optimierung der Informationsflüsse
- Beachtung des Wirtschaftlichkeitsprinzips (Krcmar 2005, 51).

Bei diesen Aspekten ist die Grenze zwischen Informationsmanagement und Wissensmanagement fließend. Beide Managementformen stehen miteinander in Verbindung. Abgrenzen lassen sich beide Theorien wie folgt: Das Informationsmanagement ist für die Bereitstellung von Informations- und Kommunikationstechnologien zuständig, wohingegen das Wissensmanagement für den Inhalt verantwortlich ist. Dem entgegengesetzt steht eine Theorie, die den Standpunkt vertritt, dass es Wissensmanagement als solches nicht geben kann. Der Vertreter dieses Ansatzes spricht davon, dass es ausschließlich Informationsmanagement geben kann, da Wissen nicht gemanagt werden kann. Der Terminus „Informationsmanagement" von Mujan wird nicht nur technisch eingeordnet, sondern es wird davon ausgegangen, dass das Management von Informationen bereits in dem Ausdruck enthalten ist.
Der Wortlaut deutet bereits darauf hin, dass nicht Informationstechnik sondern Information an sich zum Gegenstand des Managements wird (Mujan 2006, 17).
Dieser Ansatz ist haltbar, allerdings sollte der Begriff des Wissensmanagements nicht ganz vernachlässigt werden, denn er wird in der wirtschaftlichen Praxis genutzt. Es wird normalerweise klar zwischen Wissensmanagement und Informationsmanagement unterschieden.[1]
Wissensmanagement besteht aus insgesamt drei Elementen, aus denen sich das so genannte TOM-Modell zusammensetzt (Abbildung 2).

1) *Technik:* Es geht um Informations- und Kommunikationstechnologien, die ins Wissensmanagement einbezogen werden können, aber nicht zwangsläufig einbezogen werden müssen. Entscheidet man sich für eine Wissensdatenbank, so hat das Unternehmen die Chance zu wählen. Zum einen gibt es die Möglichkeit, die Datenbank in das bereits bestehende

1 Im Folgenden wird in der Regel von Wissensmanagement gesprochen; nur wenn es um die Informationstechnologie geht, werde ich den Terminus Informationsmanagement verwenden.

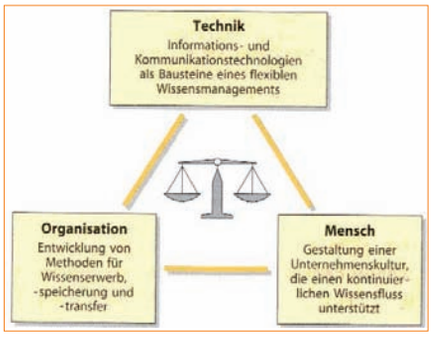


Abbildung 2: TOM-Modell (Quelle: Lucko & Trauner 2005, 24).

System einzubinden, oder man kauft Fertiglösungen. Diese Fertiglösungen werden in verschiedenen Formen angeboten und können auch bis zu einem gewissen Grad auf das jeweilige Unternehmen zugeschnitten werden. Beide Möglichkeiten haben Vor- und Nachteile. Einen Vorteil bringt meines Erachtens die Implementierung einer Datenbank in das bereits vorhandene System mit sich. Die Mitarbeiter kennen das System und sind nicht angehalten, sich eine neue Technik anzueignen.

2) *Organisation*[2]: Wissensmanagement muss in das Tagesgeschäft eingebunden werden. Es darf nicht etwas Besonderes sein, sondern es sollte als selbstverständlich angenommen und wie alle Vorgänge verinnerlicht werden. Das kann nur die Organisation bzw. das Unternehmen mit Hilfe der Unternehmenskultur den Mitarbeiter vermitteln.

3) *Mensch*: Natürlich wird die Unternehmenskultur von Menschen, nämlich bevorzugt von der Unternehmensleitung, festgelegt. Sie tragen Sorge dafür, dass Wissensmanagement bestehen kann und angenommen wird. Wissen muss in der Unternehmenskultur eine große Rolle spielen. Ein wichtiger Faktor ist somit auch der Mitarbeiter. Er muss bereit sein, sein Wissen zur Verfügung zu stellen und aktiv mit ihm zu arbeiten. Um dies zu gewährleisten, kann ein Unternehmen Anreizsysteme zur Hilfe nehmen. Jedoch verhindern diese Anreizsysteme es indirekt, dass Wissensarbeit zu etwas alltäglichem wird. Es bleibt immer etwas Besonderes. Letztlich ist der Mensch der wichtigste Faktor, da dieser die Ressource Wissen besitzt und nur er den kontinuierlichen Wissens- und Informationsfluss unterstützen kann (Lucko & Trauner 2005, 26). Eine große Rolle spielen hierbei die so genannten *Keyplayer*. Diese sollten wenn möglich entweder nicht das Unternehmen verlassen, oder zumindest ihr Wissen vor dem Verlassen an die Firma weitergeben. Dies ist vor allem bei KMU wichtig, da es hier oftmals nicht viele *Keyplayer* gibt.

2 Der Begriff „Organisation" bezeichnet stets Einheiten sowohl aus der Wirtschaft, der Verwaltung, dem kulturellen, sozialen und wissenschaftlichen Bereich, als auch Non-Profit Organisationen (Brücher 2004, 5).

3 Polanyi: Implizites Wissen

Zwei Arten des Wissens, die bewusst im Zusammenhang mit der Begriffsklärung nicht angesprochen wurden, sind das implizite und explizite Wissen. Dennoch dürfen diese beiden Wissensarten nicht vergessen werden. Im Gegenteil, gerade für ein Unternehmen ist das implizite Wissen besonders wichtig. Hier liegt letztlich auch die Schwierigkeit, Wissen zu bewahren und abrufbar zu machen.
Polanyi war in der Folge der Unterscheidung von Know-how und Know-that von Gilbert Ryle (1946), einer der ersten, der das Phänomen des impliziten und expliziten Wissens erfasste. Bereits 1967 sprach er von der *Tacit dimension* im Zusammenhang mit Wissen. Er stellte folgende These auf: Wir wissen mehr als wir zu sagen wissen (Polanyi 1985, 14).

Dieser Ansatz des impliziten Wissens oder auch *tacit knowledge* wurde von Wissenschaftlern verschiedener Fachrichtungen weiterentwickelt. An dieser Stelle soll zunächst die Theorie Polanyis näher betrachtet werden und dann Entwicklungen, die sich auf das Wissensmanagement beziehen. Das implizite Wissen ist das Wissen, das Mitarbeiter in den Köpfen besitzen, welches aber nicht ohne weiteres weitergegeben werden kann. Es kann nicht einfach dokumentiert werden, da es sich um Wissen aus z. B. langjähriger Erfahrung oder subjektiver Intuition handelt. Das Sprechen einer Fremdsprache könnte man als ein Beispiel für implizites Wissen nennen.

> Implizites Wissen besteht in der Fähigkeit, Distinktionen und Selektionen weitgehend intuitiv, laufend zu treffen sowie diese in praktische Handlungen umzusetzen (Schwaninger 2000, 4).

Es geht um Dinge, die nicht übersetzt und erfasst werden können. Implizites Wissen manifestiert sich in dem, was wir tun. Es geht weniger darum, stilles Wissen zu explizieren, als darum, dieses bei sozialen Interaktionen anzuwenden (Tsoukas 2002,1).

Im Gegensatz dazu steht das explizite Wissen. Hier geht es um Wissen, das ohne Probleme geäußert werden kann. Solch ein Wissen könnte z.B. der Inhalt eines Fachbuchs sein. Es versteht sich von selbst, dass das explizite Wissen auch wichtig ist, aber mit konsequenter Dokumentation und Pflege greifbar gemacht werden kann.

> Explizites Wissen umfasst Inhalte, die aufgrund bereits getroffener Unterscheidungen (Distinktionen) und durch Auswahlvorgänge (Selektionen) zustandegekommen sind (Schwaninger 2000, 3).

Polanyi versteht unter implizitem Wissen diejenigen Gegebenheiten, die man nicht ausdrücken kann, weil sie nicht bewusst gewusst werden. Er vertritt die Meinung,

> „dass jeder unserer Gedanken Komponenten umfasst, die wir nur mittelbar, nebenbei, unterhalb unseres eigentlichen Denkinhalts registrieren und dass alles Denken aus dieser Unterlage – die gleichsam ein Teil unseres Körpers ist – hervorgeht" (Polanyi 1985, 10).

Zur Verdeutlichung soll hier ein Beispiel angeführt werden: Beobachtet man einen Dieb bei einem Einbruch, so wird es einem mit hoher Wahrscheinlichkeit später schwer fallen, das Aussehen des Diebes in Worte zu fassen. Trotzdem ist man in der Lage, diesen Einbrecher wieder zu erkennen.

Es gibt umgekehrt auch den Fall, dass man praktisch etwas kann, es aber nicht theoretisch verbalisieren kann. Insbesondere im Zusammenhang mit verschiedenen Sportarten, wie z. B. Skilaufen, lässt sich dieses Phänomen beobachten. Man kann die Bewegung und wird sie auch nicht mehr verlernen. Es ist jedoch nicht möglich zu beschreiben, was genau getan wird. Das Ausüben der Bewegung ist zu einem unbewussten Vorgang geworden bzw. wurde automatisiert. Dem *Tacit knowledge* ist letztlich vieles von dem, was wir nicht bewusst tun, zuzuordnen.

> We engage in tacit knowing through virtually anything we do: we are normally unaware of the movement of our eye muscles when we observe, of the rules of language when we speak, of our bodily functions as we move around. Indeed, to a large extent, our daily life consists of a huge number of small details of which we tend to be focally unaware (Tsoukas 2002, 7).

Polanyi selbst spricht davon, dass Wissen im Allgemeinen immer zwei Aspekte besitzt. Zum einen handelt es sich um etwas Praktisches (Know-how) und zum anderen um etwas Theoretisches (Know-that) (Polanyi 1985, 16). Im Bezug auf das praktische Wissen geht es um die Anwendung, z.B. weiß ein Handwerker, wie er mit einem Hammer umgehen kann. Wenn man vom theoretischen Wissen spricht, so erkennt man einen Hammer und könnte im Idealfall eine Abhandlung über diesen schreiben, kann aber trotz des theoretischen Wissens nicht mit ihm umgehen.

Das implizite Wissen wird von Polanyi in eine zweigliedrige Grundstruktur eingeteilt. Er unterscheidet zwischen proximalem und distalem Term (ebd., 19). Zusammen ergeben die beiden Terme die funktionale Struktur des impliziten Wissens. Unter dem proximalen Term versteht er das Wissen, das uns zwar nah, aber nicht artikulierbar ist. Es handelt sich hier um einen Teil des impliziten Wissens, den wir eben nicht bewusst erfahren.

> Es ist dann der proximale Term, von dem wir ein Wissen haben, das wir nicht in Worte fassen können (Polanyi 1985, 19).

Der distale Term hingegen ist für uns weiter weg und sozusagen eine Folge des proximalen Terms. Er ist letztlich eine Konsequenz, die wir erfahren haben. Polanyi argumentiert ebenfalls, dass „wir bei einem Akt impliziten Wissens unsere Aufmerksamkeit von etwas auf etwas anderes verschieben, genauer gesagt: Vom ersten auf den zweiten Term jener stummen Relation“ (Polanyi 1985, 19). Der distale Term, welcher dem zweiten Term entspricht, steht im Zentrum, und zieht die ganze Aufmerksamkeit auf sich. Der proximale Term wird hingegen durch den distalen vernachlässigt.

Auch hier soll die Beziehung zwischen proximal und distal anhand eines Beispiels verdeutlicht werden. Angenommen ein Turmspringer springt eine Schraube vom Fünf-Meter-Brett. Er konzentriert sich insbesondere auf die Schraube, nicht aber auf einzelne Bewegungen, die er tätigt, damit die Schraube eine Schraube wird. Bleibt man bei Polanyis Worten, so wandert die Aufmerksamkeit von der einzelnen Bewegung auf die Schraube. Hier erkennt man also klar, dass die Bewegungen dem proximalen Term entsprechen, wohingegen die Schraube an sich dem distalen entspricht. Die Bewegungen sind uns näher und die Schraube an sich ist weiter von uns entfernt. Dadurch, dass die Durchführung der Schraube im Mittelpunkt steht, ist es nicht möglich die Bewegungsabläufe zu artikulieren.

> Allgemein lässt sich sagen, dass wir den proximalen Term eines Aktes impliziten Wissens im Lichte seines distalen Terms registrieren; wir wenden uns von etwas her etwas anderem zu und werden seiner im Lichte dieses anderen gewahr. Wir können dies die *phänomenale Struktur* des impliziten Wissens nennen (Polanyi 1985, 20).

Das heißt, dass wir den proximalen Term zwar wahrnehmen, aber letztlich erst in der Handlung bzw. durch den distalen Term. Wichtig ist festzuhalten, dass der proximale Term nicht bewusst wahrgenommen wird.

Für das Wissensmanagement ist von Bedeutung, wie man implizites Wissen sichtbar machen bzw. vermitteln kann. Nach Polanyi muss man sich dazu in einen anderen Menschen hineinversetzen.

> Die Einfühlung, (…) ist ein viel genauer definierter Akt als Empathie und ist die Grundlage aller Beobachtungen, einschließlich aller bisher als Einfühlung beschriebener Beobachtungen (Polanyi 1985, 24).

Bei Weltanschauungen sollten proximale Terme so sehr verinnerlicht werden, dass sie für einen selbst ebenfalls proximal werden. Implizites Wissen kann nur durch Üben und Imitationen entwickelt werden, nicht ausschließlich über Beschreibungen.
Die Sichtweise des Wissensträgers (Lehrer) kann auf zwei verschiedene Arten von einer weiteren Person (Schüler) absorbiert werden. Zum einen auf der körperlichen Ebene und zum anderen auf der geistigen. Polanyi spricht von einer Person, die ihren Körper einsetzt und von einer, die den Geist einsetzt (Polanyi 1985, 34). Bei beiden Möglichkeiten muss der Lernende sich in den Lehrenden hineinversetzen. Für eine Vermittlung auf geistiger Ebene führt Polanyi einen Schachspieler mit seinem Schüler an.
Um in den Geist des Lehrers „einzudringen“ bzw. sein Handeln nachvollziehen zu können, müssen Anhaltspunkte gesammelt werden. Eine solche Untersuchung besteht darin, Anhaltspunkte zu sammeln, also Merkmale, die mutmaßlich mit etwas in Zusammenhang stehen, auf dessen Dasein sie hindeuten (Polanyi 1985, 34).

Diese Anhaltspunkte bleiben in der Regel unspezifisch, werden jedoch trotzdem aufgenommen. Auch wenn, wie oben bereits geschildert, die betriebswirtschaftliche Sicht der Dinge eine klare Trennung zwischen explizitem und implizitem Wissen vorsieht, so ist sie im Ansatz von Polanyi nicht haltbar. Letztlich fußt alles auf impliziten Wissensstrukturen. Das explizite Wissen kann nicht ohne das implizite existieren. Um handlungswirksam zu werden, bedarf es der Kombination beider Strukturen

> Die Geschicklichkeit eines Fahrers lässt sich durch keine noch so gründliche Schulung in der Theorie des Kraftfahrzeugs wettmachen (Polanyi 1985, 27).

Dies ist auch in der Mathematiktheorie der Fall. Sie wird nur verstanden, wenn sie auch praktisch angewendet werden kann. Sie wird vom Lernenden eingeordnet, jedoch mit bestimmten Hintergründen. Laut Polanyi ist es anders nicht möglich.

So erweist sich das Ideal einer umfassenden mathematischen Theorie der Erfahrung, in der alles implizite Wissen verschwände als selbstwidersprüchlich und logisch anfechtbar (Polanyi 1985, 28).
Im Gegenzug ist jedoch die Existenz des impliziten ohne das explizite Wissen möglich. Nicht alles, was man praktisch anwenden kann, muss mit theoretischem Wissen untermauert sein. Man kann z.B. einen Herd bedienen, auch wenn man nicht weiß, wie er funktioniert und physikalisch aufgebaut ist.
Wie der Ansatz von Polanyi gut zeigt, kann implizites Wissen eben nicht ausschließlich über eine Datenbank vermittelt werden. Es geht hier auch um zwischenmenschliche Zusammenarbeit. An diesem Punkt sind die Teamfähigkeit eines Mitarbeiters und seine *Softskills* gefragt.
Janik entwickelte Polanyis Theorie folgendermaßen weiter (Abbildung 3).

Things not put into word	Things inexpressible in words
Trade secrets	Knowledge by acquaintance or familiarity, e.g. sounds, smells
Things overlooked e. g. craft knowledge/skill	the open textured character of rule-following” acquired through practice
presuppositions	

Abbildung. 3: Verschiedene Typen des Tacit Knowledge (Quelle: Janik 1988, 56).

Er teilt das „stille Wissen“ – ebenso wie Polanyi – in zwei Aspekte auf, jedoch unter anderen Gesichtspunkten. Auf der linken Seite der Abbildung wäre es theoretisch möglich, die Dinge beim Namen zu nennen. Janik erklärt, dass Mitarbeiter entweder Wissen nicht weitergeben wollen oder nicht wissen, dass es sich dabei um essentielles Wissen für eine andere Person handelt. Bei den Dingen, die sich nicht artikulieren lassen, gibt es zwei Möglichkeiten. Entweder wird etwas wirklich nicht bewusst gewusst oder es gibt keinen Weg dieses bestimmte Wissen zu beschreiben. Es wird z.B. schwierig sein darzustellen, wie Kaffee riecht. Man muss wissen, wie er riecht, andernfalls kann man mit der Beschreibung nichts anfangen. Streng genommen entspricht nur das Wissen auf der rechten Seite der Abbildung den Vorstellungen Polanyis.
Die Unterscheidung, die in Abbildung 4 gemacht wird, ist nachvollziehbar, weil beide Arten des stillen Wissens nicht geäußert werden und somit nicht weitergegeben werden. Allerdings würde ich die Dinge, die nicht gesagt werden sollen, als explizites, stilles Wissen beschreiben.

Es könnte theoretisch ausgedrückt werden, wird aber nicht in Worte gefasst, um einen gewissen Vorteil gegenüber Anderen zu erhalten. Dieses explizite Wissen, welches Janik als implizites Wissen beschreibt, kann ggf. mittels Anreizsystemen dem Einzelnen entlockt werden. Implizit ist es in sofern, als dass dem Unternehmen dieses Wissen nicht zur Verfügung steht.
Ein weiterer betriebswirtschaftlicher Ansatz unterscheidet klar zwischen implizitem Wissen und *Tacit Knowledge*.

> Implizites Wissen: umfasst Formeln von Wissen (…) zur Lösung eines anliegenden Problems, die von Personen getragen werden, aber im Gegensatz zu explizitem Wissen nicht artikuliert sind.
>
> Tacit Knowledge: umfasst als eine Teilmenge des impliziten Wissens Formen von Wissen, die nicht artikuliert sind, weil entweder die Problemrelevanz der Wissenskomponenten nicht bewusst ist, Artikulation keine effiziente Form der Übertragung ist oder es sowohl an Problembewusstsein als auch an Artikulationseffizienz mangelt (Rüdiger & Vanini 1998, 472).

In dieser Argumentation wird das implizite Wissen als ein Oberbegriff des *Tacit knowledge* verstanden. Bei implizitem Wissen geht es darum, dass etwas nicht artikuliert werden kann. Beim *Tacit knowledge* kommt noch das fehlende Bewusstsein für die Wichtigkeit der Artikulation hinzu. Der im Zusammenhang mit Wissensmanagement geläufigere Ansatz ist aber, dass implizites Wissen und *Tacit Knowledge* synonym behandelt werden.

> In the KM literature the term implicit knowledge is often treated as synonymous with Michael Polanyi's term tacit knowledge, but then both are further dealt with as a form of private knowledge that is treated as „informal" and even, in a sense, "unconscious" knowledge (Day 2005, 630).

In einem Unternehmen ist implizites Wissen nicht ausschließlich in den Köpfen der Mitarbeiter zu finden, sondern es befindet sich auch in Arbeits- und anderen organisationalen Routinen (Wagner, o. J.). Deswegen kann auf dem Weg der Imitation bzw. Zusammenarbeit viel gelernt und vermittelt werden. Letztlich wird es aber nie möglich sein, das komplette implizite Wissen explizit zu machen. Es wird immer bis zu einem gewissen Teil an bestimmte Personen gebunden sein. Hier ist es für das Unternehmen wichtig, Schlüsselpersonen zu lokalisieren und sie im Betrieb zu halten. Im Idealfall bilden diese dann durch enge Zusammenarbeit aktiv die Nachfolger aus.

4 Der Ansatz von Nonaka und Takeuchi

Insbesondere in den 1990er Jahren ist das implizite Wissen im Zusammenhang mit betrieblichen Aspekten sehr bekannt geworden. Der vermeintlich bekannteste Ansatz, der versucht, die Theorie Polanyis praktisch umzusetzen, ist der von Ikujiro Nonaka und Hirotaka Takeuchi. Im Gegensatz zu westlichen liegt der Fokus bei asiatischen Unternehmen nicht auf dem expliziten Wissen, sondern auf dem impliziten.
Wie Polanyi unterscheiden auch Nonaka und Takeuchi zwischen implizitem und explizitem Wissen. Das implizite Wissen verfügt in ihrer Theorie über weitere zwei Dimensionen. Zunächst gibt es die informelle, technische Dimension, die sich mit schwer definierbaren Fertigkeiten beschäftigt. Sie steht sozusagen für das Know-how eines jeden. Des Weiteren gibt es noch die kognitive Dimension. Hier geht es um mentale Modelle und Vorstellungen, die vom Menschen für selbstverständlich gehalten werden (Nonaka & Takeuchi 1997, 19). Mittels dieser Modelle ist es dem Menschen möglich, sich in der Welt zurechtzufinden (Nonaka & Takeuchi 1997, 72). Implizites und explizites Wissen werden in ihrem Ansatz als weitgehend komplementär angesehen.

Implizites Wissen (subjektiv)	Explizites Wissen (objektiv)
Erfahrungswissen (Körper)	Verstandeswissen (Geist)
Gleichzeitiges Wissen (hier und jetzt)	Sequentielles Wissen (da und damals)
Anloges Wissen (Praxis)	Digitales Wissen (Theorie)

Abbildung 4: Die verschiedenen Arten des Wissens nach Nonaka und Takeuchi (1997, 73).

Das Ziel von Nonaka und Takeuchi ist es, implizites in explizites Wissen zu transformieren und umgekehrt. Dies geschieht mit Hilfe der so genannten Wissensspirale.

> Our dynamic model of knowledge creation is anchored to a critical assumption that human knowledge is created and expanded through social interaction between tacit knowledge and explicit knowledge. We call this interaction "knowledge conversion" (Nonaka & Takeuchi 1995, 61).

Besonders die soziale Interaktion findet in diesem Modell Berücksichtigung. Nur durch diese ist es möglich, jemand anderem implizites Wissen zu vermitteln.
Wie in Abbildung 5 deutlich wird, kann man Wissensumwandlung in vier verschiedene Formen einteilen. Mit den abgebildeten Bausteinen arbeitet dann die Wissensspirale.

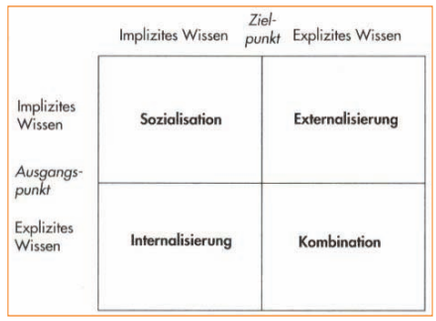


Abbildung 5: Vier Formen der Wissensumwandlung (Quelle: Nonaka & Takeuchi 1997, 75).

Bei der Sozialisation geht es darum, implizites Wissen von einem Individuum zu implizitem Wissen eines anderen umzuwandeln. Die implizite Ebene wird hierbei nicht verlassen (Schreyögg & Geiger 2004, 273). Diese Vermittlung erfolgt ohne Anwendung von Sprache, da man – wie es auch Polanyi schon festgestellt hat – implizites Wissen nicht über Sprache vermitteln kann. Die Weitergabe erfolgt über Beobachtung, Imitation und – um bei Polanyi zu bleiben – durch Einfühlung. Von Vorteil ist, wenn die Kollegen, die ihr implizites Wissen teilen sollen, über einen gewissen Rahmen gemeinsamer Erfahrungen verfügen. Ist gar kein gemeinsamer Erfahrungshintergrund vorhanden, wird die Vermittlung sehr schwierig werden.

> Without some form of shared experience, it is extremely difficult for people to share each others' thinking process (Nonaka 1994, 19).

Nonaka bezeichnet die Sozialisation als einen Prozess, der durch gemeinsame Erfahrungen ermöglicht wird (Nonaka 1994, 19). Ziel der Sozialisation ist es, gemeinsame mentale Modelle und technische Fertigkeiten zu entwickeln. Ein bekanntes Beispiel von Nonaka und Takeuchi ist das des Brotbackautomaten. Die Firma Matsushita in Osaka wollte den ersten Brotbackautomaten entwerfen, der auch für den Hausgebrauch geeignet ist. Der wichtigste Arbeitsvorgang ist der Knetprozess. Nur wenn ein optimaler Knetprozess vorhanden ist, kann auch gutes Brot gebacken werden. Es ergab sich folgendes Problem: Der Knetprozess gehörte zum impliziten Wissensvorrat des Bäckermeisters. Durch eine Analyse des Teiges mit Hilfe von Röntgenstrahlen versuchte man, den Teig mit einem aus dem Automaten zu vergleichen. Dies lieferte jedoch keine Erkenntnis. Letztlich kam die Firma auf die Idee, dass die einzige Möglichkeit, an das implizite Wissen eines Bäckermeisters zu kommen, darin bestand, bei ihm in die Lehre zu gehen. Deswegen gingen die Softwareentwicklerin Ikuko Tanaka und mehrere Ingenieure bei dem Bäcker-

meister, der das beste Brot backte, in die Lehre. Allein durch Beobachtung und Nachahmung gelang es ihnen festzustellen, dass der Teig nicht nur gedehnt, sondern auch gedreht wurde (Nonaka & Takeuchi 1997, 114 ff.).
Die Externalisierung beschäftigt sich mit der Veräußerung des erworbenen impliziten Wissens. Das durch die Sozialisation erworbene Wissen muss – um es weitergeben zu können – artikuliert werden. Da das nicht ohne weiteres möglich ist, sollen hier Metaphern und Analogien helfen. Metaphern stehen im Zusammenhang mit intuitivem Lernen und mit Vorstellungen. Eine Metapher lässt sich gut auf einen anderen Sachverhalt übertragen. So teilt Nonaka bereits 1994 mit, wie wichtig der Aspekt der Metapher ist.

