Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

Knowledge eXtended

Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten

3. Konferenz der Zentralbibliothek

2. – 4. November 2005 Jülich
Vorträge und Poster
Mit einem Festvortrag von Norbert Bolz

Bibliothek
Library

Schriften des Forschungszentrums Jülich
Reihe Bibliothek / Library Band / Volume 14

Forschungszentrum Jülich GmbH
Zentralbibliothek

Knowledge eXtended

Die Kooperation von Wissenschaftlern,
Bibliothekaren und IT-Spezialisten

3. Konferenz der Zentralbibliothek

2.– 4. November 2005 Jülich
Vorträge und Poster

Schriften des Forschungszentrums Jülich
Reihe Bibliothek / Library Band / Volume 14

ISSN 1433-5557 ISBN 3-89336-409-9

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte Bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Herausgeber und Vertrieb:	Forschungszentrum Jülich GmbH Zentralbibliothek D-52425 Jülich Telefon (02461) 61-5368 · Telefax (02461) 61-6103 e-mail: zb-publikation@fz-juelich.de Internet: http://www.fz-juelich.de/zb
Umschlaggestaltung:	Grafische Medien, Forschungszentrum Jülich GmbH
Druck:	Grafische Medien, Forschungszentrum Jülich GmbH
Copyright:	Forschungszentrum Jülich 2005
Zusammenstellung:	Roswitha Moes, Cornelia Plott, Zentralbibliothek, Forschungszentrum Jülich GmbH

Schriften des Forschungszentrums Jülich
Reihe Bibliothek / Library Band / Volume 14

ISSN 1433-5557
ISBN 3-89336-409-9

INHALTSVERZEICHNIS

Konferenzkomitees

Programmkomitee

Dr. Rafael Ball	Forschungszentrum Jülich, Zentralbibliothek
Dr. Jürgen Bunzel	Deutsche Forschungsgemeinschaft, Programmdirektor für Wissenschaftliche Literaturversorgung und Informations-systeme
Prof. Dr. Josef Herget	Hochschule für Technik und Wirtschaft, Chur, Leiter des Studiengangs Information und Dokumentation
Prof. Dr. Dr. Thomas Lippert	Forschungszentrum Jülich, Zentralinstitut für Angewandte Mathematik
Dr. Norbert Lossau	Leiter der Universitätsbibliothek Bielefeld
Prof. Dr. Heiner Müller-Krumbhaar	Forschungszentrum Jülich, Institut für Festkörperphysik, Herausgeber der Europhysics Letters

Organisationskomitee

Dr. Rafael Ball	Forschungszentrum Jülich, Zentralbibliothek (Call for Paper, inhaltliche Fragen)
Dr. Bernhard Mittermaier	Forschungszentrum Jülich, Zentralbibliothek (Konferenzplanung)
Anne Otto	Forschungszentrum Jülich, Zentralbibliothek (Sponsoring)
Edith Salz	Forschungszentrum Jülich, Zentralbibliothek (Presse)

Vorwort

Die wissenschaftliche Informationsversorgung von heute steht vor einem qualitativen Sprung. Längst ist die Vision der digitalen Bibliothek Realität geworden und die medienbruchfreie Informationsversorgung gerade im Bereich von Science, Technology und Medicine (STM) fast vollständig umgesetzt.

Die nächste Stufe der Informationsversorgung kann nicht allein durch Bibliothekare und Informationsspezialisten geleistet werden, sondern nur durch die Zusammenarbeit mit anderen Akteuren, den IT-Spezialisten und den Wissenschaftlern als Produzenten und Nutzern von wissenschaftlichen Inhalten selbst. Durch die immer leistungsfähigere und intelligentere IT-Infrastruktur sind heute der Transport und das Prozessieren von gewaltigen Datenmengen problemlos möglich geworden. Virtuelle Informationswelten und künstliche Intelligenz rücken wieder in den Blickpunkt, nachdem sie bereits vor zwanzig Jahren aufgrund der begrenzten Leistungsfähigkeit der IT-Strukturen in eine Sackgasse geraten waren. In Verbindung mit neuen Ansätzen eines „e-Science“ (enhanced Science oder electronic Science) erwarten die Wissenschaftler heute eine umfassende Versorgung mit Literatur und Information, aber auch die volle und kompetente Unterstützung bei der Archivierung und Prozessierung ihrer wissenschaftlichen Ergebnisse und Primärdaten. Nur die intensive Zusammenarbeit aller Beteiligten kann diese neuen Herausforderungen gelingen lassen.

Die Konferenz „Knowledge eXtended: Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten“ ist bereits die dritte Konferenz der Zentralbibliothek zu wichtigen und aktuellen Themen im Umfeld von Wissenschaft, Publikation und Bibliothek. DataMining als Methode für die Generierung neuer Inhalte aus bestehenden Datenmengen und Publikationen, semantische Netze, die geeignet sind, qualitativ hochwertige und tiefgehende Informationen zu generieren und Inhalte aufzufinden, GridComputing als eine Möglichkeit, leistungsfähige Systeme durch verteilte IT-Ressourcen zu schaffen und Open Access als Ansatz für eine neue Wissenschaftskommunikation sind die Schwerpunkte im Konferenzband. Darüber hinaus enthält der Band auch Beiträge, die in der Postersession vorgestellt worden sind, sowie die Festrede von Norbert Bolz, in der er die Wissensgesellschaft als Stupid Stuff oder Serious Business auf die Probe stellt.

Danken möchte ich all jenen, die zum Gelingen der Konferenz, sei es als Vortragende, Moderatoren, Organisatoren, im Programmkomitee oder als Teilnehmer beigetragen haben. Besonderer Dank gilt dem Forschungszentrum Jülich, das die Konferenz in dieser Form erst möglich gemacht hat.

Dr. Rafael Ball
Forschungszentrum Jülich, Zentralbibliothek

Festvortrag

Die Wissensgesellschaft - Stupid Stuff oder Serious Business?

Norbert Bolz, Berlin

Globalisierung, Vernetzung, Weltkommunikation - da kommt man leicht ins Träumen. Als ob alles zusammenwachsen würde zur Einen Welt, zum digitalen Weltdorf. In der Tat fallen ja allerorten die Grenzen: zwischen den Staaten Europas, zwischen den Geschmäckern der Klassen, zwischen den Ebenen der Unternehmen. Überall strahlen die Sterne von Hollywood, überall gibt es McDonalds, überall herrscht Microsoft. Und zumindest in der Chance, arbeitslos zu werden, sind wir fast alle gleich. Doch wenn man aus diesem Millennium-Traum aufwacht, kann man schon recht deutlich neue Grenzen erkennen. Die schärfste Grenzlinie hat Benjamin Barber markiert: Dschihad gegen McWorld, also der Heilige Krieg des Fundamentalismus gegen eine westliche Netzwerk- und Medien-Kultur, die eben gar kein Fundament mehr hat. Ähnlich verläuft die Demarkationslinie zwischen den Angeschlossenen und den Ausgeschlossenen, den "linked" und "linked-nots". Wer keine IP-address hat, fällt durch die Maschen des Weltnetzes. So weit, so einfach.

Schwieriger ist es schon, die Grenze zwischen Programmierern und Programmierten zu sehen und die Folgen dieser neuen Klassenspaltung abzuschätzen. Das ist vor allem deshalb schwierig, weil die Software-Industrie alles daransetzt, den User als lernunfähigen Konsumenten zu umgarnen. Der ideale Kunde soll nichts von der logischen Tiefe des Computers wissen - "nur vom Fachmann zu öffnen..."

Und auch durch den Cyberspace zieht sich heute eine klare Grenzlinie, nämlich zwischen Kapitalismus und Anarchie. Dazu liefert eine Anzeige von Lotus wunderbaren Klartext: "The great invisible guiding hand of capitalism has just smacked the Internet upside the head - now what?" Das Internet ist also ein anarchistischer Augias-Stall, angefüllt mit Geschwätz, Bildchen, Witzen und Wahn. Der wird nun endlich gereinigt und den Sicherheitsstandards des "serious business" unterworfen - alles andere ist "stupid stuff", Blödsinn.

Wie gesagt: das ist Klartext. Dagegen klingt es doch wie ein romantisches Märchen, wenn Jon Katz im Cyberkultmagazin *Wired* die Geburt der Digitalen Nation aus dem Geist des Internet beschwört. Die Bürger der Digitalen Nation sind jung, gebildet, ehrgeizig; sie arbeiten in den Universitäten, Telekommunikationsgesellschaften, Medienhäusern und Banken; sie sind tolerant, vernünftig, medienkompetent und - nein, nicht unpolitisch, sondern "postpolitisch". Diese Netzbürger haben nur eine

Leidenschaft: den freien Fluss der Information. Jeder soll sprechen und gehört werden können.

Doch das ist zu schön, um wahr zu sein. Gesellschaft an sich schließt nämlich absolute Informationsfreiheit aus - man denke nur ans Top Secret von Militär und Politik, ans wirtschaftliche Betriebsgeheimnis und Copyright, aber auch an die nur durch Verschlüsselungsfreiheit garantierte Privatsphäre. Deshalb wird es immer Firewalls, Gateways, Zugangsbeschränkungen und geheime Informationen geben. Und deshalb wird es wohl auch zu einer Verdoppelung des Internet kommen. Denn als radikaldemokratisches Kommunikations-medium ist es für den Kapitalismus uninteressant. Internet I wird dann Tools für "serious business" bereitstellen - gegen gutes Geld. Und Internet II lässt den Rest der Welt Surfen im "stupid stuff". Eine neue Grenze.

Business als Netzwerk - Business im Netzwerk

Wenden wir uns also dem "serious business" zu. Wie sieht die Wirtschaft der Multimedia-Gesellschaft aus? Man kann sie als eine Ellipse beschreiben, die um zwei Brennpunkte konstruiert ist: die Produktivkraft Kommunikation und die kommunikative Lust. Es geht also nicht nur um Information als Aufklärung und Daten-Prozess, sondern auch - und das ist etwas ganz anderes! - um Kommunikation als Faszination. Von Kommunikation als Faszination der Multimedia-Gesellschaft war gerade die Rede. Benennen wir jetzt die objektiven Faktoren, die jeden Unternehmer und Top-Manager heute zwingen, die Produktivkraft Kommunikation zur Chefsache zu machen:

- Globalisierung der Wirtschaft (Global Players, Placeless Society)
- Virtualisierung der Arbeitsverhältnisse (Telecommuter; Kommunikations-nomaden)
- Heterarchie der Organisationen (das Unternehmen als Konversationsnetz)
- Immaterialisierung der Produkte (wachsender Beratungsbedarf).

Für die Wirtschaft des Unsichtbaren ist Herbert Marshall McLuhans Vision Wirklichkeit geworden: das dezentrale elektronische Weltdorf, die geschrumpfte Welt der Satellitenkommunikation, in der räumliche Distanzen unwichtig sind, solange man ans Netzwerk angeschlossen ist. Es hat deshalb einen guten Sinn, wenn William Knoke die vernetzte Gesellschaft als "Placeless Society" charakterisiert. In der Wirtschaft kann das leicht zu einer emotionalen Abkopplung der Firmen vom Standort führen - zumal dann, wenn die Firmen an ihrem alten Standort auf Regulierungssüchtige treffen. Das meinte wohl auch Martin Herzog (VDA) mit seinem lapidaren Satz: "Wir sind zwar Patrioten, aber keine Idioten."

Unternehmen sind komplexe Systeme. In einer turbulenten Welt stabilisieren sie ihre Identität gerade durch ihre Offenheit für Umwelteinflüsse. Mit anderen Worten: Komplexe Systeme wie etwa Firmen müssen sensibel für Irritationen sein. Das setzt voraus, dass man einen positiven Begriff von Störungen entwickelt. So lässt sich die Selbsterhaltung komplexer Organisationen, also etwa das Überleben einer Firma, als Kommunikationsprozess in einem Netz von Rückkopplungsschleifen beschreiben.

Man kann prinzipiell sagen: Wenn eine Organisation sich auf sich selbst bezieht, wird sie Kommunikation. Im Feedback-Prozess werden die Irritationen und Turbulenzen verarbeitet. Das setzt voraus, dass ein Unternehmen das Chaos des Marktes nicht als Ausnahmezustand, sondern als Norm begreift. Und das gilt für alle Organisationen der Gegenwart, die in enger Markt- und Technologiebindung ständig zu raschen Entscheidungen gezwungen sind.

Die Stabilität und Flexibilität eines Systems zeigt sich an seiner Kommunikationsfreudigkeit. Sollte der Manager eines solchen Unternehmensnetzwerks aber versuchen, Führungsstärke durch Befehl und Hierarchie zu beweisen, so wird er allenfalls erreichen, dass ihm seine klugen Mitarbeiter vorspielen, die Geführten zu sein. Je mehr sich Wirtschaftsunternehmen in flache Netzwerke und nichthierarchische Rückkopplungssysteme verwandeln, um so mehr verlagern sich die entscheidenden Machtprozesse auf die Ebene der Angestellten selbst. Damit definiert sich aber die Aufgabe des Managers völlig neu: Er muss sich als Trainer und zugleich als Schiedsrichter im Machtkampf der Untergebenen verstehen. Der Manager ist selbst ein Element des Systems, das er steuert.

Man könnte deshalb mit Winograd und Flores das "Management als Sorge um die Artikulation und Aktivierung eines Netzwerkes wechselseitiger Verpflichtungen" definieren. Eine souveräne Führungs-persönlichkeit wird in Zukunft nur noch einen Rahmen definieren, innerhalb dessen sich Prozesse evolutionärer Selbstorganisation vollziehen können - Führung zur Selbstführung, könnte man sagen. Hierarchie ist der Gegensatz von Kommunikation. Der technische Standard der Netzwerke erzwingt also auch tiefgreifende Veränderungen im Entscheidungs-Prozess. Der Befehlsfluss verläuft nicht mehr von der Spitze zur Basis, sondern in kleinen Schleifen - Stichwort: Heterarchie. Und so wie Netzwerke kleiner Rechner die Dinosaurier der Mainframes zum Aussterben bringen, so wird eine Modularisierung der Betriebe in Zukunft einen neuen Organisationsstil fordern - fraktales Management.

Auch dieses Zauberwort ist rasch erklärt. Das Großunternehmen der Zukunft zerfällt in flexible, "selbstähnliche" Unternehmensmodule, die quasi-autonom operieren. Die Außenbeziehungen des Unternehmens nehmen den Charakter der Telekooperation an, seine Mitarbeiter werden zunehmend Telecommuter - mit dem Grenzwert eines

virtuellen Arbeitsplatzes. Schon heute gibt es virtuelle Unternehmen, die überhaupt nur projektgebunden im Datennetz existieren. Ist das Projekt abgeschlossen, löst sich das Unternehmen in nichts - genauer gesagt: in vollkommen voneinander unabhängige Module auf.

Wer im Business der Zukunft Erfolg haben will, muss deshalb Medienkompetenz und einen anspruchsvollen Begriff von Kommunikation haben. Denn Firmen und Unternehmen sind Organisationen, und "Organisationen existieren als Netze aus Direktiven und Kommissiven." [Winograd/Flores]. Das klingt kompliziert, ist aber ganz einfach zu verstehen. Es geht ja bei jedem Geschäft um Angebote und Rückfragen, um Versprechen und Zusagen. Das Organisationsnetz ist also ein Konversationsnetz. Halten wir deshalb fest: Die nicht weiter auflösbaren Letzelemente von Business-Netzwerken sind Kommunikationen. Der Guru der Managementtheorie, Tom Peters, sagt deshalb zurecht: "Knowledge output and the poetry of networks are part of a dramatic shift in the way we think about work - work as conversation".

Wenn aber Arbeit und Geschäft heute ein Gespräch sind, ist es von entscheidender Wichtigkeit, diesen Konversations-Prozess durch fortgeschrittene Medien zu stützen. Darin liegt die eigentliche Bedeutung der aktuellen Diskussion um Multimedia, Hypermedia und Hypertext.

Netzwerkabhängigkeit: Business to Business

Für Friedrich von Hayek ist der Markt der Ort, an dem alle wichtigen Informationen der Wirtschaft konzentriert sind. Diese Theorie gewinnt für die Informationsgesellschaft unserer Tage einen neuen, prägnanten Sinn. Marketing ist die kommunikative Steuerung des Unternehmens vom Markt aus. Allerdings genügt heute nicht mehr die Information der Preise und das "Entdeckungsverfahren" des Wettbewerbs, um sich ökonomisch zu orientieren. An die Information der Konsumenten pirscht man sich bekanntlich mit den Techniken der Trendforschung und des computergestützten Direkt-Marketing heran. Doch wie muss sich das Marketing im Business-to-Business-Bereich einstellen?

Sicher genügt es nicht, auf der Homepage des Unternehmens "links" zu den Geschäfts- und Marktpartnern anzubieten. Aber das Internet ist ein bedeutsames Medium für die Optimierung des Business-to-Business-Marketing. Und Business-to-Business ist selbst ein unvergleichliches Medium des Vergleichs zwischen Unternehmen, der zur Entdeckung von Marktnischen, zum Out-Sourcing dessen, was andere besser können, zu Joint Ventures mit intelligenten Partnern oder auch nur zur Veränderung der eigenen Organisations-struktur führen kann. Die Unter

nehmen stehen also keineswegs nur in Konkurrenzverhältnissen, sondern sie ergänzen sich und bilden Netzwerke.

Interorganizational Networks unterlaufen die Unterscheidung von Unternehmen und Markt. Immer häufiger kommt es in der vernetzten Welt zu Hybridbildungen und wechselseitigen Durchdringungen zwischen Markt und Organisation - "symbiotische Kontrakte", wie Erich Schanze sagt. Das ist einfacher zu verstehen als es klingt; man denke nur an Joint Ventures oder das Franchising. Antony Dnes sagt zurecht: "Franchising is more like an integrated business than a set of independent firms." Jeder Knoten im Netz arbeitet gleichzeitig autonom für sich und für das Netz. Hier handelt es sich nicht mehr um reine Organisationsstrukturen, aber auch nicht um bloße Marktkontrakte, sondern um eigentümliche Mischgebilde, die für das Business der Zukunft charakteristisch sind. Sie sind rigider, also verlässlicher als der Markt, aber flexibler als die Organisation.

Digital vernetzte Organisationen lassen sich nicht mehr sinnvoll in einer Befehlshierarchie darstellen oder als klar abgegrenzte "Körperschaft" identifizieren. Ein Unternehmen ist heute nichts anderes als der Inbegriff seiner internen und äußeren Beziehungen, die im wesentlichen als Informationsprozesse gestaltet sind. Tom Peters nennt das "the intangibilizing of everything": Weder die Organisation, noch die Arbeit oder das Produkt lassen sich handgreiflich "fassen". Kevin Kelly spricht in diesem Zusammenhang sogar von einem "neuen Spiritualismus" der Netzwerk-Ökomomie.

Die Wettbewerbsposition eines Unternehmens wird also vor allem durch das Beziehungsgefüge bestimmt, in dem es zu anderen Unternehmen steht. Die Firmen treten gleichzeitig als Kunden, Konkurrenten und Partner zueinander in Beziehung. Alvin Toffler spricht in diesem Zusammenhang von "power-mosaics", in denen es nur noch eine Form von Kontrolle gibt - nämlich Kommunikation. Der Wettbewerb wird zunehmend zum "partnering process". Und Marktabhängigkeit heißt heute konkret: Netzwerkabhängigkeit.