> As a method of perception, metaphor depends on imagination and intuitive learning through symbols, rather than on the analysis or synthesis of common attributes shared by associated things (Nonaka 1994, 21).

Eine Analogie wird von Nonaka und Takeuchi im Gegensatz dazu wie folgt beschrieben:

> Eine Analogie erlaubt im Vergleich zu einer Metapher eine viel strukturierter Differenzierung. Sie klärt, wo die Ähnlichkeiten und Unterschiede zwischen Ideen oder Gegenständen liegen. In diesem Sinne erscheint die Analogie als ein Stadium zwischen reiner Phantasie und logischem Denken (Nonaka & Takeuchi 1997, 24).

Betrachtet man diese beiden „Werkzeuge", kann man von einem zweistufigen Prozess sprechen. Metaphern werden durch freie Assoziationen gefunden und ermöglichen einen ersten sprachlichen Zugang. Anschließend müssen Analogien zwischen den entstandenen Metaphern und bereits bekanntem Wissen hergestellt werden. Das implizite Wissen kann so in explizites Wissen transformiert werden (Schreyögg & Geiger 2004, 274).
Bei der Kombination geht es um die sinnvolle Zusammenfügung von verschiedenen expliziten Wissensbestandteilen. Das durch die Externalisierung entstandene Wissen wird mittels verschiedener Medien, wie z.B. Datenbanken, Dokumente oder auch Gesprächen und persönliche Treffen, kombiniert.

Der letzte Baustein ist die Internalisierung. Ihr Ziel ist es, explizites zu implizitem Wissen zu transformieren. Nach einer Versprachlichung kann es von einem weiteren Mitarbeiter vollständig verinnerlicht werden. Erst durch den alltäglichen Umgang kann explizites in implizites Wissen umgewandelt werden. Solch eine Internalisierung ist das Ziel eines jeden Lernprozesses.
Wichtig an dem Modell der Wissensspirale (s. Abbildung 6) ist, dass die Bausteine in Zusammenhang miteinander stehen.

> While each of the four modes of knowledge conversion can create new knowledge independently, the central theme of the model of organisational knowledge creation proposed here hinges on a dynamic interaction between the different modes of knowledge conversion (Nonaka 1994, 20).

Die Spirale lässt sich in insgesamt fünf Phasen einteilen (Tsoukas 2002, 10):

1. : Teilen von implizitem Wissen durch Mitglieder der Organisation,
2. : Ermittlung von Konzepten,
3. : Rechtfertigung der Konzepte,
4. : Bildung einer Urform: Das bereits für sinnvoll erachtete Konzept wird in den Archetyp umgewandelt,
5. : *Cross-Level* des Wissens: Der Archetyp wird anderen Organisationsteilen oder auch außerhalb der Organisation zur Verfügung gestellt. Es können neue Wissensspiralen entstehen.

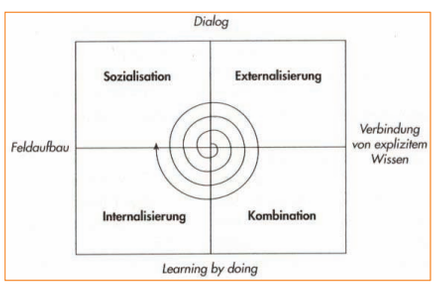


Abbildung 6: Die Wissensspirale (Quelle: Nonaka & Takeuchi 1997, 84).

Dieser Spiralprozess sieht eine Kollektivierung vor. Das implizite Wissen, welches zunächst von einem Einzelnen getragen wird, soll weiteren Mitarbeitern zur Verfügung gestellt werden. Nonaka und Takeuchi sprechen von einem Wissensnetz des Unternehmens, das erschaffen werden soll.

> Wissensschaffung im Unternehmen muss daher als Prozess verstanden werden, der das von einzelnen erzeugte Wissen verstärkt und es im Wissensnetz des Unternehmens verankert (Nonaka & Takeuchi 1997, 71).

Auf den ersten Blick klingt der Ansatz plausibel und durchführbar. Allerdings ist fraglich, ob sich implizites Wissen wirklich artikulieren bzw. sich in explizites Wissen umwandeln lässt. Polanyi macht deutlich, dass eine Externalisierung nicht möglich ist. Ob Metaphern und Analogien dabei wirklich helfen, bleibt meiner Meinung nach fraglich. Implizites Wissen kann nicht in explizites überführt werden, sondern ist höchstens ein Teil des letzteren.
Im Hinblick auf die Sozialisation und die Internalisierung sollte es keine großen Schwierigkeiten geben. Der Ansatz dieser beiden Bausteine lässt sich gut mit dem von Polanyi vereinbaren. Auch eine Kombination ist ohne weiteres möglich und wird in westlichen Unternehmen meist schon angewandt.

5 Der Ansatz von Gilbert Probst et al.

In dem Werk „Wissen managen" liegt das Augenmerk auf explizitem Wissen und seiner Konservierung. Diese Theorie basiert auf dem typischen betriebswirtschaftlichen Regelkreis, der Ziele, Umsetzung und Kontrolle beinhaltet. Um diese Aufteilung noch zu verstärken, teilen Probst et al. ihre Art Wissen zu managen in insgesamt acht Wissensbausteine auf. Diese Bausteine gliedern sich in einen inneren und einen äußeren Wissenskreislauf. Sechs davon liegen innerhalb, wohingegen zwei außerhalb liegen. Insgesamt ergeben die sechs Kern- und zwei Außenprozesse einen interdependenten Managementregelkreis (Rohleder 2004).
Der Ansatz ist sehr praktisch orientiert; daher bietet das Werk Managern von Unternehmen einen Leitfaden zur Einführung von Wissensmanagement.
Ausgangspunkt ist die Überlegung, dass ein Unternehmen aus individuellem und kollektivem Wissen besteht, welches zusammengefügt werden sollte. Individuelles Wissen und Fähigkeiten stellen für das kollektive Wissen die Basis dar.

> Die individuellen Fähigkeiten von Wissensarbeitern sind eine grundlegende Basis für das erfolgreiche Agieren von Unternehmen. Darüber hinaus hängt das Gelingen vieler Projekte und Strategien jedoch entscheidend davon ab, ob verschiedene Wissensbestandteile und Wissensträger effizient kombiniert werden können (Probst et al. 2006, 20).

Abbildung 7 zeigt die einzelnen Wissensbausteine, die für die Optimierung des Wissens- und Informationsvorrats in einem Unternehmen sorgen sollen.

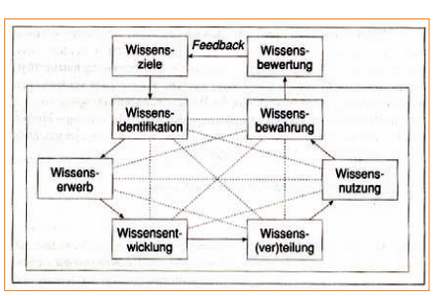


Abbildung 7: Verschiedene Wissensbausteine nach Probst et al. (Quelle: Probst et al. 2006, 32).

Die Wissensbausteine „Wissensziele“ und „Wissensbewertung“ liegen außerhalb des inneren Kreislaufs. Ein Unternehmen muss zunächst Wissensziele haben, um sein Wissen optimieren zu können. Diese müssen klar formuliert werden. Im Hinblick auf diese Ziele unterscheidet man zwischen normativen, strategischen und operativen Wissenszielen. Erstgenannte beziehen sich auf die Gesamtunternehmensebene (*corporate strategy*) und die Geschäftsbereichsebenen (*business strategy*), wozu beispielsweise die Erschließung von Zielgruppen oder auch von Märkten gehört.

Normative[3] Wissensziele stehen für die unternehmenspolitischen und -kulturellen „Leitbilder“. Hier muss zunächst die aktive Arbeit mit Wissen beschlossen und von den Chefetagen vorgelebt werden. Grundlegende Voraussetzung für ein an Wissenszielen orientiertes Management ist die Grundeinstellung, dass Wissen eine zentrale Größe für den Unternehmenserfolg darstellt (Probst et al. 2006, 41). Die Festlegung der normativen Wissensziele erfolgt durch die Führungskräfte und ist Voraussetzung für die Etablierung von Wissensmanagement bzw. für die Ziele im strategischen und operativen Bereich.

Die nächste Ebene stellt die strategischen[4] Wissensziele dar. Sie stehen letztlich für die Erstellung eines in der Zukunft angestrebten Fähigkeitsportfolio und liefern damit eine inhaltliche Bestimmung des organisationalen Kernwissens (Probst et al. 2006, 49). Als letzten Aspekt bieten sie eine strategische Orientierung von Organisationsstrukturen und Managementsystemen. Mittels dieses Wissens kann Kernwissen entwickelt werden und Wissen als Basisfaktor genutzt werden.

Die letzte Ebene der Wissensziele ist die operative[5]. An dieser Stelle wird das Wissensmanagement aktiv umgesetzt und die normativen und strategischen Ziele werden in konkrete operationalisierbare Teilziele umgewandelt.

Der zweite Außenprozess ist die Wissensbewertung. Hier geht es darum, die Effizienz und den Erfolg der eingeführten Maßnahmen zu überwachen. Durch die Wissensmessung soll sichtbar gemacht werden, was sich wohingehend verändert hat und was die entstandenen Vorteile sind. Letztlich dient dieser Baustein dazu, festzustellen, inwiefern das angewandte Wissensmanagement zielführend ist. Da Wissen keine messbare Ressource ist, erweist sich eine Bewertung als schwierig. Dennoch gibt es verschiedene Methoden, die die Messung von normativen, strategischen und operativen Wissenszielen ermöglichen. Durch den Baustein „Wissensbewertung“ werden Ziele unter Umständen umstrukturiert bzw. verändert.

Die anderen sechs Bausteine gehören zu den Kernprozessen der Theorie. Bei der Wissensidentifikation geht es darum, interne und externe Transparenz von vorhandenem Wissen zu schaffen. Diese Transparenz bezieht sich zunächst auf die personelle Seite. Das bedeutet, dass im Idealfall das Wissen eines Einzelnen zunächst identifiziert und im nächsten Schritt mehreren Personen zur Verfügung gestellt wird. Dadurch hat dieser Personenkreis einen Nutzen aus dem Wissen gezogen. Eine gewisse Transparenz muss immer feststellbar sein, da mangelnde Transparenz zu Ineffizienz und zu Fehlentscheidungen seitens der Mitarbeiter führt. Durch die entstandene Offenheit wird der einzelne Mitarbeiter in seinem Handeln unterstützt und Wissenslücken können geschlossen werden. Mittels dieses Bausteins wird deutlich, welches Know-how bei wem in der Firma liegt, aber auch, wo Informationen fehlen. Häufig stellen Unternehmen fest, dass nicht bekannt ist, welches Wissen im Unternehmen verfügbar ist. Unklarheit herrscht häufig darüber, welche Person welche Fähigkeit besitzt und welche Erfahrung wo schon gemacht wurde (Probst et al. 2002, 4). Abschließend lässt sich sagen, dass alle Maßnahmen folglich darauf ausgerichtet sind, dem Unternehmen einen besseren Überblick über interne und externe Wissensbestände zu verschaffen. Dadurch kann eine Grundlage für weitere Fortschritte in den übrigen Bausteinen geschaffen werden (Probst et al. 2002, 4). Durch die Wissensidentifikation kommt es also bis zu einem gewissen Grad zum Wissenserwerb. Der Baustein „Wissenserwerb“ steht klar im Zusammenhang damit, wo bestimmtes Wissen ermittelt wird und wo man fehlende Information bekommen kann. Wissensmärkte gewinnen an Bedeutung, da ein erheblicher Teil des Wissens auch aus Quellen kommt, die außerhalb des Unternehmens liegen. Durch Beziehungen zum Kunden, aber auch durch Wissen über Konkurrenten ist es möglich, einen klaren Wettbewerbsvorteil zu erlangen. Es geht bei diesem Baustein um den Erwerb von Wissen externer Wissensträger, von Wissen anderer Firmen, von Stakeholderwissen oder von Wissensprodukten (Probst et al. 2006, 93).

Abbildung 8 zeigt einige mögliche Quellen, über die Wissen für das Unternehmen kommen kann. Es ist selbstverständlich, dass dieses Wissen erworben werden muss. Deswegen sollte auch die Zusammenarbeit mit externen Partnern gefördert bzw. neue Informationsquellen aufgetan werden.

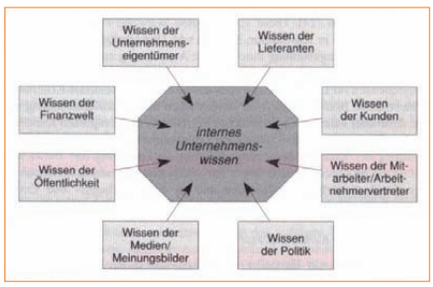


Abbildung 8: Verschiedene Wissensquellen eines Unternehmens (Quelle: Probst et al. 2006, 103).

Der Baustein „Wissensentwicklung“ ist für die Schließung von Wissenslücken zuständig, aber auch für die Entwicklung neuen Wissens. Hierbei handelt es sich um einen komplementären Baustein des „Wissenserwerbs“. Insbesondere im Hinblick auf Innovationen ist dieser Baustein sehr wichtig. Eine Erneuerung des Produkts erfordert neues Wissen. Oftmals ist es jedoch so, dass sich Wissen z.B. durch die Entwicklung eines Produktes automatisch weiterentwickelt und sich kaum steuern lässt. Meist ist es jedoch möglich, dem Ganzen zumindest eine Richtung zu weisen. Es geht bei diesem Baustein neben der eigentlichen Wissensentwicklung auch darum, gewisse Instrumente oder auch Möglichkeiten den Mitarbeitern bereitzustellen. Dort sollen sie verschiedene Ansätze und neues Wissen austauschen und miteinander beratschlagen, was sinnvoll ist und was nicht. Ein Beispiel hierfür ist die Errichtung einer Kaffee-Ecke. Auch durch zunächst informelle Gespräche kann ein reger Ideenaustausch stattfinden. Bei den Bausteinen „Wissen erwerben und entwickeln“ geht es in erster Linie darum, neue Fähigkeiten, neue Produkte, bessere Ideen und leistungsfähigere Prozesse zu gewinnen (Probst et al. 2002 ,4).

Um die Wissensentwicklung voll ausnutzen zu können, muss geklärt werden, wie Wissen sinnvoll an die jeweils zuständigen Mitarbeiter gebracht werden kann. Dafür ist der Baustein „Wissens(ver)teilung“ verantwortlich. Hier geht es um die sinnvolle Teilung des Wissens, also um die Bestimmung, wer welches Wissen benötigt und wer nicht. Neben dem so genannten „Verteilen“ spielt aber auch das „Mitteilen“ von Wissen unter Individuen bzw. im Rahmen von Teams und Arbeitsgruppen eine große Rolle (Probst et al. 2006, 142). Es muss einerseits die zentrale Verteilung sichergestellt werden, andererseits auch der Wissensaustausch durch Zusammenarbeit gegeben sein. Vor allem wenn eine Firma bevorzugt in Gruppen oder Teams arbeitet, ist dieser Baustein wichtig.

> Die Aufgaben der Wissens(ver)teilung lassen sich schematisch in drei Gebiete einteilen. (1) Die Multiplikation von Wissen durch rasche Verteilung

3 Normativ bedeutet allgemein normgebend, also steht es für das Aufstellen von Regeln und Normen.

4 Eine Strategie ist ein längerfristig ausgerichtetes planvolles Anstreben eines Ziels.

5 Operativ steht in diesem Zusammenhang für die konkrete Durchführung.

auf eine Vielzahl von Mitarbeitern. (2) Die Sicherung und Teilung vergangener Erfahrung und (3) den simultanen Wissensaustausch, der in die Entwicklung neuen Wissens mündet (Probst et al. 2006, 170).

Der Baustein „Wissensnutzung" stellt die Art und Weise der Anwendung sicher. Oft wird die Nutzung fremden Wissens durch Barrieren beschränkt, muss aber trotzdem sichergestellt werden. Eine mögliche Barriere kann z.B. das Misstrauen gegenüber fremden Wissens sein. Nur wenn dieser Baustein erfolgreich angewandt wird, ist es sinnvoll, Wissen zu erwerben, zu entwickeln und anzusammeln. Generell kann die Wissensnutzung durch die nutzungsorientierte Gestaltung der Arbeitssituation verbessert werden. Wissen muss an dieser Stelle in konkrete Resultate umgewandelt werden. Die Wissensnutzung ist sozusagen das Kernelement des Wissensmanagements. Die anderen Bausteine benötigt man zur Optimierung dieses Bausteins.
Der letzte Punkt, der zu den Kernprozessen gehört, ist das Bewahren von Wissen. Hier geht es um die zielführende Speicherung von Wissen. Abbildung 9 zeigt die entscheidenden Schritte, die unmittelbar im Zusammenhang mit diesem Baustein stehen.

Abbildung 9: Hauptprozesse der Wissensbewahrung (Quelle: Probst et al. 2006, 193).

Das Selegieren steht für die Auswahl von Wissen. Es geht darum, welches Wissen benötigt wird und welches nicht. Durch diesen Punkt kann eine Informationsflut vermieden werden. Nicht jeder kleinste Bericht ist für die Firma wichtig und muss abgespeichert werden. Fragen, die man sich im Zusammenhang mit der Wissensbewahrung stellen muss, sind:

- Was passiert mit den Erfahrungen eines Mitarbeiters, wenn er von heute auf morgen die Firma verlässt?
- Wer findet die zentralen Dokumente oder Präsentationen auf der mehr oder weniger gut organisierten Festplatte wieder?
- Sind die wichtigen Kontaktpersonen und Prozesse der Position dokumentiert? (Probst et al. 2006, 193f)

Bei dem Aspekt des Speicherns wurde bereits entschieden, welche Information gespeichert werden soll. Damit aber die Wissensbasis auch immer aktuell gehalten wird, muss sie in regelmäßigen Abständen aktualisiert werden. Gewisse Informationen veralten und werden dadurch unbrauchbar. Diese Informationen können gelöscht oder zumindest archiviert werden. Durch einen sinnvollen Aufbau der Wissensbewahrung kann es zur Bildung eines so genannten kollektiven Gedächtnisses kommen. Im Idealfall entstehen kaum Probleme, wenn Mitarbeiter des Unternehmens unerwartet die Firma verlassen.
Probst et al. machen in ihren Ausführungen deutlich, dass Wissensmanagement mit Humanresources unweigerlich zusammenhängt. Aufgrund dessen vertreten sie den Standpunkt, dass Wissensmanagement nicht ausschließlich etwas mit Informationstechnologien zu tun hat, sondern auch eine zentrale Schnittmenge mit dem Personalmanagement besitzt. Wissensmanagement kann Personalmanagement in zentralen Punkten ergänzen (Probst et al. 2002, 1).

Zusammenfassung

In diesem Artikel wird ein grober Überblick über das Wissensmanagement im Allgemeinen geliefert. Insbesondere wichtige Begrifflichkeiten und verschiedene Theorien, wie z.B. die von Nonaka und Takeuchi und Probst et al. werden näher erläutert.
Hier wird noch nicht zwischen KMU und Großunternehmen unterschieden. Die erfolgt erst im zweiten Teil des Beitrags.

Literatur

Brücher, Heide (2004). Leitfaden Wissensmanagement. Von der Anforderungsanalyse bis zur Einführung. Zürich: vdf Hochschulverlag.

Dani et al. (2006). A methodology for best practice knowledge management. In: Proceedings of the Institution of Mechanical Engineers/Part B, Journal of engineering manufacture 220(10), S. 1717-1728.

Davenport, Thomas H. & *Prusak*, Laurence (1998a). Wenn ihr Unternehmen wüsste, was es alles weiß..., Das Praxisbuch zum Wissensmanagement. Landsberg: Verlag moderne Industrie.

Davenport, Thomas & *Prusak*, Laurence (1998b). Working Knowledge. How organization manage what whey know. Boston: Harvard Business School Press.

Day, Ronald E. (2005). Clearing Up "Implicit Knowledge": Implications for Knowledge Management, Information Science, Psychology, and Social Epistemology. In: Journal of the American Society for Information Science and Technology 56(2005)6, S. 630-635.

Janik, A (1988). Tacit knowledge, working life and scientific method. In: *Göranzon*, B. & *Florin*, M. (Hrsg.): Knowledge, Skill and artificial intelligence. London: Springer S. 53-63.

Krcmar, Helmut (2005). Informationsmanagement. Berlin: Springer, 4. Auflage.

Krcmar, H. & *Rehäuser*, J. (1996). Wissensmanagement einer Unternehmung. In: *Schreyögg*, Georg & *Conrad*, Peter (Hrsg.): Managementforschung 6, Berlin, New York: De Gruyter, S.1-40.

Kuhlen, Rainer (1995). Informationsmarkt. Chancen und Risiken der Kommerzialisierung von Wissen. Konstanz: UVK (Schriften zur Informationswissenschaft; 15).

Linde, Frank (2005). Ökonomie der Information. Göttingen: Universitätsdrucke.

Lucko, Sandra & *Trauner*, Bettina (2005). Wissensmanagement. München: Hanser, 2. Auflage.

Mujan, Dzemal (2006). Informationsmanagement in Lernenden Organisationen. Berlin: Logos.

Nonaka, Ikujiro (1994). A Dynamic Theory of Organizational Knowledge Creation. In: Organization Science, Vol. 5(1994)1 S. 15-37.

Nonaka, Ikujiro & *Takeuchi*; Hirotaka (1995). The knowledge creating company. New York: Oxford University Press.

Nonaka, Ikujiro & *Takeuchi*, Hirotaka (1997). Die Organisation des Wissens. Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen. Frankfurt am Main: Campus.

Polanyi, Michael (1985). Implizites Wissen. Frankfurt am Main: Suhrkamp.

Probst, Gilbert; *Gibbert*, Michael & *Raub*, Steffen (2002). Wissensmanagement. In: *Oechsler*, W.; *Weber*, W. (Hrsg.): Handwörterbuch des Personalwesens, Stuttgart: Schaeffer-Poeschel.

Probst, Gilbert, *Romhardt*, Kai & *Raub* Steffen (2006). Wissen managen. Wiesbaden: Gabler, 5. Auflage.

Richard C. *Hicks* et al. (2007). A metaphor for knowledge management: explicit islands in a tacit sea. In: Journal of Knowledge Management, 11(2007)1, S. 5-16.

Rüdiger, Matthias & *Vanini*, Sven (1998). Das Tacit knowledge-Phänomen und seine Implikationen für das Innovationsmanagement. In: Die Betriebswirtschaft (1998)4, S. 467-480.

Ryle, Gilbert (1946). Knowing how and knowing that. In: Proceedings of the Aristotelian Society 46, S. 1-16.

Schneider, Ursula (2001). Die 7 Todsünden im Wissensmanagement. Kardinaltugenden für die Wissensökonomie. Frankfurt am Main: Frankfurter Allgemeine Zeitung Verlagsbereich Buch.

Schwaninger, Markus (2000). Implizites Wissen und Managementlehre: Organisationskybernetische Sicht. Verfügbar unter: www.ifb.unisg.ch/org/ifb/ifbweb.nsf/SysWebRessources/beitrag41/$FILE/DB41_ImplizitesWissen_def.pdf [04.06.07].

Schreyögg, Georg & *Geiger*, Daniel (2004). Kann man implizites Wissen konvertieren? Die Wissensspirale auf dem Prüfstand. In: *Frank*, U. (Hrsg.), Wissenschaftstheorien in Ökonomie und Wirtschaftsinformatik, S. 269-288.

Stock, Wolfgang, G. (2000). Informationswirtschaft. München: Oldenbourg .

Stock, Wolfgang, G. (2007). Information Retrieval. Informationen suchen und finden. München: Oldenbourg.

Tsoukas, Haridimos (2002). Do we really understand tacit knowledge? In: Paper presented at Knowledge Economy and Society Seminar, LSE Department of Information Systems, London. Verfügbar unter: www.lse.ac.uk/collections/informationSystems/events/ESRCSeminars/tsoukas.pdf [04.06.07].

Wagner, Hubert (o.J.). Wissen des Menschen. Verfügbar unter: www.hubert-wagner.de/wissen_des_menschen.htm [04.06.07].

Wagner, Michael P. (1995). Groupware und neues Management. Braunschweig: Vieweg.

Wissensmanagement, Informationsbedarfsanalyse, Kleine und mittlere Unternehmen, KMU, implizites Wissen, tacit knowledge, Polanyi, Probst, Nonaka, Takeuchi, Wissensspirale, Wissensbaustein

Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen

Teil 2: Wissensmanagement in KMU

Sonja Gust von Loh, Düsseldorf

Bei KMU ist die Implementierung von Wissensmanagement anders gelagert als in Großunternehmen. Auch wenn durch kürzere Wege die Entscheidung für ein Wissensmanagementsystem einfacher ist, so stehen dem Unternehmen wesentlich weniger Ressourcen zur Verfügung. Sowohl personell wie auch finanziell sind Grenzen gesetzt. Auf Grund dessen sollte man in KMU nur einzelne Bausteine umsetzen und nicht gleich ein umfassendes Wissensmanagementsystem einführen. Wir führen vor, mit welchen Tools diese Prozesse unterstützt werden können und welche Bausteine für kleine und mittelständische Unternehmen sinnvoll sind. Im Fokus dieses Teilartikels stehen zum einen die Wissensmanagementprozesse in KMU, aber auch zum anderen Informationsbedarfsanalysen, die für eine sinnvolle Einführung von Wissensmanagement notwendig sind. Vor allem in kleinen und mittelständischen Unternehmen können diese richtungweisend für die Implementierung einzelner Bausteine sein. Informationsbedarfsanalysen lassen sich in Abteilungen gezielt einsetzen, um Anhaltspunkte zu erlangen, wo die Kommunikation intern, wie aber auch extern krankt. Daraufhin arbeitet man gezielt an einer Optimierung. Wir erläutern detailliert, welche Methoden es zur Durchführung einer solchen Informationsbedarfsanalyse gibt und welche besonders sinnvoll in KMU anzuwenden sind.

Knowledge Management and Information Needs Analysis in Small and Medium Sized Enterprises. Part 2: Knowledge Management in SME

In SMEs the implementation of knowledge management is different from big companies. On the one hand there are shorter ways for making a decision, but on the other hand there are fewer resources, fewer human resources and fewer financial resources as well. Therefore only single components of a knowledge management system should be implemented. The article explains which tools support the knowledge management working process in small and medium sized enterprises. Another focus of this article is the analysis of information needs and demands, which is a necessary precondition of a useful implementation of knowledge management. Particularly in the small and mediums sized companies such an analysis can give a direction for the implementation of single tools of knowledge management. Analyses of information needs can be conducted in different departments of a company for providing a basis of information about the internal and external communications and their failures. Thereupon knowledge management can be optimized. The article shows in detail which methods of information needs analyses exist and which are especially expedient in SMEs.

Einleitung

Aufbauend auf den ersten Teil (Gust von Loh 2008) beschäftigt sich der zweite Teilartikel mit der Problematik des Wissensmanagements in kleinen und mittelständischen Unternehmen (KMU). Es wird hier etwas spezieller und vor allem praktischer. In KMU gewinnt Wissensmanagement zusehends an Bedeutung. Zum einen ist es womöglich einfacher, Wissensmanagement einzuführen, da die Entscheidungswege in der Regel kürzer sind, zum anderen ist jedoch das finanzielle Budget meist nicht in einem solchen Rahmen vorhanden, wie es bei großen Unternehmen ist. Auf Grund dessen sollte bei KMU zunächst an einem bestimmten Punkt angesetzt werden, um dann das Wissensmanagement weiter ausbauen zu können. Zur ersten Orientierung dient eine Informationsbedarfsanalyse. Durch sie kann eine bestimmte Anzahl an Mitarbeitern mit in Entscheidungsprozesse einbezogen werden, so dass neu eingeführte Wissensmanagement-Tools besser angenommen werden. Wie solch eine Informationsbedarfsanalyse in konkreten Fällen aussehen kann und was für Möglichkeiten es allgemein im Wissensmanagement gibt, soll der vorliegende Artikel klären.

1 Wissensmanagement in KMU

In kleinen und mittelständischen Unternehmen (KMU) gewinnt Wissensmanagement zunehmend an Bedeutung. Es ist auch hier kein Selbstzweck, sondern dient als Instrument zur Förderung der Wettbewerbsfähigkeit eines Unternehmens (Schauer & Frank 2007, 150). Dennoch muss Wissensmanagement in KMU anders etabliert werden als in einem großen Unternehmen. Letztere verfügen in der Regel über andere technische und vor allem personelle Ressourcen (Groegger 2001, 2) und stellen andere Ansprüche an das Wissensmanagement. Auch der finanzielle Hintergrund ist anders.
Durch den internationalen Background und durch die hohe Mitarbeiteranzahl bei großen Unternehmen sind andere Instrumente des Wissensmanagements sinnvoll als bei KMU. Meist implementieren Mittelständler ausschließlich einzelne Bausteine des Wissensmanagements (s. Probst) und nicht den vollständigen Kreislauf (Abele et al. 2003, 375).
Schwerpunkte für die Einführung sind ähnlich wie die bei großen Unternehmen, müssen aber meist anders gesetzt werden. In der Regel liegt der Fokus bei kleinen und mittelständischen Unternehmen insbesondere auf der Zusammenarbeit, der Analyse, dem Suchen, dem Veröffentlichen von Informationen und auf der Aufgabenverfolgung (Schiller & v. Guretzky 2004, 10). Deswegen kann man auch sagen, dass der Mitarbeiter im Mittelpunkt steht und weniger die Implementierung eines leistungsfähigen Wissensmanagementsystems (Abele et al. 2003, 375). Das sollte ausschließlich unterstützende Wirkung haben, aber keinesfalls im Zentrum des Geschehens stehen. Generell lässt sich sagen, dass die Erwartung an ein Wissensmanagementsystem auch abhängig von der jeweiligen Position eines Mitarbeiters ist. Ein Sach-

bearbeiter stellt natürlich andere Ansprüche als ein Geschäftsführer.
Ein großer Vorteil für Wissensmanagement – wenn nicht sogar unabdingbar – ist ein offenes Betriebsklima. Nur so kann der Wissensfluss zwischen den verschiedenen Hierarchieebenen sichergestellt werden. Dies sollte nicht nur von unten nach oben der Fall sein, sondern Informationen sollten auch von oben nach unten fließen (Schiller & v. Guretzky 2004, 10). Außerdem muss bereits in der Unternehmenskultur die Wichtigkeit von Wissen verankert sein und den Mitarbeitern auch vermittelt werden.

> Unternehmenskultur, Wissen zu sammeln, zu verteilen und zu nutzen muss dem Unternehmen „in Fleisch und Blut" übergehen: Angefangen bei der Geschäftsleitung, die dazu ermuntert, bis hin zu den Mitarbeitern, die ihre Kommunikation über Erfahrungen und Kenntnisse pflegen müssen (v. Guretzky 2002).

Die größten Probleme treten im Hinblick auf die Organisation auf. Die Technik ist meist unproblematisch zu implementieren. Die Schwierigkeit liegt darin, soziale und kulturelle Barrieren zu überwinden (Pan et al. 2007, 404) und die Mitarbeiter so zu einem regen Wissensaustausch zu motivieren.
Schon auf der Ebene der Geschäftsführung muss die tragende Rolle des Wissens vermittelt werden. Der Wissensfluss sollte deswegen nach unten optimiert werden. Um diesen Grundsatz mit Überzeugung vertreten und vermitteln zu können, muss jedoch für alle Beteiligten ein klarer Nutzen zu sehen sein. Deswegen sollte man bei KMU bei den operativen Zielen ansetzen. Für eine strategische Zielsetzung mangelt es häufig an Zeit, da diese vollkommen vom Tagesgeschäft beansprucht wird. Die Frage: „Welche Prozesse im Unternehmen können durch den systematischen Umgang mit Wissen effizient gestaltet werden?" (TBS NRW 2003, 41) sollte gestellt und geklärt werden.
Ein weiterer, ebenfalls sehr wichtiger Faktor der Unternehmenskultur ist die Vertrauensbasis zwischen den Sachbearbeitern untereinander, aber auch zwischen Vorgesetzten und Sachbearbeitern. Dieses Vertrauen sollte im Idealfall abteilungsübergreifend vorhanden sein. Fehlt diese Basis, wird es sehr schwierig sein, Mitarbeiter zu motivieren, ihr Wissen zu teilen bzw. anderen zur Verfügung zu stellen.
Im Hinblick auf das Wissensmanagement ist der wichtigste Unternehmenskultur-Faktor die Offenheit sämtlicher Mitarbeiter, Wissen (nicht nur Information) rasch und transparent zur Verfügung zu stellen und/oder abzufragen. Entscheidend ist also eine Kultur der gegenseitigen Hilfsbereitschaft, die es aufzubauen gilt (Specht 2007, 36).
Dieses Vertrauen kann unter anderem durch Kommunikation hergestellt werden. Auch das Risiko von Konflikten wird durch eine gesunde Kommunikation vermindert (Southon et al. 2002, 1052).
Neben der Offenheit muss dem Mitarbeiter natürlich auch eine gewisse Zeit eingeräumt werden, um sein Wissen zu dokumentieren, zu pflegen bzw. anderen mitzuteilen. Zeitmangel und Mangel an Bereitschaft zur Weitergabe des eigenen Wissens sind meist die Probleme, an denen Wissensmanagement scheitert (Weissenberger-Eibl & Kelm 2005, 29f).
Im Zusammenhang mit Wissensmanagement in KMU sollte man auch nicht die Bausteine Probsts ausser Acht lassen. Insbesondere die Bausteine „Wissensschaffung", „Wissensspeicherung", „Wissensverteilung" (...) sind wichtig (Fink & Ploder 2005, 26).
Wissensschaffung ist ein Aspekt, mit dem ein Wettbewerbsvorteil erschaffen werden kann. Diese dient dazu, neue und interessante Informationen zu ermitteln. Durch Strukturierung von Wissen und Information kann bereits vorhandenes Wissen für andere Mitarbeiter als Wissensschaffung dienen. Die Erstellung eines Lieferprogramms wäre hier als Beispiel anzuführen.
Die „Wissensverteilung" beschäftigt sich mit der sinnvollen Vergabe von Wissen. In der Regel benötigen mehrere Mitarbeiter ähnliches Wissen, so dass es durchaus Sinn macht, gleiches Wissen einer bestimmten Anzahl von Mitarbeitern zur Verfügung zu stellen. Sie können vom Wissen des anderen profitieren. Hier sollte jedoch ein klares Geben und Nehmen erkennbar sein. Es darf nicht sein, dass gewisse Mitarbeiter ausschließlich profitieren, aber kein eigenes Wissen bereitstellen. Es geht darum, das Unternehmen für alle transparenter zu machen und dadurch zu erreichen, dass Mitarbeiter miteinander arbeiten. Laut Betty Zucker entwickeln Gruppen durch gemeinsame Alltagsarbeit Kompetenzen, welche die Fähigkeiten der Einzelnen erst voll zum Tragen bringen (Zucker 2000, 49).
Abbildung 1 zeigt das Zusammenspiel zwischen einzelnen Individuen, dem Team und der gesamten Organisation.
Es lässt sich erkennen, dass im Idealfall ein kontinuierlicher Austausch in beide Richtungen zwischen Individuum, Team und Organisation erfolgt. Die Koordination soll verbessert werden und einzelne Unternehmensbereiche dürfen nicht gegeneinander sondern müssen miteinander arbeiten (Schneider 2001, 19).
Wissensspeicherung ist ebenfalls ein sehr wichtiger Aspekt, der in der Regel der Wissensverteilung vorangehen sollte. Durch die Wissensspeicherung lässt sich ein Generationenwechsel besser vollziehen. Neben der Wissensablage in eine Datenbank ist es aber durchaus nützlich, dass bereits im Vorfeld ein Nachfolger

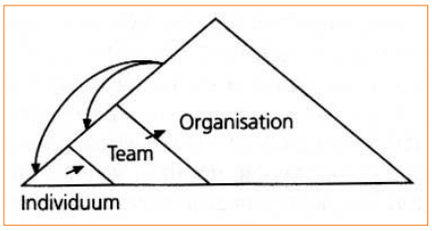


Abbildung 1: Vom Individuum zum Team zur Organisation und zurück (Quelle: Zucker 2000, 49)

ausgesucht wird, der dann durch enge Zusammenarbeit mit den Seniorchefs lernen kann. In japanischen bzw. asiatischen Unternehmen ist diese Praxis alltäglich, wohingegen sie in westlichen Ländern eher selten vorzufinden ist.
Natürlich kann auch das Wissen der einzelnen Sachbearbeiter festgehalten werden. Dieser Aspekt ist interessant, da die Fluktuation der Mitarbeiter auf Sachbearbeiterebene wahrscheinlich höher ist. Ersetzbar werden Mitarbeiter jedoch dadurch trotzdem nicht, da jeder von ihnen über einen gewissen Erfahrungshintergrund verfügt, der nicht ersetzt werden bzw. nicht in Worte gefasst werden kann. Nur das so genannte Schlüsselwissen kann durch die Wissensspeicherung abgelegt werden.
Gerade für KMU ist die Einführung von Wissensmanagement durch kurze Entscheidungswege vergleichsweise einfach; mit verhältnismäßig geringem Aufwand lassen sich deutliche Erfolge bei der wissensfreundlichen Gestaltung von Geschäftsprozessen erzielen (TBS NRW 2003, 23).
Wichtig ist, dass die einzelnen Wissensziele aufeinander abgestimmt sind. Abbildung 2 zeigt eine Zielpyramide, mit Hilfe derer die wichtigsten, zu berücksichtigenden Ziele aufgeführt sind. Zuerst müssen von der Unternehmensleitung Unternehmensziele definiert werden, welche dann mit Wissenszielen in Beziehung gesetzt werden. Im letzten Schritt können dann die Projektziele entsprechend angepasst werden.

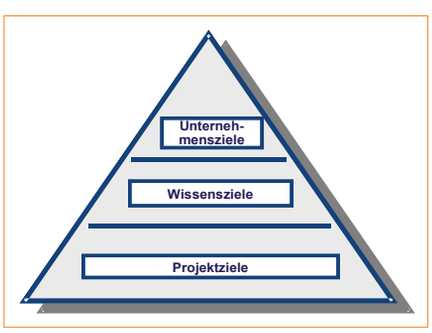


Abbildung 2: Zielpyramide (Quelle: TBS NRW 2003, 42).

Abbildung 3 zeigt noch einmal verschiedene Fragestellungen, die im Zusammenhang mit Wissensmanagement in einem KMU auftauchen können. Auch verschiedene Teilbereiche, die mit dem Wissens-

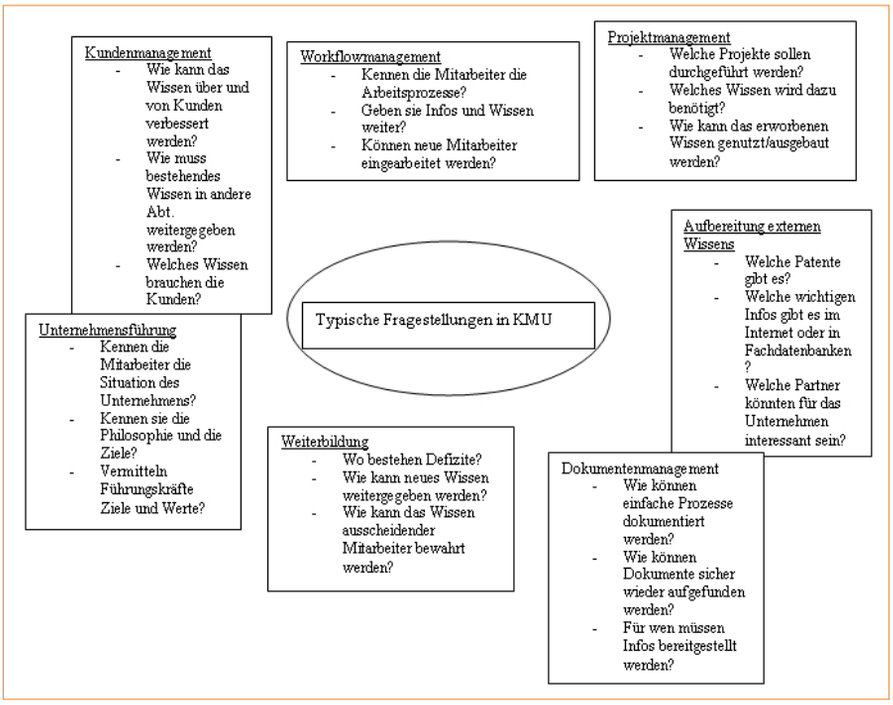


Abbildung 3: Typische Fragestellungen in KMU (In Anlehnung an Abele et al.2004, 415).

management Berührungspunkte haben, werden hier aufgezeigt.

2 Wissensmanagement-Tools und -Techniken für KMU

Für die Anwendung von Wissensmanagement gibt es diverse Tools, die speziell für den Einsatz im Unternehmen bestimmt sind. Zunächst benötigt ein Unternehmen eine Basis, wo die verschiedenen kleineren Tools integriert werden können. Dazu kann ein bestehendes Datenbankmanagementsystem (DMS) dienen. Vor allem in KMU wird angeraten, bereits angewandte Systeme für die zentrale Ablage von Daten zu nutzen. Hier können die Tools meist genauso eingefügt werden, wie es bei gekauften Lösungen der Fall ist. Der Nachteil gekaufter Wissensmanagementsysteme ist der hohe Preis. Meist ist dieser für KMU nur mit Schwierigkeiten tragbar. Eine einfache Portallösung (mit virtueller Community, Intranet, Extranet, Erfahrungsdokumentation etc.) kostet um die 200.000 Euro. Der Programmierungsaufwand, die Pflege und die Akzeptanzkosten sind in diesem Preis noch nicht inbegriffen (Hesseler 2005, 20). In einem solchen Paket sind wahrscheinlich auch Elemente dabei, die nicht genutzt werden, oder sogar nicht genutzt werden können, da einfach keine Verwendung für sie besteht. Die verschiedenen Tools lassen sich am besten nach den oben angesprochenen Bausteinen von Probst einteilen. Angesprochen werden hier die wichtigsten Tools. Einige sind für mehrere Bausteine zuständig, so sind z.B. die *Communities of practice* nicht nur für die Wissenserzeugung nützlich, sondern auch für die Wissensverteilung. Diese Praktiken werden jeweils nur einmal näher erläutert.

Wissen identifizieren

1. *Expertenverzeichnisse/ Yellow Pages*: Es handelt sich um eine relativ einfache Methode. Hier werden die einzelnen Fähigkeiten von Mitarbeitern eingetragen, z.B. Fremdsprachen, bestimmtes technisches Know How etc. Dies ermöglicht dem Unternehmen die Kenntnis über die Fähigkeiten verschiedener Mitarbeiter zu erlangen. Vor allem für international agierende Unternehmen bzw. für Großkonzerne ist diese Art der Wissensidentifikation eine einfache aber effiziente Möglichkeit, Wissensträger zu ermitteln. Zu beachten ist jedoch, dass, wenn personengebundene Informationen in ein System aufgenommen werden, dies an die Zustimmung der Mitarbeiter bzw. des Betriebsrats gebunden ist.
2. *Wissenskarten:* Wissenskarten sind graphische Verzeichnisse von Wissensträgern, Wissensbeständen, Wissensquellen, Wissensstrukturen oder Wissensanwendungen. Die Transparenz in einem Unternehmen wird durch sie erhöht und die Auffindung von bestimmten Experten gesichert. Sie erleichtern das Einordnen von neuem Wissen in bestehendes. Es gibt verschiedene Wissenskarten, die sich in ihrer Struktur unterscheiden (Probst et al. 2006, 67).

Wissen erwerben/entwickeln

1. Debriefing: *Debriefing* bedeutet im Englischen Nachbesprechung. Da dies ein eher allgemein gehaltener Ausdruck ist, kann *Debriefing* in verschiedenen Zusammenhängen angewendet werden. Im Wissensmanagement ermöglicht es zum einen die Entwicklung von Wissen, aber auch die Speicherung von Wissen. Die Ermittlung von Wissen wird deswegen erreicht, da durch Besprechungen von Mitarbeitern andere Mitarbeiter eigenes Wissen aufbauen und somit für sich selbst Wissen entwickeln können. Ein interessanter Aspekt ist, dass auch beim Ausscheiden eines Mitarbeiters aus dem Unternehmen mit dieser Technik das Fehlen des Mitarbeiters kompensiert werden kann. Das *Debriefing* ist eine Art Interview, in dem das Wissen bestimmter Experten ermittelt und mit einem Programm kodifiziert wird. Es wird letztlich in Datenbanken abgespeichert, die für alle zugänglich sind. An dieser Stelle kann natürlich eine Benutzerbeschränkung eingestellt werden, damit nicht alle auf dieses Wissen zugreifen können. Es geht hier letztlich um die Ermittlung von implizitem Wissen und um die anschließende Überführung des impliziten in explizites Wissen. Das in dem Interview erhobenen Wissen kann mittels spezieller Software (z.B. PowerKNOW) in explizites Wissen aufbereitet werden. Dieses Wissensmanagement-Tool ist noch verhältnismäßig jung, gewinnt aber stetig an Bedeutung. *Debriefing* sollte nicht zu häufig angewendet werden, da sonst die Gefahr besteht, dass zu viel Informationsballast entsteht. Ähnlich wie das *Debriefing* sind auch die Aufgaben der Mentoring-Beziehungen gelagert. Hier geht es um die Vermittlung von wichtigen Informationen für neue Mitarbeiter, aber auch um Vermittlung von Informationen von bald ausscheidenden *Keyplayern* an potenzielle Nachfolger (Schütt 2006, 11).

2. Customer Knowledge Management: Das Wissen der Stakeholder wird ins Unternehmen geholt und genutzt, um das Unternehmen weiterzubringen. Ideen und Wünsche werden berücksichtigt, so dass es zu einer höheren Kunden- und Lieferantenzufriedenheit kommt. Hierbei profitiert ganz klar das Unternehmen.

3. Customer Relationship Management (CRM): CRM beschäftigt sich mit der Pflege von Kundendaten und berücksichtigt auch Wünsche der Kunden. Es geht hier letztlich um die Kundenfreundlichkeit. Insbesondere profitiert hier der Kunde, aber indirekt natürlich auch das Unternehmen, da es durch die Zufriedenheit von Kunden und Partnern zu einer stabilen Kunden- bzw. Partnerbindung kommt.

4. Wikis
Das Web 2.0 ist ein breit diskutiertes Thema; Einsätze in Firmen werden als Unternehmen 2.0 diskutiert (Peters, Stock

2007). Der Begriff des Web 2.0 steht für eine aktive Teilnahme einzelner Personen an der Erstellung verschiedener Inhalte im Web. In diesem Zusammenhang gibt es verschiedene Tools, wie z.B. Blogs oder Wikis. Für KMU ist die Social Software von Interesse, denn sie kann auch ins Intranet eingebunden werden. Der wohl bekannteste Vertreter von einer Wiki-Anwendung ist die freie Enzyklopädie Wikipedia[1].

Wikis wurden 1995 von Ward Cummingham entwickelt (Weiss 2005, 21). Es handelt sich hierbei um ein Nachschlagewerk, dass von allen Mitarbeitern erstellt wird. Benutzer des Wiki erhalten Schreib- und Leserechte, so dass die Artikel in dieser Datenbank von jedem verändert, ergänzt oder korrigiert werden können. Vor allem diese Interaktivität macht Wikis so interessant. Neben dem einfachen Nachschlagewerk ist eine zusätzliche Wissensablage und -strukturierung mittels Indices, Suche und der Möglichkeit des Up- und Downloads von zugehörigen Dateien möglich (Bendel 2007, 95). Durch die gemeinsame Arbeit an einem Wiki entsteht eine vernetzte Wissensumgebung, an der im Idealfall alle Mitarbeiter aktiv teilnehmen. Wissen wird auf diese Art dauerhaft gespeichert und wiederauffindbar gemacht. Dadurch, dass Kategorien definiert werden können, entsteht eine interne Struktur und die Informationen lassen sich besser wiederfinden. Die Artikel sind untereinander verlinkt, so dass man automatisch über einen Artikel zum nächsten Artikel, der ebenfalls etwas mit dem Thema zu tun hat, kommt. Außerdem können auch externe Links ins Intranet-Wiki integriert werden, so dass direkt passende Informationen im Internet gefunden werden.

Unternehmenswikis erfreuen sich wachsender Beliebtheit. So sprechen das *Wall Street Journal* und die *Business Week* von einer aufstrebenden Technologie, die für die Kommunikation innerhalb eines Unternehmens aber auch über Unternehmensgrenzen hinweg einsetzbar ist.

> Wall Street Journal and Business Week have identified wikis as an up-and-coming technology to support collaboration within and between firms (Majchrazak et al. 2006, 99).

Speziell für kleine und mittlere Unternehmen sind sie effizient. Sie setzen genau an dem zentralen Punkt der meisten KMU an, nämlich der Wissensdokumentation und Wissensbewahrung (Fischer 2007). Außerdem handelt es sich bei Wikis um eine sehr kostengünstige Möglichkeit, Wissen zu strukturieren, Kommunikation, Kooperation und Interaktion zwischen Mitarbeitern zu fördern, da die Software eine *Open Source Software* (wie z.B. Wikimedia) und damit frei zugänglich ist. Die verhältnismäßig einfache Bedienbarkeit des Tools ermöglicht darüber hinaus einen breiten Zugang. Dadurch, dass die Artikel immer weiter geschrieben bzw. verändert werden können, kommt es zur aktiven gemeinsamen Zusammenarbeit der Mitarbeiter.

Jede Veränderung lässt sich namentlich nachvollziehen. Schon alleine deswegen werden die Mitarbeiter motiviert sein, ausschließlich seriöse Daten einzutragen (Schütt 2005, 14). Die Qualität der Artikel steht und fällt natürlich mit den Fähigkeiten der Mitarbeiter. Ein Vorteil, der nicht unerwähnt bleiben sollte, ist, dass durch die aktive Einbeziehung der Mitarbeiter ungeahnte Potenziale geweckt werden können.

> Mitarbeiter lassen sich aus einer Reserve locken und mehr für das Unternehmen begeistern. Statt Arbeit nach Plan bekommt das Unternehmen plötzlich, echte aktive Mitarbeit (ebd. S. 14).

5. *Researchteams, Communities of practice:* Sie bieten ein Mittel der Aufhebung von Kommunikations- und Motivationsbarrieren (Schütt 2006, 11). *Communities of Practice* sind ebenfalls für die Wissensverteilung günstige Möglichkeiten. Streng genommen handelt es sich hier nicht um ein Tool, sondern um eine Technik. Zur Anwendung werden ausschließlich Mitarbeitergruppen, deren Mitglieder ein ähnliches Interessengebiet haben, zusammengestellt.

Diese Mitglieder können durchaus an unterschiedlichen Orten arbeiten. Sie sollten Interesse daran haben, ihr eigenes Wissen auszubauen und vor allem auch auszutauschen (Probst et al. 2001, 38). Wichtig ist, dass die Teilnahme freiwillig ist und kein Mitarbeiter dazu gezwungen werden sollte. Derjenige, der gezwungen wird, wird höchstwahrscheinlich nicht bereit sein, sein eigenes Wissen preiszugeben. Die Themen der einzelnen Wissensgemeinschaften (*Communities of practice)* haben natürlich etwas mit den Themenschwerpunkten des Unternehmens zu tun. Bei *Communities of Practice* können insgesamt vier Dimensionen unterschieden werden:

1. Die einzelnen Personen bzw. Mitglieder der *Community of practice,*
2. Art der Interaktion der Personen,
3. die daraus resultierende Wissenstransformation,
4. organisatorische Verankerung der Wissensgemeinschaft in die Gesamtorganisation.

6. *Interaktion:* Diese kann in Form von jeglichen Kommunikationsmitteln oder durch direkte Gespräche geschehen.

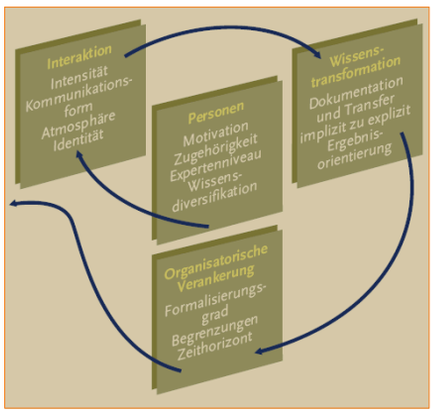


Abbildung 4: Gestaltungsdimensionen von Wissensgemeinschaften (Quelle: Probst et al. 2001, 39).