Gunther Teubner hat in diesem Zusammenhang versucht, einen trennscharfen Begriff des Business-Netzwerks zu entwickeln: "Von Netzwerk sollte man dann und nur dann sprechen, wenn ein Handlungs-system sich zugleich als formale Organisation und als Vertragsbeziehung zwischen autonomen Akteuren formiert." Diese klare, durchaus weiterführende Definition ist allerdings noch zu sehr auf Handlung fixiert und verliert dabei aus dem Auge, dass beim neuen Geschäft als Gespräch die nicht weiter dekomponierbare Letzteinheit eben Kommunikation ist.

Gunther Teubner diskutiert das Business-Netzwerk als Handlungssystem vor allem wohl auch deshalb, weil es ihm um die Frage des "Kollektivakteurs" geht. Kann man

etwa ein Franchising-Netzwerk als handelndes Subjekt begreifen? Wie soll man Handlungen auf Netzwerke zurechnen? Diese Frage ist nicht nur für die Jurisdiktion, sondern gerade auch für die Gewerkschaften von allergrößtem Interesse. Denn der Arbeitskampf wird ja sinnlos, wenn Unternehmensentscheidungen nur noch Emergenzphänomene des Netzwerks sind. Elektronische Verknüpfungen verwischen die Grenzen, Kampf- und Konkurrenzlinien in und zwischen Organisationen.

Das Büroleben im Cyberspace

Wer von Globalisierung der Wirtschaft redet, muss auch von lokaler Selbstorganisation in den Betrieben reden; denn beide Prozesse sind komplementär. Deshalb fasziniert heute das Internet nicht nur als neue Infrastruktur der Weltkommunikation, sondern auch als Metapher für spontane Ordnung. Und beides hat massive Konsequenzen für das Büroleben. Man gewinnt den Eindruck, dass hierarchische Autorität zunehmend durch Kommunikation ersetzt wird. Früher war ja Information in Autorität fundiert - der Chef hat es gesagt. Heute ist Autorität auf Information fundiert. Und man begreift allmählich, dass sich die Effektivität einer Organisation nur durch den Wettbewerb der Informationsquellen steigern lässt. Das nennt man auch Heterarchie, zu Deutsch: Die Organisation ist ein Konversationsnetz.

So entfaltet sich heute - gegen den Strich der soziale Hierarchie des traditionellen Bürolebens - eine "flache" Netzwerk-Kultur. Technisch implementiert wird das durch eine Software, die man Meetingware nennt; sie untergräbt Hierarchien. Doch die Ersetzung von hierarchischer Autorität durch Kommunikation hat ihren Preis: Konsens kostet Zeit - während eben umgekehrt ein Befehl Zeit spart. Horizontale Kommunikation ist zeitraubend. Man kennt das ja von der Face-to-face-Kommunikation: sie erzwingt Termine.

Und ein weiteres Problem des heterarchischen Bürolebens kommt hinzu. Kommunikation zu fördern heißt nicht auch schon, die Informationsverarbeitung zu fördern. Im Gegenteil: Geselligkeit tendiert zum Geschwätz. Aber gerade dieses Geschwätz, die sogenannte informelle Kommunikation im Büro, ist für dessen Funktionieren ja von unschätzbarer Wichtigkeit. Deshalb orientiert sich auch das Software-Design immer entschiedener an Kommunikation. Der IMB-Mainframe war der Inbegriff klassisch-autoritärer Informationsverarbeitung; der Personal Computer versprach dann jedem einzelnen "information at your fingertips"; und heute zielt man auf einen Interpersonal Computer, der das Büroleben nicht mehr mit einer Information, sondern mit einer Beziehung beginnen lässt.

Das Büroleben in der Netzwerk-Kultur steht zwar im Zeichen lokaler Selbstorganisation, macht damit aber Management nicht überflüssig. Die Frage ist nur, ob

Management noch Sache von Managern sein wird. Unter dem Titel Artificial Intelligence wird heute die Anwendung von Taylors Scientific Management auf den Geist des Managements erprobt. Man könnte von einem permanenten Turing-Test für Manager sprechen:

Welchen Teil ihrer Leistung kann man durch Software ersetzen? Wer diesen Test übersteht, d.h. nicht durch ein "intelligent agent based system" ersetzt wird, ist ein Meister der Paradoxien. Das Management steht heute nämlich vor allem vor dem Problem von Dezentralisierung und Kontrolle. Die Paradoxie liegt darin, dass man nur noch erfolgreich Kontrolle ausüben kann, wenn man die Kontrolle aus der Hand gibt. D.h. die Kontrolle der "core values" ermöglicht Dezentralisierung und *loose coupling*. Um es wiederum mit einer Faustformel zu sagen: Wer in einem Unternehmen die Hierarchie des Warum retten will, muss gerade die Heterarchie des Wie fördern.

Solchen komplexen Sachverhalten ist mit der naiven Vorstellung, der Computer sei ein neues Werkzeug der Büroarbeit, nicht beizukommen. Von den Netzwerken wie Inter- und Intranet kann man lernen, dass der Computer gerade kein Werkzeug ist; die Bürokratie ist ja auch kein Werkzeug, sondern objektiver, geronnener Geist. Dass man heute so gerne mit der Netzwerk-Metapher operiert, hat also einen guten Sinn: man stellt damit auf Medium um. Um es in aller Deutlichkeit zu sagen: Der Computer als Hypermedium ist kein Werkzeug, sondern der universale Arbeitsplatz. In diesem präzisen Sinne muss man den Cyberspace als Workspace verstehen. Er ist das elektronische Nervensystem der Weltwirtschaft.

"Ich bin ein Business"

Der Computer auf dem Schreibtisch des Büros - das ist ein zwar vertrautes, aber viel zu einfaches Bild vom Büroleben in der neuen Medienwirklichkeit. Um den Paradigmenwechsel prägnant benennen zu können, ist es hilfreich, sich noch einmal daran zu erinnern, dass Büro traditionell dreierlei meint:

- Trennung von Leben und Arbeit;
- Aktenförmigkeit;
- Betriebscharakter.

Die Gewohnheit wird hier zum Eigenwert. Das Verselbständigte ist das Selbstverständliche, der Betrieb, "es läuft". Und je besser es läuft, desto geringer wird die Fähigkeit, sich an das Unvorhergesehene anzupassen. Mit anderen Worten: Die traditionelle Welt des Büros ist rational, stabil und verlässlich - aber eben deshalb auch unflexibel und innovationsfeindlich.

Genau dagegen richtet sich heute das Konzept des One Person Office. Technisch konkret wird hier das Modem zum Widersacher des traditionellen Büros. Der Teleworker sagt: Mein Büro ist, wo mein Modem ist. Solche Kommunikationsnomaden erscheinen zumeist auch als Kommunikationsmonaden. Ob es der Laptop im Flugzeug oder das Handy im Intercity-Großraumwagen ist - ad hoc entsteht das One Person Office und der Rest der Welt versinkt. Gadgets wie der Nokia Communicator zeigen heute schon, wohin die Reise geht: Telefon, Fax, Computer, Internetanschluss - man trägt das Büro in der Hand.

Auch hier greift man zu kurz, wenn man Gadgets wie dieses als Werkzeug bezeichnet. Mindestens eben so sehr ist es ein faszinierendes, intelligentes Spielzeug. Und genau das ist für das Büroleben im Cyberspace charakteristisch: Arbeitsplatz und Spielplatz überschneiden sich. Das sei auch all jenen Chefs zum Trost gesagt, die ihre Angestellten beim Computerspielen erwischen. Problematisch ist nicht die Lust am Spiel, sondern etwas ganz anderes: Die Funktionslust und die Auslöserwirkung des Geräts führen dazu, dass wir im Umgang mit Computern beginnen, diejenigen Aufgaben zu bevorzugen, die sich mit Computern lösen lassen.

Der Computerfreak Peter Glaser hat zu unserem Thema einmal sehr schön bemerkt: "Man fährt nicht mehr zur Arbeit und kommt erledigt nach Hause, sondern die Arbeit kommt nach Hause und fährt erledigt in die Firma zurück." Die Arbeit emanzipiert sich vom Arbeitsplatz - das klingt nach Freiheit. Aber sie hat ihren Preis: Die soziale Umwelt der Face-to-face-Interaktion schrumpft. Und hinzu kommt, dass derartige Jobs keinen "Beruf" mehr ausmachen, sondern nur noch Strategien zur Lösung von Organisationsproblemen darstellen. Doch nicht nur der Beruf zerfällt, sondern auch das Unternehmen. Philosophen könnten von einer Dekonstruktion der Firma sprechen: Jeder wird ein Unternehmer. "Ich bin ein Business." Das ist der logische Grenzwert der Entorganisierung.

Das Layout des Büros bestimmt die Kommunikationsstruktur. Auf welche Kommunikationsform ist es angelegt? Face-to-face, Aktengang, Telefon und Email sind nämlich nicht äquivalent. Email, Anrufbeantworter und Fax sind ganz neue Formen, eine Nachricht zu hinterlassen. Sie sind dringlich und ersparen doch den Druck der Unmittelbarkeit und Präsenz. Ganz anders funktioniert der Brief oder - auf der entgegengesetzten Seite des Kommunikationsspektrums - das Telefon. Und mit all diesen technisierten Formen von Kommunikation wiederum unvergleichbar ist, was Erving Goffman face-work genannt hat. Übrigens ist, entgegen einem weit verbreiteten Vorurteil, gerade Face-to-face-Kommunikation nicht einfach, sondern hochkomplex "multimedial".

Deshalb braucht ein Manager heute Medienkompetenz, d.h. Sinn für den richtigen Medienmix im Büro. Management ist Kommunikationsdesign und das Büro ist ein Labor der neuen Kommunikationsmedien. Die Experimente, die dort am "lebendigen Objekt" vorgenommen werden, beantworten Fragen wie die folgenden: Wie verändert sich die sozial basierte informelle Kommunikation unter neuen Medienbedingungen? Wie ertragen es Menschen, dass alle mit ihnen kommunizieren können? Kann man sich mit einem virtuellen Unternehmen identifizieren?

Wir können vermuten, dass sich im Büro der Zukunft die Schere zwischen Data processing und People processing immer weiter öffnen wird. Unsere Gesellschaft macht gewaltige Fortschritte in Sachen Informationsverarbeitung, tritt aber "sozial" auf der Stelle. Und dieses Problem lässt sich durch kein "Management by xy" lösen. Vielleicht bildet es die Schlussfigur des Bürolebens.

Vernetzung und Interaktivität zielen ja auf eine Potenzierung des Zusammenhangs und der wechselseitigen Abhängigkeit in Systemen, auf höhere Verknüpfungsdichte also. Doch die Interdependenz in Organisationen findet ihre prinzipiellen Grenzen in der Kapazität individueller Informationsverarbeitung. Internet und Email führen uns sehr rasch an diese Grenzen. Denn dass jeder mit jedem kommunizieren kann, überlastet die Aufmerksamkeit. Daraus können wir für das Büroleben im Cyberspace lernen, dass Aufmerksamkeit die knappste Ressource ist. Management ist deshalb immer "attention management" (Simon): Selektion, Filterung, Design.

Damit sind wir aber wieder bei der Notwendigkeit hierarchischer Strukturen, die der heilige Geist der Informationsanarchie doch auflösen wollte. Gegen Information Overload hilft nämlich nur: delegieren. In der Datenflut der Multimedia-Gesellschaft kann "Mehrwert" nur heißen: weniger Information. Aber *wie* weniger? Man kann Informationen, die da sind, *nicht nicht* prozessieren. Hier haben wir wieder den Punkt erreicht, wo deutlich wird, dass nur Menschen etwas können, what computers can't do. Was Informationen nämlich brauchbar und bedeutsam macht, ist die spezifisch menschliche Form des Information processing: "Vergessen".

Das Unbehagen an der Information

Es gibt ein Unbehagen an der Selbstbeschreibung unserer Kultur als "Informationsgesellschaft". Das wäre ja eine Gesellschaft, die nicht mehr von Prozessen der Materie und Energie, sondern von Unterschieden, letztlich Sequenzen von 0 und 1, angetrieben würde. Denn Information ist nach Gregory Batesons berühmter Definition nichts als ein Unterschied, der Folgen hat - nämlich einen Unterschied macht. Dann würde alles Wesentliche an Handgreiflichkeit verlieren. Und in der Tat spricht der Management-Theoretiker Tom Peters vom "intangibilizing of everything".

Wer heute also von Wissensgesellschaft redet, artikuliert ein Ungenügen an der modischen Formel von der Informationsgesellschaft. Information unterscheidet nicht zwischen Sinn und Unsinn. Sie ist kein Maß für den Wert einer Botschaft. Der amerikanische Dichter Donald Hall kann deshalb sagen:

"Information is the enemy of intelligence." Weniger polemisch formuliert: Die digitale Information hat nichts mit der Welt der Intentionalität zu tun, also mit dem, was wir meinen, wenn wir von "Sinn" sprechen.

Herbert Simon hat einmal zurecht bemerkt: "information doesn't have to be processed just because it is there." Doch genau diese simple Einsicht hat unsere Multimedia-Gesellschaft aus dem Blick verloren. Unter dem Druck neuer Informationstechnologien neigt man dazu, alle Probleme als Probleme des Nichtwissens zu deuten. Doch Sinnfragen lassen sich nicht mit Informationen beantworten: "The problem is confusion, not ignorance." (Karl Weick)

Informationsübertragung hat nur sehr wenig mit dem zu tun, was menschliche Kommunikation ausmacht. Hinzu kommt, dass die Instantaneität der Datenprozesse uns keine Zeit des Nachdenkens mehr einräumt. Man könnte sagen: Instantaneität entmutigt die Reflexivität. Diese Unterscheidungen stecken wohl auch hinter der berühmten Formel des Philosophen Jürgen Mittelstrass, wir Menschen der westlichen Welt seien "Informationsriesen und Wissenszwerge zugleich". Um Informationen nutzen (und: genießen) zu können, braucht man Vorinformationen: das Wissen der Kultur, die Redundanz der Bildung. Um es auf eine einfache Formel zu bringen: Information ist nicht Wissen, sondern Wissen ist Form im Medium Information.

Die ultimative Ressource Wissen

Daniel Bell hat über die postindustrielle Gesellschaft gesagt: "Theoretisches Wissen ist zur Matrix der Innovation geworden." Eine naheliegende, aber zu simple Lesart dieses Satzes würde lauten: Wissen ist Macht. Die Formel macht erst Sinn, wenn man den gemeinten Sachverhalt temporalisiert. Nicht Wissen ist Macht, denn Wissen ist universal. Aber Vorsprungswissen ist Macht. Effektives Wissen hat heute einen Zeitindex. Ich komme darauf unter dem Titel "Halbwertzeit" des Wissens wieder zurück.

Die Rede von der Wissensgesellschaft ist - im wortwörtlichen Sinn: "grenzenlos" optimistisch. Denn Wissen ist die Ressource, die sich scheinbar nie erschöpft - ja die sich durch Gebrauch sogar vermehrt. Die traditionellen Produktivitätsfaktoren (Grundbesitz, Kapital, Arbeit) sind demgegenüber heute nur noch "constraints" der

einzigen Wohlstandsquelle: Wissen. Man kann das auf innovativen Märkten bereits beobachten. Das Produkt der Zukunft hat einen Intelligenz-Kern und eine Service-Hülle. Und daraus folgt auch: Je wichtiger die Produktivkraft Intelligenz wird, desto mehr konvergieren Wirtschaft und Bildung. So findet man neuerdings Probleme der Wissenschaft ganz analog in der Wirtschaft, z.B. Gaston Bachelards "obstacles épistémologiques" - also die Blockade durch vergangene Erfolge.

Ressource Wissen; Konvergenz von Wirtschaft und Bildung - das heißt konkret: Jetzt wird erst eigentlich die Produktivität der geistigen Arbeit entdeckt. Die Wirtschaft des Unsichtbaren hat es vor allen Dingen auch mit unsichtbaren Kosten zu tun:

- Forschung und Entwicklung
- Lizenzen, Patente
- Marketing, Service.

All das sind Formen des Wissens. Robert B. Reich, der ehemalige Arbeitsminister der USA, spricht in diesem Zusammenhang von symbolanalytischen Dienstleistungen; gemeint ist der Service des Sinns, den Leute bieten, die mit Problemen handeln und Daten manipulieren. Der Job der Info-Elite besteht im Wissensdesign. Und der Begriff Info-Mapping signalisiert in diesem Zusammenhang, dass es heute v.a. darum geht, zu wissen, wo das Wissen ist. Das Zugangsproblem hat sich von den Gütern auf das Wissen verschoben.

Paradoxien des Wissens

In *The Education of Henry Adams* heisst es sehr schön: "The more he was educated, the less he understood." Etwas positiver formuliert: Je mehr man gelernt hat, um so mehr muss man noch lernen. In der Moderne machen wir die enttäuschende Erfahrung, dass die Wissenschaft die Unwissenheit erweitert. Mit den präzisen Worten von Daniel Bell: "More and more we know less and less." Je mehr einige Leute wissen, desto ignoranter wird der Rest. Der Soziologe Niklas Luhmann hat deshalb eine "Berufsrisikobereitschaft bei der Aneignung von Wissen" gefordert. Wer Zukunftssicherheit will, muss hohe Fremdselektion akzeptieren - das Unternehmen, in dem er arbeiten möchte, kann vorschreiben, was er zu lernen hat. Individualität durch Selbstselektion heißt demgegenüber: Unsicherheit auf dem Markt - ich bestimme selbst, was ich lernen und wissen will, riskiere aber damit, mich am Markt vorbei zu qualifizieren.

Diesem Problem ist nicht durch ein Mehr an Information beizukommen - im Gegenteil. Je mehr Information, desto größer die Unsicherheit und desto geringer die Akzeptanz. So zwingt uns die moderne Welt zur Kompensation des steigenden

Nichtwissens durch Vertrauen - und Vertrauen heißt ja, die Information, die man von oder über jemanden hat, zu überziehen. Vor allem vertraut man dem Selbstvertrauen der Experten, also jener wenigen, die in dem Maße mehr wissen, als wir ignoranter werden.

Neben die Nötigung zum Vertrauen in das Wissen anderer tritt die Nötigung zum Black Boxing des eigenen Wissens. Ich ziele mit dieser Formulierung auf eine Unterscheidung zwischen Strukturwissen und Funktionswissen, also zwischen Erkenntnis und Know how. Sich *auf* eine Sache zu verstehen, heißt nämlich noch nicht: eine Sache zu verstehen. Nun gilt für unsere moderne Welt: Es wächst das Wissen, das man nicht versteht und doch benutzen muss. Der Arbeitsteilung entspricht also eine Wissensteilung. Heute gilt das "divide et impera" auch im Bereich des Wissens. Und genau diese Arbeitsteilung des Wissens nennt man Black Boxing. Schon Georg Simmel hat das genau beobachtet: "Die ungeheure Ausdehnung des objektiv vorliegenden Wissensstoffs gestattet, ja erzwingt den Gebrauch von Ausdrücken, die eigentlich wie verschlossene Gefäße von Hand zu Hand gehen".

Dass es heute mehr lebende als tote Wissenschaftler gibt, ist der prägnanteste Ausdruck für den Big Bang des Wissens, der unsere postindustrielle Gesellschaft von allen früheren Gesellschaftsformationen trennt. Und kompensatorisch zur Wissensexplosion hat dann die Einfachheit und Naivität der Weisheit Konjunktur. "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?", fragt T.S.Eliot in seinem Gedicht "The Rock". Diese schön formulierte Naivität macht deutlich, dass sich die Wissensgesellschaft in einer Zeit des Wissensüberdrusses formiert, den man nur aus der Tatsache heraus verstehen kann, dass die Wissenschaft ihren Gegner verloren hat. Und wie immer verwandelt sich der besiegte Feind in eine Hypothek.

Lernen des Lernens

Was heute gilt, kann morgen Schnee von gestern sein. Schüler, Auszubildende und mehr noch die Pädagogen müssen begreifen: Man kann nicht mehr "für das Leben lernen" sondern muss mit Wissenshalbwertzeiten kalkulieren. Deshalb übernimmt der Arbeitsgeber immer häufiger die Rolle des Lehrers. Weil Erfahrung im Sinne von Tradition nicht mehr zählt, wird das Leben ein Prozess permanenten Umlernens. Erwachsenen-bildung heißt im Klartext: das ganze Leben ist Medium der Erziehung.

Im Konzept "lebenslanges Lernen" wird Erziehung sich selbst zum Endzweck. Und dieses lebenslange Lernen wird wieder situationsangepasst und zeitelastisch erfolgen. So wenig wie ich heute wissen kann, was ich morgen wissen muss, kann der Unternehmer wissen, was die Märkte von morgen fordern. Wie für das Lernen gilt

auch für die Produktion der Zukunft: Just in time. Auf allen Märkten - auch auf dem Markt der Bildung - entscheidet die Schnelligkeit des Wandels, die Sensibilität für Zeitdifferenzen.

Vormodern wollte die Pädagogik perfekte Menschen schaffen; modern dann nur noch "gebildete". Die Wohlstandsgesellschaft sprach schon nicht mehr von Bildung, sondern wollte - modernitätskritisch - "kritische" Menschen erziehen. Mit anderen Worten: Die moderne Gesellschaft ersetzte Bildung durch die neue Schlüsselqualifikation der "Kritikfähigkeit". Doch "Kritik" ist dann selbst dogmatisch geworden und hat sich in den Paradoxien der Aufklärung sowie in den Stereotypien der Protestkommunikation verfangen. Seither, postmodern, propagiert unsere Gesellschaft Lernfähigkeit. Doch diese Entwicklung liegt ganz in der Linie des "Bildungsgedankens" selbst. Schon Schleiermacher und Humboldt forderten nämlich das "Lernen des Lernens" - eine autologische Formel also, in der ein Begriff auf sich selbst angewandt wird. Bildung wird damit selbstbezüglich und Lernen auf Dauer gestellt. Es geht also schon seit Schleiermacher um die Erlangung der "Fertigkeit, Fertigkeiten zu erlangen". Was das Lernen beim Lernen von etwas eigentlich lernt ist: sich selbst. Das meint wohl auch Dennetts Definition von Lernen als *Self-Design*.

Was soll man eigentlich lernen: das Wissen oder das Lernen? Heute muss man wohl vor allem das Umlernen lernen, das positive Verlernen. Statt Bildung: das Management von Enttäuschungen im Medium der Lernfähigkeit. Luhmann und Schorr haben das einmal so formuliert: "Zu lernen ist die Dauerbereitschaft, Neuem durch Änderung von bereits gelernten Erwartungsmustern zu begegnen." Also eigentlich: Entlernen.

Es gibt offenbar kein Lernen mehr ohne Verlernen, also Revision, also Enttäuschung. Fritz B. Simon kann deshalb sogar behaupten: "Wissen und Lernen sind Gegensätze. Wo Wissen bewahrt wird, wird Lernen verhindert. [...] Lernen zerstört Wissen, indem es verhindert, dass alte Unterscheidungen weiter vollzogen werden. [...] Wissen macht lernbehindert, Erfolg macht lernbehindert. Ohne Misserfolg - sei er aktuell oder nur für die Zukunft befürchtet - kein Lernbedarf."

Deshalb vertragen sich Lernen und Macht so schlecht. Denn der Wille will sich durchsetzen - also nichts lernen. Aber auch mit der Ethik verträgt sich das postmoderne Lernen schlecht, denn Lernbereitschaft ist das Gegenteil von normativer Erwartung. Und das heißt positiv: Lernen als Umlernen und Entlernen ist ein Training der Unsicherheitsabsorption und Erziehung zur Kontingenz. Alles Lernen verarbeitet nämlich Irritationen, indem es das Störende, Neue in Redundanzen und vertraute Muster einspinnt.

Und hier scheint mir vor allem eine Unterscheidung wichtig, nämlich die zwischen Lehre und Training. Während das Training aus dem Menschen eine triviale Maschine macht, die auf einen bestimmten Input mit einem ebenso bestimmten Output reagiert, dürfte Lehre eigentlich nur heißen, was an nicht-triviale Maschinen gerichtet ist - z.B. offene Fragen. Abfragbares Wissen ist ja kein hinreichendes Kriterium von Intelligenz. Deshalb muss - wie das Lernen - auch die Pädagogik "autologisch", also selbstbezüglich angelegt sein; sie muss das Lehren lehren - und zwar im Unterschied zum Training. Das Lehren vom Dressieren zu unterscheiden, war ja schon ein zentraler Impuls der Aufklärung, und schon Kant hat in diesem Sinne den "Informator" vom Führer unterschieden.

Unglückliche Bildung und das Ärgernis Selektion

"Bildung" war einmal die Formel eines Abwehrzaubers der Pädagogen gegen Politik und Wirtschaft. Doch heute ist das Bildungssystem unglücklich über seine Autonomie; deshalb ruft sein Establishment immer wieder nach "Politik" oder "Ethik". Die Konjunktur der Ethik ist aber nur ein Symptom für die Hilflosigkeit der Pädagogik, die offenbar selbst nicht lernen will. Und fehlende Lernbereitschaft äußert sich meist als moralische Sollensforderung an die Realität. Normen sind ja lernunwillige Erwartungen, die gegen die Tatsachen durchgehalten werden. Und beim Kontakt von Politik und Bildung kommt es nicht zur Inspiration, sondern allenfalls zur Subvention. Staatliche Subventionen verzögern dann das Unvermeidliche. Damit betäuben sie die Betroffenen.

Ein weiteres probates Mittel zur Verdeckung der Bildungsparadoxien ist die Reformdiskussion, genauer gesagt: die Flucht ins Reformprogramm. Die Bildungsreformdiskussion scheint aber selbst die Krankheit zu sein, deren Heilung sie verspricht. Denn was Pädagogen dabei gerne - und zwar mit dem Begriff "Bildung" - verdrängen, sind Selektion und Zwang. Die Notwendigkeit der Selektion ist das Ärgernis der Bildungsanstalten. Denn alle Erziehung ist analog, d.h. sie fordert und beruht auf Allmählichkeit. Selektion dagegen ist digital; letztlich läuft sie auf Feststellungen hinaus wie "bestanden" oder "durchgefallen". So kann man im Rückblick auf das gutgemeinte Projekt der Gesamtschule heute sehen: Politisch gewollt war Differenzierung ohne Selektion - und das konnte nicht gutgehen.

Das Dilemma der Politik ist klar: Einerseits sind die Ansprüche an Ausbildung genau wie die an Gesundheitsvorsorge unbegrenzt. Andererseits heißt jemanden fördern immer: andere zurücksetzen. Man kann nicht werten, ohne jemanden zurückzusetzen. Und deshalb gilt: Selektion schließt Konsens aus. Und deshalb ist die Verlockung fast unwiderstehlich, Selektion als "Schuld" der "Gesellschaft" zuzurechnen. Das war bekanntlich eine Spezialität der 68er.

In den Bildungsanstalten heute scheut man die Information der Selektion - und ruft deshalb nach "Beratung". Doch nüchtern betrachtet sind Bewertungen, Lob, Tadel und Zensuren die einzigen Gewißheitsäquivalente des Bildungssystems. Und jeder Universitätsprofessor (zumindest der Geisteswissenschaften) wird die Beobachtung von Luhmann und Schorr bestätigen können, dass ohne die Digitalität der Selektion die Prüfung "zu einer Prüfung der Prüfer im Hinblick auf ihre Fähigkeit, Unfähigkeit zu tolerieren", entartet.

Vielleicht gibt es für das Bildungssystem hier aber einen Weg zurück in die Zukunft. Niklas Luhmann hat eine interessante Apologie des mittelalterlichen Systems von Trivium (als Kommunikationstheorie) und Quadrivium (als Welterkenntnis) angedeutet. "Es verzichtet auf ein direktes Hineinkopieren von Unter-schieden der sichtbaren, erfahrbaren Welt in den Schulunterricht. Es nutzt die Möglichkeiten der Distanz" - und entlastet den Unterricht von Erziehung. Erst mit Beginn des 19.Jahrhunderts hat sich die Pädagogik dann mit der Idee des erziehenden Unterrichts überfordert. Daraus könnte man lernen, dass Schule und Universität, die sich - in den USA wegen des Rassismus, in Deutschland wegen der Nazis - zu lange als Sozialagentur verstanden, von sozialen Zielen zu Lernzielen zurückfinden müssen.

Doch das Ärgernis der Selektion verstört nicht nur die Pädagogen, sondern auch die Lernenden. "Bildung" wollte ja immer schon "autonome", verantwortliche Menschen erziehen, und Verantwortung heißt im Kontext unserer Fragestellung eben Selbstselektion. Ein gebildetes Individuum hat sich im Prozess der Bildung selbst selektiert - und kann deshalb nicht sicher sein, ob das angeeignete Wissen im Berufsleben gerade nachgefragt wird. Wer dagegen große Zukunftssicherheit will, muss hohe Fremdselektion akzeptieren. Und diese Option unterstützt unser gegenwärtiges Bildungssystem schon dadurch, dass es das Selektions-problem der Wirtschaft überlässt. Doch je unsicherer die berufliche Zukunft für alle wird, desto riskanter ist eine "praxisnahe" Ausbildung. Die an Neulinge adressierte Standardformel ist bekannt: "Jetzt vergiss erst einmal alles, was du an der Uni gelernt hast...". Das klingt zunächst einleuchtend. Der Vorrang der Praxis ist nämlich eine Folge von Überkomplexität; sie zwingt zur Verkürzung, d.h. zum Handeln "als ob" man Sicherheit hätte. Doch der Praxis-Schock fördert gerade nicht Lernfähigkeit. Im Gegenteil. Die Praktiker versteifen sich meist auf die eigenen Erfahrungen. Praxis ist dann oft nur ein Deckbegriff für Begriffslosigkeit. Hier sollte man sich nicht einschüchtern lassen. Praxis ist - gerade auch an Universitäten - der Lieblingsbegriff der Begriffslosen und "Praxisrelevanz" das Wort, mit dem man heute die theoretische Neugierde wieder in ein Laster verwandelt.

Vom Umgang mit Black Boxes

Bildung musste ihre Formen immer schon in zwei ganz unterschiedliche Medien einprägen: einmal in die lose gekoppelten Vorstellungen des Lernenden, sodann in die Daten der technischen Medien. Unter Computer-bedingungen stellt sich heute die Frage nach dem Verhältnis zwischen pädagogischer Technik und Medien-technik mit neuer Schärfe. Medientechnik hat es mit dem "data processing" zu tun - die Pädagogik stellt sich dem Problem des "people processing". Beim people processing gibt es kein spezielles Kommunikations-medium. Das Gelingenskriterium ist hier nicht Anschlusskommunikation, sondern die Veränderung von Menschen, z.B. in der Erziehung oder in der Therapie. Und wir wissen: Bildung und Heilung kann man nicht automatisieren.

Der Pädagoge steht heute zwischen der Black Box des Lernenden und der Black Box des Computers. Der Lernende ist für den Lehrer eine Black Box, denn wer weiß, kann nicht mehr wissen, wie es war, nicht zu wissen. Deshalb muss der Lernende von der Pädagogik immer wieder neu "erfunden" werden. Durchsichtiger als der Lernende ist der Rechner. Der Computer als Black Box fasziniert durch Komplexität. Er kann der Pädagogik als konzeptionelle Brücke vom Einfachen zum Komplexen dienen. Der Computer ist nicht so simpel wie eine Uhr, aber auch nicht so undurch-sichtig wie ein Mensch.

Es ist heute wohl unstrittig, dass man triviale Lernprozesse am besten mit Computern stützt. Doch Computer stellen schon als technische Medien eine lernfreundliche Umgebung dar; sie nehmen einem - Stichwort Benutzerfreundlichkeit - die Angst, etwas falschzumachen. Aaron Wildavsky sagt: "Fear of failure inhibits learning." Und genau da setzt computergestütztes Lernen an. Der Computer als pädagogisches Medium - Stichwort Hypertext - schenkt uns "freedom from fear of failure" (Ben Sneiderman). Doch natürlich fällt es den Bildungsanstalten schwer, daraus Konsequenzen zu ziehen und Abschied von der Mensch-zu-Mensch-Pädagogik zu nehmen. So konnte Reinhard Kahl (Die Zeit, 26.3.98) sarkastisch bemerken: "Langsam wird es Zeit, die Medienphobie der Pädagogen als Berufskrankheit anzuerkennen."

Wissensmanagement

Wir müssen uns mit dem Gedanken vertraut machen, dass es keine vorgegebenen "Bildungswege", kein stabiles Wissen mehr gibt. Daraus ergibt sich als Hauptaufgabe der Bildungspolitik: einen wahrhaft freien und kompetenten Zugang zu den Archiven und Data-Pools zu organisieren. Das Zugangsproblem hat sich von den Gütern auf das Wissen verschoben. Zurecht spricht man in diesem Zusammenhang von

Medien-kompetenz; allerdings hat dieser Begriff in der öffentlichen Diskussion noch keine angemessene Komplexität. Wir müssen die Kunst des Fragens wieder lernen. Uns fehlt es ja nicht an Wissen, sondern wir suchen die Fragen, auf die unser Wissen eine Antwort sein kann.

Kenneth Boulding hat das zentrale Paradoxon des Wissens einmal so formuliert: "We have to know what we want to know before we can start looking for it." Mit anderen Worten, wir suchen die Fragen, auf die unser Wissen eine Antwort sein kann. Und soweit Philosophie sich als die Kunst des Fragens versteht, hat sie in der Wissens-gesellschaft eine Zukunft als Technik der Komplexitätsreduktion und als Metadesign von Wissen. Denn bloßes Operations Research hilft hier nicht weiter. Operations Research optimiert ja lediglich Problem-lösungen. Aber war das Problem richtig gestellt? Hier können Philosophen einhaken - als Theoriedesigner.

Statt Bildung also: Wissensmanagement. Wie konnte es dazu kommen? Der Startschuss des abendländischen Wissenschaftsprozesses fiel mit der listigen Formel des Sokrates: Ich weiß, dass ich nichts weiß. Über zweitausend Jahre später war die entscheidende Frage immer noch die nach den Grenzen des Wissens (gegenüber dem bloßen Meinen) - in Kants Formulierung: Was können wir wissen? Erst der Historismus besann sich dann auf das Wozu des Ganzen: Was wollten wir wissen? Doch unsere postmoderne Situation lässt sich mit diesen Thesen und Fragen nicht mehr fassen. Wir müssten heute - am Gegenpol von Hegels "sich wissendem Wissen" angelangt und Sokrates parodierend - eher sagen: Wir wissen nicht, was wir wissen.

Die Anwendung des Wissens auf Arbeit hat vor hundert Jahren die Produktivität entdeckt. Heute arbeiten wir an der Selbstanwendung des Wissens. Wissen wird auf Wissen angewandt - und hier zeigt sich die Produktivität der geistigen Arbeit. Die eigentliche intellektuelle Leistung der Zukunft liegt also im Wissens-design. Und je wichtiger die Produktivkraft Intelligenz wird, desto mehr konvergieren Wirtschaft und Bildung. Wissen ist die ultimative Ressource der zukünftigen Kultur. Und deshalb wird "Research into Knowledge" nicht mehr nur die Sache des an Erkenntnis interessierten Wissenschaftlers bleiben.

Robert B. Reich, der ehemalige Arbeitsminister der USA, hat in diesem Zusammen-hang von symbol-analytischen Dienstleistungen gesprochen. Gemeint ist der Service des Sinns, den Leute bieten, die mit Problemen handeln und Daten manipulieren. Der Job der Info-Elite besteht im Wissensdesign! Und der Begriff Info-Mapping signalisiert, dass es heute v.a. darum geht, zu wissen, wo das Wissen ist. Die symbolanalytischen Dienstleister bieten in der Welt des Information Overload den Luxus "Sinn".

Es handelt sich hier vor allem um das Problem der "intelligent discrimination": Was wird nicht erforscht? Was kann ich vernachlässigen? Welche Bücher muss ich wirklich lesen? Das wertvollste Wissen ist heute: zu wissen, was man nicht zu wissen braucht. Nützlichkeit ist aber ein anthropozentrischer Begriff, der sich nicht mathematisch formalisieren lässt.

Wir sind davon ausgegangen, dass Bildung in der Postmoderne mit Wissenshalbwertzeiten kalkulieren muss; d.h. man kann nicht mehr für das Leben lernen. Schon Schleiermacher hat eben deshalb das "Lernen des Lernens" ins Zentrum der "Bildung" gerückt. Das ist aber erst heute technisch implementierbar. Hypermedien präsentieren ein Wissen, das sich dem Lernenden anpasst. Sie ermöglichen erstmals einen interaktiven und multimedialen Wissenszugriff. Hypermedien bringen die Simultanpräsentationsleistungen des Bewusstseins inmitten der sequentiellen Informationsverarbeitung von Kommunikation zur Geltung.

"Man könnte Argumente genauso visualisieren wie Daten, indem man alle Standpunkte zu einem Problem als mehrdimensionale Figur darstellt und über Hypertext verbindet.", so Esther Dysons konkrete Utopie gleichsam begehbarer Wissensstrukturen. Heißt das: Bildung multimedial? Humboldts "Kosmos" im Cyberspace? Ich vermute eher: Der Computer ist das Holzpferd der Griechen im Troja der Bildung.

Neue Universitätsmythen

Es gibt immer noch Professoren und Lehrer, die mit einer Art trotzigem Stolz darauf bestehen, von Computern keine Ahnung zu haben. Das kann als Kultmarketing des Geisteswissenschaftlers durchaus funktionieren: Er stilisiert seine technische Inkompetenz als philosophische Besonnenheit. Gegen Internet und Cyberspace bringt er Einsamkeit und Freiheit in Stellung. Und diese Attitüde ist in einer Gesellschaft, die wieder nach dem "Sinn" sucht, durchaus attraktiv. Dienst am Subjekt - das bietet das Serviceunternehmen "Geisteswissenschaften". Doch die heroische Nachdenklichkeit als Pfeiler im Datenstrom hat die Zeitlogik der modernen Gesellschaft gegen sich.