7. *Lessons learned:* Hier geht es darum, aus Erfahrung zu lernen. Insbesondere Projekte werden häufig abgeschlossen, ohne das reflektiert wird, was gut und was schlecht gelaufen ist. An dieser Stelle könnte man jedoch aus alten Fehlern oder aus Techniken, die sich in anderen Projekten bereits bewährt haben, lernen. Mit den Worten von Probst evaluiert *lessons learned* systematisch einen Projektverlauf (Probst et al. 2001, 42). Es sollten Bewertungen erstellt werden, die auch an andere Projektteams weitergegeben werden, damit auch diese aus bestimmten Techniken oder auch aus Fehlern lernen. Hier spielt natürlich wieder das gegenseitige Vertrauen der Mitarbeiter eine tragende Rolle. Vertrauen sich die Mitarbeiter nicht, oder haben sie Angst vor übler Nachrede, wenn sie in einem weniger effizienten Team waren, wird es kaum möglich sein, *lessons learned* sinnvoll anzuwenden. Für Probst ist ein Projekt erst dann abgeschlossen, wenn die Projektarbeit bewertet und Erfahrungen systematisch gelernt und geteilt worden sind (ebd., 42). Die Nachbearbeitung eines Projekts sollte somit nicht vernachlässigt werden. Jedes Unternehmen kann sich den für sich selbst richtigen Weg ausdenken. Ein Beispiel, wie es vorbildlich wäre, sieht wie folgt aus: Zunächst sollte das Projekt noch einmal kurz zusammengefasst und bestimmte Regeln für die Evaluation festgehalten werden. Wichtig ist hier, dass die Kritik konstruktiv sein muss. Es geht nicht um Schuldzuweisungen oder sonstige negative Aspekte. In der nächsten Phase wird ein *Brainstorming* durchgeführt, wie das Projekt abgelaufen ist. Dann wird dieses bewertet und Reflektionen, warum das Projekt so gelaufen ist, werden angestellt. Welche Schritte gut und welche Schritte weniger gut waren, sollte festgehalten werden. In der letzten Phase wird der Lerneffekt des Projekts zusammengefasst. Abschließend wird ein Report geschrieben, in dem die wichtigsten Aspekte festgehalten werden (Probst et al. 2001, 42). Abbildung 5 zeigt, wie mit *lessons learned* gearbeitet werden kann.

1 www.wikipedia.de

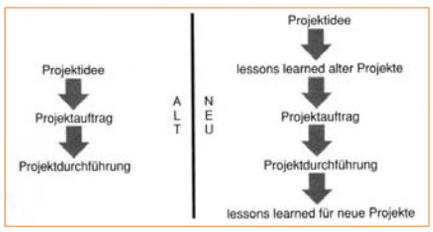


Abbildung 5: Überarbeitung von Projekterfahrung (Quelle: Probst et al. 2006, 133).

Mit zunehmendem Einstieg in die Wissensgesellschaft wird die Entscheidungsbefugnis immer mehr direkt an die Wissensarbeiter delegiert, die natürlich hierzu befähigt sein müssen. Das erfordert auch einen direkten Zugriff auf wesentlich mehr Informationen als früher, was wiederum nicht zu komplex ausfallen darf (Schütt 2006, 12).

8. Fachinformationsdienste: Diese bieten die Möglichkeit, sich vor allem zur Branche und zum alltäglichen Wirtschaftsleben Informationen zu beschaffen. Man recherchiert spezielle Themen und erwirbt gezielt Artikel. Einige der wichtigsten Wirtschaftsinformationsdienste sind z.B. Factiva, Genios, Lexis Nexis, GBI und die Bundesagentur für Außenwirtschaftsinformation (Stock & Stock 2003, 29f).

Wissensverteilung

1. Wikis

2. *Dokumentation, Erfahrungsberichte*: Um eine gewisse Transparenz zu erreichen, sollten Dokumentationen bzw. Erfahrungsberichte verfasst werden. Diese können zum Beispiel geschrieben werden, wenn es einen Besuch beim Kunden gab oder sonstige Geschäftsreisen getätigt wurden. Wichtig ist, dass diese Berichte zumindest abteilungsintern zur Verfügung gestellt und nicht einfach zur Seite gelegt werden, ohne dass sie jemand gelesen hat. Ein Ablagesystem im Intranet wäre denkbar, um den Bericht zu verteilen.

3. *Groupwareprogramme* (wie Lotus Notes): Sie bieten die Möglichkeit, dass verteilte Dateien abgerufen und überprüft werden können. Neben der klassischen Funktion der Dokumentenbereitstellung stellen sie auch noch einen Kalender, der abteilungsübergreifend gepflegt werden kann zur Verfügung. So ist es jedem Mitarbeiter möglich zu sehen, wer im Haus ist und wer nicht.

4. *Interne Foren*: Sie verbessern und beschleunigen den Informationstransport (Weichbrodt 2001, 15). Man greift bewusst auf das Wissen anderer zurück und sucht Hilfe bei ihnen. Auch hier ist wieder die Vertrauensbasis wichtig.

5. Communities of practice

6. Speicherung in schriftlicher Form: Hier handelt es sich um eine Art Dokumentation, die entweder nüchtern oder auch etwas aufgelockerter verfasst werden kann. Beinhalten sollte sie größere Projekte mit Ansprechpartnern und Produkte im Allgemeinen inklusive Bildern.

7. Storytelling: Streng genommen ergänzt diese Technik bereits vorhandene Wissensmanagementansätze. Es handelt sich dabei um Storys, die in der Regel erzählt werden. Durch den direkten Kontakt zwischen Mitarbeitern verschiedener Abteilungen kann die Vertrauensbasis gut aufgebaut und die Kommunikation verbessert werden (Erlach & Thier 2004, 213). Es sollten nicht nur trockene Fakten besprochen werden, sondern im Idealfall werden diese ausgeschmückt. Durchgeführt werden kann *Storytelling* im

82 von 569 wunderbaren Argumenten für Pleuser:

Unbeschränkte Haftungseigenschaften (>100 Jahre), testiert vom »Institut für Papierfabrikation der TU Darmstadt« • Optimal geeignet für die Außen- und Innenbeschriftung von Büchern und Archivalien • Weichmacherfestes Papier • Weichmacherfreie Klebstoffe • Deutsches Bundespatent Nr. 4023539 • pH-neutral • Alterungsbeständig • Lichtecht • Formate, Farbigkeit und Gestaltung nach Wunsch • Modernste Technikelemente: Barcode-Druck • Sicherheits-Etiketten für Buchsicherheitsanlagen • Absolute Serviceorientierung • Kleiner Auszug aus unserer Kundenliste: • Hochschulbibliothek der RWTH Aachen • Universitätsbibliothek Augsburg • Landschaftsbibliothek Aurich • Universitätsbibliothek Bamberg • Universitätsbibliothek Bayreuth • Kreis- und Stadtbibliothek Bergisch Gladbach • Staatsbibliothek zu Berlin • Zentral- und Landesbibliothek Berlin • Bibliothek des Otto-Suhr-Instituts für Politikwissenschaft Berlin • Universitätsbibliothek der Technischen Universität Berlin • Universitätsbibliothek der Universität der Künste Berlin • Bibliothek des Auswärtigen Amtes Berlin • Bibliothek des Deutschen Bundestages Berlin • Bibliothek der Fachhochschule für Technik und Wirtschaft Berlin • Bibliothek des Max-Planck-Instituts für Wissenschaftsgeschichte Berlin • Kunstbibliothek der Staatlichen Museen zu Berlin • Universitätsbibliothek Bochum • Stadtbücherei Bochum • Universitäts- und Landesbibliothek Bonn • Hessische Landes- und Hochschulbibliothek Darmstadt • Universitätsbibliothek Dortmund • Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden • Universitätsbibliothek Duisburg • Johannes a Lasco Bibliothek Große Kirche Emden • Universitäts- und Forschungsbibliothek Erfurt-Gotha • Universitätsbibliothek Erlangen-Nürnberg • Hochschul- und Landesbibliothek Fulda • Niedersächsische Staats- und Universitätsbibliothek Göttingen • Bibliothek des Hamburgischen Weltwirtschafts-Archivs Hamburg • Universitätsbibliothek Heidelberg • Universitätsbibliothek Ilmenau • Thüringer Universitäts- u. Landesbibliothek Jena • Badische Landesbibliothek Karlsruhe • Bibliothek des Bundesgerichtshofs Karlsruhe • Stadtbibliothek Kiel • Universitäts- u. Stadtbibliothek Köln • Deutsche Zentralbibliothek für Medizin Köln • Landesbüchereistelle Rheinland-Pfalz Koblenz • Universitätsbibliothek Leipzig • Universitätsbibliothek Mannheim • Universitätsbibliothek Marburg • Stadtbücherei Mülheim • Bayerische Staatsbibliothek München • Universitätsbibliothek Technische Universität München • Bibliothek des Bundesfinanzhofs München • Zentralbibliothek für Medizin Münster • Universitäts- u. Landesbibliothek Münster • Diözesanbibliothek Münster • Regionalbibliothek Neubrandenburg • Stadtbibliothek Osnabrück • Universitätsbibliothek Paderborn • Universitätsbibliothek Regensburg • Universitätsbibliothek Rostock • Saarländische Universitäts- u. Landesbibliothek Saarbrücken • Stadtbibliothek Schifferstadt • Universitätsbibliothek Siegen • Universitätsbibliothek Trier • Kommunikations- u. Informationszentrum der Universität Ulm • Bibliothek der Hochschule Vechta • Herzogin Anna Amalia Bibliothek Weimar • Universitätsbibliothek Würzburg • Universitätsbibliothek Wuppertal • Universitätsbibliothek Graz • Universitätsbibliothek Innsbruck • Universitätsbibliothek Klagenfurt • Österreichische Nationalbibliothek Wien • Bibliothek der Technischen Universität Wien • Universitätsbibliothek für Bodenkultur Wien • Universitätsbibliothek Wien • Bibliothek der Veterinärmedizinischen Universität Wien • Universitätsbibliothek Amsterdam • Universitätsbibliothek Maastricht • Universitätsbibliothek von Wales Aberystwyth • Königliche Nationsbibliothek Stockholm • Nationalbibliothek Madrid • Universitätsbibliothek Politecnica Valencia • F. Hoffmann-La Roche AG Basel • Theologisches Seminar Basel • Thurgauische Kantonsbibliothek Frauenfeld • Biblioteca Cantonale Lugano • ETH-Bibliothek Zürich • Kunsthaus Zürich •

Die restlichen 487 senden wir Ihnen gern persönlich, selbstverständlich mit Mustern und allem Drum und Dran. Einfach anrufen oder mailen.

Sie werden sehen, manchmal lohnt es sich, auch das Kleingedruckte studiert zu haben.

pleuser® etiketten

Seit 1872

Bernhard Pleuser GmbH
Otto-Hahn-Straße 16 • 61381 Friedrichsdorf • Deutschland
Telefon: +49 (0) 6175 7982727 • Fax: +49 (0) 6175 7982729
info@pleuser.de • www.pleuser.de

Rahmen von Workshops. Durch das lebhafte Erzählen können vorherrschende Strukturen in den einzelnen Abteilungen vermittelt werden und bei Gesprächen über Projekte, die gut gelaufen sind, positive Techniken von anderen Abteilungen übernommen werden. Man spricht auch von Geschichtenmanagement.

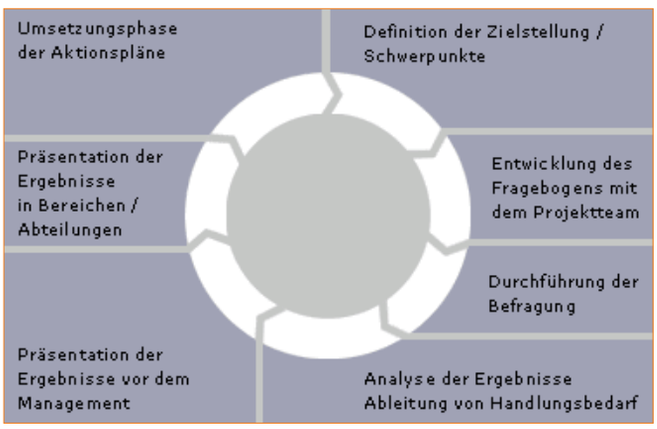


Abbildung 6: Schematischer Ablauf einer Mitarbeiterbefragung (Quelle: http://www.lead-gmbh.de/Mitarbeiterbefragung.htm).

Geschichtenmanagement kann als Methodik benutzt werden, um verborgenes Wissen aufzuspüren, um Wissensflüsse zu verstehen, um Wissen zu vermitteln, ja sogar um die Unternehmenskultur zu verändern oder um Verständnis für andere Kulturen zu wecken (Schütt 2000, 11).
Geschichten werden häufig gerade vom Management als etwas Negatives gesehen. Aber es handelt sich hier nicht um fiktive Geschichten, sondern um Geschichten, die auf Erfahrungen beruhen.
Durch *Storytelling* werden verschiedene Notwendigkeiten abgedeckt. Neben dem Vertrauen, welches aufgebaut werden kann, kommt es zur Wissensweitergabe und zum Profit für andere. Insbesondere für den Transfer des impliziten Wissens ist dies eine geeignete Möglichkeit (Reinmann-Rothmeier & Vohle 2001, 297). Diese Art der Wissensverteilung kann mit Blogs kombiniert werden. Blogs sind Web-Tagebücher, die durch Verweise auf andere Einträge und ggf. Webseiten das Wissen optimieren sollen. Auch Unternehmen versprechen sich von ihnen einen effektiven Distributionskanal (Peters & Stock 2006, 40).

Wissen nutzen

Das Wissen kann nur genutzt werden, wenn es in Form von Information nutzergerecht aufbereitet ist. Wichtig ist, dass eine gewisse Strukturierung und Übersichtlichkeit vorhanden ist. Ein Tool wird nur dann angenommen und genutzt, wenn es für den Nutzer intuitiv zu bedienen ist. Um eine Struktur herzustellen, sollten Dokumente geordnet werden (Abele et al. 2003, 378). Dies kann mittels normierten Vokabularien (Nomenklaturen, Klassifikationen, Thesauri, Ontologien) geschehen (Stock & Stock 2008).

Wissen bewahren

1. *Datenbankmanagementsystem:* Ein Datenbankmanagementsystem dient der Strukturierung von Daten. Hier können gezielt Daten von mehreren Personen abgerufen werden. Es handelt sich in der Regel um die Informationen, die zur alltäglichen Arbeit notwendig sind.

2. *Intranet/Extranet:* Das Intranet ist eine Informationsquelle für Mitarbeiter. Hier sollten aktuelle Entwicklungen und nützliche Informationen – wie z.B. Normen oder Patente – abgelegt werden. Zugriff sollte jeder Mitarbeiter haben. Auch Kranken- und Urlaubsscheine, sowie die Zeiterfassung der Mitarbeiter können über das Intranet zur Verfügung gestellt bzw. erfasst werden.
Das Extranet hingegen richtet sich an Lieferanten und Kunden und deren Wünsche. Es handelt sich um Seiten im Web, die aber nur für einen bestimmten Kreis zugänglich gemacht wird. Hier werden für den Kunden wichtige Informationen veröffentlicht, aber auch Feedback seitens des Kunden ist erwünscht. Dadurch ist es dem Unternehmen möglich, neues Wissen aufzubauen.

3. *Dokumentation:* Besuchsberichte , Erfahrungsberichte

4. *Debriefing*

5. *Story telling*

6. *Retrieval System:* Ein Retrievalsystem dient dazu Dokumente nutzergerecht und geordnet abzulegen und wiederauffindbar zu machen.

3 Informationsbedarfsanalysen in KMU

Das Ziel einer Informationsbedarfsanalyse ist es zu zeigen, wo Stärken und Schwächen in der Kommunikation eines Unternehmens liegen. Vor allem der subjektive Informationsbedarf soll verbessert und dem objektiven angeglichen werden. Brücher spricht von einer Evaluierung der Bedürfnisse der Mitarbeiter (Brücher 2004, 21). Nach Koreimann hingegen geht es nicht nur um den subjektiven Aspekt, sondern auch um einen allgemeineren.

> Unter Informationsbedarfsanalyse subsumieren wir jene Verfahren und Methoden, die geeignet sind, diejenigen Informationen zu ermitteln, die für die Lösung konkreter betriebswirtschaftlicher Aufgaben im Rahmen eines Unternehmens erforderlich sind (Koreimann 1976, 65).

Eine solche Analyse sollte im Idealfall immer durchgeführt werden, bevor Wissensmanagementsysteme eingesetzt werden. Sie lässt darauf schließen, wo der Schwerpunkt im Unternehmen zu legen ist. Gerade für KMU ist eine solche Analyse wichtig, da diese häufig nicht die Möglichkeit haben, Wissensmanagement als Ganzes einzuführen. Meist werden nur einzelne Bausteine umgesetzt. Durch die Informationsbedarfsanalyse wird gezeigt, wo beim Wissensmanagement angesetzt werden kann. Zusätzlich lässt sich in der Regel so auch abschätzen, welche Systeme oder Tools von Mitarbeitern angenommen werden würden. Außerdem ist aus der Sicht des Wissensmanagements das Wissen über den Bedarf der Nutzer genauso ausschlaggebend wie das Wissen in den Köpfen der Nutzer (Ursel 2006, 2).
Die Durchführung einer Informationsbedarfsanalyse erfolgt in verschiedenen Schritten. Abbildung 6 zeigt, welche Schritte durchgeführt werden müssen:
Im Zusammenhang mit Informationsbedarfsanalysen stehen einige Fachtermini, die zunächst an Hand einer Skizze erklärt werden müssen. Abbildung 7 zeigt das Zusammenspiel von verschiedenen

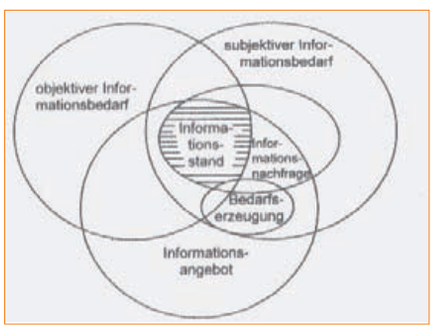


Abbildung 7: Venn-Diagramm zum Informationsbedarf (Quelle: Mujan 2006, 33).

Aspekten, die den Informationsstand in einem Unternehmen bestimmen.
Ausschlaggebend sind folgende drei Elemente: objektiver Informationsbedarf, subjektiver Informationsbedarf und Informationsangebot. Der subjektive Informationsbedarf wird durch das Informationssubjekt (von Spiegel 1994, 1), also den Mitarbeiter, bestimmt. Abhängig ist dieser Bedarf von dem jeweiligen informations- und risikopolitischen Verhalten des Mitarbeiters (Szyperski 1980, 907). Erfahrungen und die Art und Weise, wie man sich in ein Thema einarbeitet, spielen eine große Rolle (Mujan 2006, 26).

Bei dem objektiven Informationsbedarf hingegen geht es in erster Linie um Informationen, die aufgabenbezogen und aus der wertorientierten Unternehmenszielgröße ableitbar sind (Krause & Schmitz 2006, 356). Letztlich entsteht diese Art von Informationsbedarf also durch die Aufgaben, die im Unternehmen anfallen, und ist damit personenunabhängig. Zum objektiven Informationsbedarf zählen u.a. Kundeninformationen, Produktinformationen und Brancheninformationen.

Diese strikte Trennung ist jedoch problematisch. Ein objektiver Informationsbedarf lässt sich nie unabhängig vom subjektiven betrachten (Picot & Franck 1998). Im Idealfall decken sie sich. In der Realität ist dies aber so gut wie gar nicht möglich, da meist eine mangelnde Aufgabenstrukturiertheit oder auch personelle Gründe dagegen sprechen (Welter 2006, 73). Eine Annäherung ist nur dann möglich, wenn der Mitarbeiter über bestimmte Informationsquellen, die der objektive Informationsbedarf fordert, bescheid weiß. Verfügt er nicht über diese Kenntnis, so kann es nicht zu einer Deckung kommen (Mujan 2006, 24). Zu einer Übereinstimmung kommt es in der Regel nur bis zu einem gewissen Grad (Bahlmann 1982, 44).

Unter dem Informationsangebot versteht man die Summe der zu einem bestimmten Zeitpunkt verfügbaren Informationen (Koreimann 1976, 61). Hier geht es nicht nur um das Angebot, welches man innerhalb der Firma auffindet, sondern um das allgemeine Angebot, was sich auf dem Markt befindet. Das können auch Informationen sein, die das Unternehmen sehr sinnvoll einsetzen könnte, aber von denen es gar nicht weiß, dass es sie gibt. Das Diagramm zeigt, dass letztlich nur ein Bruchteil des bestehenden Angebots von der Firma genutzt wird. Des Weiteren ist zu erkennen, dass ein Großteil des Angebots sich ausschließlich mit dem objektiven Informationsbedarf deckt. Die Mitarbeiter scheinen nicht über das Informationsangebot informiert zu sein und können es somit auch nicht für das Tagesgeschäft nutzen. Aus subjektiver Sicht besteht kein Bedarf für die Nutzung des Angebots. Es gibt auch den Fall, dass sich das Informationsangebot mit dem subjektiven, nicht aber mit dem objektiven Informationsbedarf deckt. Hier hat das Angebot jedoch keinen Nutzen für das Unternehmen, sondern es handelt sich meist um persönliche Interessen des Informationssuchenden (ebd., 23).

Neben den bereits erläuterten Schnittmengen im Venn-Diagramm gibt es drei weitere. Der Informationsstand bezeichnet den Teil des Informationsangebots, der sowohl vom objektiven, wie auch vom subjektiven Informationsbedarf gefordert und gefunden wird. Diese Schnittmenge bezeichnet das Ideal und sollte vergrößert werden. So wird eine Optimierung der Informationsversorgung sichergestellt.

Die Informationsnachfrage steht im klaren Zusammenhang mit dem subjektiven Informationsbedarf. Informationen werden immer von bestimmten Individuen erfragt und sind somit auch individuell abhängig von Wünschen, Vorlieben etc. Im Idealfall deckt sich die Informationsnachfrage mit dem Informationsstand.

Das ursprüngliche Venn-Diagramm (Bahlmann 1982) beinhaltet den von Mujan neu eingeführten Aspekt der Bedarfserzeugung nicht. Dieser muss sich mit dem objektiven Bedarf decken, damit er ein Nutzen für das Unternehmen hat. Durch Informationsmarketing und Sensibilisierung von Arbeitnehmern für bestimmte Informationen kann Bedarf erzeugt werden (Mujan 2006, 33). Mit z.B. einem Angebot von Fachinformation kann der Mitarbeiter an die Nutzung herangeführt werden. Auch eine bessere Strukturierung, die das intuitive Suchen ermöglicht, kann zur Bedarfserzeugung beitragen. Neben dem Angebot von bestimmten Informationen kann auch bereits durch die Informationsbedarfsanalyse Bedarf erzeugt werden. In meinem Fall war es z.B. bei einigen Mitarbeitern so, dass sie nicht alle Funktionen des eigens entworfenen Datenbankmanagementprogramms kannten. Durch Fragen zum Lieferprogramm erfuhr der eine oder andere Mitarbeiter, dass es überhaupt die Möglichkeit gibt, ein Lieferprogramm anzulegen. Insbesondere die Mitarbeiter, die noch nicht so lange im Unternehmen waren, kannten diese Funktion nicht. Zu ihrer Verteidigung muss man jedoch sagen, dass dieser Punkt in dem Unternehmen auch noch nicht aktiv genutzt wird. Dies soll sich jedoch in Zukunft ändern, so dass jeder von der Existenz wissen sollte. In diesem Fall wurde also ein Bedarf erzeugt, der ohne das Interview eventuell erst viel später aufgedeckt worden wäre. Es ist also wichtig, bereits im Interviewleitfaden Informationsangebote aufzunehmen, von denen man – als Fachexperte für Wissensmanagement und für Informationsdienste – annehmen kann, dass diese den objektiven Informationsbedarf decken können, wiewohl man überzeugt ist, dass nicht jeder Interviewpartner (subjektiv) einen entsprechenden Bedarf verspürt. So kann direkt im Interview durch die Besprechung bislang unbekannter Informationsangebote Bedarf daran geweckt werden.

Möchte man in einem Unternehmen eine Informationsbedarfsanalyse durchführen, so muss zunächst geklärt werden, bei welchem Bereich begonnen werden soll. Welche verschiedenen Bereiche es gibt, zeigt Abbildung 8.

Im nächsten Schritt wird überlegt, welches Verfahren man für die Durchführung

Abbildung 8: Mögliche Bereiche einer Informationsbedarfsanalyse (Quelle: www.lead-gmbh.de).

|a|S|tec|
angewandte Systemtechnik GmbH

aDIS/Wissen –
für ein effektives Informations- und Wissensmanagement

- Thematisches Gruppieren beliebiger Dokumente und Informationsquellen
- Automatisierte Wissensakquisition
- Einfache Gestaltung der Wissenspräsentation

Weitere Produkte:
aDIS/BMS
aDIS/Archiv

|a|S|tec| GmbH
Paul-Lincke-Ufer 7c
10999 Berlin

Tel.: (030) 617 939-0
Fax: (030) 617 939-39
info@astecb.astec.de

http://www.astec.de

	Standardisiertes Interview	Halbstandardisiertes Interview	Nicht standardisiertes Interview
Formulierung der Fragen	feststehend	teils feststehend - teils frei	frei
Reihenfolge der Fragen	feststehend	Grundgerüst steht fest	frei
Inhaltliche Vorkenntnisse	hoch	mittel	gering
Terminologie	einheitlich	weitgehend einheitlich	uneinheitlich
Kenntnisse des Interviewers	gering	mittel bis hoch	hoch

Abbildung 9: Verschiedene Interviewformen (Quelle: Fank 2001, 249).

benutzt. Die verschiedenen Methoden lassen sich in deduktive und induktive Verfahren einteilen. Erste sind aufgabenbezogene, sachlogische Verfahren, wobei der Informationsbedarf systematisch ermittelt wird. Das Augenmerk liegt hier auf dem objektiven Informationsbedarf. Beispiele hierfür sind: Strategieanalysen und Prozessanalysen. Letztere sind personenbezogen und ermitteln den subjektiven Informationsbedarf (Krause & Schmitz 2006, 352). Unter induktive Verfahren fallen offene Befragungen der Mitarbeiter, Beobachtungen, Dokumentenanalysen und Organisationsanalysen. Des Weiteren gibt es noch die gemischten Verfahren, die eine strukturierte Befragung durchführen und kritische Erfolgsfaktoren ermitteln (Krcmar 2005, 62).
Eine Informationsbedarfsanalyse kann mit verschiedenen Techniken durchgeführt werden. Typische Instrumente zur Ermittlung des Informationsbedarfs sind: Fragebogen, Interviews, Beobachtung, Dokumentenanalyse und Selbstaufschreibung (Fank 2001, 245). Möglichkeiten der Interviewtechnik zeigt Abbildung 9.