Das ist rasch erklärt: Das einsame Nachdenken eilt nicht. Deshalb verliert es immer mehr an gesellschaftlichem Wert. Was nicht unbedingt jetzt gemacht werden muss, wird zurückgestellt. Und das heißt letztlich: Was nicht dringlich ist, disqualifiziert sich selbst. Dagegen führt alles Dringliche eine Wertvermutung mit sich. Deshalb spricht alle Welt von Teamgeist und Vernetzung. Denn Kooperation impliziert Terminierung, diese erzeugt Dringlichkeit - und diese impliziert eben Wichtigkeit. Die terminbestimmte Zeitstückelung verunmöglicht Nachdenklichkeit. Gedacht wird nur noch, was in bestimmten Fristen zu Ende gedacht werden kann.

Seit die Pathosformel von Einsamkeit und Freiheit nur noch für das Marketing der Geisteswissenschaften taugt, schreibt die Universität denn auch an ganz neuen Mythen:

1. Praxisnähe, also Fremdselektion. Und das heißt im Klartext: Andere (vor allem natürlich aus der Wirtschaft) entscheiden, was wissenswert ist.
2. Teamgeist - statt Einsamkeit und Freiheit. Hier entfaltet sich der sanfte Wahn, irgendeine mysteriöse Eigenschaft der "Gruppe" könne beim Denken helfen oder "motivieren". Ich denke, Gegenindikationen wären leichter zu erbringen.
3. Betreuung, also Mensch-zu-Mensch-Pädagogik. Studenten und Politiker scheinen sich einig, dass es die Professoren an Beratung, Betreuung und persönlicher Zuwendung mangeln lassen. Das ist die wohl unausrottbare "Der Mensch im Mittelpunkt"-Ideologie, mit der man Probleme der Technik, Selektion und Finanzierung unsichtbar macht.
4. Dienstleistung statt people processing.

Dieser letzte, jüngste Universitätsmythos verdient besondere Aufmerksamkeit. Der Lehrberuf wird heute ganz selbstverständlich als symbolanalytische Dienstleistung begriffen. Und rollenkomplementär dazu versteht sich der Student als König Kunde. So halten die zentralen Marktmaximen der Benutzerfreundlichkeit und Kundenorientierung Einzug in die Universität. Doch hinter diesen schönen Formeln verbirgt sich wieder eine handfeste Paradoxie: Eine Routine soll als Nicht-Routine erscheinen. Der Professor soll sich zum Studenten verhalten wie der Arzt zum Patienten und der Pfarrer zum Sünder. Wie der studentische Wunsch nach "Betreuung" nährt das Selbstverständnis des Professors als Dienstleister die Illusion der persönlichen Zuwendung - als ob es keinen Zeitdruck gäbe; als ob es ein "Eingehen" auf den anderen geben könnte.

Inkompetente honorieren eher Performanz als Kompetenz. Deshalb sind Professoren beliebt, deren Bürotür offen steht und um die der Duft frisch gebrühten Kaffees ist - körperlich präsent und stets zu einem Gespräch bereit. Doch die Universität ist keine große Familie. Menschenfreundlichkeit macht hier die Probleme unbenennbar - und letztlich unsichtbar. Und eben deshalb ist ständig von Praxisnähe, Teamgeist, Betreuung und Service die Rede. All diese neuen Mythen verdunkeln das Selektionsproblem.

The Navigator of Ignorance

In der Welt von Forschung und Lehre gibt es weder Technologie noch Erfolgskriterium. Man weiß nicht, warum nicht mehr Wahrheiten anfallen. Und man weiß auch

nicht, warum Studenten keinen Bock haben. Ersatzrationalisierungen lauten dann: kein Geld, zu große Seminare, faule Professoren. Vor allem der periodisch wiederkehrende Vorwurf, Professoren seien faul, hätten fünf Monate Urlaub und müssten nur acht Stunden in der Woche arbeiten, macht ein Grundproblem intellektueller Arbeit deutlich: sie ist weitgehend unsichtbar. Und gerade deshalb dreht sich in der akademischen Öffentlichkeit alles um Publikationsliste, Zitationsindex und Reputation. Man muss die eigene Rolle wirkungsvoll dramatisieren, um den sozialen Rang und die Unkosten der eigenen Leistung sichtbar zu machen. Vor allem Professoren, die eben heute als intellektuelle Dienstleister angesehen werden, haben das Grundproblem, dass der Kunde, also die Studenten, aber auch die Beobachter (Politiker und Journalisten) die laufenden Kosten des akademischen "Service" nicht sehen können. Man muss also das, was man leistet, zusätzlich vorführen, dramatisieren. Ich werde zitiert, also bin ich.

Nach dem zweiten Weltkrieg war vor allem ein akademischer Selbstdramatisierungsstil erfolgreich: Man war "kritisches Bewusstsein", das in Studenten und Gesellschaft kritisches Bewusstsein "produzierte". Und das war durchaus eine Folgelast des Bildungsgedankens. Denn seit Parsons kennt man das Problem: Je mehr "Bildung", desto stärker stehen die Menschen im Bann der unrealistischen Interaktionstypik des Unterrichts. Die Universität mit ihrem psychosozialen Moratorium, also der stabilen Möglichkeit, das Erwachsenwerden zu verweigern, war der Ort der unschuldigen Beobachtung von außen - gegen Vater und Staat. An diesem archimedischen Ort konnte man "Entlarvung" trainieren. Die unkritische Selbstbezeichnung als kritisches Bewusstsein verwandelte "Kritik" in ein Ornament der Jugendkultur und die wissenschaftliche Methode in ein Initiationsritual: "Welchen Ansatz hast du?", fragte man damals. Und gemeint war:

Welcher Sekte gehörst du an?

Doch wie könnte es anders sein? Ich denke, es würde sich lohnen, einmal über Douglas Hagues Konzept einer neuen Universität nachzudenken. Dort gäbe es:

1. Star-Akademiker, die sich ganz auf die Forschung konzentrieren, aber allen Universitäten als Vortragende zur Verfügung stehen;
2. Medienberater, die für die jeweiligen Lehrinhalte und Lernprozesse die angemessene (heute natürlich: multimedial) technische Implementierung sicherstellen;
3. akademische Impressarios, deren Kompetenz in der Umsetzung von Forschungsergebnissen in lehrbares Wissen besteht;
4. "educational consultants", die Studenten in allen Studienfragen beraten.

Über den Reflexionsstil, den eine solche Universität fördern würde, kann man natürlich nur Mutmaßungen anstellen. Schon organisatorisch stellt sie eine narzisstische Kränkung der Einen Vernunft dar. Auch die Universität muss lernen, dass es in einer hochkomplexen Gesellschaft nur arbeitsteilige Rationalität gibt. Das könnte zu einer Kultur der ironischen Vernunft führen. Ihr "Geist" wäre bestimmt von souveränem Eklektizismus und organisierter Ignoranz. Ich meine das im Sinne von Henry Adams, der sich am Ende seines Bildungsweges als "the navigator of ignorance" beschrieb.

Print im Medienmix

Und das Buch? Und das Lesen? Und die abendländische Kultur? Vor allem die deutsche Mediendiskussion geht von der eigentümlichen Voraussetzung aus, dass hinter einer Zeitung immer ein kluger Kopf, vor dem Bildschirm aber immer ein Dummkopf sitzt. Ganz selbstverständlich unterstellt man den Printmedien und der Kulturtechnik Lesen eine Affinität zu "Kultur" und Intelligenz. Deshalb scheint das Abendland bedroht, wenn Statistiken nachweisen, dass die Deutschen weniger lesen und sich stattdessen von den neuen Medien faszinieren lassen. Doch wer sich von derartigem Kulturpessimismus nicht kopfscheu machen lässt, kann etwas ganz anderes vermuten: Für Print als Medium stellen die neuen Medien keine tödliche Gefahr dar, zwingen es aber zur Neupositionierung, zur Besinnung auf die eigenen, printspezifischen Stärken. Printprodukte haben unersetzbare Materialqualitäten, die man optimieren kann: bequem zu handhaben, gut zu lesen, rascher Überblick, Tastbarkeit, man kann sie wegwerfen. Das werden die Zeitung und Zeitschrift der Online-Welt immer voraushaben - und lässt sie überleben. Dem "kritischen Bewusstsein" der "Aufklärung", das sich in der ökologischen Nische von Buch und Zeitung so erfolgreich etabliert hatte, läutet allerdings das Totenglöckchen.

Das Internet ist ein sanfter Imperialist. Sich ihm zu entziehen, wäre tödlich. Ich bin versucht, zu sagen: Man kann nicht nicht online gehen. Bei Zeitungen und Zeitschriften kommt hinzu, dass sie ihre neue Funktion im Medienmix selbst bestimmen sollten - im Idealfall als Kommunikationskontinuum zwischen traditioneller Hard Copy und Online-Auftritt.

Und natürlich im Blick auf die eigenartigen Kulturtechniken der Jugendlichen. Navigation, Surfen und Channel-Hopping sind neue "Lesegewohnheiten", die sich vom traditionellen, aufmerksamen, linearen Lesen radikal unterscheiden. Die Rezeptionsweise der Jugendlichen ist zerstreut, mehrdimensional, mosaikartig, folgt dem Lustprinzip und steht unter Zeitdruck. Aufmerksamkeit ist ihre knappste Ressource.

Das Überleben der Printmedien ist auch dadurch gesichert, dass die neuen Medien die alten brauchen, um zu sich zu kommen. Die Geschichte zeigt, dass neue Medien zunächst einmal alte Medienangebote wiederholen müssen, bis sie zu ihren eigentlichen, medienspezifischen Inhalten durchdringen. Das ist eine Frage der Evolution - und darum kaum zu beschleunigen. Der Inhalt eines Mediums ist immer ein anderes Medium, wusste schon Marshall McLuhan. Insofern ist es ganz sinnvoll, im uralten Medium über das brandneue zu berichten - so hat jede Zeitung heute ihre Medien- oder Software-Seite. Allerdings setzt das voraus, dass die Autoren nicht nur informativ, sondern vor allem auch gut schreiben können - und da hapert es in Deutschland. *Konra@d* etwa ist eine schöne Zeitschrift, der man Glück wünschen möchte. Aber es fehlt ihr doch die Kultqualität, die Mondo 2000 und Wired zum Faszinosum der Cyber-Freaks hat werden lassen.

Zukunftsmärkte in der Bleiwüste

Der Erfolg von Zeitungen wie *USA Today* oder Nachrichtenmagazinen wie *Focus* hat gezeigt, dass wir in eine neue Ära der Printmedien eingetreten sind. Der Vormarsch der bunten Bilder ist nicht mehr aufzuhalten. Und wer seit Jahren nicht mehr die *Times* gelesen hat, wird überrascht sein, welcher Blumenstrauß von Bildern und bunten Grafiken hier kredenzt wird. Besonders lehrreich war natürlich, wie der asketische *Spiegel* durch die Markteinführung des *Focus* (der ja weniger Fakten, Fakten, Fakten als vielmehr Bilder, Bilder, Bilder bietet) unter Anpassungsdruck geriet. Kein Zweifel: Der visuellen Kommunikation gehört die Zukunft. Und dennoch: Die klassischen Printmedien haben auch eine Zukunft auf dem Markt der Medien. Hier sind die Gründe.

Printmedien haben den Bonus der Glaubwürdigkeit - das kann man nur historisch verstehen. "Bleiwüste" lautet das böse, aber genaue Wort derer, die den klassischen Medien der Aufklärung Bilderfeindlichkeit vorwerfen. Der Vorwurf ist berechtigt, denn in der Tat haben die Geister der Aufklärung und Gesellschafts-kritik in den Bildern immer nur Verblendung, Schein und Ablenkung gesehen. Gegen das magische Bild stand der Text als Werkzeug der Analyse - das ist noch heute das Pathos "seriöser" Zeitungen wie *Le Monde* oder *FAZ.* Ein Foto der AEG-Werke sagt nichts über die Wirklichkeit der AEG-Werke, lautete ein kritisches Bonmot von Bertold Brecht. Und in der Tat konnte für Aufklärer nur das Schwarz auf Weiß des Gedruckten zum Medium der Wahrheit werden. Aus dieser Zeit stammt noch der Bonus der Glaubwürdigkeit, den bilderlose Texte genießen. Auch wer nicht (mehr) liest, vermutet die Wahrheit im Gedruckten.

Aber es gibt noch einen viel massiveren Grund, warum traditionelle Printmedien gerade auch in der Multimedia-Gesellschaft optimistisch in die Zukunft blicken können. Heute ist es zwar technisch möglich, Zeitungen auf Individuen zuzuschneiden, also maßgeschneiderte Special-Interest-Magazine zu vertreiben.

Doch das setzt voraus, dass die Leser nur an den eigenen Interessen interessiert sind und dass sie sie zu benennen wissen. Aber das ist die Ausnahme. Im allgemeinen weiß ich nämlich nicht, was ich wissen will. Und diese Situation wird mit dem Anschwellen der Datenflut immer dramatischer. Wir haben kein Informationsproblem, sondern ein Orientierungsproblem. Was wir brauchen, ist eine tägliche Arche Noah in der Sintflut des Sinns - und genau das ist die klassische Zeitung. Anders gesagt: Der Zeitungsredakteur ist der Vorreiter der zukünftigen Wissensarbeiter. Das sind menschliche Informationsprozessoren; ihre spezifische Dienstleistung ist das Infomapping - sie wissen, wo das Wissen ist. Informationen allein helfen uns bei Problemen nämlich nicht weiter. Sie müssen erst gefiltert, konfiguriert und strukturiert werden. Um Information intelligent zu machen, braucht man eben Wissensdesigner, Redakteure, Journalisten. Wie der Wissenschaftler, der Regisseur, der Marketing-Experte, der Finanzberater oder der Dichter gehört der Zeitungsmensch zu jenen Leuten, die mit Problemen handeln und Daten manipulieren. Robert B. Reich nennt sie "symbolanalytische Dienstleistunger". Sie alle handeln mit Sinn und verkaufen Orientierung.

Die gegenwärtige Diskussion über die Zukunft der Printmedien steckt deshalb in einer Sackgasse, weil sie sich von einer technischen Möglichkeit computergestützter Kommunikation faszinieren und blenden lässt: der Interaktivität. Auch Interaktivität ist aber eine Aktivität, und nichts spricht psychologisch oder soziologisch dafür, dass Menschen immer aktiv sein wollen. Weder strebt der Mensch von Natur aus nach Wissen, wie uns Aristoteles einreden wollte, noch braucht der Mensch ständig information at his fingertips, wie uns Bill Gates einreden will. Ähnlich wie das vollständig passive Medium Fernsehen ist auch das Zeitunglesen ein Ritual. Man braucht die Zeitung wie den Morgenkaffee. Man sollte sich also vom Phantom des (inter-)aktiven Konsumenten, das die Software-Industrie beschwört, nicht den Blick für die trivialen Realitäten des Alltags verstellen lassen. Zeitung lesen ist nur in den seltensten Fällen Information Retrieval, fast immer aber: lustvolles Blättern; man lässt sich von den Neuigkeiten aus aller Welt berieseln.

Trivial aber entscheidend ist auch, dass es unterschiedliche Bequemlichkeitsstufen des Lesens gibt. Und natürlich ist auf unabsehbare Zeit das Lesen am Bildschirm ein ergonomisches Ärgernis. Die Love Story gehört ins Taschenbuch, die Hintergrundreportage ins Nachrichtenmagazin, der Sportbericht in die Zeitung. Diese Medien sind billig und handlich. Mit einem Griff entledige ich mich der Kleinanzeigen oder des Feuilletons; den *Focus* lasse ich im Intercity liegen; das Taschenbuch erträgt

klaglos den Sand am Strand. Das flüchtige Blättern ermöglicht im übrigen auch problemlos den Effekt, der im Internet als Surfen gepriesen wird - dass man nämlich zufällig auf Interessantes trifft. Serendipity nennen das die Amerikaner. Wie gesagt: Ich weiß nicht, was ich wissen will. Und was mich nicht interessiert, landet im Papierkorb. Nichts ist deshalb ferner als die Utopie der Maus-Erfinders Douglas Engelbart: a life without hard copy. Gerade um den Übergang in die neue, digitale Welt ertragen zu können, brauchen wir Vertrautes aus der alten Welt, das wir in die neue hinüberretten. Darin liegt die Zukunft des gedruckten Wortes.

Zeit zum Lesen

Das Land, das die größte Buchmesse der Welt ausrichtet, meint seit Jahren Grund zu der Klage zu haben, seine Schriftsteller leisteten keinen Beitrag mehr zur Weltliteratur. Vielleicht haben uns die Goethe-Institute, die Pina Bausch, Günther Grass und Jürgen Habermas als Exportschlager der deutschen Kultur um die Welt schickten, aber schon seit Jahrzehnten in eine wohltätige Illusion gehüllt - dass man sich nämlich auch außerhalb Mitteleuropas für deutsches Denken und Dichten interessiere. In Zeiten knapper Kassen fallen dann die Hüllen. Was die Leute in Sofia und Litauen an deutscher Kultur interessiert, ist nicht Sprachkunst sondern Wirtschaftsdeutsch. Wir sind als Handelspartner attraktiv, nicht als Kulturnation.

Martin Heidegger durfte noch glauben, dass man, um zu verstehen, was die Welt im Innersten zusammenhält, Griechisch und Deutsch können muss. Heute würden wir wohl sagen: Englisch und Computerchinesisch. Und wenn die Amerikaner, die die Weltsprache Englisch durchgesetzt haben, heute verstärkt Spanisch lernen, dann nicht, um Cervantes im Original zu lesen, sondern um die hispanischen Eigen-kulturen der amerika-nischen Metropolen in Schach zu halten und das Idiom der *emerging markets* zwischen Mexiko City und Buenos Aires zu verstehen.

Die legitimste Frage eines Lesers lautet: Welche Texte machen Lust, welche Autoren gönnen mir Spaß am Lesen? Doch Lust wird immer mehr zum Zeitproblem. Ich bin ein Berufsleser - wie wohl auch einige Leser dieser Zeilen. Wir kultivieren einen Ausnahmezustand. Und das macht uns betriebsblind. Normal ist aber, dass Menschen mit Büchern gerade dann konfrontiert werden, wenn sie erschöpft sind. Am Abend im Bett, am Wochenende nach dem Rasenmähen, zwischen brüllenden Kindern. Und man wird fragen dürfen: Ist das, was unter solchen Umständen möglich ist, überhaupt noch lesen? Wer hat noch Zeit zum Lesen, wer kann noch lesen? Hinzu kommt das Problem: Wer kann noch mithalten mit dem, was angesagt ist? Jochen Hörisch hat dieses Problem schon vor Jahren in einem Band über das schnelle Altern der neuesten Literatur exponiert. Zentral steht dabei die Erfahrung, dass, was veraltet ist, nicht alt sein muss. Die Beschleunigung des Veraltens heißt

für unsereins vor allem: Man sieht mit den eigenen literarischen Erfahrungen schnell alt aus!

Uninteressant für unser Thema sind die ungelesenen Bücher, die aus Prestigegründen in den Ikea-Regalen gesammelt werden. Viel aufschlussreicher sind die Bücher, deren erste Seiten gelesen wurden - aber eben: nur die ersten Seiten. Der Leser war zwar geneigt, aber dann musste er doch vor der Arbeit des Lesens kapitulieren. Der Mann ohne Eigenschaften, Johnsons Jahrestage, von Ulysses ganz zu schweigen - einmal aufgebrochen und dann für alle Zeiten geschlossen. So schmücken sie als weithin sichtbare Zeichen kultureller Resignation unsere Wohnzimmer. Deshalb können uns die Statistiken der Marktforschung und die Verkaufszahlen der Verlage nichts über die Lage der deutschen Literatur sagen. Man kann zwar feststellen, was gekauft wird, aber nicht, was gelesen wird. Mit dem Kauf signalisiert man vor allem die kulturelle Wertschätzung des Mediums Buch. Man denke auch an die Bücherstapel auf den Schreibtischen der Berufsleser; was man eigentlich alles lesen müsste - aber natürlich dann doch nicht liest. Die *Niemandsbucht* von Handke zum Beispiel.

Was wir alle also dringend brauchen, ist Orientierungshilfe. Und hier spielt jene Marktforschung nun doch eine entscheidende Rolle. Denn wenn niemand weiß, was wirklich zählt, zählen die Zahlen. Das ist zutiefst demokratisch. Die Bestsellerlisten sind gerade auch für denjenigen wichtig, der gar nicht liest. Er weiß dann immerhin, was man lesen müsste. Man kann das entsprechende Buch mit großer Verhaltenssicherheit kaufen und verschenken. *Das Parfüm* etwa, oder *Sofies Welt*. Kurzum: Bestseller entlasten die Urteilskraft. Und wir können an dieser Stelle schon festhalten, dass die heutige Klage über den Niedergang der deutschen Literatur im Klartext lautet: Die Bestseller schreiben die anderen.

Autoren und Kritiker als Kultmarken

Dass es Bestseller und Buchmessen gibt; dass wir jeden Herbst mit der Überraschung des literarischen Herbstes rechnen dürfen; dass seriöse Feuilletons von Kultbüchern berichten; dass es alle Jahre wieder zur Peinlichkeit des Literaturnobelpreises kommt; dass Kulturfunktionäre humanistische Literaturpolitik zu betreiben glauben, indem sie dem durch nichts mehr zu erschütternden Publikum Lyrik aus den Anden verschreiben - das alles sind deutliche Hinweise darauf, dass Literatur nicht auf dem Schauplatz der Kunst erscheint.

Deutlich erkennbar sind vielmehr die Grundlinien dessen, was man literarisches Kultmarketing nennen könnte, nämlich:

- Autorschaft ist Marketing;
- Bestsellerautoren sind Kultmarken.

Für literarische Kultmarken gilt wie für andere Marken auch, dass sie kaufbare Sicherheit präsentieren. Man kann nicht falsch liegen, wenn man einen Grünbein oder Goetz kauft. Und wie andere Kultmarken auch bilden Bestsellerautoren um sich herum Kommunikationswelten - vom sensiblen Hintergrundinterview in der *Zeit* bis zum skandalösen Auftritt in der Talkshow.

Nun wird man sich fragen müssen, wie ein Schriftsteller heute Kultstatus erreicht, wenn denn zutrifft, dass die Leute nicht mehr lesen können? Wenn die gesellschaftliche Anerkennung als Kulturheros nicht mehr über das Lesen läuft, muss man - das legt die Logik der Massenmedien nahe - provozieren. Mit einem *Bocksgesang* (Botho Strauss) zum Beispiel oder einem *Lob der Serben* (Peter Handke). Nun haben aber Provokationen in der Postmoderne das prinzipielle Problem, durchgespielt zu sein, also nicht mehr zu provozieren. Deshalb sind die gerade genannten Texte für Handke und Strauß naiv provokativ. Man könnte allenfalls sagen, dass sie das Bedeutsamkeitspotential des Reaktionären ausschöpfen. Wahrhaft zeitgemäß ist demgegenüber ihre Selbststilisierung: der Autor als Held der selbstgewählten Einsamkeit. Von der Publikumsbeschimpfung führte ihr Weg zur kultivierten Öffentlichkeitsscheu. Handke und Strauß sind die Medienpräparate des "Unzeitgemäßen" - und nichts ist in einem Kultursystem, das die Massenflucht vor dem Mainstream inszeniert, zeitgemäßer.

Wer schreibt, möchte einen Bestseller schreiben, und die meisten, fast alle, müssen sich dann fragen, warum es nicht geklappt hat. Man kann sich zurecht damit trösten, dass der Literaturmarkt Qualität und Erfolg entkoppelt hat - obwohl natürlich Erfolglosigkeit auch hier kein Qualitätssiegel ist. Doch wer besiegelt den Erfolg? Wer herrscht im Reich der deutschen Literatur? Um einen prägnanten Satz Carl Schmitts zu missbrauchen: Souverän ist, wer über den Bestseller entscheidet. Und das ist zum Beispiel Marcel Reich-Ranicki, der Kritiker als Star. Und wenn Kritik in einem parasitären Verhältnis zur Literatur steht, dann muss man vom sichtbaren Wohlbefinden des Kritikers auf einen guten Gesundheitszustand der deutschen Literatur schließen dürfen.

Die kulturelle Funktion des Großkritikers besteht schlicht darin, gangbare Wege im Dschungel der Buchmessen, Frühjahrskataloge und Literaturbeilagen zu bahnen. In der Sintflut des Sinns, den die deutsche Literatur über uns ergießt, zeigt uns der Großkritiker die rettende Arche Noah: Das ist Prosa! Das literarische Urteil fasziniert, weil es unvorhersehbar ist und mit einem Zauberschlag Ordnung schafft. Hilflos wären wir geneigten, aber überforderten Leser vor der Lyrik Ulla Hahns ohne das Orakel aus dem Fernsehen.

Das Werturteil ist die argumentative Form einer Illusion, und mit ihr zaubert der Dezisionismus des Kritikers: Das ist ein schlechtes Gedicht! Es geht hier nicht um Kompetenz, sondern um Performanz. So gewinnt die wunderbare Inszenierung des Literarischen Quartetts ihren dezisionistischen Reiz vor allem auch daraus, dass mit Frau Löffler und Herrn Karaseck abwägende Leser mit in der Runde sitzen, die die erstrebte Klarheit des Urteils kunstvoll unterdrücken - bis Reich-Ranicki zum apodiktischen Rundumschlag ausholt. Hier kann man in Sekundenschnelle berühmt werden. Und auch wer durchfällt, trägt seinen Teil zum Gelingen der großen Literaturshow bei. Wie Kandidaten für Gameshows rekrutiert das Fernsehen deutsche Dichter als Crash Test Dummies fürs Literarische Quartett.

Das funktioniert offenbar deshalb so reibungslos, weil Kritik wichtiger ist als Literatur. So sind die Rezensionen oft besser als die Bücher, die sie besprechen. Und so war es auch schon in der Frühromantik oder zu Walter Benjamins seligen Zeiten. Um es auf eine einfache Formel zu bringen: Literatur ist der Eigenwert der Literaturkritik.

Dem Literarischen Quartett verdanken wir eine Art Reindarstellung dieser Autopoiesis der Literaturkritik. Das Werk dient hier nur noch als Auslöseereignis, zur erfrischenden Irritation einer völlig selbstbezüglichen Rede. Bei Gelegenheit von... wird die Entscheidung des Kritikers zum Ereignis. Reich-Ranickis Runde verwirklicht das unendliche Gespräch des Novalis. Und auch ein zweites Grundelement des literarischen Kultmarketing ist frühromantischen Ursprungs. Was Friedrich Schlegel "objektive Willkür" des Kunstwerks genannt hat, ist jetzt das Betriebsgeheimnis der Literaturkritik: beliebiger Anfang und Selbstkonditionierung. In den Massenmedien verschreibt der Großkritiker Literatur. Und am nächsten Morgen geht man in die Buchhandlung wie in eine Apotheke, um das einzig wirksame Medikament gegen "Sinnlosigkeit" zu kaufen.

Natürlich gibt es Menschen, die verstehen, was sie lesen. Doch hohe Literatur wirkt auch, wenn man nicht versteht - zumindest beruhigend. Man kann weitgehend Verstehen durch Kaufen ersetzen und sich die Namen merken.

Wie andere Formen der Kunst auch ist hohe Literatur ein Placebo, das bekanntlich auch dann wirkt, wenn man weiß, dass es sich um ein Placebo handelt. Erkennbar ist sie für einen nichtprofessionellen Leser daran, dass sie keinen Spaß macht - wer hätte je seinen Spaß an *Finnegan's Wake* gehabt? - und dass das Lesen Arbeitscharakter annimmt. Das lässt sich natürlich auch einfacher sagen: Dass man sich langweilt, ist eine Kultqualität moderner Literatur. Wir schließen dann von Unverständlichkeit auf Tiefe und von Langeweile auf Bedeutsamkeit.

Die Klagen über den Niedergang der deutschen Literatur sind nur das Symptom dafür, dass wir in eine neue Etappe der Medienevolution eingetreten sind. Den massivsten Beleg für diese These sehe ich darin, dass die Klagen über den Verfall als Medienereignis inszeniert werden. Und über den Tod des Romans kann man viel aufregenderes schreiben und sagen, als über untote Romane. Vor allem wird es dann möglich, die Totsagung unter Hinweis auf die putzmunteren Autoren des eigenen Verlags zu leugnen. Nicht die Romane belegen dann diese Gegenthese, sondern der Auftritt ihrer Autoren in einem anderen Medium.

Gerade auch für den Autor gilt: Medienpräsenz ersetzt den Ruhm. Die Massenmedien ermöglichen dem Schriftsteller den Auftritt als "unerschrockener Intellektueller" (Grass) oder als Dichter gegen den Strom der Zeit (Strauß), als Selbst- (Sloterdijk) oder doch wenigstens als Querdenker (Franz Alt). Und gerade dort geben sie oft ihr Bestes. Nicht nur *pro domo* sondern auch zurecht hat Marcel Reich-Ranicki das Fernsehen als rhetorische Schule der prägnanten Formulierung gewürdigt.

Das Medium ist die Botschaft, und der Inhalt eines Mediums ist immer ein anderes Medium - diese Kernsätze Marshall McLuhans finden hier eine einfache Anwendung. Literatur im Fernsehen ist nur die auffälligste Form eines neuen kulturindustriellen Angebots: Literatur im Medienverbund. Und gerade deshalb - nämlich um das eben dadurch provozierte *reproduction antique feel* zu bedienen - kann man dann auch die *Geschichte des Bleistifts* (natürlich wieder: Peter Handke) schreiben.

Wenn es überhaupt einmal eine bürgerliche Öffentlichkeit mit dem Anspruch gesellschaftlicher Allgemeinheit gab, so ist sie unter Bedingungen des neuen Medienverbunds längst in "kulturelle Kasten" (Enzensberger) zerfallen. Das ermöglicht ein Parallelprozessieren von Hochkultur, Popkultur und dem unterhaltsamen White Trash zwischen Glückspirale und Musikantenstadl. Die Hochkultur hält sich dabei durch eine geschickte Ausbeutung unseres schlechten Gewissens am Leben. Hochkultur ist das, was mich eigentlich interessieren sollte (second order desire); und als eine Art Ablass für die Sünden der Trivialität zahle ich dann gerne die Steuern, mit denen die Schillertheater und Goethe-Institute dieser Welt subventioniert werden.

Auftakt

Open Access in der Deutschen Medizin - das Projekt "German Medical Science"

Ludwig Richter, Köln

Abstract

Die hochwertigen Forschungsergebnisse der deutschen medizinischen Wissenschaft werden – da oft in deutschsprachigen Fachzeitschriften veröffentlicht – international bislang deutlich zu wenig wahrgenommen. Um diesen Sachverhalt zu ändern, haben sich die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF), das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI) und die Deutsche Zentralbibliothek für Medizin (ZB MED) in dem DFG-geförderten Projekt „German Medical Science„ zusammengeschlossen. Hauptziel des ambitionierten Vorhabens ist die Gründung von wissenschaftlich hochrangigen E-Journals (www.egms.de), in dem die Forschungsergebnisse der deutschen Medizin in ihrer ganzen Bandbreite präsentiert werden.

Mit German Medical Science ist am 1. Juli 2003 ein alternatives interdisziplinäres Publikationsportal eröffnet worden, in dem auf einer öffentlichen, kostenlos zugänglichen sowie dauerhaft archivierten und damit zitierfähigen Basis Publikationen unterschiedlichster Art den Forschern im Volltext zur Verfügung stehen. Den Erfolg belegen die monatlichen Nutzungszahlen von ca. 135.000 eindeutig. Neben einem weiteren Ausbau liegt der Schwerpunkt im Jahr 2005 vor allem auf der Entwicklung eines tragfähigen Geschäftsmodells in Zusammenarbeit mit den 155.000 in der AWMF zusammengeschlossenen Fachwissenschaftlern.

Das Projekt umfasst grundsätzlich die folgenden Publikationsbereiche:

1. German Medical Science als elektronische Zeitschrift publiziert hochrangige interdisziplinäre Original- und Übersichtsarbeiten mit Peer-Review aus dem Gesamtspektrum der Medizin in einer interdisziplinären Ebene,
2. German Medical Science publiziert in einer weiteren Ebene elektronische Journale einzelner Fachgesellschaften unter deren eigenem Titel mit wissenschaftlichen Original- und Übersichtsarbeiten aus dem jeweiligen Fachgebiet.

Die AWMF und ihre Fachgesellschaften stellen denn Editor-in-Chief und das Editorial Board sowie Gutachter für das Peer-Review-Verfahren; bei der ZB MED sind Redaktion, Marketing und Standardisierungsmaßnahmen, beim DIMDI die Leitung der Softwareentwicklung, die technische Implementierung sowie Betrieb und Archivierung der Journale angesiedelt.

German Medical Science wird getragen von der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF), dem Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) und der Deutschen Zentralbib

liothek für Medizin (ZB MED). Zielgruppe von German Medical Science sind alle professionell Tätigen aus allen Bereichen der medizinischen Wissenschaft, Forschung und Versorgung.
German Medical Science ist nicht nur als elektronische Zeitschrift, sondern auch als Portal der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) konzipiert, in der 153 Fachgesellschaften mit 151.000 Mitgliedern zusammengeschlossen sind. In der Endausbaustufe umfasst gms vier Publikationsbereiche: gms selbst publiziert hochrangige interdisziplinäre Original- und Übersichtsarbeiten aus dem Gesamtspektrum der Medizin in einer interdisziplinären Ebene. Zudem werden elektronische Journale einzelner Fachgesellschaften unter deren eigenem Titel mit wissenschaftlichen Originalarbeiten aus dem jeweils spezifischen Fachgebiet veröffentlicht. Ein dritter Bereich dient zur öffentlichen und internen Kommunikation der Fachgesellschaften, aber auch zur Veröffentlichung von Kongressabstracts u.ä., in einem vierten werden komplexe Forschungsprimärdaten zur Verfügung gestellt.

Die AWMF stellt das Editorial Board und das Scientific Committee für das elektronische Journal gms und organisiert die Begutachtung der eingereichten Artikel in einem „Peer-Review“ . Das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI) ist der zweite Kooperationspartner bei German Medical Science. Das Deutsche Institut für Medizinische Dokumentation und Information ist eine nachgeordnete Behörde des Bundesministeriums für Gesundheit und Soziale Sicherung. Zu seinem Aufgabenbereich gehört es, der fachlich interessierten Öffentlichkeit aktuelle Informationen aus dem gesamten Gebiet der Medizin einfach und schnell zugänglich zu machen. Im Rahmen von gms ist beim DIMDI die Leitung der Softwareentwicklung, die technische Implementierung sowie Betrieb und Archivierung der Journale angesiedelt. Praktisch bedeutet dies, das im DIMDI neben den gms – Datenbanken das gesamte Publikationssystem von gms liegt. Die ZB MED , die grösste Medizinbibliothek in Europa, betreibt die gesamte Redaktion von gms. Das bedeutet, dass sie alle eingehenden Artikel und Kongressabstracts registriert, an den Editor-in-chief und die Peer-Reviewer weiterleitet und alle Korrespondenzen, Korrekturvorgänge und die Zeitabläufe überwacht. Schließlich werden die Artikel und Abstracts veröffentlichungsreif editiert und zur Publikation freigegeben.
Mittels des webbasierten und zentral beim DIMDI vorgehaltenen und gepflegten MOPS können die Fachgesellschaften rasch und kostengünstig den gesamten Reviewprozess – je skalierbar nach ihren spezifischen Anforderungen (doubleblind Review, single blind, offen) – selbst in die Hand nehmen. Gleichzeitig ermöglicht das MOPS aber auch neue Transparenz für Autoren: Können sich diese doch mittels eines besonderen Zugangs auf das MOPS einloggen und dort z.B. sehen, in welcher Phase des Begutachtungsprozesses sich ihr Artikel gerade befindet. Zentraler Vorteil der von Ihnen noch angesprochenen „My gms“ Komponente ist eine weitgehende

Individualisierbarkeit; so kann der interessierte Kunde dort z.B. eigene Such- bzw. Interessensprofile hinterlegen, in einem Archivbereich seine Suchen speichern und spezielle Print-on-Demand Dienstleistungen nutzen.

Hintergrund des Projektes war die sich besonders im STM-Bereich immer mehr verschärfende Zeitschriftenkrise, die es Bibliotheken, aber eben auch den einzelnen wissenschaftlichen Fachgesellschaften immer schwerer macht, eine adäquate Informationsversorgung sicherzustellen.
In erster Linie ist German Medical Science also die Antwort der deutschen wissenschaftlichen medizinischen Fachgesellschaften auf die ungebremst steigenden Zeitschriftenpreise, in zweiter Linie spielt natürlich auch der Gedanke des Open-Access, so wie er in der „Berliner Erklärung“ verstanden wird, eine nicht zu unterschätzende und gewichtige Rolle. Seit dem September 2002 wird gms von der DFG finanziert; ein Verlängerungsantrag wurde im September 2004 bewilligt. Für die Abrundung des beim DIMDI aufliegenden Publikationssystem hat zudem das Bundesministerium für Gesundheit und Soziale Sicherung (BMGS) Mittel zur Verfügung gestellt. Darüber hinaus stellen alle drei Projektpartner in erheblichem Maße Eigenleistung zur Verfügung.

Wie andere Initiativen (BioMedCentral, PloS) hat auch gms sich den unbeschränkten Zugang zu wissenschaftlichem Wissen auf die Fahnen geschrieben. Allerdings ist die Position von gms zu Veröffentlichungsgebühren eine grundlegend andere als bei zahlreichen anderen Initiativen: Als Initiative der wissenschaftlichen medizinischen Fachgesellschaften gegründet, überlässt es gms den einzelnen Fachgesellschaften selbst, ob sie für die von Ihnen publizierten Journals Autorengebühren erheben wollen. Für die Zeitschrift gms selbst stehen Gebühren pro eingereichtem Artikel derzeit nicht zur Debatte – hier wird ein Geschäftsmodell favorisiert, das eine Kostendeckung mittels verschiedensten Mehrwegdienstleistungen (wie etwa Print on Demand und Abstractmanagement) erreichen will.

Zudem will und muss gms den einzelnen wissenschaftlichen Fachgesellschaften einen Full-Service rund um das Publizieren bieten: das erstreckt sich vom Branding der eigenen Zeitschrift über Logoerstellung, Beantragung einer ISSN (bzw. einer ISBN), redaktioneller Starthilfe und Betreuung bis hin zur intensiven Beratung, welche Vorteile des neuen Mediums konkret die Fachgesellschaft genutzt werden sollten. Sie können sich sicher vorstellen, dass aufgrund der Heterogenität der Fachgesellschaftsanforderungen hier flexibel auf die Kundenwünsche reagiert werden muss. Ziel muss es vor allem sein, die Fachgesellschaften in jeder Phase des Publikationsprozesses redaktionell und konzeptionell zu unterstützen – und dies geht natürlich nicht allein mit Hilfe einer (wie auch immer gearteten) Software; hier ist vielmehr die Qualifikation der Projektmitarbeiter als Ansprechpartner der Fachgesellschaften gefragt.