Günstig ist nach eigenen Erfahrungen ein halbstandardisiertes Interview zu wählen. Das Grundgerüst wird durch den, dem Interviewer vorliegenden Fragebogen, gestellt. Diese feststehenden Fragen können aber – je nach Gesprächsverlauf – teilweise umformuliert werden.
Um den Fragebogen auf Tauglichkeit zu testen, werden zunächst *Pretests* durchgeführt. Der Vorteil eines Interviews ist, dass auf den Einzelnen konkret (bis zu einem gewissen Maße) eingegangen werden kann. Rückfragen sind auf beiden Seiten möglich. Bei einer Informationsbedarfsanalyse mit einem Fragebogen kann es passieren, dass Unsicherheiten durch nicht verstandene Fragen oder auch Antworten entstehen. Ein weiteres Problem der Fragebögen ist die Rücklaufquote, die meist sehr gering ist.
Ein persönliches Interview besitzt neben der kommunikativen Komponente auch eine persönliche: Neben der eigentlich formulierten Antwort kann auch die Reaktion beobachtet, interpretiert und dokumentiert werden (Fank 2001, 258). Neben diesen Vorteilen gibt es auch Nachteile. Die wahrscheinlich größte Gefahr ist, dass die Mitarbeiter bei einem Interview Angst haben, sich frei zu äußern. Auch wenn es anonymisiert ist, weiß doch der Interviewer, wer was gesagt hat. So besteht häufig die Befürchtung, dass etwas Gesagtes nach außen dringt. Auf Grund dessen ist es bei Interviews vielfach zu beobachten, dass die Mitarbeiter so antworten, wie sie denken, dass es von ihnen erwartet wird.

Die Aufgabe des Interviewers ist es, Fragen neutral zu stellen und eine emotional freie Ebene herzustellen. Färbungen, Meinungen, Beschönigungen und Wertungen sollten vermieden werden. Ein sachliches Eingreifen des Interviewers ist möglich, jedoch sollten die Antworten nicht gelenkt werden. Um die Auswertung genauer und einfacher zu machen, wird nach Absprache mit den Interviewten ein Diktiergerät eingesetzt. So kann eine exakte Reproduktion gewährleistet werden.
Wie viele Leute für eine Informationsbedarfsanalyse befragt werden sollten, ist individuell verschieden. Die Anzahl der Befragten sollte jedoch repräsentativ sein. Ich beschränkte die Interviews auf die Verkaufsabteilungen. Es gab vier Abteilungen, die für mich von Interesse waren. In diesen vier Abteilungen arbeiteten ca. 45 Mitarbeiter, interviewt wurden 16. Für ein Interview sollte in einem KMU ein fachkundiger Wissensmanager beschäftigt werden. Mit Aushilfskräften dürfte es schwierig werden, die Interviews durchzuführen. Natürlich ist eine Informationsbedarfsanalyse zunächst ein Projekt. Im Gegensatz zu Großunternehmen, die über eigene Wissensmanagementabteilungen verfügen, ist bei einem KMU die Beschäftigung des Wissensmanagers zeitlich begrenzt. Verantwortlich für die Umsetzung sind die Firma bzw. insbesondere die Vorgesetzten. Diese müssen als erstes entscheiden, was umgesetzt wird. Ist diese Entscheidung gefallen ist es wichtig, den Mitarbeiter von der Wichtigkeit der Umsetzung zu überzeugen, da ohne seine Mitarbeit nichts funktioniert.

Fazit

Der Artikel verdeutlicht, dass Wissensmanagement nicht nur für große Unternehmen oder Konzerne wichtig ist, sondern durchaus mittlerweile auch kleine und mittelständische Unternehmen Interesse daran finden. Der große Unterschied zwischen KMU und Großunternehmen ist darin zu sehen, dass zum einen weniger finanzielle Mittel vorhanden sind und auch meistens nicht eigens für das Wissensmanagement ausgeschriebene Stellen, wie etwa „Wissensmanager", zur Verfügung stehen. Damit eine Umsetzung von Wissensmanagement in KMU nicht an Zeitmangel der Mitarbeiter scheitert, muss gut abgewogen werden, wo in einzelnen Untenehmen angesetzt werden sollte. Dafür eignet sich eine Informationsbedarfsanalyse, die zwar mit einem gewissen Zeitaufwand verbunden ist, trotzdem aber aufzeigen wird, bei welchen Wissensbausteinen die Implementierung von Wissensmanagementsystemen angefangen werden sollte.
Ein empfehlenswertes Verfahren zur praktischen Durchführung einer Informationsbedarfsanalyse ist ein halbstandardisiertes Interview mit einem Interviewer, dem ein Fragebogen vorliegt. Durch die gezielte Befragung einzelner Mitarbeiter lässt sich herausfinden, welche Arbeitsroutinen Unterstützung durch Wissensmanagements gebrauchen können. Es können in einzelnen Abteilungen Informationsbedarfsanalysen durchgeführt werden, so dass schon vorher eine grobe Richtung festgelegt werden kann, wo als erstes innerhalb der Firma angesetzt werden soll.

Literatur

Abele, Eberhard; *Kuhn*, Sven Martin & *Liebeck*, Tobias (2003). Wissensmanagement in KMU, Entwicklung eines Wissensmanagement Werkzeugkastens für kleine und mittelständische Unternehmen. In: Zeitschrift für wirtschaftlichen Fabrikbetrieb, 98(7-8), S. 375-379.

Bahlmann, Arnold (1982). Informationsbedarfsanalyse für das Beschaffungsmanagement. In: Betriebswirtschaftliche Schriften zur Unternehmensführung, Bd. 41, Betriebliche Logistik, Gelsenkirchen: Manhold.

Bendel, Oliver (2007). Social Software als Mittel des Wissensmanagements in KMU. In: *Belliger*, A. & *Krieger*, D. (Hrsg.): Wissensmanagement für KMU. Zürich: vdf Hochschulverlag, S. 93-110.

Erlach, Christine & *Thier*, Karin (2004). Mit Geschichten implizites Wissen in Organisationen heben. In: *Wyssusek*, B. (Hrsg.), Wissensmanagement komplex, Perspektiven und soziale Praxis, Berlin: Erich Schmidt.

Fank, Matthias (2001). Einführung in das Informationsmanagement. München: Oldenbourg, 2. Auflage.

Fink, Dieter (2001). Knowledge Management in small and medium sized professional accounting firms: progress and challenges. In: Wirtschaftsinformatik 43 (5), S. 487-496.

Fink, Kerstin & *Ploder*, Christian (2005). Wissensmanagement muss nicht teuer sein. In: Wissensmanagement 07/05, S. 25-27.

Fischer, Hajo (2007). Wissensmanagement für KMU: Wikis als Instrumente des Wissensmanagements – eine Einführung. Verfügbar unter: www.wm-impulse.net/archive/2007/02/000082.html [04.06.07].

Groegger, Bernd (2001). Vorsprung durch Wissen; Beratung für kleine und mittlere Unternehmen. In: TBS Schnittpunkt. Verfügbar unter: www.faktorwissen.de/material [27.02.07].

Guretzky von, Bernhard (2002). Wissensmanagement in kleinen und mittleren Unternehmen. Verfügbar unter: www.community-of-knowledge.de/cp_artikel.htm?artikel_id=134 [04.06.07].

Güttel, Wolfgang H. & *Dietrich*, Angelika(2001). Aspekte der externen Informationsbeschaffung im Rahmen des Wissensmanagements in kleinen und mittleren Beratungsunternehmen. In: Wirtschaftsinformatik 43 (5), S. 477-486.

Gust von Loh, Sonja (2008). Wissensmanagement und Informationsbedarfsanalyse in kleinen und mittleren Unternehmen. Teil 1: Grundlagen des Wissensmanagement. In: Information – Wissenschaft und Praxis 59(2008)2, S. 118-126.

Hesseler, Michael (2005). Die wirtschaftlichen Vorteile von Wissensmanagement einschätzen – aber wie? In: Wissensmanagement 1/05, S. 20-23.

Jörs, Bernd (2006). Wissensmanagement- zwischen Anspruch und Wirklichkeit- eine informationswirtschaftliche Bestandsaufnahme und Orientierung für KMU. In: REFA- Nachrichten, 2, S. 30-39.

Koreimann, Dieter (1976). Methoden der Informationsbedarfsanalyse. Berlin: De Gruyter.

Krause, Herbert & *Schmitz* ,Uwe (2006). Entwicklung einer Methode für werttreiberbezogene Informationsbedarfsbestimmung. In: Controlling 7/06, S. 351-361.

Krcmar, Helmut (2005). Informationsmanagement. Berlin: Springer, 4. Auflage.

www.lead-gmbh.de [04.06.07].

Lim, Kwang K. & *Klobas*, Jane (2000). Knowledge management in small enterprises. In: Electronic Library 18(6), S. 420-432.

Majchrazak, Ann et al. (2006). Corporate wiki users: results of a survey. In: Proceedings of the 2006 international symposium on Wikis. WikiSym '06, S. 99-104.

Moffett, Sandra & *McAdam*, Rodney (2006). The effects of organizational size on knowledge management implementation: Opportunities for small firms? In: Total Quality Management & Business Excellence 17(2006)2, S. 221-241.

Mujan, Dzemal (2006). Informationsmanagement in Lernenden Organisationen. Berlin: Logos.

Pan, Shan L.; *Newell*, Sue; *Huang*, Jimmy & *Galliers*, Robert D. (2007). Overcoming knowledge management challenges during ERP implementation: The need to integrate and share different types of knowledge. In: Journal of the American Society for Information Science and Technology 48(2007)3, S. 404-419.

Peters, Isabella & *Stock*, Wolfgang G. (2006). Corporate Blogs im Wissensmanagement. In: Wissensmanagement 8(2006)6, S. 40-41.

Peters, Isabella & *Stock*, Wolfgang G. (2007). Web 2.0 im Unternehmen. In: Wissensmanagement 9(2007)4, S. 22-25.

Picot, Arnold & *Franck*, Egon (1988). Die Planung der Unternehmensressource Information (II). In: Das Wirtschaftsstudium 17(1988)11, S. 608-614.

Prescher, Claudia (2006). Wissensmanagement in KMU : eine exemplarische Studie zum Umgang mit Wissen in sächsischen Unternehmen. In: *Bergmann* & *Baerbel* (Hrsg.): Kompetent für die Wissensgesellschaft. Münster: Waxmann, S. 405-424.

Probst, Gilbert; *Wiedemann*, Christian & *Armbruster*, Heidi (2001). Wissensmanagement umsetzen: drei Instrumente in der Praxis. In: New Management Nr. 10, S. 37-43.

Probst, Gilbert, *Romhardt*, Kai & *Raub*, Steffen (2006). Wissen managen. Wiesbaden: Gabler, 5. Auflage.

Reinmann-Rothmeier, Gabi & *Vohle*, Frank (2001). Was Schiedsrichter, Manager und Rotkäppchen gemeinsam haben: Mit Geschichten Wissen managen. In: Zeitschrift für Führung und Organisation 5, S. 293-300.

Schauer, Hanno & *Frank*, Ulrich (2007). Methoden des Wissensmanagements. In: *Belliger*, A. & *Kriger*, D. (Hrsg.): Wissensmanagement für KMU. Zürich: vdf Hochschulverlag, S. 147-166.

Schiller, Gabriele & *Guretzky*, Bernhard von (2004). Bausteine für eine innovationsorientierte, wissensbasierte Unternehmenskultur. Verfügbar unter: www.community-of-knowledge.de/cp_artikel.htm?artikel_id=134 [04.06.07].

Schneider, Ursula (2001). Die 7 Todsünden im Wissensmanagement. Kardinaltugenden für die Wissensökonomie. Frankfurt: Frankfurter Allgemeine Zeitung Verlagsbereich Buch.

Schütt, Peter (2000). Die Macht der Geschichten. In: Wissensmanagement 2(2000)5, S. 10-14.

Schütt, Peter (2005). Blogs und Wikis: Mehr Mitarbeit wagen. In: Wissensmanagement 7(2005)7, S. 14-16.

Schütt, Peter (2006). Die Werkzeugkiste des Wissensmanagements. In: Wissensmanagement 8(2006)8, S. 10-12.

Southon, Gray; *Tod*, Ross & *Seneque*, Megan (2002). Knowledge management in three organizations: An exploratory study. In: Journal of the American Society for Information Science and Technology 53(12), S. 1047-1059.

Specht, Rolf (2007). „Soft Factors" – Die Relevanz von psychologischen Faktoren im Wissensmanagement. In: Belliger, A. & Kriger, D. (Hrsg.): Wissensmanagement für KMU. Zürich: vdf Hochschulverlag, S. 37-41.

Spiegel, Josephin von (1991). Methoden der Informationsbedarfsanalyse – Beschreibung einer Vorgehensweise zur Bestimmung des Bedarfs an externen Informationen im Management. In: *Schiefer*, G. (Hrsg.): Informationsmanagement in Agrar- und Ernährungswirtschaft, Bericht A-91/1, Bonn: ILB.

Stock, Mechtild & *Stock*, Wolfgang G. (2008). Wissensrepräsentation. Informationen auswerten und bereitstellen. München: Oldenbourg.

Stock, Mechtild & *Stock*, Wolfgang G. (2003). Dialog, Factiva, GBI und Genios im Qualitätsvergleich. Online Hosts für Wirtschaft und News auf dem deutschen Informationsmarkt. Eine komparative Analyse. In: Password 7&8/2003.

Szyperski, Norbert (1980). Informationsbedarf. In: *Grochla*, Erwin (Hrsg.): Handwörterbuch der Organisation, Stuttgart: Poeschel, 2. Auflage, S. 904-913.

TBS NRW zur Gestaltung von Arbeit, Technik, Organisation und Gesundheit (2003). Handlungshilfe Wissensmanagement für KMUs. Verfügbar unter: www.faktorwissen.de/material [04.06.07].

Ursel, Doussou (2006). Nutzungs- und Bedarfsanalyse im Wissensmanagement. Projekt: Wissensmanagement, Universität des Saarlandes, Informationswissenschaft (Hausarbeit).

Weichbrodt, Rainer (2001). Das Wissensportal – Einstieg in Wissensmanagement bei KMU. In: Wissensmanagement 3(2001)5, S. 12-15.

Weiss, Aaron (2005). The power of collective intelligence. In: netWorker 9 (2005)3, S. 16-23.

Weissenberger-Eibl, Marion & *Kelm*, Alexander (2005). Wissensmanagement in KMU – ein Umsetzungsproblem? In: Wissensmanagement 7(2005)6, S. 29-31.

Welter, Markus (2006). Der Informationsbedarf von Dienstleistungsunternehmen. Besonderheiten und Ansatzpunkte für ein dienstleistungsspezifisches Informationsmanagement. In: Information Management & Consulting 21(2006)1, S. 72-78.

Wong, Kuan Yew (2005). Critical success factors for complementing knowledge management in small and medium enterprises. In: Industrial Management & Data Systems 105(2005)3-4, S. 261-279.

Zaunmüller, Hanna (2005). Anreizsysteme für das Wissensmanagement in KMU: Gestaltung von Anreizsystemen für die Wissensbereitstellung der Mitarbeiter. Wiesbaden: Deutscher Universitätsverlag.

Zucker, Betty (2000). Wissen gewinnt: Innovative Unternehmensentwicklung durch Wissensmanagement. Düsseldorf: Metropolitan Verlag.

Wissensmanagement, Informationsbedarfsanalyse, KMU, Kleine und mittelständische Unternehmen, Wissensmanagement-Tools, Web 2.0, Unternehmenswiki, Informationsbedarf, Informationsangebot, Informationsnachfrage, Informationsbedarfserzeugung, Mitarbeiterbefragung, Interview

Die Autorin

Sonja Gust von Loh, B.A., M.A.

(geb. Weber) ist seit August 2007 wissenschaftliche Mitarbeiterin der Abteilung für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf. Sie schloss ein B.A.-Studium in den Fächern Romanistik und Informationswissenschaft und ein M.A.-Studium in Informationswissenschaft und Sprachtechnologie ab. 2002 hat sie einen Sprachaufenthalt in Florenz sowie ein Praktikum beim Röhrenproduzenten Dalmine (nahe Bergamo), 2003 ein Praktikum beim Italienischen Kulturinstitut in Köln absolviert. Sie lebte und studierte sechs Monate in Rom. Ihre Masterarbeit schrieb sie in einem mittelständischen Unternehmen zum Thema „Wissensmanagement". Seit Juni 2007 ist sie Mitglied im erweiterten Vorstand der DGI.

Institut für Sprache und Information
Abteilunformationswissenschaft
Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1
40225 Düsseldorf
Telefon 0211 81-10803
Telefax 0211 81-12917
gust-von-loh@phil-fak.uni-duesseldorf.de
www.phil-fak.uni-duesseldorf.de/infowiss/

On Click Fraud

Simone Soubusta, Düsseldorf

Internet advertising is getting more and more important. As in every growing line of business, criminals try to profit from that growth. They do so by manipulating Pay-Per-Click (PPC) adverts with artificially generated clicks, a process which is known as Click Fraud (CF). CF is not difficult to commit and, consequently, it could threaten the existence of companies like Google. At present, there seems to be no way to either detect or prevent CF. It can, however, be altogether avoided by switching from PPC to other advertising schemes like Cost-Per-Action.

Über Klickbetrug

Internetwerbung wird immer wichtiger. Wie in jeder Wachstumsbranche versuchen auch hier Kriminelle von diesem Wachstum zu profitieren. Dies ist durch die Manipulation von Pay-Per-Click (PPC) Werbeanzeigen mithilfe künstlich produzierter Klicks möglich – eine Vorgehensweise, die als Klickbetrug (Click Fraud) bekannt ist. Klickbetrug ist nicht schwierig und könnte daher die Existenz von Firmen wie Google bedrohen. Zur Zeit scheint es keine Möglichkeit zu geben, Klickbetrug zu entdecken oder zu verhindern. Man kann Klickbetrug jedoch vermeiden, indem man PPC Werbung durch eine andere Berechnungsgrundlage, wie z. B. Cost-Per-Action, ersetzt.

1 Introduction

1.1 Internet Advertising

It is a truth universally acknowledged that the World Wide Web (WWW) has revolutionised our planet. Today the web allows us to communicate with people nearly anywhere in the world – and, indeed, with billions of people at once. It has changed the way we think, as well as the way we do business. But the web is not only a means of communication, it also contains a wealth of knowledge. Since the rise of the search engines, huge amounts of information have become readily accessible. The known web alone spans several billion pages.

Most of the information and services on the web come at no cost to the user. The publisher or webmaster, on the other hand, has to pay to provide them. In order to profit from their complimentary services, webmasters rent out space on their websites to advertisers. This is the main means of covering the costs of providing online services nowadays. The most famous example of this phenomenon is probably Google. The search engine accumulated considerable wealth by auctioning off advertising space next to their search results.

Advertising plays an important role for every company. "Most businesses operate with an advertising budget of 2-5 percent of their previous year's gross sales" (Egelhoff, n.d.). Since its inception, advertising has thrived on mass media like newspapers and TV, where an ad could reach a wide audience. It comes as no surprise, then, that it has found the web a suitable playground.

1.2 Cyber Crime

The rise of the internet has also brought a completely new kind of crime: cyber crime. The most well-known type of cyber crime is *Phishing,* which means stealing sensitive information online. This can be anything from user name/password combinations to credit card information and the troublesome social security number in the U.S. The criminal gains this information by masquerading as an authority. In order to gain access to the PIN/TAN combination of a user's bank account, for example, the criminal might write an e-mail to the client, in which he pretends to be the bank itself. He presents the user with a link that allegedly leads to the bank's online banking login page, where the user is asked to log in for 'maintenance reasons' or something similar, but in reality the link leads to the phisher's website (which imitates the look and feel of the bank's site), allowing him to steal the information and, therefore, the user's money. Two other types of cyber crime are *credit card fraud,* which the internet has made infinitely easier to commit, and *spam,* which doesn't warrant further explanation.

However, there is another kind of cyber crime that the general public is not well acquainted with, but which is nonetheless detrimental to many people: so-called *click fraud,* which this paper will examine in detail.

1.3 Structure of the Article

Chapter 2 introduces the reader to the concept of click fraud against the background of Pay-Per-Click advertising, followed by a short look at the legal ramifications. Chapter 3 expounds how click fraud is committed. An explanation of how a click can be faked is followed by two different techniques that are used to avoid detection: using a network of hijacked computers (a so-called botnet) to commit click fraud, and making visitors of a website unintentionally assist in the fraud. Chapter 4 presents different ways to detect click fraud, as well as ways to prevent click fraud altogether. Finally, chapter 5 concludes the article with a look at its result. (see Fig. 1)

2 What is Click Fraud?

In the past, online advertising used the Cost-Per-Impression model to charge for advertisements. The cost per impression is often measured in Cost Per Mile (CPM), that is, the cost of one thousand impressions of the ad. This advertising model was based on traditional TV and print advertising, where the advertiser is also charged on the basis of the number of times that the ad is viewed. Cost-Per-Impression is the preferred model of publishers (webmasters), because they are paid regardless of the effectiveness of an ad. However, search engines such as Google have given rise to the Pay-Per-Click (PPC) model of online advertising. In this type of arrangement, advertisers pay a certain amount of money to the publisher for every click on their ad (which leads to the advertiser's website). That is to say, the costs are performance-dependent. Consequently, the model is favoured by advertisers, who get an interested visitor to their website for the money they pay.

In particular, Google's AdWords and AdSense (see Fig. 1 and 2) services have added to the widespread use of PPC. With AdWords, advertisers bid in an auction for the placement of their ad next to the search results for a chosen keyword (or a combination of keywords). This form of advertising is also known as sponsored

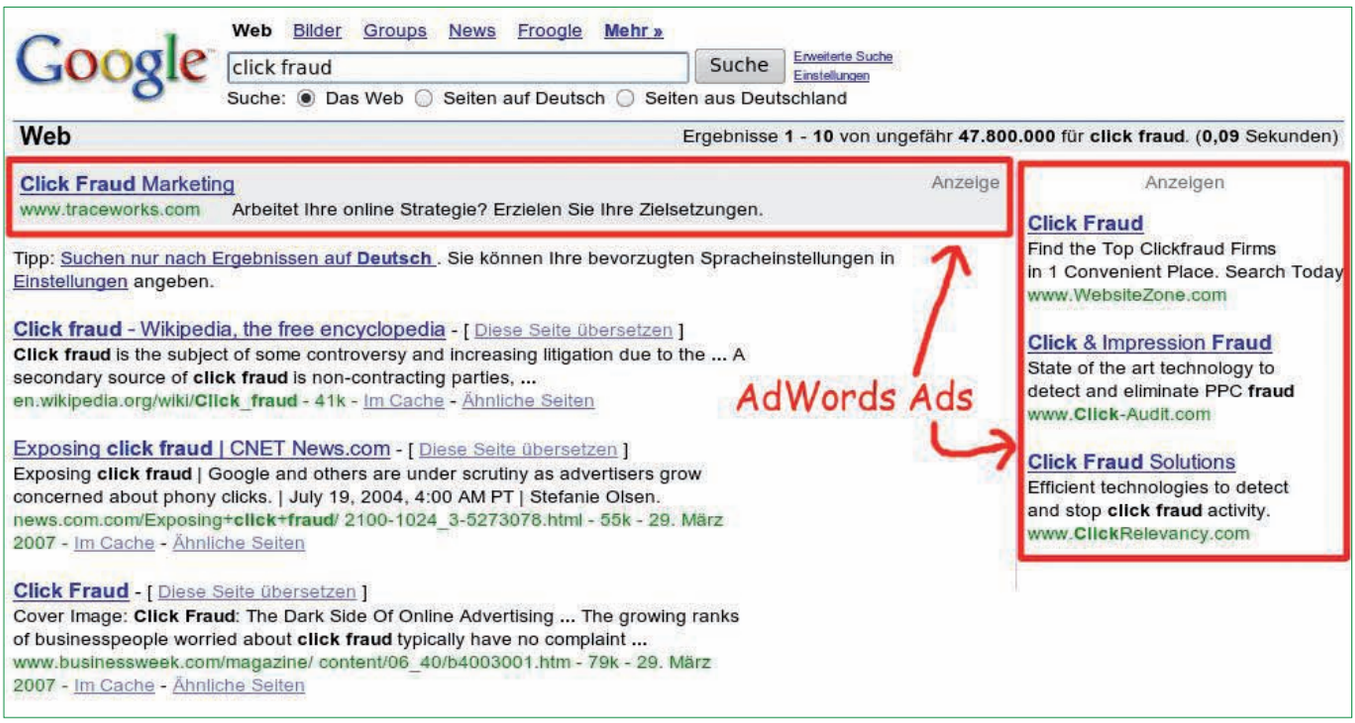


Figure 1. Google AdWords.

search. The obvious advantage for advertisers is that the user who sees the advertisement is already interested in the topic that is connected to the keyword (since he searched for it), hence there is a high chance that he is also interested in the product or service that is advertised next to it. A user who searches for the term *camera*, for example, might be shown an advertisement for a camera by Canon. In addition, the user also profits from this *targeted advertising*, since he is only shown ads for products he likely has an interest in.

AdSense, on the other hand, is an advertising network in which Google acts as a commissioner. Similar to AdWords, advertisers bid on certain keywords. Google then automatically analyses the content of websites that have signed up as publishers for AdSense and places the ad on those websites that contain content which is closely related to the keyword. Every time someone clicks on such an ad, the website owner gets a percentage of the proceeds which Google receives from the advertiser, and the rest is Google's commission.

From this point of departure, we can distinguish two types of click fraud: competitor click fraud and publisher click fraud.

Competitor click fraud is not committed for any monetary gain (and, in fact, there is none), but rather in order to harm one's competitor. A prerequisite for the fraud is that the competitor in question has signed up as advertiser in a PPC scheme (e.g. Google AdWords). The fraudster, knowing that every click on the ads costs the competitor good money, clicks repeatedly on his ad to cause harm. He might perform this task himself or use more sophisticated means (see below).

Publisher click fraud is, as the name indicates, performed by the publisher of an ad, that is, by the owner of the website on which the ad (or, indeed, multiple ads) is displayed. It is committed solely for monetary gain. Since the publisher is paid for every click on an ad, it is in his interest to receive a high number of clicks. Thus, the fraudster artificially increases the number of clicks in order to make more money. Harmless dilettantes do this by clicking on the ads themselves or by asking their friends to do so. Professional fraudsters, on the other hand, either outsource the job to armies of 'clickers' in India (Vidyasagar, 2004) or China, or they automate the process entirely. This paper will concern itself primarily with the latter option.

According to Noogie C. Kaufmann from the University of Münster, *competitor click fraud*, as described above, is illegal under German law since it violates § 4 Nr. 10 UWG (purposeful hindrance of competitors) and § 826 BGB (immoral deliberate damnification) (Kaufmann, n.d.). Unfortunately, though, there hasn't been an original precedent yet to confirm this assumption. Nonetheless, competitor click fraud is a big problem for the harmed party, and several advertisers who participate in Google's AdWords program have sued Google "for their failure to prevent click fraud" (Hadjinian, 2006). Their point of attack is the contract between the two parties which states that only *actual clicks* will be charged, which enabled them to sue Google for breach of contract when they noticed that they were charged for fraudulent clicks as well as for actual ones (Hadjinian, 2006).

Publisher click fraud is, in most cases, simply a breach of contract. The Terms of Service of Google's AdSense program, for example, state: "You shall not, and shall not authorize or encourage any third party to: (i) directly or indirectly generate queries, Referral Events, or impressions of or clicks on any Ad, Link, Search Result, or Referral Button through any automated, deceptive, fraudulent or other invalid means, including but not limited to through repeated manual clicks, the use of robots or other automated query tools and/or computer generated search requests, and/or the unauthorized use of other search engine optimization services and/or software; [...] (iii) frame, minimize, remove or otherwise inhibit the full and complete display of any Web page accessed by an end user after clicking on any part of an Ad ("Advertiser Page"), any Search Results Page, or any Referral Page; (iv) redirect an end user away from any Advertiser Page, Search Results Page, or Referral Page; provide a version of the Advertiser Page, Search Results Page, or Referral Page that is different from the page an end user would access by going directly to the Advertiser Page, Search Results Page, or Referral Page; intersperse any content between the Ad and the Advertiser Page, between the page containing the Search Box and the Search Results Page, or between the Referral Button and the Referral Page; or otherwise provide anything other than a direct link from an Ad to an Advertiser Page, from the page containing the Search Box to the Search Results Page, or

from the Referral Button to the Referral Page" (*Google AdSense Online Standard Terms and Conditions*, n.d.)
Based on these terms of service Google won a court battle against a participant of its AdSense program, Auctions Expert, in 2004 (Hadjinian, 2006). Although this case could be seen as establishing a precedent, it does not completely solve the publisher click fraud problem from a legal perspective, since the fraudster might not fall under the jurisdiction of the U.S. law.

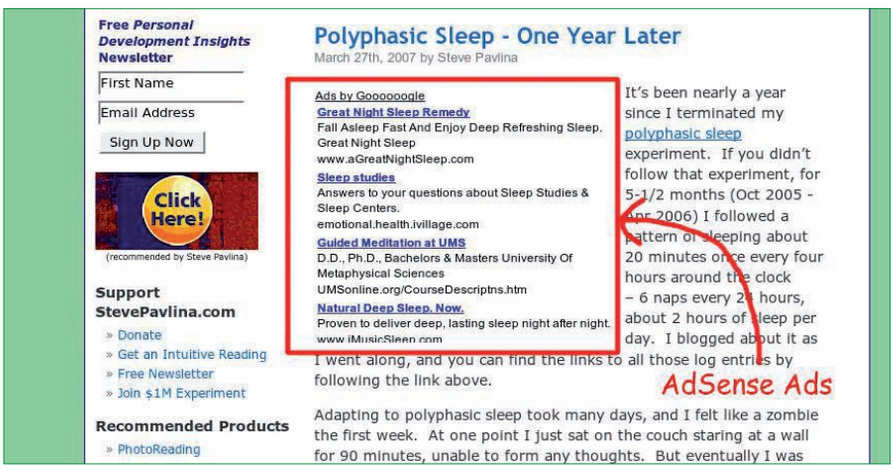


Figure 2. Google AdSense.

3 Why Does Click Fraud Matter?

Google can easily be used to exemplify the explosiveness that click fraud presents today: This unquestionably famous billion-dollar concern makes 99 per cent of its turnover through Pay-Per-Click advertising. If click fraud is not countered in the near future, the backbone of its business threatens to collapse. Google's advertising revenues have risen from 6.07 billion US dollars in 2005 to 10.49 billion dollars in 2006. In both years (as well as in 2004) the advertising revenues accounted for 99 per cent of the overall revenues. In 2006, 60 per cent of the revenues (i.e. 6.29 billion dollars) were made through Google AdWords, a system that is susceptible to competitor click fraud, while the remaining 40 per cent (i.e. 4.2 billion dollars) were made through Google AdSense, a system that is susceptible to publisher click fraud (*Google Annual Report 2006,* 2007).
But Google is not the only one to profit from internet advertising. In the year 2005 the total revenue of internet advertising amounted up to 12.5 billion dollars. In 2006 it rose by 35 per cent to 16.9 billion dollars (*IAB Internet Advertising Revenue Report,* 2007). And the turnover in the first quarter of 2007 was an all-time high of 4.9 billion dollars (*Internet Advertising Revenues Soar*, 2007). Even though the numbers might look small compared to TV advertising revenues of 48.35 billion dollars in 2006 (*2006 TV Ad Revenue Figures,* 2007) and print advertising revenues of 46.6 billion dollars (*Mind the Gap*, 2007), the trend is clearly in favour of online advertising: TV advertising revenues recorded only a rise of 5.3 per cent and revenues of print advertising actually fell 1.7 %.
Although, as we will see later, it is not (yet) possible to accurately estimate the number fraudulent clicks, studies suggest that about 14 percent of all clicks aren't proper (Mills, 2006a; Mills 2006b). Consequently, companies like Google have two alternatives to avert the danger that click fraud presents for their business: Firstly, they can open up new business segments in order to put an end to their total dependency on their advertising customers. Secondly, they can stop the fraudsters before they cause any damage. However, in order to counter click fraud, one needs an understanding of how it works, which the following chapter will provide.

4 How it Works

4.1 Simulating a Click

At the heart of (automated) click fraud lies the simulation of a click on an advertisement. It follows, then, that in order to understand how click fraud works, we must first gain an understanding of the technology behind the advertisement. "Typical online advertisement services [...] work by providing webmasters a snippet of JavaScript code to add to their pages. This code is executed by the web browser of a visitor to the site, and downloads ads from the advertiser's server at that time. The ad download triggers a rewrite of the frame in which the JavaScript appears, replacing it with the HTML code necessary to display the ads. When a user clicks an advertisement link, they "click through" the ad provider's server, giving the ad provider the opportunity to bill the client for the click. The user is then taken to the ad client's homepage" (Gandhi et al., 2006).
It is evident that a program which simulates clicks has to perform many of the same tasks as a browser. First, it has to execute the JavaScript code that retrieves the HTML code of the ad, as mentioned above. Then it has to parse this HTML code for links (indicated by a leading `href="` and a trailing ") and, lastly, it has to send an HTTP (Hypertext Transfer Protocol) request to the web server at the given URL (Uniform Resource Locator).

4.2 Distributed Click Fraud with Botnets

Simulating a click is, in effect, rather easy. However, detecting this kind of click fraud is just as easy. The reason for that is simple: when the program sends an HTTP request to the advertiser's (or commissioner's) server, the IP address of the computer making the request is transmitted in order to establish a connection between client and server. Thus, all it takes to detect the fraudster is for the advertiser to take a look at his log files, where a high number of requests from one IP address provides sufficient evidence of the fraud.
To avoid being discovered and to increase the efficiency of the fraud, the fraudster can distribute the program so that it does its work from all over the internet, with the help of a so called *botnet*. A botnet is a network of thousands or even millions (Keizer, 2005) of hacked computers, all of which do the bidding of the owner of the botnet.

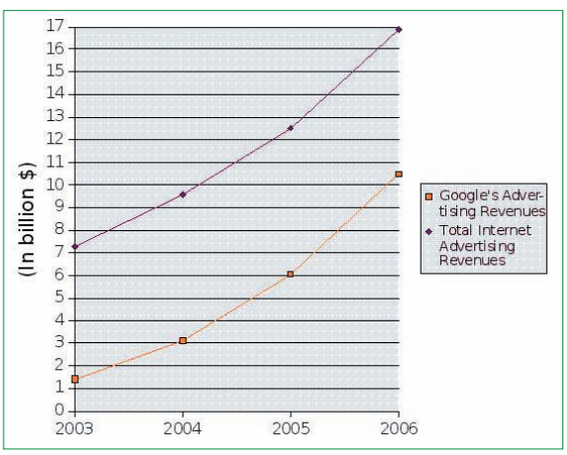


Figure 3. Web Advertising Revenues 2003 – 2006. Sources: IAB Internet Advertising Revenue Report, 2007; Google Annual Reports.

4.2.1 Taking over a Computer

The targeted computer(s) can be compromised by exploiting security holes and vulnerabilities. The program which does the exploiting is commonly referred to as an exploit. Attackers either write these exploits themselves or, more commonly, use exploits for known security holes that are available on the internet. Most exploits utilise buffer overflows (Override, 2001).

"A buffer overflow occurs when a program or process tries to store more data in a buffer (temporary data storage area) than it was intended to hold. Since buffers are created to contain a finite amount of data, the extra information – which has to go somewhere – can overflow into adjacent buffers, corrupting or overwriting the valid data held in them. Although it may occur accidentally through programming error, buffer overflow is an increasingly common type of security attack on data integrity. In buffer overflow attacks, the extra data may contain codes designed to trigger specific actions [...]" (*Buffer Overflow*, n.d.).

Once the attacker has decided upon a specific exploit, he begins scanning (IP) address blocks for systems which fulfil the requirements of the exploit, i.e. that are running a certain version of an operating system as well as the vulnerable program. One rather popular way to do these scans is to use the open source program Nmap, which "uses raw IP packets in novel ways to determine what hosts are available on the network, what services (application name and version) those hosts are offering, what operating systems (and OS versions) they are running, what type of packet filters/firewalls are in use" (*Nmap Security Scanner*, n.d.).

Once a vulnerable system has been detected, the attacker uses the exploit to gain remote access to that computer. He then deposits a program which will contact the botnet's command and control centre, allowing the machine to be controlled from a central point.

4.2.2 Command & Control

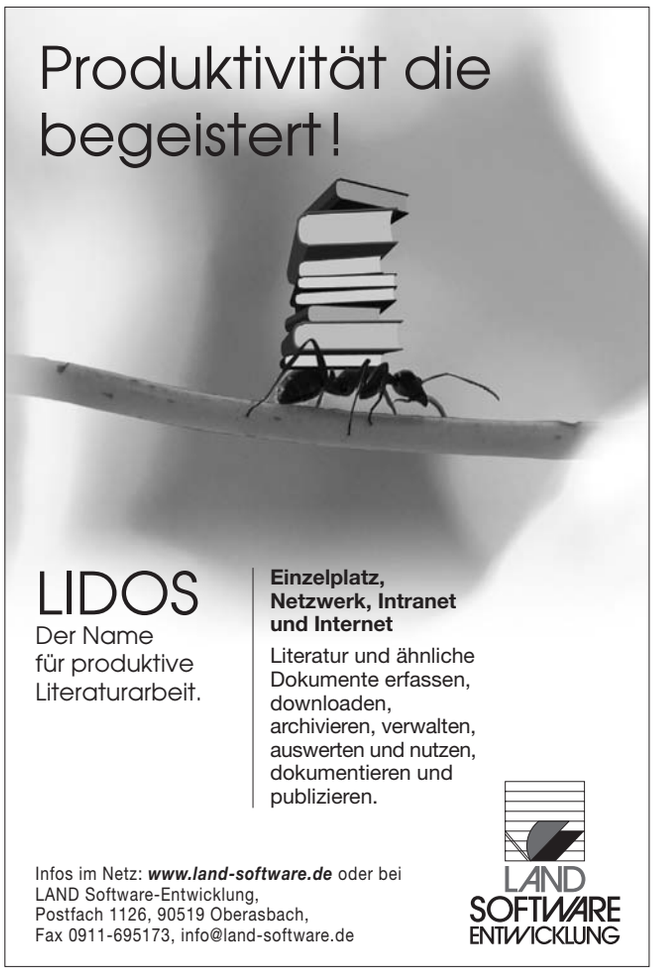
Most of the time, an internet relay chat (IRC) is used as command and control centre (C&C) for a botnet. An IRC consists of one or more servers which relay messages and/or commands to the connected clients. That way the botnet owner can centrally command the clients to download and execute a program which will commit click fraud on the owner's website(s) (or, respectively, the competitor's ads).

An alternative way of controlling the bots is to use a web interface to which the clients connect. However, this method requires the client to continually request updates from the C&C, whereas, when IRC is used, the server can just relay commands. Thus, using a web interface causes more traffic to and from the clients, increasing the chances of a bot being exposed and incriminating the botnet (Ianelli & Hackworth, 2005).

In general, a command and control centre will have to include a commonly-used protocol with little overhead and bandwidth usage, in order to remain inconspicuous.

4.3 Referrer Click Fraud

Anupam et al. introduce another, as-yet undetectable, method of committing click fraud: the deceitful publisher puts a script on his website that is automatically downloaded onto a visitor's computer when said visitor goes to the publisher's website. The script then imitates a click onto the advertisement leaving the visitor none the wiser. The log files of the advertiser (and, if applicable, of the commissioner) will thus show the visitor's client ID and IP address, which are unique for everyone. This simple setup, however, is still detectable. All it takes in order to uncover the fraudulent publisher is for the advertiser to visit the publisher's website and check afterwards whether his IP address and client ID have been registered as having performed a click. This can be remedied by a slight modification. The solution requires two versions of the publisher's website: one which includes the aforementioned script and one which does not. Internet users who visit the websites are always presented with the latter version and, consequently, if the advertiser or the commissioner checks the website, he cannot detect any sign of fraudulent activity. The publisher then needs a second website, which is wholly unconnected to the first one, over which he has complete control. This can either be a site he owns or he could work together with the owner for the scheme to work. He puts a script onto this second website which automatically loads the first, fraudulent, version of the first website in the background whenever the second website is accessed (Anupam, Mayer, Nissim, Pinkas, & Reiter, 1999).



5 Proposed Solutions

In effect, click fraud costs advertisers and – as a result of lawsuits – advertising network operators (Hadjinian, 2006) millions of dollars a year. It should not come as a surprise, then, that a lot of people have worked (and are still working) on the problem, coming up with several more or less effective solutions. A few of these solutions shall be reviewed in this chapter.

5.1 Cost-Per-Action (CPA)

In the Cost-Per-Action model, advertisers don't pay for clicks, but rather for specific actions that are performed on the advertiser's page *after* the click. These actions might, for example, be making a purchase, filling out a form, or registering.

"Such systems are used by Amazon, for example, to sell books on web pages: a service provider, say Expedia, can list an Amazon ad for a travel guide with the understanding that, should a user purchase the product advertised, then the service provider will receive a payment" (Immorlica et al., 2005).

Although CPA does prevent publisher and competitor click fraud, it leaves room for advertiser fraud. Since the publisher (and the commissioner, if he exists) have no way to confirm whether or not a specific action has taken place they depend on the advertiser to report the customer's action truthfully. Additionally, the publisher has to rely on the advertiser's ability to produce both efficient advertisements and – if the action in question is making a purchase – worthwhile products on the target side in order to make a profit. This means that the publisher does not profit directly from the advertising space and traffic he provides anymore. Furthermore, a user might click on an ad on the publisher's website without making a purchase on the target site, only to return to the target site and make the purchase later – thus robbing the publisher of his well-deserved commission. For the advertiser, on the other hand, the CPA model is very advantageous: since he only has to pay when he actually does make a sale, there is virtually no risk on his side.

Security expert Bruce Schneier says: "It's a hard model to make work – [the publisher] would become more of a partner in the final sale instead of an indifferent displayer of advertising – but it's the right security response to click fraud: change the rules of the game so that click fraud doesn't matter" (Schneier, 2006).

However, it is still a model that is based on trust – the only difference being that, this time, it is the other party that has to trust – and as such it is a model that publishers might be wary of adopting. If they do, they will likely favour big companies like Amazon and Ebay over small and possibly unreliable ones.

5.2 Pay-Per-Impression

An alternative to the PPC model of advertising is the old Pay-Per-Impression model. It ''remains popular on major Internet portals, such as yahoo.com, msn.com, and aol.com'' (Edelman et al., 2005). In this model the advertiser pays a specific amount of money for each time that the ad is displayed. This is the most appealing form of advertising for publishers, since they aren't reliant on the effectiveness of the advertiser's ads, but are paid for the space and traffic they provide. Unfortunately, though, Pay-Per-Impression is not fraud resistant either. The technical methods that make click fraud possible can be easily adapted to so-called impression fraud. Instead of simulating a click, though, the script repeatedly requests the website on which the ad is displayed and consequently artificially increases the number of impressions.

5.3 Pay-Per-Percentage of Impressions

Pay-Per-Percentage of Impressions is an alternative to Pay-Per-Click that was suggested by J. Goodman (2005). In his paper he describes this model thusly: "In this system, an advertiser picks a keyword, e.g. 'cameras' and purchases, perhaps through bidding, a certain percentage of all impressions for that keyword. For instance, an advertiser might pay $1.00 to MSN Search. In return, the advertiser might receive 10% of all impressions for "camera" for 1 week. What does this mean? It means that for 1 week, one out of ten times that someone searches for the word 'camera', they will see the ad." The costs of advertising are thus fixed and do not depend on whether or not the ad is clicked. They do not even depend on the number of impressions: "if there are R real impressions over the week, and F fake impressions, that the advertiser will receive $.1 \times R$ real impressions and $.1 \times F$ fake ones" (Goodman, 2005).

Consequently, this system is not susceptible to competitor fraud of any kind (neither click nor impression fraud). However, evaluating how much a certain percentage of impressions on a specific website is worth remains problematic. If such an evaluation is based on the average number of impressions of a site it remains vulnerable to impression fraud.

Goodman (2005) himself admits that "the pay-per-percentage model is not appropriate for all kinds of affiliate advertising; in particular, it is most appropriate for high volume sites". He recommends using a rating company to estimate the traffic on these sites, basing the cost-per-percentage on the estimates.

For the publisher this model is just as advantageous as Pay-Per-Impression, since his payment depends solely on the services he provides. Although the advertiser is safe from competitor fraud, this scheme leaves him susceptible to publisher Impression fraud if he does not advertise on a trustworthy site. Furthermore, the model is worse for him than Pay-Per-Click because the price is not based on the effectiveness of the ad.

5.4 Duplicate Detection

Metwally, Agrawal and Abbadi have proposed detecting fraudulent clicks through a method they call "Duplicate Detection in Click Streams". In order to differentiate between authentic and fraudulent clicks, the advertising commissioner "tracks individual customers by setting cookies. Duplicate clicks within a short period of time, a day for example, raise suspicion on the commissioner's side" (Metwally et al., 2005a).

Duplicate Detection can, no doubt, detect *amateur* click fraud, where the fraudster operates from one or a handful of computers. However, it is clearly inadequate when it comes to detecting distributed click fraud – where millions of computers simulate clicks on an advertisement from all over the world – since it relies on cookies (text files which the commissioner stores and accesses on the user's computer) for detection. In the case of distributed click fraud, every computer in the attacker's vast network will have its own individual cookie. Moreover, an attacker who knows what he is doing will just delete the cookies after each 'click', leaving no duplicates to be detected.

5.5 Association Rules

Metwally, Agrawal and Abbadi have also proposed a solution to the *referrer click fraud* discussed in section 3.3. They propose encouraging ISPs (Internet Service Providers) to provide the data stream necessary to detect this kind of click fraud. This data stream would contain the HTTP requests to page P, which might or might not be fraudulent. They would devise an algorithm to detect associations between one or more sites that refer to P very frequently, and clicks on an ad on P. If strong associations are found, it is very probable that P is using one or more 'decoy' websites in order to commit undetected click fraud (Metwally et al., 2005b).

6 Conclusion

We have seen that it is frighteningly easy to commit click fraud. You can simply set up two websites to implement *referrer click fraud* or, with sufficient funds, buy control of a botnet. You could even relatively effortlessly build a botnet with enough time, starting small and expanding in imitation of a website that grows more and more popular with time.
Neither botnet click fraud nor referrer click fraud can be satisfactorily detected. Of the alternatives presented, only Pay-Per-Percentage of Impressions and Cost-Per-Action are noteworthy, since the Pay-Per-Impression model is as vulnerable as Pay-Per-Click, as well as being less effective. However, both alternatives are far from perfect: Pay-Per-Percentage of Impressions is only a solution to competitor click fraud, since a dishonest publisher can still artificially increase the number of impressions of the ad and, consequently, increase both his income and the costs for each advertiser, since the overall price of advertising on the page in question will increase relative to the number of impressions. Furthermore, if the click-through-rate of the ad is taken into account, he can simply increase that too by commit-

ting click fraud. Hence, we can see that the system is inherently vulnerable.
Cost-Per-Action, on the other hand, is the best alternative to Pay-Per-Click that has been put forward to date. The advantages of this model are quite obvious: not only is it much more effective for the advertiser, who only has to pay a sort of commission when he actually sells something, but it is also nearly impossible to fake *buying* something. (It is actually possible, but that is *credit card fraud* and is an entirely different problem that has little to do with advertising.) However, this model is not only less profitable for the (honest) publisher, but also susceptible to *advertiser* fraud: since the publisher has no way of knowing whether a click on his advertisement has actually led to a sale (or another significant action), it is quite easy for the advertiser to lie about the number of sales generated by an ad and shortchange the publisher. Nonetheless, since the advertisers are probably going to have the last word on the matter, it seems inescapable that Cost-Per-Action will at least play a big role in the future. Indeed, Google is already taking first steps into that direction (Parker, 2006).

References

2006 TV Ad Revenue Figures. (2007). http://tvb.org/rcentral/adrevenuetrack/revenue/2006/ad_figures_1.asp [07.07.2007].

Anupam, V., Mayer, A., Nissim, K., Pinkas, B., & Reiter, M. K. (1999). On the security of pay-per-click and other web advertising schemes. In WWW '99: Proceeding of the eighth international conference on world wide web (pp. 1091–1100). New York, NY, USA: Elsevier North-Holland, Inc.

Buffer overflow. (n.d.). http://searchsecurity.techtarget.com/sDefinition/0,290660,sid14gci549024,00.html [05.06.2007].

Edelman, B., Ostrovsky, M., & Schwarz, M. (2005). Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. http://www.seo.com.ph/pdfguide/RP1917.pdf [06.06.2007].

Egelhoff, T. (n.d.). How to plan your advertising budget strategy. http://www.smalltownmarketing.com/adbudget.html [31.05.2007].

Gandhi, M., Jakobsson, M., & Ratkiewicz, J. (2006). Badvertisements: Stealthy click-fraud with unwitting accessories. Journal of Digital Forensic Practice, 1.

Goodman, J. (2005, June). Pay-per-percentage of impressions: An advertising method that is highly robust to fraud. Presented at the ACM E-Commerce Workshop on Sponsored Search Auctions, June, 2005.

Google AdSense Online Standard Terms and Conditions. (n.d.). https://www.google.com/adsense/static/en US/Terms.html [05.06.2007].

Google Annual Report 2005. (2006). http://investor.google.com/ pdf/2005_Google_AnnualReport.pdf [07.07.2007].

Google Annual Report 2006. (2007). http://investor.google.com/ pdf/2006_Google_AnnualReport.pdf [07.07.2007].

Hadjinian, D. L. (2006). Clicking away the competition: The legal ramifications of click fraud for companies that offer pay per click advertising services. Shidler J. L. Com. & Tech. at http://www.lctjournal.washington.edu/Vol3/a005Hadjinian.html, 3.

IAB Internet Advertising Revenue Report. (2007). http://www.iab.net/resources/adrevenue/pdf/IAB_PwC_2006_Final.pdf [07.07.2007].

Internet Advertising Revenues Soar Again, Near $5 Billion in Q1 07. (2007). http://www.iab.net/news/pr_2007_06_06.asp [07.07.2007].

Ianelli, N., & Hackworth, A. (2005). Botnets as a vehicle for online crime. Retrieved May 31, 2007, from www.cert.org/archive/pdf/Botnets.pdf.

Immorlica, N., Jain, K., Mahdian, M., & Talwar, K. (2005). Click fraud resistant methods for learning click-through rates. Lecture Notes in Computer Science, 3828.

Kaufmann, N. C. (n.d.). Click-spamming – ein Fall fuer das reformierte UWG? http://rsw.beck.de/rsw/shop/default.asp?sessionid=80CB3C7120844459E1FFCDA11953749&docid=138085&docClass=NEWS&from=mmr.root [10.05.2007].

Keizer, G. (2005). Dutch botnet suspects ran 1.5 million machines. http://www.techweb.com/wire/security/172303160 [10.05.2007].

Metwally, A., Agrawal, D., & Abbadi, A.E(2005a). Duplicate detection in click streams. In Proceedings of the 14th WWW International World Wide Web Conference, 12-21.

Metwally, A., Agrawal, D., & Abbadi, A E. (2005b). Using association rules for fraud detection in web advertising networks. In Vldb '05: Proceedings of the 31st international conference on very large data bases (pp. 169–180). VLDB Endowment.

*Metwally, A., Agrawal, D., &Abbadi, A.E.*2007). Detectives: detecting coalition hit inflation attacks in advertising networks streams. In Proceedings of the 16th international Conference on World Wide Web, 241-250.

Mills, E. (2006a). Study: Click fraud could threaten pay-per-click model. http://www.news.com/Study-Click-fraud-could-threaten-pay-per-click-model/2100-1024_3-6090939.html [05.10.2007].

Mills, E. (2006b). Click fraud increasing, study finds. http://www.news.com/Click-fraud-increasing,-study-finds/2100-1030_3-6095074.html [05.10.2007].

Mind the Gap. (2007). http://recoveringjournalist.typepad.com/recovering_journalist/2007/03/mind_the_gap.html [07.07.2007].

Nmap security scanner. (n.d.). http://insecure.org/nmap/ [05.06.2007].

Override, C. (2001). Exploits. http://www.gcf.de/papers/exploits.htm [01.06.2007].

Parker, P. (2006). Google tests cost-per-action. http://www.clickz.com/showPage.html?page=3615476 [25.06.2007].

Richardson, M., Dominowska, E., and Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international Conference on World Wide Web, 521-530.

Schneier, B. (2006). Google's click-fraud crackdown. http://www.wired.com/politics/security/commentary /securitymatters/2006/07/71370 [25.06.2007].

Vidyasagar, N. (2004, May 3). India's secret army of online ad 'clickers'. The Times of India.

Betrug, Werbung, Suchmaschine, Google, Benutzung, Fehler

Click fraud, Web advertising, Pay per Click, Google AdWords, Google AdSense, Competitor click fraud, Publisher click fraud, Botnet, Referrer click fraud, Pay per action, Advertiser fraud

The Author

Simone Soubusta, B.A.

has studied Information Science and Language Technology at Heinrich-Heine-University Düsseldorf. This article is a revised and shortened version of her bachelor's thesis.

simone@soubusta.de

24. Oberhofer Kolloquium vom 10. bis 12. April in Magdeburg

Im Jubiläumsjahr 2008 der DGI steht das traditionsreiche Oberhofer Kolloquium unter dem Rahmenthema „Informationskompetenz 2.0. Zukunft von qualifizierter Informationsvermittlung."

Die Veranstalter, DGI und VDI, freuen sich, dass der bekannte Informatiker und Medienkritiker Joseph Weizenbaum sein Kommen zugesagt hat. Auch andere Persönlichkeiten aus Informationswissenschaft und -praxis haben angekündigt, sich der Diskussion mit den Newcomern und Quereinsteigern im Bereich Informationsvermittlung stellen zu wollen. Das Tagungshotel bietet einen idealen Rahmen für das traditionelle Motto der Tagung: „Tagen und Wohnen unter einem Dach". Die Kommunikation unter den Teilnehmern, die informell auch in die Abendstunden fortgesetzt wird, gehört zu den wesentlichen Elementen des Kolloquiums. Bei den Vorträgen stehen berufsbezogene praktische Erfahrungen im Vordergrund. Alles über die Tagung und das detaillierte Programm finden Sie unter www.dgi-info.de/oberhofer.aspx.