Es ist sicher als Erfolg zu bezeichnen, dass es gms in sehr kurzer Zeit gelungen ist, sich als Publikationsalternative für die wissenschaftlichen medizinischen Fachgesellschaften in Deutschland zu positionieren. Um gms jedoch nachhaltig zu einem Erfolg zu führen, muss neben einem kontinuierlichen und professionellen Management vor allem das Angebot von gms arrondiert werden: Dies bezieht sich auch – aber nicht nur - auf die Softwarekomponenten, sondern umfasst vor allem den Aus- und Aufbau von gms zu einem Publikationsdienstleistungsportal für wissenschaftlichen medizinischen Fachgesellschaften. Dauerhaft wird sich gms in diesem hartumkämpften Markt nur behaupten können, wenn es seine größte Stärke sinnvoll einzusetzen und bedacht zu nutzen weiß: Den engen Kontakt zu den Fachwissenschaftlern.

Fallbeispiel: Einsatz eines Enterprise Content Management Systems für Verwaltung und Forschung

Doris Wochele, Rainer Kupsch; Karlsruhe

Abstract

Ein gemeinsames Ablage- und Informationssystem für Verwaltung, Wissenschaftler und Internetauftritt stellt sehr große Herausforderungen an ein Managementsystem. Informationsaufbereitung und -verteilung müssen ebenso flexibel sein wie die Berechtigungs- und Zugriffsstruktur. Die Anforderungen reichen von einem revisionssicheren Archiv bis zur "adhoc" Platform in der Projektarbeit. Ebenso muss die Anbindung diverser vorhandener Datenquellen berücksichtigt werden. In diesem Vortrag werde ich Ihnen einen Überblick geben über Aufbau, Einführung und Betrieb des Stellent 'Universal Content Management' im Forschungszentrum Karlsruhe und unsere erfolgreiche Verbindung von DMS und CMS.

1. Das Institut für Wissenschaftliches Rechnen im Forschungszentrum Karlsruhe

Das Forschungszentrum Karlsruhe ist eine der größten natur- und ingenieurwissenschaftlichen Forschungseinrichtungen in Europa und wird von der Bundesrepublik Deutschland und dem Land Baden-Württemberg gemeinsam getragen. Sein Forschungs- und Entwicklungsprogramm ist eingebettet in die übergeordnete Programmstruktur der Hermann von Helmholtz-Gemeinschaft Deutscher Forschungszentren und gliedert sich in die fünf Forschungsbereiche:

- Struktur der Materie
- Erde und Umwelt
- Gesundheit
- Energie
- Schlüsseltechnologien

Das Forschungszentrum Karlsruhe setzt sich aus 40 Organisationseinheiten (Instituten, Hauptabteilungen, Stabsabteilungen) zusammen mit insgesamt ca. 3500 Mitarbeitern. Zum Forschungsbereich Schlüsseltechnologien gehört auch das *Institut für Wissenschaftliches Rechnen (IWR)*, das als junges Institut aus dem Rechenzentrum des Forschungszentrums hervorgegangen ist und nun neben dem IT-Service für das Zentrum zunehmend Forschungsthemen im Bereich des „Grid-Computing“ bearbeitet.

Grid-Computing soll der Forschung, Industrie und Gesellschaft zukünftig jederzeit schnell und standortunabhängig den Zugriff auf weltweit verteilte Daten und andere

IT-Ressourcen ermöglichen. Grundlage für den Zugriff und die effiziente Nutzung den Globus umspannender verteilter Ressourcen wird dabei eine global vernetzte Informationsinfrastruktur mit komplexen Zugangs- und Administrationsmechanismen sein. Das Forschungszentrum verfolgt das umfassende Ziel, „Methoden, Werkzeuge und Standards zur Nutzung von Grid-Architekturen“ zu entwickeln und zu implementieren. Das IWR stellt IT-Anwendungen und die geeignete „Grid-Systemtechnik“ für den ständig wachsenden wissenschaftlichen Bedarf aus Medizin, Umweltforschung, Elementarteilchen- und Astrophysik zur Verfügung und entwickelt diese stetig weiter. Die Vision des „World Wide Grid“ ist der Zugang von jedem Ort zu allen Rechnern weltweit. In diesem Sinne beteiligt sich das IWR an den internationalen Projekten LCG (**L**arge Hadron Collider **C**omputing **G**rid) und EGEE (**E**nabling **G**rids for **E**-scienc**E**) und der deutschen Grid Initiative D-Grid. Im Rahmen von LCG/EGEE werden im IWR im Jahr 2007 etwa 4500 Prozessoren, ca. 3 Petabyte Online-Speicher und ca. 3 Petabyte Archivdaten für Verfügung stehen. In einem weiteren Vorhaben, dem Projekt „Campusgrid“, sollen die heterogenen Rechnerarchitekturen im Forschungszentrum in einem lokalen Grid zusammen gebracht und wie eine Ressource nutzbar gemacht werden. Die Entwicklung und Installation einer Middleware, eines Ressource Brokers, eines Globalen Filesystems und einer gemeinsamen Benutzerverwaltung stellen hier zum Teil erst ansatzweise gelöste Herausforderungen dar.
Der Servicebereich im IWR unterstützt Datenmanagement, Systemüberwachung, Informationsmanagement und zentrale Datenbanken, Netzdienste, Windows, Unix, Compute-Server, Office-Produkte sowie die Bürokommunikation.

2. Wie alles begann

Mitte 2002 wurde von Vorstand und „Wissenschaftlich Technischem Rat“ (Gremium zur Mitbestimmung) eine Neugestaltung des Internetauftritts des Forschungszentrums angeregt. Die Vorgaben beinhalteten ein einheitliches Erscheinungsbild (Corporate Design) für alle Organisationseinheiten des Zentrums, einen zielgruppenorientierter Einstieg, Benutzerfreundlichkeit, sowie den Einsatz eines zeitgemäßen Content Management Systems. Daraufhin wurde aus der Stabsabteilung Marketing und Patente (MAP), der Öffentlichkeitsarbeit, dem IWR und der Hauptabteilung Bibliothek und Medien eine Projektgruppe mit dem Namen FIND (Forschungszentrum Internet Development) zusammengestellt. Der erste Arbeitsschritt bestand in der Auswahl eines geeigneten Content Management Systems (CMS) auf Grund der vorhandenen Anforderungen. Die Wahl fiel auf das CMS von Stellent. Nach nur neun Monaten harter Projektarbeit konnte der neue Internet Auftritt des Forschungszentrums freigeschaltet werden. Die neue Navigationsstruktur, das einheitliche Konzept mit zentraler Administration, die Corporate Identity, die Mehrsprachigkeit, einfache Publikationsprozesse und die Darstellung der Inhalte erfüllten die

Erwartungen. Leider erkannte man schon kurz nach der Einführung, dass der neue Webauftritt mit der zunehmenden Anzahl von Seiten nicht ausreichend skalierte. Das lag in erster Linie an dem Konzept, statische Seiten im Web zu publizieren.
Die ersten Anfänge der Suche nach einem zentrumsweiten Dokumentenmanagementsystem (DMS) gehen auf das Jahr 2001 zurück. Eine Umfrage im Forschungszentrum ergab den Bedarf von einigen Organisationseinheiten an einem DMS. Ebenfalls 2002 wurde vom Vorstand in einem Projekt die Auswahl eines geeigneten Systems in Auftrag gegeben. Schon bei der Beschaffung des CMS von Stellent wurde darauf geachtet, dass die Software auch eine DMS-Komponente hatte. Somit war Stellent auf dem Weg zum ECM die erste Wahl. Es erfolgte eine Evaluierung in einer Pilotanwendung, die die gute Verwendbarkeit von Stellent demonstrierte. Anschließend wurde ein zentrumsweites Konzept für die Einführung eines DMS erstellt, welches die Grundlage für eine Beschaffung im April 2004 bildete.
Mit dieser Entscheidung waren die Randbedingungen für eine Verbesserung der in FIND erzielten Ergebnisse geschaffen. Ein neues Projekt unter der Federführung von MAP mit Beteiligung des IWR und eines externen Partners schloss sich an. Besondere Beachtung verdient die Tatsache, dass in diesem Projekt durch die Verschmelzung des Internetauftritts mit dem Dokumentenmanagementsystem, beides auf der Basis von Stellent, eine moderne ECM-Plattform innerhalb der Helmholtz-Gemeinschaft geschaffen werden konnte. Mit der Verwaltung der in Word geschriebenen „Web-Rohdokumente“ im DMS und der dynamischen Generierung der entsprechenden Webseiten beim Aufruf wurde größtmögliche Flexibilität geschaffen und die Performance Probleme des ersten FIND-Ansatzes beseitigt. Im Juni 2005 konnte die zweite Phase des Internetprojektes FIND erfolgreich abgeschlossen werden. Durch die einfache Pflege der in Word geschriebenen Webseiten im DMS ergibt sich der Vorteil sehr guter Retrievalmöglichkeiten ohne Kenntnis der Webstrukturen, einfacher Versionierung und sehr granularer und flexibler Zugriffsrechte. Mit der Integration des Webauftritts als einer Applikation innerhalb des DMS ist der erste wesentliche Meilenstein auf dem Weg zu einem zentrumsweiten ECM erreicht worden.

3. Das ECM-System Stellent

Stellent® Universal Content Management[1] bietet eine flexible und skalierbare Lösung für die sichere, benutzerdefinierte Verwaltung von Dokumenten in ihrem gesamten Lebenszyklus sowie die Publikation von Inhalten in einer Webpräsenz im Corporate Design (CD). Die offene Architektur erlaubt es, eigene, spezielle Anforderungen über einfache Programmiertechniken einzubringen und über Connectoren, wie Web-Services, kann das System in die bestehende IT-Infrastruktur eingebunden werden.

[1] www.stellent.com

Mit dem Stellent Content Server und dem Stellent Dynamic Converter waren es letztlich nur zwei Produkte die für die CMS- und DMS-Gesamtlösung benötigt wurden. Somit wird auch der administrative Aufwand minimiert. Synergien ergeben sich auch durch die Verwaltung interner Dokumente gemeinsam mit den Webseiten, da die annähernd 100'000 Webseiten sowohl im Intranet als auch im Internet ohne Zeitverzögerung wahlweise Mitarbeitern und Partnern des Forschungszentrums über die mehrsprachige Oberfläche zur Verfügung gestellt werden können.
Das Stellent-Framework verfügt über alle gängigen DMS-Funktionen wie Revisionierung, Expiration, Workflow etc. Diese wurden durch die Integrationspartner um „Sammelmappen", „Laufzeitsuche" (z.B. Gültigkeitsdauer bei Verträgen) und „Suche in Metadaten" erweitert. Die große Zahl von möglichen Formatkonversionen in webtaugliche Formate (z.B. PDF) ermöglicht es, spezielle Dokumentformate wie CAD-Zeichnungen für den betriebsystemunabhängigen Zugriff über den Browser automatisch aufzubereiten.
Das Sicherheitskonzept ordnet Dokumente in Konten und erlaubt den rollenbasierten Zugriff über Gruppen. Das ermöglicht mit der Anbindung an das Active Directory des Forschungszentrums eine flexible Berechtigungsstruktur. Über „closed user groups" können Webseiten und Dokumente für autorisierte Personen veröffentlicht werden.
Die Stellent Installation wurde speziell für das Forschungszentrum konfiguriert und programmiertechnisch angepasst. So wurde der gesamte Publikationsprozess der Webseiten über eine eigene Komponente definiert, um CD- und Navigationsanforderungen umzusetzen. Eine weitere Komponente ermöglicht die dezentrale Administration der Webseitenstruktur, Metadaten und Zugriffsrechte für die Organisationseinheiten.
Durch die Benutzer werden laufend Wünsche eingebracht, die, sofern sie in das Betriebskonzept passen, in Praktikumsarbeiten realisiert werden können, da Erweiterungen in JSP[2] und dem proprietären „Idoc" technisch einfach sind.

[2] **J**ava **S**erver **P**ages

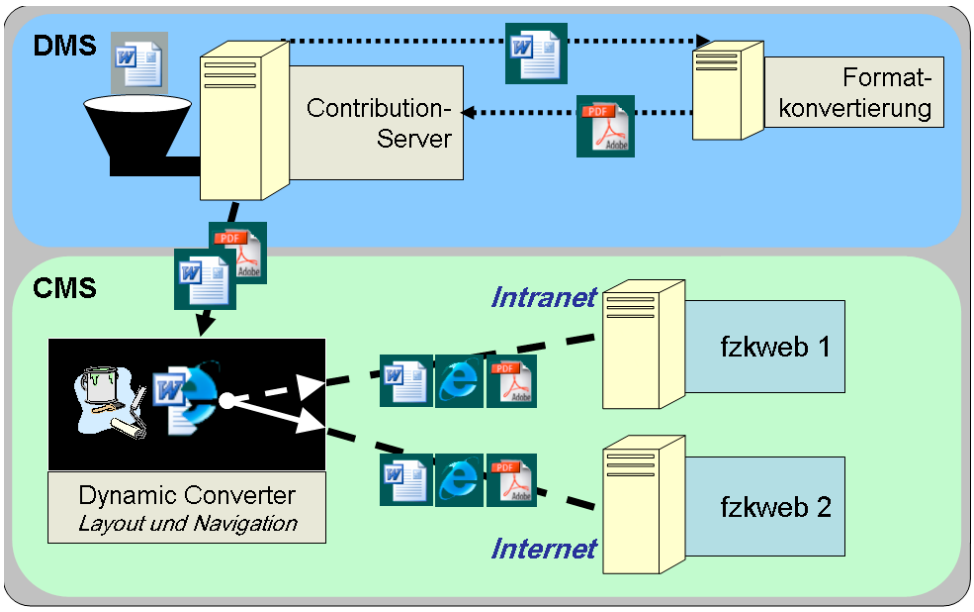


Abbildung 1: Schematische Darstellung des Stellent-Systems im FZK. Stellvertretend für alle Arten von Content wird hier gezeigt wie das Word-Dokument im *e*DMS in PDF konvertiert und danach ins CMS repliziert wird. Dort wird das Layout erzeugt und die HTML-Seite in die Navigation eingebettet. Als Ergebnis kann die Webseite in drei Formaten auf dem Intranet oder Internet abgerufen werden.

4. Dokumentenmanagement in einer Forschungseinrichtung

Ein Dokumenten Management System dient als eines von vielen Werkzeugen für das Wissensmanagement (Probst 2003). Dabei übernimmt es die Rolle Wissen zu wahren, zu teilen und zu nutzen nur dann, wenn es damit gelingt, möglichst effizient entscheidungsrelevante Informationen zu identifizieren und für eigenes Handeln zu nutzen. Forschung und Entwicklung ist ein höchst dynamisches, sehr informationslastiges und informationsabhängiges Geschäftsfeld. Nach R.Dippold (Dippold 2005) ist ein umfassendes Informationsmanagement ein kritischer Erfolgsfaktor im internationalen Wettbewerb, dazu zählt insbesondere die rechtzeitige Bereitstellung von Information mit hoher Datenintegrität. Das Dokumenten Management System stellt sicher, dass immer auf die aktuelle Version zugegriffen wird und bietet gleichzeitig den Überblick über alle Veränderungsprozesse.

Dokumentenmanagement in einer Forschungseinrichtung kann in zwei Bereiche eingeteilt werden. Das sind zum einen wissenschaftliches Arbeiten, Forschung, Entwicklung und Lehre (Präsentation von Ergebnissen, Zusammenarbeit in der wissenschaftlichen community und Wissenskontinuität und –weitergabe), zum anderen

umfasst es das „Verwaltungshandeln“ (Gremienarbeit, Anträge, interne Verwaltungsprozesse). In beiden Bereichen bestehen Anforderungen für ein geregeltes Informationsmanagement, wozu ein DMS einen entschiedenden Beitrag leisten kann. Durch die Berliner Erklärung[3] über offenen Zugang zu wissenschaftlichem Wissen wird gefordert, Publikationen im Internet öffentlich zugänglich zu machen. Der Verwaltungsbereich im öffentlichen Dienst ist durch die DOMEA[4] Initiative des Bundes angehalten ihre Geschäftsprozesse auf digitale Aktenbearbeitung umzustellen. Die Geschäftsführung des Forschungszentrums hat erkannt, dass von einer Effizienzsteigerung in Verwaltung und Management das gesamte Unternehmen profitiert und die Vereinfachung der Verwaltungsvorgänge auch den Wissenschaftler entlastet. Die wissenschaftlichen Bereiche und Träger von Querschnittsaufgaben sehen für Ihre Arbeit den unmittelbaren Nutzen eines DMS .

5. Einführungsstrategie

Durch den Auftrag des Vorstandes im Forschungszentrum Karlsruhe künftig ein einheitliches System für DMS und CMS einzusetzen wird den Organisationseinheiten lediglich der Entscheidungsprozess für die Beschaffung abgenommen und die Verantwortung für die Intranet und Internetseiten der Internetredaktion (MAP) und dem IWR übertragen. Für institutsinterne Webseiten oder Dokumentenmanagement sind Insellösungen vorhanden und die Bereitschaft diese zugunsten einer einheitlichen Lösung aufzugeben muss erarbeitet werden.
Ein Dokumentenmanagementsystem alleine führt noch nicht zum Übergang auf die Entwicklungsstufe „Datenmanagement“ eines Unternehmens, dazu bedarf es zunächst einer übergreifenden Datenstandardisierung und einer gemeinsamen Datenbasis (Dippold 2005). Ein strategisches Vorgehen um ein einheitliches Daten- oder Informationsmanagement zu erreichen ist eine nahezu unlösbare Aufgabe. Dazu sind im Forschungszentrum die Arbeitgebiete zu breit gefächert und die zahlreichen Forschungsgruppen arbeiten innerhalb verschiedener Programme mit diversen Kooperationspartnern und Auftraggebern von EU, Bund und Land auf verschiedenste Arten zusammen. Aussichtsreicher ist daher ein taktisches Vorgehen mit Ausrichtung auf Kosten-Nutzen-Überlegungen mit kurz- oder mittelfristigen Zielen, um damit eine von den DMS-Anforderungen aus den Fachabteilungen angetriebene Standardisierung und einheitliche Modellierung zu erreichen.
Soll sich das *e*DMS im Zentrum etablieren, muss es in den Arbeitsprozess der beteiligten Fachbereiche integriert werden können und durch die neuen oder verbesserten Funktionen einen „Return-On-Investment (ROI) erkennen lassen. Vielfach wird zunächst nur der erhöhte Aufwand bei der Ablage der Dokumente gesehen.