Die Anzahl der Zimmer ist beschränkt, nehmen Sie Ihre Anmeldung möglichst rasch vor.

Die Digitale Kluft: eine Analyse der Informationsinfrastruktur in Bosnien-Herzegowina

Violeta Trkulja, Düsseldorf

Eines der meistdiskutierten Themen bezüglich der Informationsgesellschaft ist das der Digitalen Kluft. Die Informations- und Kommunikationstechnologien (IKT) sind Grundbausteine einer jeden Informationsgesellschaft und bieten große Chancen für die soziale und ökonomische Entwicklung eines Landes. Im Vergleich zu den Teilstaaten des ehemaligen Jugoslawiens wird der aktuelle Stand der Informationsinfrastruktur in Bosnien-Herzegowina aufgezeigt. Auch wenn die IKT in den letzten Jahren eine dynamische Entwicklung erfahren haben, leidet das Land noch immer an den Folgen des Bürgerkrieges. Die komplizierten politischen Verhältnisse verzögern den Ausbau hin zu einer Informationsgesellschaft.

The Digital Divide: an Analysis of the Information Infrastructure in Bosnia and Herzegovina

One of the most frequently discussed issues regarding the information society is the so called Digital Divide. Information and communication technologies (ICT) build the foundation of any information society and offer great opportunities for the social and economic development of a country. The current level of information infrastructure within Bosnia and Herzegovina is depicted and compared to the other states of the former Republic of Yugoslavia. Although ICT has developed dynamically, the country still suffers from the aftermath of civil war. The complicated political circumstances retard the development and expansion of an information society.

Digitalna podjela društva: analiza informacione infrastrukture u Bosni i Hercegovini

Jedna od najčešće pominjanih tema u vezi sa informacijskim društvom je takozvana digitalna podjela (Digital Divide) društva na one koji imaju direktan pristup digitalnoj tehnologiji i na one koji taj pristup nemaju. Informacijska i komunikacijska tehnologija (ICT) predstavlja osnovu svakog informacijskog društva i nudi velike šanse za socijalni i ekonomski razvoj zemlje. Nivo razvijenosti tehnologije informacijske infrastrukture u Bosni i Hercegovini se uporedjuje sa situacijom u drugim zemljama bivše Jugoslavije. Iako je razvoj ICT-a posljednjih godina ubrzan, vidljivo je da zemlja još pati od posljedica nedavnog gradjanskog rata. Komplikovana politička situacija dodatno koči razvoj i širenje informacijskog društva.

Informationsgesellschaft

Als Informationsgesellschaft bezeichnet man Wirtschafts- und Gesellschaftsformen, in denen Informations- und Kommunikationstechnologien (IKT) eine herausragende Bedeutung einnehmen. Die Arbeits- und Lebensbedingungen der Menschen erfahren durch die von diesen Technologien gewährleistete neue Art der Gewinnung, Speicherung, Verarbeitung, Vermittlung und Nutzung von Information und Wissen eine tiefgreifende Veränderung. Der Stellenwert der Informations- und Kommunikationstechnologien steigt stetig, und langfristig gesehen bieten sie große Chancen für den Fortschritt in verschiedenen gesellschaftlichen Bereichen (Haushalten, Unternehmen, Bildung, Verwaltung etc.). Schon heute sind Mobilfunk, Computer und Internet aus unserer Gesellschaft nicht mehr wegzudenken.

Kennzeichen der Informationsgesellschaft ist eine Gesellschaft,

- „deren Basisinnovationen von der Ressource Information getragen werden [...],
- in der Informationsinhalte aller Arten überall und jederzeit zur Gänze zur Verfügung stehen [...] und auch intensiv genutzt werden [...],
- deren Mitglieder bevorzugt Telematikgeräte zur Information und Kommunikation benutzen [...]“ (Stock, 2000, 7).

Diese telematischen Geräte und Verfahren (Computer, Netzwerke, Software, Mobiltelefone usw.), die den Zugang zur sowie die Weiterverarbeitung der Ressource Information ermöglichen, sind die Grundbausteine moderner Informationsgesellschaften. Der Ausbau dieser Informationsinfrastrukturen schreitet mit hohem Tempo voran. Seit der Überwindung der Krise im Jahr 2000 befindet sich die Branche in den OECD-Ländern im Umbruch (OECD, 2007, 13 f.). Nach wie vor ist die Sprachtelefonie einer der Hauptmotoren für die Telekommunikationsmärkte, deren Einnahmen mittlerweile 1 Bill. USD erreicht haben. Technologien wie Voice over Internet Protocol (VoIP) drücken auf die Preise für Sprachdienste, was deutlich in den Kosten für Festnetzgespräche sowie in Paketangeboten zu Pauschalpreisen zum Ausdruck kommt, die von den Telekommunikationsunternehmen angeboten werden. Auf die mobilen Dienste entfallen mittlerweile rund 40 Prozent der gesamten Telekommunikationseinnahmen, und die Anzahl der Mobilfunkkunden übersteigt die der Festnetzkunden in einem Verhältnis von 3 zu 1. Hauptgrund für die schnelle Verbreitung der Internettelefonie ist die steigende Zahl von Hochgeschwindigkeits-Internetverbindungen. Die Breitbandtechnologie ist die vorherrschende Technologie für den Zugang zum Internet, und im OECD-Raum verfügen bereits 60 Prozent der 256 Millionen Internetkunden über einen Breitbandanschluss (OECD, 2007, 13 f.).

Die USA bilden nach wie vor den größten IKT-Einzelmarkt, gefolgt von Japan und Deutschland. Verantwortlich für das weltweite Wachstum sind allerdings die so genannten BRIC-Länder Brasilien, Russland, Indien und China mit Wachstumsraten über 20 Prozent (Graumann & Wolf, 2007, 17 ff.). Den größten Anteil am IKT-Markt (54,4 Prozent) nimmt auch im internationalen Vergleich die Telekommunikation ein, wobei in technologisch

59(2008)2, 142-148 Information Wissenschaft & Praxis

fortgeschrittenen Ländern ein Rückgang von Festnetzanschlüssen zu verzeichnen ist, was mit einem Anstieg der Mobilfunkanschlüsse und einer zunehmenden Nutzung von Breitbandtechnologien und den damit alternativen Telekommunikationsmöglichkeiten wie VoIP einhergeht (Graumann & Wolf, 2007, 115 ff.).

Neben der Computer-, Internet- und Telekommunikationsbranche hängen heute die meisten Industrieproduktionen vom Einsatz moderner IuK-Technik und elektronischer Systeme ab. Sie sind Grundlage der wirtschaftlichen Leistungsfähigkeit von Industrienationen, und ihre Schlüsseltechnologien stoßen Innovationen an, erzeugen Wirtschaftswachstum und schaffen dadurch Arbeitsplätze (BMBF). Die IKT hat sich damit zu einer treibenden Kraft für die Weltwirtschaft insgesamt entwickelt:

> *„The digital and information revolution has changed the way the world learns, communicates, does business, and treats illness. New information and communication technologies offer vast opportunities for progress in all walks of life in all countries – opportunities for economic growth, improved health, better service delivery, learning through distance education, and social and cultural advances"* (WDI, 2007, 307).

Informationsungleichheit

Eines der meistdiskutierten Themen bezüglich der Informationsgesellschaft ist das des *information gap*, der Kluft zwischen den so genannten „information haves" (information rich) und „information have-nots" (information poor). Darunter versteht man das unausgewogene Verhältnis zwischen Bevölkerungsschichten mit geringerem und höherem sozialen Status im Hinblick auf den Besitz und die Nutzung von Informationen und Kommunikationsmitteln wie Telefon, Fernsehen oder das Internet (Kostov, 2006, 4). Zurückzuführen ist diese Theorie auf die in den 1970er Jahren an der Minnesota University entwickelte Wissensklufthypothese von Philipp J. Tichenor, George A. Donohue und Clarice N. Olien:

> „As the infusion of mass media information into a social system increases, segments of the population with higher socioeconomic status tends to acquire this information at a faster rate than the lower status segments, so that the gap in knowledge between these segments tends to increase rather than decrease" (Tichenor, Donohue, & Olien, 1970, 159).

Mitte der 1990er Jahre wurde die Diskussion um den *Information Gap* nahezu vollständig von der des *Digital Divide* verdrängt. In der deutschsprachigen Literatur wird der gebräuchliche englische Terminus technicus mit „Digitaler Spaltung", „Digitaler Graben", „Digitales Gefälle", „Digitaler Kluft" oder „Digitaler Trennung" übersetzt. Ich verwende im folgenden die Bezeichnung „Digitale Kluft."

Unter *Digitaler Kluft* verstand man zunächst schlicht die Tatsache, dass die Welt unterteilt werden konnte in Menschen die Zugang und Menschen die keinen Zugang zu modernen Informations- und Kommunikationstechnologien besitzen. Daneben konzentriert sich die wissenschaftliche Diskussion verstärkt auf die sogenannte *Informationskompetenz* (Information Literacy). Die Befähigung, Informationen zu finden, abzurufen, diese zu analysieren, zu verarbeiten und für spezifische Aufgaben nutzbar zu machen, ist von enormer Bedeutung. Das internetgestützte Lernen verändert nicht nur die Art und Weise der Ausbildung, die erforderlich ist, um im Internet zu arbeiten, sondern auch die Lernfähigkeit in einer auf dem Internet basierenden Wirtschaft und Gesellschaft (Castells, 2005, 272). In diesem Zusammenhang geht es bei der Diskussion um die Digitale Kluft verstärkt auch um die sogenannte *Inhaltliche Kluft* (Content Divide), denn „eine Vielzahl der webbasierten Informationen entsprechen den wirklichen Bedürfnissen der Menschen nicht. Nahezu 70 Prozent der weltweiten Internetseiten sind in englischer Sprache verfasst und übertönen so von Zeit zu Zeit lokale Stimmen und Sichtweisen" (Annan, 2003).

Digitale Kluft

Der bloße Zugang zum Internet kann das Problem der Digitalen Kluft nicht lösen. Um letztlich eine Verbesserung des Lebensstandards zu gewährleisten, müssen IKT zum Zweck persönlicher, sozialer und ökonomischer Entwicklung eingesetzt werden. Die IKT und allen voran das Internet spielen auf vielen Gebieten des privaten, gesellschaftlichen, wirtschaftlichen und politischen Lebens eine zentrale Rolle. Personen, die keinen Zugang zu diesen Technologien oder nicht die Befähigung haben, diese adäquat nutzen zu können, werden marginalisiert (Castells, 2005, 261).

In ähnlicher Weise wie der Knowledge Gap existiert die Digitale Kluft zwischen gebildeten und ungebildeten, einkommensstarken und einkommensschwachen Bevölkerungsschichten sowie weltweit zwischen wirtschaftlich hoch- und unterentwickelten Staaten. Manuel Castells (2005, 262 ff.) hat die Dimensionen der Digitalen Kluft aufgezeigt und analysiert:

- Income Divide: Wohlhabende sind häufiger online als Ärmere.
- Geographic Divide: Stadtbewohner sind häufiger online als Landbewohner.
- Ethnic Divide: Weiße sind häufiger online als Angehörige anderer Ethnien.
- Education Divide: Hochqualifizierte sind häufiger online als niedrig Qualifizierte.
- Gender Divide: Männer sind häufiger online als Frauen.
- Age Divide: Junge Menschen sind häufiger online als alte.
- Technological Divide: technologisches Gefälle zwischen ISDN-, Breitband-Zugängen einerseits und analogen Zugängen andererseits.
- Global Divide: In Entwicklungs- und Schwellenländern sind dramatisch weniger Menschen online als in Industrienationen.

Neben der ausreichenden Infrastruktur, die notwendig für einen Zugang zum Internet ist, sind zwei weitere Aspekte von enormer Bedeutung (Servon, 2002, 7): Zum einen die „IT literacy" worunter man die Fähigkeit versteht „to use IT for a range of purposes, and the knowledge of how and why IT can be used as a key resource." Zum anderen die Betrachtung des digitalen Content, welchen die Nutzer in Entwicklungs- und Schwellenländern vorfinden: „When disadvantaged groups do log on, they often find that there is no content there. The kind of information they seek – information that is directly related to their lives and communities and cultures – does not exist."

Dieser Typologie kann zusätzlich der Aspekt der *Demokratischen Kluft* hinzugefügt werden: „[…] within the online community, the democratic divide signifies the difference between those who do, and do not, use the panoply of digital resources to engage, mobilize, and participate in public life" (Norris, 2001, 4 f.). Er berücksichtigt nicht nur die nationale und internationale Ebene, sondern auch die Machtverhältnisse innerhalb der Internet-Gemeinschaft.

Auf dem ersten Weltgipfel zur Informationsgesellschaft (WSIS) 2003 in Genf nahmen sich die Vereinten Nationen des Themas im größeren Stil an. An diesem Gipfel nahmen erstmalig auch Vertreter von Nichtregierungsorganisationen sowie von Wirtschaftsverbänden teil. Neben den industriellen Interessen und den Forderungen der NGO stand für den damaligen UN-Generalsekretär Kofi Annan die Überwindung der Digitalen Kluft zwischen Industrie- und Entwicklungsländern im Vordergrund:

"The new information and communication technologies are among the driving forces of globalization. They are bringing people together, and bringing decision-makers unprecedented new tools for development. At the same time, however, the gap between information "haves" and "have-nots" is widening, and there is a real danger that the world's poor will be excluded from the emerging knowledge-based global economy" (Annan, 2002).

Neben der Grundaussage der WSIS-Deklaration (WSIS, 2003), dass es das Recht eines jeden Menschen sei, Wissen zu schaffen, zu benutzen und auszutauschen, stehen vor allem die Armutsbekämpfung und gesellschaftliche Entwicklung mithilfe eines verbesserten Zugangs zu IuK-Technologien im Vordergrund. Unter den 175 Teilnehmerstaaten des WSIS war auch Bosnien-Herzegowina. Das Land ist wie der Kongo ein Nachkriegsland, dass seine Infrastruktur erst aufbauen muss, während Schwellenländer wie Indien eine sich schnell entwickelnde IKT-Infrastruktur verzeichnen können (Finlay, 2007, 99). Auch wenn die Bedeutung von IKT als einschneidende und strategische Angelegenheit für die soziale und ökonomische Entwicklung erkannt wurde, ist es Bosnien-Herzegowina bis heute nicht gelungen, die aus dem Gipfel hervorgegangen Projekte erfolgreich umzusetzen (Pellizzer, 2007, 115). Dies hängt nicht nur mit dem Bürgerkrieg und der Nachkriegszeit zusammen, die das Land weit hinter die wirtschaftliche und gesellschaftliche Entwicklung der anderen Balkanstaaten zurückgeworfen haben. Insbesondere durch das komplizierte politische System des Landes wird der Aufbau einer Informationsgesellschaft in Bosnien-Herzegowina beträchtlich verzögert (Pellizzer, 2007, 115).

Bevor auf den Entwicklungsstand der IKT-Infrastruktur eingegangen wird, soll zunächst ein Einblick in die Folgen des Krieges und die aktuelle politische und wirtschaftliche Situation des Landes gegeben werden.

Die Hinterlassenschaften des Jugoslawienkrieges in Bosnien-Herzegowina

Bosnien-Herzegowina, einer der Teilstaaten des ehemaligen Jugoslawien, ist mit einer Gesamtfläche von 51.197 km² und knapp vier Millionen Einwohnern ein vergleichsweise kleines Land. Die Bevölkerung des Landes setzt sich aus Angehörigen verschiedener Nationalitäten und Religionen zusammen, die bis zum Krieg in heterogen zusammengesetzten Siedlungsgebieten wohnten. Der Bürgerkrieg von 1992 bis 1995 zerstörte nicht nur die physische Infrastruktur, sondern vor allem auch die einzigartige multiethnische Gesellschaftsstruktur des Landes (Calic, 1996). Während die Bevölkerungsverteilung einstmals durch die naturräumlichen Bedingungen nachhaltig beeinflusst war, veränderte sie sich durch den Krieg vollständig. Die Fluchtbewegungen und Vertreibungen (allein bis Mitte Juli 1992 waren zwei Millionen Menschen und bis November 1994 mehr als 2,7 Millionen auf der Flucht) führten zu neuen ethnischen Siedlungsmustern (vgl. die Aufteilung in Abb. 1). Als der Krieg Ende 1995 militärisch zu Ende ging, war die Infrastruktur des Landes zerstört und das Bruttoinlandsprodukt um drei Viertel reduziert.

Abb. 1: Bosnien-Herzegowina (Quelle: www.bosnien-herzegowina.info).

Die Bevölkerung setzt sich zu 48 Prozent aus Bosniaken, 37 Prozent Serben und 14 Prozent Kroaten zusammen, wobei 40 Prozent der Bosniaken muslimischen, 31 Prozent der Serben orthodoxen und 14 Prozent der Kroaten römisch-katholischen Glaubens sind (CIA Factbook). Im alten Jugoslawien waren alle Einwohner des Landes *bosanci*. Man verwendete die Bezeichnung *Muslim* (großgeschrieben) zur Kennzeichnung der Nationalität sowie *muslim* (kleingeschrieben) zur Konfessionsbezeichnung. Um Verwirrungen mit der Religionszugehörigkeit zu vermeiden, wurde die Bezeichnung Muslim durch Bosniake ersetzt. Die Landessprachen sind bosniakisch, serbisch und kroatisch. Bis auf die Nutzung des kyrillischen Alphabets im Serbischen unterscheiden sie sich kaum.

Mit Unterzeichnung des Dayton-Abkommens am 14. Dezember 1995 wurde der Staat Bosnien-Herzegowina geschaffen, der sich aus zwei Entitäten, der Republika Srpska (RS), der Föderation Bosnien-Herzegowina (FBiH) sowie dem Distrikt Brčko zusammensetzt. Neben der Verfassung des Gesamtstaates (Schwarz, 2003) verfügen die zwei Teilstaaten über eigene Verfassungen mit umfassenden Gesetzgebungskompetenzen (Schwarz, 2004a, Schwarz, 2004b).

Problematisch ist dabei, dass zentrale Staatsorgane keine Macht haben, während die regionalen über etwas mehr und kantonale Organe fast über die gesamte Macht verfügen. Das bedeutet, dass sich die Staatsregierung nicht in die Arbeit der Entitäten einmischen darf und im Grunde für fast nichts im Lande zuständig ist. Es wäre für Bosnien-Herzegowina entscheidend gewesen, die Macht der Nationalisten auf allen Seiten einzuschränken (Oschlies, 2006, 740 f.). Weil dies nicht gelang und bis heute immer noch nicht vollständig gelingt, ist das Mandat des Hohen Repräsentanten in Bosnien-Herzegowina von zentraler Bedeutung. Durch seine umfassenden Kompetenzen, die die der Entitäten und

des Gesamtstaates überlagern, ist er in der Lage, das komplizierte Gebilde zusammenzuhalten (Schwarz, 2003, 132 f.). So gelang es den Hohen Repräsentanten bisher, zahlreiche politische Blockaden zu überwinden und integrative Elemente in der Staatenorganisation zu verankern. Der derzeitige Hohe Repräsentant Miroslav Lajčák, der zugleich auch Sonderbeauftragter der EU ist, schränkte zuletzt die Vetomöglichkeiten der Volksgruppen im Parlament und in der Regierung ein, um das Land und die Institutionen funktionsfähiger zu machen.

Wirtschaftslage und Entwicklungsstand

Neben dem Bürgerkrieg, der immer noch einen langen Schatten auf die Wirtschaft Bosnien-Herzegowinas wirft, hat das durch das Daytoner Abkommen geschaffene komplizierte politische Gebilde dazu geführt, dass das Land neben Mazedonien zu den ärmsten Nachfolgestaaten des ehemaligen Jugoslawiens zählt. Der volkswirtschaftliche Vergleich verdeutlicht den schlechten Entwicklungsstand des Landes (vgl. Tab. 1). Neben einem niedrigen Bruttoinlandsprodukt von 11,5 Mrd. USD und einem Pro-Kopf-Einkommen von 2.945 USD weist das Land die geringste Kaufkraftparität[1] von 6.621 PPP-Dollar sowie die zweithöchste Inflationsrate auf.

Tabelle 1: Volkswirtschaftlicher Vergleich der Länder Ex-Jugoslawiens 2006 (nach: EIU, 2007).

	Bosnien-Herzegowina	Serbien und Montenegro[a]	Kroatien	Slowenien	Mazedonien
BIP (Mrd. USD)	11,5	31,2	43,0	37,4	5,9
BIP pro Kopf (USD)	2.945	4.209	9.426	18.581	2,899
BIP je Einwohner (USD/PPP)	6.621	7.012	13.114	24.672	7.456
Inflationsrate (%)	7,4	12,7	3,2	2,5	3,2
Leistungsbilanz (Mrd. USD)	-1,3	-3,7	-3,2	-1,1	-0,1
Leistungsbilanz (in % des BIP)	-11,4	-11,7	-7,4	-3,0	-1,1
Exporte (Mrd. USD)	3,4	6,5	10,6	21,2	2,4
Importe (Mrd. USD)	-7,6	-12,7	-21,1	-22,7	-3,7
Auslandsschulden (Mrd. USD)	6,6	19,6	33,9	21,4	2,2

[a] ohne Kosovo

Die Arbeitslosenquote liegt bei 44 Prozent, und knapp 18 Prozent der Menschen leben unterhalb der Armutsgrenze. Deutlich wird der Entwicklungsstand Bosnien-Herzegowinas auch anhand des Ländervergleichs des Human Development Index der Vereinten Nationen, welcher das Land auf Platz 66 von 177 Ländern aufführt (mit einem HDI-Wert von 0,803). Dieser Ländervergleich berücksichtigt nicht nur das Bruttoinlandsprodukt, sondern auch die Lebenserwartung, den Bildungsgrad und den Lebensstandard unter Berücksichtigung der Kaufkraftparität sowie des Einkommens. Der HDI unterteilt die Länder in drei Kategorien: Länder mit hoher menschlicher Entwicklung ≥ 0,8, mit mittlerer menschlicher Entwicklung < 0,8 und ≥ 0,5 sowie Länder mit geringer menschlicher Entwicklung < 0,5. Zum Vergleich: Cuba steht auf Platz 51 im Ländervergleich, obwohl es ein geringeres Bruttoinlandsprodukt hat als Bosnien-Herzegowina.

Damit liegt Bosnien-Herzegowina knapp über der Grenze der Länder mit mittlerer menschlicher Entwicklung. Eines der wichtigsten Ziele des UN-Entwicklungsprogramms ist daher die Reduzierung der Armut in Bosnien-Herzegowina (UNDP, 2007a; UNDP, 2007b).

Schattenwirtschaft

Ein großes Problem in den Ländern des ehemaligen Jugoslawiens ist die Schattenwirtschaft, die rund 30 Prozent (Schneider, 2004) des Bruttoinlandproduktes ausmacht (vgl. Tab. 2). Wie alle Transitionsländer durchlaufen auch diejenigen des ehemaligen Jugoslawiens einen Transformationsprozess: „a growing shadow economy is one of the 'collateral effects' of a shift from a command economy to a market economy. The size of the shadow economy affects the impact of any economic policy and should be seen in the context of socioeconomic relations and shadow activities" (Dell'Anno & Piirisil, 2007, 626).

Bosnien-Herzegowina leidet am stärksten unter der Schattenwirtschaft, was auf die schlechtere Lage des Arbeitsmarktes, die höhere Arbeitslosigkeit und die verheerenden Konsequenzen des Krieges zurückzuführen ist (Nastav & Bojnec, 2007, 29). Der in der Nachkriegszeit zunächst fehlende Wiederaufbau schuf die Basis für das Aufkommen der Schattenwirtschaft. Darüber hinaus stieg der Anteil der in der Schattenwirtschaft Beschäftigten an, begünstigt durch die bis zum Jahre 2000 in Kraft gewesenen, äußerst restriktiven Gesetze hinsichtlich des Beschäftigungsschutzes. Die Mehrheit der gemeldeten Arbeitslosen ist nicht wirklich arbeitslos, sondern arbeitet in der Schattenwirtschaft (Nastav & Bojnec, 2007, 35).

Tabelle 2: Ausmaß der Schattenwirtschaft in den Ländern Ex-Jugoslawiens (in % des BIP) (Quelle: Schneider, 2004).

	1999 –2000	2001 – 2002	2002 – 2003
Bosnien-Herzegowina	34,1	35,4	36,7
Kroatien	33,4	34,2	35,4
Mazedonien	34,1	35,1	36,3
Serbien und Montenegro	36,4	37,3	39,1
Slowenien	27,1	28,3	29,4

Es gibt keine offiziellen Daten über Software-, Musik- und Filmpiraterie in Bosnien-Herzegowina. Obwohl die gesetzlichen Grundlagen für ein Verbot bestehen, ist es üblich, und das gilt nicht nur für Bosnien-Herzegowina, Filme, Musik und Software auf der Strasse bzw. auf dem Markt zu kaufen (vgl. Abb. 2). Die CDs bzw. DVDs mit Unterhaltungsinhalt kosten im Schnitt zwischen zwei und fünf Euro, Software ist meist etwas teurer.

Abb. 2: Verkauf von Video-DVDs in Banja Luka.

Positiv ist jedoch die wirtschaftliche Entwicklung der letzten Jahre zu bewerten. Trotz schleppender politischer Reformen ist die Wirtschaft in 2006 um 6,2 Prozent gewachsen. Dies ist auf eine robuste Industriekonjunktur, auf den anziehenden Außenhandel sowie eine rege Investitionstätigkeit zurückzuführen. Steigende Löhne und der nach wie vor anhaltende

1 Die Kaufkraftparität (PPP - Purchasing Power Parity) ist ein Umrechnungsfaktor, der es ermöglicht, die in nationalen Währungen dargestellten Bruttoinlandsprodukte ohne Verfälschung durch zum Teil stark schwankende Wechselkurse zu konvertieren und zu vergleichen. Dadurch können Vergleiche der realen Wirtschaftsleistung von Volkswirtschaften (Volumenvergleiche) durchgeführt werden, die z. B. für Armuts- und Reichtumsvergleiche herangezogen werden.

Nachholbedarf sorgen zudem für eine gute Entwicklung des privaten Verbrauchs (BFAI, 2007).