[3] www.mpg.de/pdf/openaccess/BerlinDeclaration_dt.pdf

[4] **Do**kumenten**m**anagement und **e**lektronische **A**rchivierung im IT-gestützten Geschäftsvorgang der öffentlichen Verwaltung

Benutzerfreundlichkeit, Stabilität und hoher Nutzen durch zielgerichtete Anpassungen stehen deshalb im Vordergrund um den Benutzern den Mehrwert insbesondere in der strukturierten Ablage und dem besseren Retrieval überzeugend nahe zu bringen. Behutsam müssen Geschäftsprozesse umgestellt werden, damit bestehende Insellösungen aufgegeben werden können.
Derzeit betreiben wir einen Contribution Server und zwei Server für das Intranet und Internet mit zwei Konvertierungsrechnern an einem Oracle „Fail-over-Cluster“ und dazu die entsprechenden Backup- und Entwicklungs-Systeme, um einen 24x7 Betrieb und geregeltes Configurationmanagement zu gewährleisten.
Momentan sind zwei Verwaltungsabteilungen, und zwei wissenschaftliche Abteilungen produktiv im *e*DMS tätig. Vier weitere Abteilungen mit Querschnittsaufgaben sind in der Projektphase. Zusammen mit den Webseiten haben wir ca. 14600 Dokumente mit 7500 Revisionen und ca. 200 User im *e*DMS.

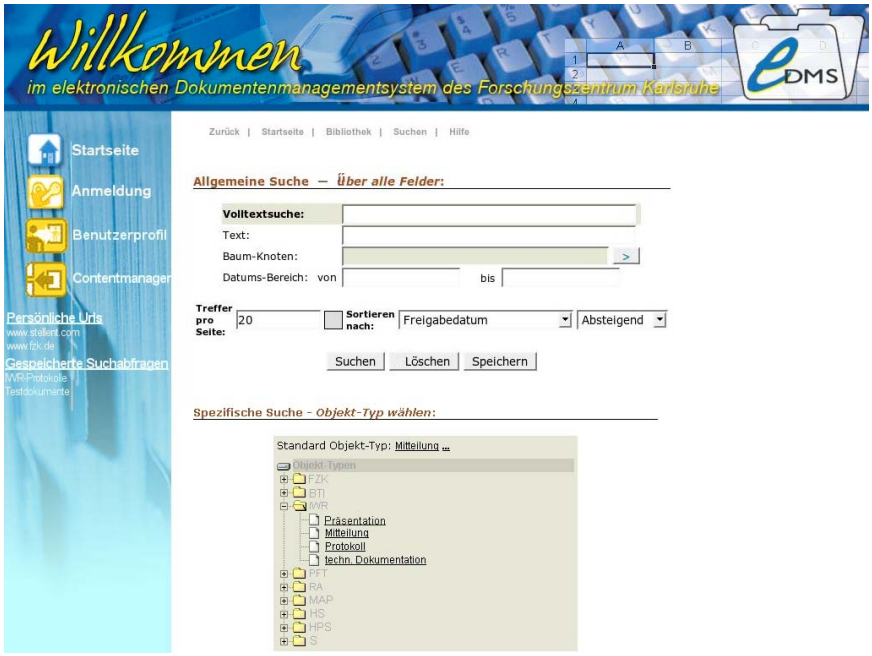


Abbildung 2: Suchseite im mit „*e*DMS“ bezeichneten Contribution Server des Forschungszentrums.

Die Einführung und Integration von CMS in DMS wurde von zwei kompetenten Partnern begleitet. Die Fa. Netway Solutions AG (CH) ist Technologiepartner von Stellent und verfügt über fundiertes Produkt Know How. Nur durch Ihr gutes Gespür für Nutzerfreundlichkeit und Ihre hohe Flexibilität bei der Projektrealisierung konnten unsere speziellen Anforderungen realisiert werden. Herr Michael Riester hat als freier

Stellent Consultant das Projekt begleitet und stand der Projektleiterin Frau Dr. Aida El-Kholi als technischer Berater zur Seite. Ohne diese Partner mit ihrer Erfahrung für das customizing des Stellent-Frameworks und den guten Kontakt zu Stellent Deutschland, die sich stets unserer Probleme und Schwierigkeiten angenommen haben, wären wir kaum so rasch zu einer solch umfassenden, zufrieden stellenden Lösung gelangt.

6. Aufbau der Dokumentstruktur

„Ich möchte eine Publikation ablegen“ – ausgehend von der Arbeitsweise des Menschen wurde der „Dokumenttyp“ (Begriffe nach Götzer 2004) als Ausgangspunkt für Ablage und Recherche im Dokumentenmanagementsystem des Forschungszentrums Karlsruhe gewählt. Für jedem Dokumenttyp wird ein fester Satz semantischer Metadaten definiert, der die Daten inhaltlich beschreibt und es dem Benutzer ermöglicht, die Relevanz der Daten festzustellen. Durch eine Analyse des Geschäftsprozesses werden die für jeden Arbeitsablauf benötigten Dokumenttypen erarbeitet und definiert. Durch diese festen Vorgaben bei der Zuweisung der Metadaten soll es ermöglicht werden, die Dokumenttypen zur Zusammenarbeit und zum Dokumentaustausch für das gesamte Forschungszentrum zu nutzen. Anzahl und Detailgenauigkeit der Metadaten sind jedoch vom Arbeitsgebiet abhängig und müssen an die Verwendung angepasst werden, was die Organisationseinheiten durch Sub-Admins selbst wahrnehmen können. Als Hilfsmittel wurden Metadatenbäume (Content Trees, entsprechen in ihrer Funktion hierarchischen Auswahllisten) eingeführt. Die Zuordnung eines Dokuments zu den Ästen eines oder mehrerer Bäume bestimmt den Standort im mehrdimensionalen Metadatenraum. Obwohl das Dokumentenmanagement das Kerneinsatzgebiet für semantische Netze ist (Beier 2004), wurde bei der Einführung dieser Ansatz nicht gewählt, da die Gesamtheit der relevanten Metadaten nicht einmal ansatzweise ermittelt werden kann. Die Nutzung des Systems geschieht auf freiwilliger Basis gegen Abrechung mit der zentralen DV. Demzufolge ist der künftige „Kundenkreis“ unbekannt, und die Taxonomie orientiert sich ausschließlich an den Forderungen der derzeitigen Nutzer. Die Verknüpfung der Metadaten und der Aufbau eines semantischen Netzes bleiben uns bei der gewählten Architektur des Systems aber für die Zukunft offen.
Nach der Anmeldung im *e*DMS wählt der Benutzer „Neuen Content einchecken“ und danach den Dokumenttyp des Dokumentes welches zum Einchecken vorliegt. Jeder Benutzer kann dabei auch eigene Typ-Favoriten vorgeben. Für den gewählten Typ werden dann die definierten Metadatenfelder mit der jeweiligen Beschriftung in der Reihenfolge angezeigt wie sie vom Fachbereich gewünscht wurden. Durch die Abstraktion der Metadaten werden übergreifende Zusammenhänge gewahrt. Wird beispielsweise im Fachbereich 1 das Metadatum xtext1 im Dokumenttyp „Präsentation“ mit „Autor“ betitelt, kann es im Fachbereich 2 mit „Vortragender“ bezeichnet

sein. Es ist die Aufgabe der IT durch eine Festlegung » xtext1=Erzeuger des Content ≠ Person die eincheckt « das übergreifende Metadatenmodell sicherzustellen.
Ein CMS-Dokument ist im Sinne des Dokumentenmanagement nur ein Dokumenttyp mit speziellen, komplexen Verarbeitungskriterien. Prinzipiell ist jedes Dokument durch die Metadaten » Sprache, Publikationsort, Intranet/Internet und publizieren ja-nein « im Web publizierbar.

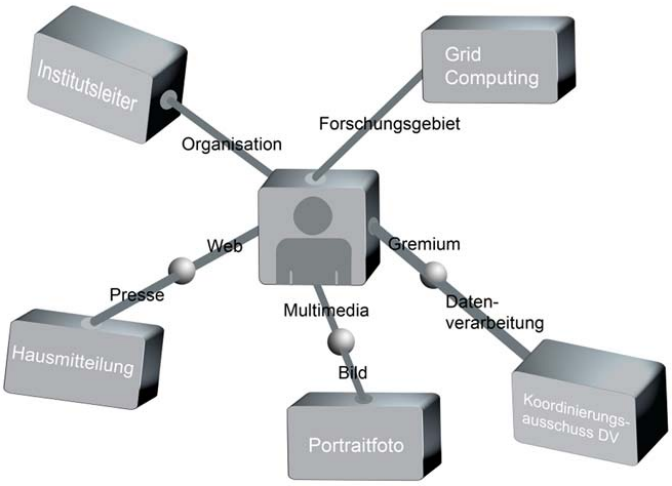


Abbildung 3: Beispielhaft der Metadatenraum für ein Foto. Es ist mit den Funktionen der Person (Klassifikation im organisatorischen Rahmen), dem Anlass der Entstehung und seinem Verwendungszweck assoziiert. Darüber hinaus existieren zum Content-Typ selbst weitere beschreibende Metadaten (Aufnahmedatum, Name, etc.)

7. Informationsretrieval

Die "Google“-Suche ist einfach und populär. So sind alle Webseiten und indizierte Contentobjekte im Volltext durchsuchbar, das Ergebnis der Suche ist jedoch unspezifisch. Eine besseres Suchergebnis lässt sich erzielen, wenn der Benutzer den Kontext angibt, in dem er nach Information sucht. Im DMS kann der Benutzer denselben Weg wie beim Ablegen des Content wählen – er gibt zu erkennen „was er sucht“ und wählt den Dokumententyp („Publikation“). Danach kann er die Suche über die spezifischen Metadaten des Typs weiter eingrenzen und auch in der Ergebnismenge die Suche verfeinern. Fehlt dem Nutzer die Terminologie oder das Wissen um den Kontext richtig vorzuwählen kann er über alle Metadaten suchen (Suche über alle Textfelder, Datumsfelder etc.), um dann die Suche weiter zu verfeinern oder danach auf einen Typ einzuschränken. Für das Intranet und Internet wurde ein Schlagwortkatalog definiert, der durch die Internetredaktion gepflegt wird. Für das

Intranet steht außerdem eine Google Search Appliance[5] zur Verfügung, die Inhalte auf den zahlreichen Webseiten der verteilten Server in den Organisationseinheiten recherchierbar macht.

8. Zugang und Personalisierung

Stellent ist mit dem Microsoft Active Directory des Forschungszentrums Karlsruhe zur Benutzerauthentifizierung gekoppelt. Die Verwaltung der Berechtigungsstrukturen und die Autorisierung erfolgt jedoch in Stellent. Um dies effizient für alle Serverinstanzen zu handhaben, wurde in einer Praktikumsarbeit eine zentrale Benutzerverwaltung auf Basis von JSP entwickelt, welche über Web-Services Benutzer und Benutzerrechte auf die Stellent-Server synchronisiert. Die Schnittstelle wird künftig auch den lokalen Administratoren der Organisationseinheiten für Ihre Benutzerverwaltung zur Verfügung stehen. Für den Zugang vom Internet auf geschützte Webseiten werden Passwörter vergeben, der Zugang zum Content-Server bleibt derzeit noch auf VPN beschränkt.
Der einzelne Benutzer kann sich die Anwendung über ein persönliches Profil an seine Bedürfnisse anpassen und Einstellungen für das Aussehen der Oberfläche, Darstellung der Suchergebnisse, persönliche URL's u.v.m. vornehmen.

9. Prozessabläufe

Dokumente werden nach vorgegebenen Regeln und Merkmalen zur Weitergabe in den Informationskreislauf eingebracht. Solche dokumentbasierte Workflows sind meist einstufig und vordefiniert (z.B. Review durch den Chefredakteur). Ein mehrstufiger, evt. formularbasierter Workflow überschreitet die Grenze zur „Collaboration-Work". Da im Forschungszentrum bereits das Workflowsystem Remedy im Einsatz ist, wird derzeit die Verbindung beider Systeme mit Web-Services evaluiert.
Standardmäßig in Stellent verfügbar ist bisher der einstufige Workflow (Review), das Abonnement, das den Leser via Mail über neue Revisionen informiert und dass Ad-hoc Versenden von Dokument-Links an beliebige Mailempfänger (Notify). Derzeit evaluieren wir die Einbindung von Suchmasken bzw. gespeicherten Suchabfragen in andere browserorientierten Informations- oder Portalsysteme wie z.B. der DV-Servicedesk, um aus den Stichworten der Hotlinedatenbank heraus gezielt Dokumente aus dem DMS abzurufen.
Papierbasierte Dokumente werden über die verteilten Scanner der Organisationseinheiten auf ein zentrales Filesystem abgelegt und über einen Batchload im 5-min-Takt regelmäßig in den Content-Server geladen. Die Zuordnung der Scanner-Ziele zu Verzeichnissen und über den Batchload zu Default-Metadaten erleichtert dem

[5] www.google.de/enterprise/

Sachbearbeiter das attributieren, da das Dokument beim Scannen per Knopfdruck mit dem richtigen Dokumenttyp eingecheckt wird. Multifunktionskopier-Scanner der Fa. RICOH haben sich im Tagesgeschäft als „Stockwerksscanner“ bewährt. Der Sachbearbeiter bestimmt dabei ob der gewählte Dokumenttyp als Bild (TIFF) gescannt und anschließend automatisch mit der Texterkenungssoftware in PDF umgewandelt wird oder vielleicht sofort als PDF eingecheckt wird, weil keine Volltextindizierung benötigt wird.

10. Fazit und Ausblick

Auf dem Weg zum ECM im Forschungszentrum Karlsruhe sind die ersten Schritte gemacht. Wir haben:

- eine skalierbare, performante Lösung für CMS und DMS in einem System
- Benutzer aus fast allen Organisationseinheiten nutzen das System, was durch die CMS-Applikation forciert wird.

Unter „Lessons learned“ lässt sich zusammenfassen

- Externer Support ist unerlässlich, doch sollte Erfahrung mit dem System vorliegen.
- Ein detailliertes Pflichtenheft und eine straffe Projektführung sind aufwändig aber unerlässlich.
- Die späteren Administratoren müssen bei allen Installationen und Modifikationen am System eingebunden werden.
- Das frühe Einbeziehen der Nutzer in die Entscheidungsprozesse schafft Akzeptanz für das Gesamtkonzept.
- Ein globales Gesamtkonzept muss erstellt werden, bevor einzelne Lösungen umgesetzt werden. Das gilt auch für CMS als eine Anwendung im DMS.
- Das Berechtigungskonzept ist als sehr kritisch einzustufen und sollte so flexibel wie möglich ausgelegt werden.
- Der Eingriff in die Geschäftsprozesse sollte behutsam angegangen werden, wobei ein Prototyp hilft, Berührungsängste abzubauen.

Derzeit befindet sich das System noch immer in einer regen Wachstumsphase. Nicht nur was die Zahl der Contributoren angeht, sondern auch was die Zahl der Funktionen betrifft. Jede Fachabteilung bringt neue Wünsche ein, die eingearbeitet werden sofern diese für das Forschungszentrum insgesamt von Bedeutung sind und sich in das Gesamtkonzept einfügen.
Die nächste große Herausforderung wird das Thema Langzeitarchivierung sein ebenso wie spezielle Storage-Systeme für revisionssichere Speicherung. Das Thema elektronische Unterschrift wird uns, wie viele Unternehmen, noch lange beschäftigen.

Nach Remedy werden wir versuchen, lokale Datenmanagementsysteme der Organisationseinheiten anzubinden und Dokumente über die SAP ArchiveLink Schnittstelle von Stellent mit SAP auszutauschen.
Für das Forschungszentrum Karlsruhe kontinuierlich eine sichere Informationsplattform voranzutreiben und der Zettelwirtschaft ein Ende zu bereiten, ist das erklärte Ziel des ECM–Teams.

Literatur

(Beier 2004) H.Beier: *Vom Wort zum Wissen, Information Wissenschaft und Praxis* 55(2004)3,133-138, DGI Frankfurt.

(Dippold 2005) R.Dippold, A. Meier, W.Schnider, K.Schwinn*: Unternehmensweites Datenmanagement*, Vieweg Braunschweig (2005).

(Götzer 2004) K.Götzer, U.Schneiderath, B.Meier, T.Komke*: Dokumenten-Management*, dpunkt Heidelberg (2004).

(Probst 2003) J.B.Gilbert Probst, S.Raub, K.Romhardt*: Wissen managen*, Gabler (2003).

Open Access – Lessons Learned

Die Auswirkungen der „Open-Access“-Initiative auf die Wertschöpfungskette und die Struktur wissenschaftlicher Kommunikation

Hans-Robert Cram, Berlin

Abstract

Die „Open-Access“-Initiative (OAI) ist eine weltweite Bewegung, die vor mehr als 10 Jahren begonnen hat und heute sowohl durch privat organisierte Wissenschaftsverbände und Non-Profit-Organisationen als auch durch ganze Bibliotheksverbände in der Realisierung begriffen ist. Ihren Ausgangspunkt genommen hat diese Bewegung durch die als „journal crisis““ bekannte Tatsache, dass in den letzten Jahren die Ladenpreise für STM-Zeitschriften weitaus stärker angestiegen sind als der Lebenshaltungsindex im gleichen Zeitraum, während auf der anderen Seite Bibliotheksbudgets stetig zurückgehen. Die Debatte um die OAI wird bisweilen sehr heftig und emotional geführt und nur allzu leicht gleiten die Argumente auch ins Unsachliche ab.

Konkretisiert hat sich die OAI in zwei Modellen: das autorenfinanzierte Modell und das Selbstarchivierungsmodell. In diesem Beitrag werden die Vor- und Nachteile und die Zukunftsfähigkeit beider Modelle diskutiert und Gründe dafür dargelegt, dass das autorenfinanzierte Modell vermutlich weniger zukunftsfähig ist als das Selbstarchivierungsmodell.

Beide Modelle verändern das bestehende System wissenschaftlicher Kommunikation, das durch die folgenden Bestandteile charakterisiert ist: *Registration, Certification, Awareness, Archiving* und *Rewarding*. Dies führt zu einer veränderten Wertschöpfungskette und einer anderen Rollenverteilung bei den beteiligten Akteuren (das sind: Wissenschaftler/Autor – Verlag – Gutachter – Bibliothek – Wissenschaftler/Leser). Eher unwahrscheinlich ist das zuweilen prophezeite völlige Verschwinden einzelner Akteure (etwa der Verlage oder gar der Bibliotheken), vielmehr wird sich der Aufgabenschwerpunkt der Beteiligten verschieben. Der Beitrag geht auf die verschiedenen möglichen neuen Wertschöpfungsketten und Kommunikationsstrukturen mit ihren jeweils veränderten Rollen und Aufgaben der Akteure ein und stellt sie einander gegenüber. Am Ende wird im Rahmen einer Synthese ein Ausblick auf die wahrscheinliche Zukunft der Wissenschaftslandschaft gegeben.

1. Hintergrund

„There is evidence to suggest that the market for STM journals may not be working well".[1]

Zu dieser ungeachtet des feinen britischen Understatements doch sehr klaren Aussage kam Großbritanniens Office for Fare Trading im September 2002. Beklagt wird in diesem Statement die als „journal crisis" bekannte Tatsache, dass in den letzten Jahren die Ladenpreise für STM-Zeitschriften weitaus stärker angestiegen sind als der Lebenshaltungsindex im gleichen Zeitraum, während auf der anderen Seite Bibliotheksbudgets stetig zurückgehen.