IKT-Infrastruktur in Bosnien-Herzegowina

Wie bereits erwähnt, löst der Zugang zum Internet allein das Problem der Digitalen Kluft nicht. Er stellt aber eine notwendige Voraussetzung dar, „um in einer Gesellschaft, in der die beherrschenden Funktionen und gesellschaftlichen Gruppierungen zunehmend um das Internet herum organisiert sind, Ungleichheit zu überwinden (Castells, 2005, 262). In Anbetracht dessen soll die Verbreitung des Internets in Bosnien-Herzegowina in Bezug auf die technologische Kluft (Technological Divide) (vgl. die Dimensionen der Digitalen Kluft bei Castells, 2005, 262), d. h. im Vergleich zu Europa[2] und den Ländern des ehemaligen Jugoslawiens[3], analysiert werden. Hierbei fällt Slowenien nicht in den Vergleich der Länder des ehemaligen Jugoslawiens, da das Land aufgrund seiner guten Wirtschaftslage mit den Industriestaaten Westeuropas vergleichbar ist.

Um den Stand der Verbreitung des Internets zu ermitteln, werden Nutzerzahlen sowie die Penetrationsrate von Internetanschlüssen herangezogen. Unter letzterem versteht man das Verhältnis zwischen der Anzahl der Einwohner und den jeweils installierten Anschlüssen (Telefon, Breitband, Kabelmodem).

Die Daten der International Telecommunication Union (ITU) basieren auf staatlich veröffentlichten Daten, und aufgrund dessen gibt es für die ITU-Statistiken keine genaue Definition von „Internetnutzer." Ein Internetnutzer kann je nach Definition jemand sein, der das Internet gelegentlich (z. B. vierteljährlich), wöchentlich oder sogar täglich nutzt. Bei den gezählten „Internetanschlüssen" ergibt sich das Problem, dass auf einen Anschluss, welcher einem Vertrag entspricht (Internet subscriber), mehrere Nutzer fallen können. Es muss deshalb in diesem Zusammenhang darauf hingewiesen werden, dass andere Studien nicht mit den Ergebnissen aus den ITU-Statistiken übereinstimmen, wobei es sich um ein grundsätzliches Problem handelt, da solche Zahlen nie exakt sind und immer nur einen Trend erfassen können.

Die Zahl der Internetnutzer in Bosnien-Herzegowina ist innerhalb von fünf Jahren auf knapp 25 Prozent gestiegen. Im europäischen Vergleich und besonders in Relation zu Deutschland fällt sie sehr niedrig aus. Anzunehmen ist, dass die hohen Zuwachsraten in unmittelbarem Zusammenhang mit der ökonomischen Entwicklung stehen, was im Vergleich zu Kroatien deutlich wird. Das Land weist die höchste Anzahl von Internetnutzern sowie das höchste Bruttoinlandsprodukt auf (Kostov, 2006, 8). Die Balkan-Länder zeichnen sich jedoch durch deutlich höhere Zuwachsraten als der europäische Durchschnitt aus (vgl. Abb. 3).
Aufgrund seiner einfachen Verfügbarkeit und der niedrigen Modemkosten ist der Festnetzanschluss nach wie vor die am weitesten verbreitete Verbindungsart zwischen Computern und Telefon. Der Nutzungsgrad der existierenden Festnetzanschlüsse ist deshalb ein wichtiger Anhaltspunkt für die Einschätzung der Internetnutzung. Die Telefondichte eines Landes ist auch ein wichtiger Faktor zur Einschätzung der gesamtwirtschaftlichen Situation sowie zum Stand der technischen Entwicklung (Graumann & Wolf, 2007, 115).

Auch in diesem Vergleich liegt Kroatien mit einer 40-prozentigen Penetrationsrate an der Spitze, gleichauf mit dem europäischen Durchschnitt (vgl. Abb. 4). Bosnien-Herzegowina hat die geringste Anzahl an Festnetzanschlüssen. Mit gerade einmal 25 Prozent liegt dieser Anteil weit unter Kroatien, Serbien und Montenegro sowie Mazedonien. Aber vor allem im Vergleich zu Deutschland, wo 65 Prozent der Einwohner über einen Festnetzanschluss verfügen, wird das Ausmaß der

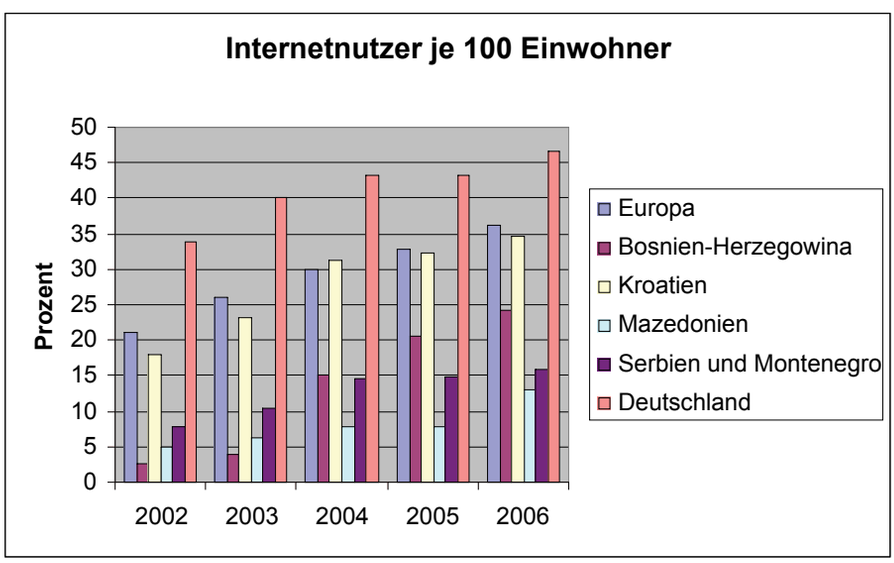


Abbildung 3: Internetnutzer je 100 Einwohner (Quelle: ITU, 2007; ITU, 2004; WDI 2006).

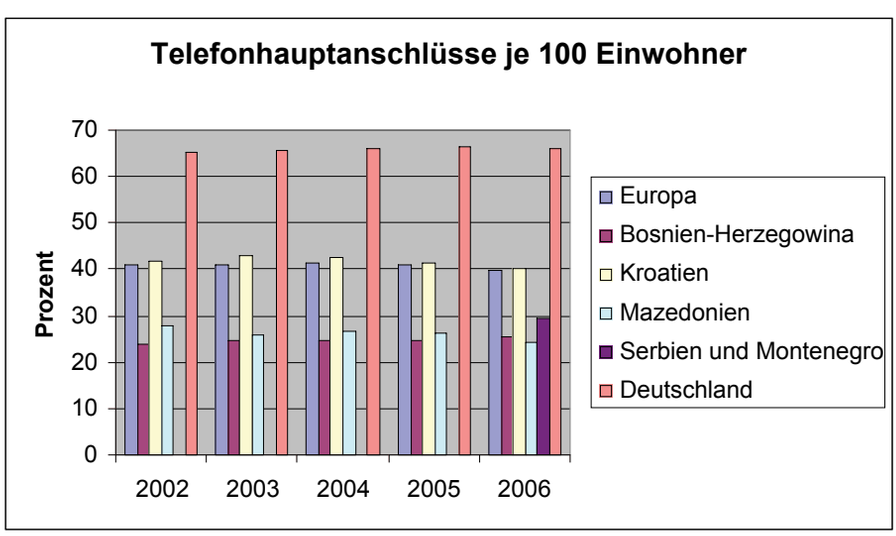


Abbildung 4: Telefonhauptanschlüsse je 100 Einwohner (Quelle: ITU 2007).

2 In der Statistik der ITU werden folgende europäische Länder aufgeführt: Albanien, Andorra, Belarus, Belgien, Dänemark, Deutschland, Estland, Färöer Inseln, Finnland, Frankreich, Griechenland, Großbritannien, Island, Irland, Italien, Kroatien, Lettland, Liechtenstein, Litauen, Luxemburg, Malta, Mazedonien, Moldawien, Monaco, Montenegro, Niederlande, Norwegen, Österreich, Polen, Portugal, Rumänien, Russland, San Marino, Schweden, Schweiz, Serbien, Slowakei, Slowenien, Spanien, Tschechische Republik, Türkei, Ukraine, Ungarn, Vatikan, Zypern.

3 Aufgrund der Trennung von Serbien und Montenegro sind die Daten für die beiden Länder in den aktuellen Statistiken der ITU zum Teil nicht aufgeführt.

Kluft ersichtlich. Der Grund für die geringe Zahl von Festnetzanschlüssen liegt in der Verstaatlichung des Telekommunikationsmarktes. Es gibt drei Anbieter in Bosnien-Herzegowina (Vockic-Avdagic, 2003). JP PTT BiH mit Sitz in Sarajewo versorgt das Territorium der Föderation Bosnien-Herzegowina und das District Brčko, während Telekom Srpske mit Sitz in Banja Luka das Territorium der Republika Srpska und das Distrikt Brčko übernimmt. HPT Mostar Ltd. mit Sitz in Mostar versorgt das gleiche Gebiet wie JP PTT BiH. Die drei Anbieter stehen untereinander nicht in Wettbewerb.

Eine Teilprivatisierung des Telekommunikationsmarktes hat im letzten Jahr durch den Verkauf der Telekom Srpske begonnen. Neuer Mehrheitseigentümer mit 65 Prozent wurde die serbische Telefongesellschaft Telekom Srbija. Es bleibt abzuwarten, wie sich dieser Verkauf auf die Weiterentwicklung des Telekommunikationsmarktes in Bosnien-Herzegowina, genauer gesagt in der Republika Srpska auswirken wird, da es sich bei der Telekom Srbija ebenfalls um ein staatliches Unternehmen handelt.

Ein großes Problem stellt immer noch die Stromversorgung in Bosnien-Herzegowina dar. Auch wenn alle Haushalte an das Stromnetz angeschlossen sind, kommt es vor allem in den ländlichen Regionen häufig zu Stromausfällen, die zum Teil mehrere Stunden, zeitweise sogar Tage anhalten können.

Die Penetrationsraten für Breitbandzugänge sind in den Ländern des ehemaligen Jugoslawien außerordentlich niedrig. Auch hier ist Kroatien wieder Spitzenreiter, allerdings mit gerade einmal knapp 6 Prozent. Im Vergleich zum Vorjahr erkennt man allerdings den enormen Zuwachs von nahezu 200 Prozent. Bosnien-Herzegowina verzeichnet lediglich eine Penetrationsrate von einem Prozent, hat seine Zuwachsrate im Vergleich zum Vorjahr allerdings um 100 Prozent gesteigert (vgl. Abb. 5).

In technologisch fortgeschrittenen Ländern zählt der Breitbandanschluss bereits zum Standard für den Internetzugang. Mit einer Penetrationsrate von 18 Prozent liegt Deutschland zwar über dem europäischen Durchschnitt von 11 Prozent, im Vergleich zu den weltweit führenden Industrie- und Technologienationen befindet es sich jedoch lediglich im unteren Bereich (Graumann & Wolf, 2007, 78 f.). Während sich folglich die westlichen Industriestaaten bereits in einem fortgeschrittenen Stadium ihrer Breitbandentwicklung befinden, stehen Bosnien-Herzegowina und die anderen Länder des ehemaligen Jugoslawiens erst am Anfang dieser Entwicklung.

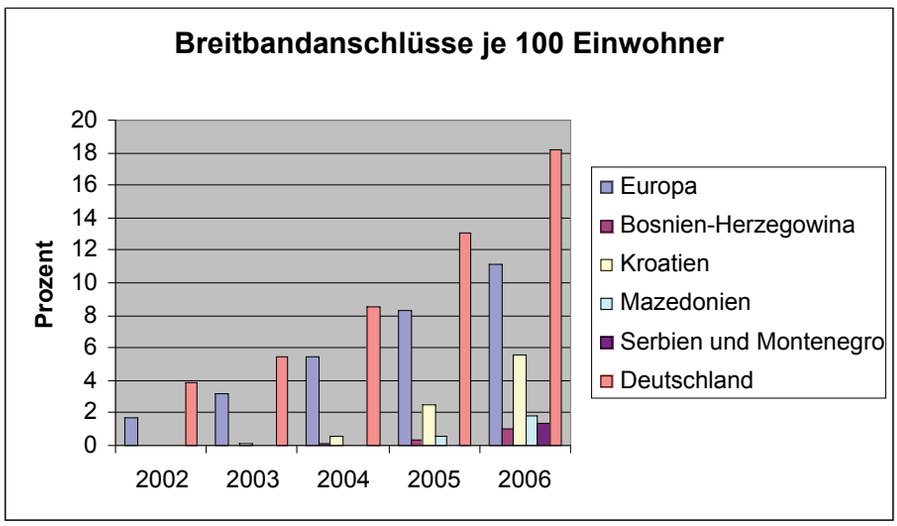
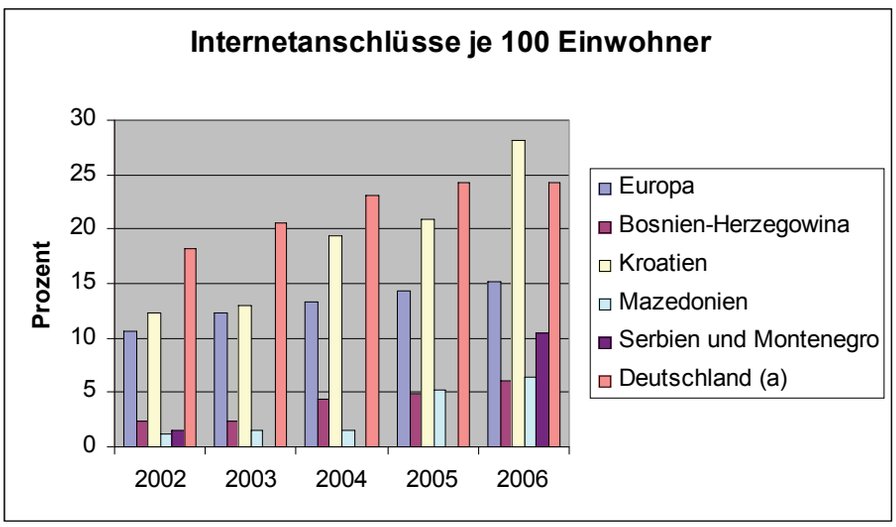


Abbildung 5: Breitbandanschlüsse je 100 Einwohner (Quelle: ITU 2007).



[a] Die Daten für Deutschland basieren auf einer Schätzung der ITU.

Abbildung 6: Internetanschlüsse je 100 Einwohner (Quelle: ITU 2007).

Der rasche Anstieg von Internetnutzern täuscht nicht über die geringe Zahl von Internetanschlüssen (Verträge in Bezug auf Breitbandzugänge, Standleitungen und Modemanschlüsse) hinweg. Die Penetrationsrate, worunter Modemverbindungen, Standleitungen und Breitbandzugänge fallen, liegt für Bosnien-Herzegowina gerade mal bei 6 Prozent (vgl. Abb. 6). Dies lässt darauf schließen, dass ein Großteil der Internetnutzer keinen eigenen Computer besitzt, sondern entweder im Internetcafe oder aber bei Freunden und Verwandten surft. Dies verwundert nicht, da sich bei einem Durchschnittseinkommen von 300 EUR die wenigsten einen eigenen Rechner leisten können. Dieser kostet ungefähr so viel wie ein PC in Deutschland, zwischen 700 EUR und 1.500 EUR. In 2004 kamen in Bosnien-Herzegowina auf 100 Einwohner 1,79 Computer. In den anderen Ländern sieht es nicht besser aus: Während die Situation für Mazedonien (1,49) und Serbien (1,34) noch schlechter aussieht als in Bosnien-Herzegowina, ist sie in Kroatien mit 2,38 Computern auf 100 Einwohner unwesentlich besser (GITR, 2006).

Ausblick

Bosnien-Herzegowina hat noch einen langen Weg bis zur Informationsgesellschaft vor sich. Wichtigste Kriterien zur Ausbildung einer solchen sind politische Stabilität sowie wirtschaftliches Wachstum. Nach mehr als einem Jahrzehnt internationaler Hilfe und Wiederaufbauprogrammen steht das Land kurz davor, die Verantwortung für seine politische und wirtschaftliche Entwicklung in die eigene Hand zu nehmen. Die Notwendigkeit von IKT für die soziale und ökonomische Entwicklung des Landes wurden zwar erkannt, aufgrund der komplizierten politischen Machtverhältnisse konnten die Projekte zur Verbesserung der IKT-In-

frastruktur jedoch bis heute nicht erfolgreich umgesetzt werden.

Im Verlauf weiterer Untersuchungen werden die Dimensionen der Digitalen Kluft in Bosnien-Herzegowina eingehender analysiert. Darüber hinaus sollen Untersuchungen der Informationsinfrastruktur in Schulen, Universitäten und Bibliotheken, die vor Ort durchgeführt werden, über den Stand der IT-Kompetenz und Information Literacy Aufschluss geben. Diese beiden Aspekte sind von grundlegender Bedeutung und haben elementare Konsequenzen für die Entwicklung hin zu einer Informationsgesellschaft. Die durch die fehlende Informationskompetenz hervorgerufene Wissenskluft erweist sich als fundamentalste Dimension aller Digitalen Klüfte, da Wissensungleichheit zu Ausgrenzung von sozialen Möglichkeiten und zu Ungleichheiten im sozialen Machtgefüge führt.

Literatur

Annan, K. (2002). On the Digital Divide. www.un.org/news/ossg/sg/stories/sg-5nov-2002.htm [10.01.2008].

Annan, K. (2003).World Summit on the information society (WSIS). www.un.org/News/Press/docs/2003/sgsm9070.doc.htm [10.01.2008].

BFAI (2007). Wirtschaftstrends Bosnien und Herzegowina Jahresmitte 2007. www.bfai.de [10.01.2008].

BMBF. Informationsgesellschaft. www.bmbf.de/de/398.php [10.01.2008].

Calic, J.-M. (1996). Krieg und Frieden in Bosnien-Hercegovina. Erw. Neuausg. Frankfurt am Main: Suhrkamp.

Castells, M. (2005). Die Internet-Galaxie. Internet, Wirtschaft und Gesellschaft. Wiesbaden: VS, Verlag für Sozialwissenschaften.

CIA. The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/index.html [10.01.2008].

Dell'Anno, R. & *Piirisild*, M. (2007). Measuring the Non-observed Economy in a Transition Economy. The Case of Bosnia-Herzegovina. Comparative Economic Studies, 49, 609-631. www.palgrave-journals.com/ces/journal/v49/n4/pdf/8100211a.pdf [10.01.2008].

EIU (2007). The Economist Intelligence Unit Limited. Country Profile 2007. Bosnia and Hercegovina.

Finlay, A. (2007). Bosnia and Herzegovina. Global Information Society Watch, 99-101. www.globalinformationsocietywatch.org/files/pdf/GISW_2007.pdf [10.01.2008].

GITR (2006). World Economic Forum (WEF). Global Information Technology Report. www.insead.edu/v1/gitr/wef/main/home.cfm [10.01.2008].

Graumann, S. & *Wolf*, M. (2007). Monitoring Informations- und Kommunikationswirtschaft: 10. Faktenbericht 2007. Eine Sekundärstudie der TNS Infratest Forschung GmbH. Monitoring Informations- und Kommunikationswirtschaft / Faktenberichte, 10. München, Berlin: TNS Infratest.

ITU (2004). International Telecommunication Union. Information technology. www.itu.int/ITU-D/ict/statistics/at_glance/Internet04.pdf [10.01.2008].

ITU (2007). International Telecommunication Union 1999 – 2006. Basic indicators; Main telephone lines; Internet indicators. www.itu.int/ITU-D/icteye/Indicators/Indicators.aspx [10.01.2008].

Kostov, Z. (2006). Global tendencies and local implications: Cyber exklusion of and within Western Balkan Countries. Discussion Paper 40. www.lse.ac.uk/Depts/global/Publications/DiscussionPapers/DP40.pdf [10.01.2008].

Nastav, B. & *Bojnec*, Š., (2007). The Shadow Economy in Bosnia and Herzegovina, Croatia, and Slovenia. The Labor Approach. Eastern European Economics, 45(1), 29-58.

Norris, P. (2001). Digital Divide. Civic Engagement, Information Poverty, and the Internet Worldwide. Communication, society and politics. Cambridge: Cambridge Univ. Press.

OECD (2007). OECD Communications Outlook 2007. http://213.253.134.43/oecd/pdfs/browseit/9307021E.pdf [10.01.2008].

Oschlies, W. (2006). Das politische System Bosnien-Hercegovinas. In: *Ismayr*, W. (Hrsg.): Die politischen Systeme Osteuropas (S. 739-768). 2., akt. und überarb. Aufl., unveränd. Nachdr. Wiesbaden: VS Verl. für Sozialwissenschaften.

Pellizzer, V. (2007). Bosnia and Herzegovina. Global Information Society Watch, 115-118. www.globalinformationsocietywatch.org/files/pdf/GISW_2007.pdf [10.01.2008].

Schneider, F. (2004). The Size of the Shadow Economies of 145 Countries all over the World. First Results over the Period 1999 to 2003. Forschungsinstitut zur Zukunft der Arbeit. Discussion Paper No. 1431. Bonn. http://ssrn.com/abstract=636661 [10.01.2008].

Schwarz, A. (2003). Die Verfassung des Gesamtstaates Bosnien-Herzegovina. Europa-Blätter Nr. 4, 128-133.

Schwarz, A. (2004a). Entitäten in Bosnien-Herzegovina. Die Verfassung der Föderation BiH. Europa-Blätter Nr. 2, 48-56.

Schwarz, A. (2004b). Entitäten in Bosnien-Herzegowina. Die Verfassung der Republika Srpska. Europa-Blätter Nr. 3, 93-101.

Servon, L.J. (2002). Bridging the Digital Divide. Technology, Community, and Public Policy. Malden, MA: Blackwell.

Stock, W.G. (2000). Informationswirtschaft. Management externen Wissens. München, Wien: Oldenbourg.

Tichenor, P.J., *Donohue*, G.A. & *Olien*, C.N. (1970). Mass Media Flow and Differential Growth in Knowledge. The Public Opinion Quarterly 34(2), 159-170.

UNDP (2007a). United Nations Development Programme: Human Development Report 2007. http://hdr.undp.org/en/media/hdr_20072008_en_complete.pdf [10.01.2008].

UNDP (2007b). United Nations Development Programme: National Human Development Report 2007. Social Inclusion in Bosnia and Herzegovina. http://hdr.undp.org/en/reports [10.01.2008].

Vockic-Avdagic, J. (2003). The Internet and the Public in Bosnia-Herzegovina. http://soemz.euv-frankfurt-o.de/media-see/newmedia/main/articles/j_avdagic.htm [10.01.2008].

WDI (2006). World Bank: World Development Indicators. States and Markets. The information age. http://devdata.worldbank.org/wdi2006/contents/Section5.htm [10.01.2008].

WDI (2007). World Bank. World Development Indicators. States and Markets. The information age. http://siteresources.worldbank.org/DATASTATISTICS/Resources/table5_11.pdf [10.01.2008].

WSIS (2003). World Summit on the Information Society. Geneva 2003 – Tunis 2005. Declaration of Principles. www.itu.int/wsis/docs/geneva/official/dop.html [10.01.2008].

Bosnien-Herzegowina, Digitale Kluft, Entwicklungstendenz, Informationsgesellschaft, Informations- und Kommunikationstechnologie, Infrastruktur, Informationsnetz, IKT, Kroatien, Mazedonien, Serbien und Montenegro

Die Autorin

Violeta Trkulja M. A.

Studium der Älteren und neueren Germanistik und Informationswissenschaft an der Heinrich-Heine-Universität Düsseldorf. Dort seit 2002 wissenschaftliche Angestellte am Institut für Sprache und Information in der Abteilung für Informationswissenschaft. Methodenkurse zu Information Retrieval, Inhaltserschließung und Konzeption und Aufbau von Webprojekten. Promotion zur Digitalen Kluft am Beispiel von Bosnien-Herzegowina. Forschungsschwerpunkte sind die Digitale Kluft, Wissensrepräsentation, mit besonderer Berücksichtigung textsprachlicher Methoden (Textwortmethode), Web-Usability und Accessibility sowie Informationskompetenz.

trkulja@uni-duesseldorf.de

www.phil-fak.uni-duesseldorf.de/infowiss

DABIS.eu

Gesellschaft für Datenbank-InformationsSysteme mbH

Ihr Partner für Archiv-, Bibliotheks- und DokumentationsSysteme

BIS-C 2000

Archiv- und Bibliotheks-InformationsSystem

DABIS.eu · alle Aufgaben · ein Team

Synergien: Qualität und Kompetenz
Software: Innovation und Optimierung
Web · SSL · Warenkorb und Benutzeraccount
Lokalsystem zu Aleph-Verbünden

Software · State of the art · Open Source

Leistung	Sicherheit
Standards	Offenheit
Stabilität	Verläßlichkeit
Generierung	Adaptierung
Service	Erfahrenheit
Outsourcing	Support
Dienstleistungen	Zufriedenheit

GUI · Web · Wap · XML · Z 39.50

Archiv	System	Bibliothek
singleUser		multiUser
Lokalsystem		Verbund
multiDatenbank		multiServer
multiProcessing		multiThreading
skalierbar		stufenlos
Unicode		multiLingual
Normdaten		redundanzfrei
multiMedia		Integration

DABIS.com
Heiligenstädter Straße 213
1190 - Wien, Austria
Tel.: +43-1-318 9 777-10
Fax: +43-1-318 9 777-15
eMail: office@dabis.com
http://www.dabis.com

DABIS.de
Herrgasse 24
79294 - Sölden/Freiburg, Germany
Tel.: +49-761-40983-21
Fax: +49-761-40983-29
eMail: office@dabis.de
http://www.dabis.de

Zweigstellen: 61350 - Bad Homburg vdH, Germany / 1147 - Budapest, Hungary / 39042 - Brixen, Italy

Lieber für GENIOS zahlen, als umsonst googeln.

Seriöse Wirtschaftsdaten sollte man sich etwas kosten lassen – es lohnt sich. GENIOS ist der größte Anbieter seriöser, deutschsprachiger Wirtschaftsinformationen. Bei uns können Sie schnell, zuverlässig und unbürokratisch jede Information kaufen – zu Firmen, Branchen oder Personen aus den Bereichen Wirtschaft, Management und Wissenschaft. Wir bieten Ihnen zum Beispiel das Wissen von ● 800 Datenbanken ● 150 Pressearchiven und ● über 400 Fachzeitschriften, außerdem ● 42 Millionen Firmeninformationen ● 10 Millionen Personeninformationen ● 6 Millionen Nachweise aus den Bereichen Wirtschafts- und Sozialwissenschaften sowie ● die Originaldaten des Bundesanzeigers. Und auf Wunsch ● individuelle, maßgeschneiderte Informationslösungen. 25.000 Nutzer profitieren bereits davon.

Was wollen Sie wissen? www.genios.de

GENIOS

German Business Information

GBI-Genios Deutsche Wirtschaftsdatenbank GmbH
Ein Unternehmen der Frankfurter Allgemeine Zeitung GmbH
und der Verlagsgruppe Handelsblatt GmbH