Im Juli 2004 wird in einem Bericht des House of Commons der britischen Regierung empfohlen, den Aufbau miteinander vernetzter und öffentlich zugänglicher Datenbanken wissenschaftlicher Einrichtungen mit öffentlichen Geldern zu unterstützen. Öffentlich geförderte Wissenschaftler sollen gezwungen werden, Kopien ihrer Artikel auf diesen digitalen Archivservern abzulegen, wo sie der Welt kostenfrei zur Verfügung stehen.[2] Zwar ist die britische Regierung in Ihrer Antwort vom 1. November 2004 diesen Empfehlungen nicht gefolgt, die Förderung des weiteren Auf- und Ausbaus frei zugänglicher Datenbanken wissenschaftlicher Institutionen mit öffentlichen Mitteln wird aber ausdrücklich unterstützt.[3]

Diese „Open Access Initiative" (OAI) genannte Bewegung, die sich in zwei unterschiedlichen, weiter unten diskutierten Modellen konkretisiert, verändert unaufhaltsam die Wissenschaftslandschaft. Das „Directory of Open Access Journals"[4] verzeichnet mittlerweile über 1.600 frei zugängliche elektronische Zeitschriften auf allen Gebieten, einschließlich Geisteswissenschaften. Parallel dazu arbeiten fast alle wissenschaftlichen Bibliotheken und wissenschaftliche Einrichtungen weltweit am Aufbau digitaler Archivserver (genannt „digital repositories"), die zum Teil auch schon online frei zugänglich sind.[5]

---

[1] UK Office for Fare Trading Report: The Market for Scientific, Technical and Medical Journals, Sept. 2002, Seite 4, http://www.oft.gov.uk/NR/rdonlyres/A56C7602-C0BD-428D-BED2-36784363243B/0/oft396.pdf.

[2] a.a.O., Abs. 212.

[3] House of Commons, Science and Technology, 14th Report, 01.11.2004, Response from the Government, http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/1200/120006.htm#a1.

[4] Stand: Februar 2005. Siehe: http://www.doaj.org.

[5] Zum Beispiel MIT's Digital Repository, http://dspace.mit.edu/index.jsp, sowie alle Universitäten der Niederlande, die allerdings noch nicht miteinander vernetzt sind; z. B. Erasmus University of Rotterdam: http://eps.eur.nl/. Siehe auch Vrije Universiteit Amsterdam: http://www.ubvu.vu.nl/index.cfm?objecttype=page&objectid=468&talen=&CFID=157724&CFTOKEN=55769447

Der STM-Markt wird beherrscht durch Reed Elsevier, Thomson, Wolters Kluwer, Springer und John Wiley. Die fast monopolartige Stellung dieser großen Verlagshäuser hat dazu geführt, dass die Ladenpreise der STM-Zeitschriften zwischen 1998 und 2003 um 58 % gestiegen sind. Für den Zeitraum von 1990 bis 2000 verzeichnen Blackwell's Periodical Price Indexes sogar einen Anstieg um 184,3% für medizinische, und von 178,3% für naturwissenschaftlich-technische Zeitschriften. Auf der anderen Seite sind die Budgets der Bibliotheken stark zurückgegangen mit der Folge, dass Bibliotheken viele Abonnements gekündigt haben, was sich in einer deutlich schlechteren Verfügbarkeit der wissenschaftlichen Inhalte niederschlägt.[6] Der Rückgang der Abonnementzahlen führt wiederum bei den Verlegern zu weiteren Preisanstiegen. Ein Ende dieses Circulus Vitiosus ist derzeit nicht zu erkennen.
Kein Wunder also, dass die Debatte um die OAI bisweilen sehr heftig und emotional geführt wird. Die großen Verlagshäuser, allen voran Elsevier als der absolute Marktführer im Bereich STM, erfahren heftige Angriffe. Man betrachtet es schon fast als Unverfrorenheit, dass Elsevier im STM Markt einen operativen Gewinn von über 30 % ausweist, während Bibliotheken an immer knapperen Mitteln leiden und sich kaum mehr die nötigsten Anschaffungen leisten können. Das Problem wird manchmal sogar so weit verallgemeinert, dass die Kommerzialisierung von Wissen überhaupt als nicht vereinbar mit den klassischen akademischen Werten betrachtet wird.[7] Hinzu kommt häufig die eher ideologischen Forderung, Wissenschaft müsse jedermann kostenlos zur Verfügung stehen. Da diese Problematik zweifellos einer der Ausgangspunkte der Open-Access-Bewegung ist, könnte man – neutral ausgedrückt – das Ziel der Bewegung darin sehen, Modelle wissenschaftlicher Kommunikation zu entwickeln, deren volkswirtschaftliche Gesamtkosten deutlich niedriger sind als bei dem herkömmlichen subskriptionsgestützten Modell.
Es gibt allerdings noch einen anderen Ausgangspunkt für die OAI, nämlich eine faktische Änderung im Forschungsverhalten. Man forscht heute zunehmend in weltweiter Zusammenarbeit, wobei alle Experten eines Gebiets weltweit einen gemeinsamen Zugriff auf riesige Mengen relevanter Daten haben, die online zur Verfügung stehen.[8] Das gegenwärtige System hat mit diesen Veränderungen nicht

[6] House of Commons Report a.a.O., Abs. 49 ff. Zu berücksichtigen ist dabei allerdings, dass auch der Umfang der Zeitschriften deutlich gestiegen ist. Wurden noch 1975 im Schnitt 85 Artikel pro Zeitschrift publiziert, waren es 2002 bereits 154 Artikel je Titel. Dennoch ist der Preisanstieg weit überproportional.

[7] So z. B. Daniel Atkins : " … the increasing tendency to proprietize knowledge, to view the output of research as intellectual property, is hostile to traditional academic values". Atkins, What is Publishing in the Future? http://books.nap.edu/html/e_journals/ch6.html.

[8] Typisches Beispiel ist das vor rund 14 Jahren begonnene „Human Genome Project“ mit dem Ziel, die rund 30.000 Gene der menschlichen DNA zu bestimmen. Mehr als 18 Länder sind an diesem Projekt beteiligt. Das Projekt hat eine ganze Industrie entstehen lassen und eine grundlegende Transformation der biomedizinischen Wissenschaft zur Folge. Vgl. House of Commons Report a.a.O., Q 200.

Schritt halten können. Allein schon die mangelnde Kompatibilität der verwendeten Datenformate (PDF, DOI, OpenURL, OAI-PMH) erschwert ein weltweites netzwerkgestütztes Arbeiten erheblich. Überhaupt gehört die Möglichkeit, wissenschaftliche Inhalte über das Internet einfach und schnell einem Maximum an Nutzern zur Verfügung zu stellen, zu den größten Vorteilen der OAI und ist auch eine ihrer wichtigsten Beweggründe.
Die Auswirkungen auf die Wertschöpfungskette und die Struktur wissenschaftlicher Kommunikation sind beträchtlich. Das gegenwärtige abonnementgestützte Publikationsmodell bei den STM-Zeitschriften sieht vor, dass der Autor seine Artikel kostenlos bei dem Verlag einreicht – zuweilen muss der Autor sogar noch so genannte „Page charges" zahlen – und dem Verlag die vollen Nutzungsrechte überträgt. Der Verlag sorgt dafür, dass die Artikel einem Begutachtungsprozess unterzogen und bearbeitet werden und in einer Zeitschrift erscheinen. Die Zeitschrift wird dann über Subscription Agents vertrieben, Hauptabnehmer sind Universitätsbibliotheken. Hauptmerkmal dieses Modells im Unterschied zu den nachfolgend beschriebenen Open-Access-Modellen ist, dass der Abonnent für die Zeitschrift bezahlt.

2. Das autorenfinanzierte Modell

Wachsende Unzufriedenheit mit dem traditionellen subskriptionsgestützten Publikationsmodell im STM-Zeitschriftenmarkt hat zu der Herausbildung eines Modells geführt, wonach der Autor selbst für die Finanzierung der Publikation seines Artikels herangezogen wird. Davon abgesehen folgen die Abläufe dem traditionellen Modell wie oben beschrieben: Der Autor reicht seinen Artikel beim Verlag ein, der einen Begutachtungsprozess einleitet. Wird der Artikel zur Publikation angenommen, so muss der Autor für die Publikation zahlen, in der Regel wird er mit 500,-bis 1.500,-$ belastet. Der Verlag sorgt dann dafür, dass der Artikel bearbeitet wird und im Rahmen einer elektronischen Zeitschrift erscheint. Die Zeitschrift steht dann allen Nutzern weltweit kostenfrei zur Verfügung. Im Unterschied zum herkömmlichen Modell überträgt der Autor dem Verlag allerdings nicht seine Nutzungsrechte. Das Copyright verbleibt vielmehr beim Autor, der sich lediglich damit einverstanden erklären muss, dass sein Artikel nach Publikation der Welt kostenlos zur freien Verfügung steht.
In der Praxis werden die Kosten aber nicht von den Autoren persönlich getragen, sondern von den Institutionen, denen sie angehören oder die ihre Forschung sponsern. Letztendlich ist es also wieder die öffentliche Hand, die für die wissenschaftlichen Inhalte bezahlt, zuweilen sogar dieselbe Institution, die nach dem klassischen Modell auch die Subskriptionen finanziert. Ob das autorenfinanzierte

Open-Access-Modell volkswirtschaftlich betrachtet insgesamt kostengünstiger ist, ist aber eher zweifelhaft.[9]

Neben BioMed Central[10] und der Public Library of Science (PLoS)[11], den prominentesten Non-profit Organizations, die sich ausschließlich dieses autorenfinanzierten Modells bedienen, experimentieren zunehmend auch kommerzielle Verlage mit diesem Modell. So führt Oxford University Press bereits einzelne Zeitschriften komplett nach dem autorenfinanzierten Open-Access-Modell. Der Springer Verlag und Blackwell hingegen lassen ihren Autoren mit den Angeboten „Open Choice"[12] bzw. „Online Open“[13] die Wahl, Artikel entweder nach dem bewährten Subskriptionsmodell zu publizieren oder für die Publikation selbst zu bezahlen, wofür der betreffende Artikel dann auf elektronischem Wege von jedermann kostenlos gelesen und heruntergeladen werden kann. Allerdings ist der Publikationspreis für den Autor mit 3.000,-$ bei Springer und 2.500,- $ bei Blackwell prohibitiv hoch.

Eine Konsequenz des autorenfinanzierten Open-Access-Modells führt dazu, dass für die Verlage – wenn sich dieses Modell weiter ausbreitete – eine gänzlich neue Wettbewerbssituation entstünde. Bleibt nach dem traditionellen abonnementgestützten Modell die Lesernachfrage weitgehend unberührt von der Höhe des Ladenpreises einer Zeitschrift, so hat im autorenfinanzierten Modell die Höhe des Publikationspreises unmittelbare Auswirkungen auf das Verhalten der Autoren. Universitätsprofessoren werden ihre Entscheidung, in welcher Zeitschrift sie publizieren, künftig davon abhängig machen, welche Zeitschrift ihres Fachgebiets den besten Impact-Factor für den besten Preis bietet. Auch wenn die Autoren das Geld nicht aus eigener Tasche bezahlen müssen, sind die Publikationskosten ein wichtiger Faktor, weil sie in das zur Verfügung stehende Forschungsbudget mit einkalkuliert werden müssen. Auf diese Weise könnten auch Verleger von „Must have“-Zeitschriften ihre Publikationspreise nicht beliebig hochschrauben, da in dieser Situation der Preis dem Rang der Zeitschrift innerhalb ihres Konkurrenzumfelds angepasst werden muss.

Nach Einschätzung des Science and Technology Committee des House of Commons überwiegen klar die volkswirtschaftlichen Vorteile des autorenfinanzierten Modells. Das Modell wird nachdrücklich unterstützt und die Research Councils werden aufgefordert, Gelder für die Publikation von Artikeln in autorenfinanzierten

[9] Vgl. die Diskussion im House of Commons Report a.a.O. Abs. 144 – 150.

[10] http://www.biomedcentral.com. BioMedCentral hat über 100 biomedizinische Open-Access-Zeitschriften in seinem Verlagsprogramm, am bekanntesten ist das „Journal of Biology“. Den Autoren werden 500,-$ je Artikel abverlangt.

[11] http://www.publiclibraryofscience.org. Die PloS führt 5 Zeitschriften, die Autoren müssen 1.500,-$ je Artikel bezahlen.

[12] Siehe http://www.springeronline.com/sgw/cda/frontpage/0,11855,1-109-2-116802-0,00.html.

[13] Siehe http://www.blackwellpublishing.com/static/onlineopen.asp.

Zeitschriften bereit zu stellen.[14] Das britische Parlament ist dieser Empfehlung allerdings nicht gefolgt und hat das Thema an die Research Councils zurück verwiesen[15]. Die Research Councils wiederum wollen das autorenfinanzierte Open-Access-Modell nicht einseitig bevorzugen und stellen für reine Publikationskosten keine Gelder zur Verfügung, es sei denn, die Publikationskosten seien schon von Anfang an Teil eines ganzen Forschungsprojektes.[16] Eine ähnliche Politik verfolgt im Übrigen seit Ende 2001 auch die Deutsche Forschungsgemeinschaft: Reine Druck- oder Publikationsbeihilfen als solche gibt es nicht mehr, die Publikationsförderung kann aber im Zusammenhang mit einem Projekt oder Stipendium beantragt werden.[17]

3. Selbstarchivierung am Institut

„Universities and research centers throughout the world are actively planning the implementation of institutional repositories. Such planning entails policy, legal, educational, cultural, and technical components, most of which are interrelated and each of which must be satisfactorily addressed for the repository to succeed".[18]
Mit dieser Äußerung hat die Budapest Open Access Initiative bereits 2001 eine Entwicklung beschrieben, die die Open-Access-Bewegung neben dem autoren-finanzierten Modell besonders deutlich vorangebracht hat. Seither ist der Aufbau digitaler Archivserver durch Universitäten und andere Forschungseinrichtungen weltweit weiter vorangeschritten. In den USA existiert bereits seit 1991 der E-Print-Server „ArXiv" für den Bereich Physik[19], daneben bauen Universitäten ihre frei zugänglichen digitalen Archive ständig aus, wie z. B. MIT[20] oder das California Institute of Technology (Caltech)[21]. In Großbritannien wurde die SHERPA-Initiative ins Leben gerufen, ein Projekt von 18 Universitäten und Forschungseinrichtungen mit dem erklärten Ziel, 13 institutionale Open Access E-Print-Archivserver aufzubauen

[14] „We recommend that the research councils each establish a fund to which their funded researchers can apply should they wish to publish their articles using the author-pays model." House of Commons a.a.O. Abs. 165.

[15] HOC, 14th Report, Response from the Government, Abs. 64; siehe http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/1200/120006.htm#a1.

[16] So hat sich Dr. Michael Jubb, der scheidende Vorsitzende der Arts and Humanities Research Councils, mir gegenüber geäußert (Januar 2005).

[17] Siehe http://www.dfg.de/aktuelles_presse/pressemitteilungen/2001/presse_2001_53.html.

[18] Budapest Open Access Initiative, http://www.soros.org/openaccess/software/.

[19] ArXiv enthält mittlerweile rund 300.000 frei zugängliche Forschungsartikel, hauptsächlich auf dem Gebiet der Physik, aber auch in der Mathematik, den Non-linear Sciences, Computer Sciences und Quantitative Biologie: http://arxiv.org/.

[20] https://dspace.mit.edu/index.jsp.

[21] http://library.caltech.edu/digital/.

und miteinander zu vernetzen.[22] In den Niederlanden wurde im Jahre 2000 die Initiative „Academical Research in the Netherlands Online“ (ARNO) gegründet, die Forschungseinrichtungen erlaubt, ihre Archivdateien digital zu managen.[23] Im September 2004 haben die Max-Planck-Gesellschaft und das Fachinformationszentrum Karlsruhe bekannt gegeben, dass das Bundesministerium für Bildung und Forschung 6,1 Mio. Euro zur Verfügung stellt, um über die nächsten fünf Jahre eine große vernetzte Plattform zu entwickeln, auf der u. a. auch alle durch die Max-Planck-Gesellschaft geförderten Publikationen frei zur Verfügung stehen sollen.[24]
In den Aufbau und die internationale Vernetzung digitaler Archivserver fließen also große Mengen öffentlicher Gelder. Um ein international einheitliches Datenformat für alle institutionalen Archivserver zu schaffen, wurde das „Open Archives Initiative Protocol for Metadata Harvesting“ (OAI-PMH) entwickelt. Alle OAI-PMH-kompatiblen Archive bilden somit ein großes virtuelles Forschungsarchiv, so dass es mittlerweile sogar möglich ist, über spezielle OAI-Serviceprovider weltweit in vielen digitalen Archiven gleichzeitig nach Inhalten zu suchen. Ziel ist der Aufbau eines großen virtuellen „Superarchivs“, in dem alle Forschungsergebnisse der Welt allen Forschern online frei zur Verfügung stehen.[25] In der Realisierung begriffen ist bereits das Archiv „OAIster“[26], das bislang aber in der wissenschaftlichen Forschung noch so gut wie keine Rolle spielt.
Das größte Hindernis zum Aufbau eines solchen digitalen Superarchivs ist derzeit noch die Copyrightfrage. Verlage lassen sich die Nutzungsrechte von den Autoren in der Regel exklusiv und umfassend übertragen und sind ausgesprochen zurückhaltend, wenn es darum geht, dass die Artikel gleichzeitig in einem digitalen Archivserver kostenfrei zur Verfügung gestellt werden sollen. Das House of Commons kommt dementsprechend auch zu der dringenden Empfehlung an die Forschungseinrichtungen, sie sollten es doch zur Bedingung für die Vergabe von Forschungsgeldern machen, dass eine digitale Kopie der Forschungsergebnisse jeweils dem Institut zur Einspeisung in den Archivserver zur Verfügung gestellt wird.[27] Die britische Regierung ist dieser Empfehlung zwar nicht gefolgt, unterstützt aber

[22] http://www.sherpa.ac.uk/.

[23] http://arno.uvt.nl/~arno/site/index.html. Die ARNO-Software wird derzeit verwendet von der University of Amsterdam, Twente University, University of Tilburg, Maastricht University, Erasmus University of Rotterdam und Fontys Hoogescholen.

[24] http://www.mpg.de/bilderBerichteDokumente/dokumentation/pressemitteilungen/2004/pressemitteilung200409061/.index.html.

[25] Siehe Jeffrey R. Young: 'Superarchives' Could Hold All Scholarly Output. Online collections by institutions may challenge the role of journal publishers, http://chronicle.com/free/v48/i43/43a02901.htm.

[26] OAIster verzeichnet über 5 Millionen Dokumente von über 400 Instituten, darunter aber auch so genannte „graue“ Literatur wie Beschreibungen einzelner Kurse, Statusberichte, NASA-Dokumente u.ä. http://oaister.umdl.umich.edu/o/oaister.

[27] House of Commons Report, a.a.O., Abs. 117.

nachdrücklich den Auf- und Ausbau institutionaler oder thematischer Archivserver auch mit öffentlichen Mitteln.[28] Zu der Haltung, Forscher zu zwingen, ihre Ergebnisse in digitalen Archivservern der Öffentlichkeit kostenlos zugänglich zu machen, neigt auch die Deutsche Forschungsgemeinschaft, und auch die wissenschaftlichen Akademien in Deutschland tendieren in diese Richtung.
In den USA ist das National Institute of Health (NIH) bereits einen Schritt weiter gegangen. Ab dem 2. Mai 2005 werden alle Forscher, deren Arbeit ganz oder teilweise von dem NIH finanziert wird, aufgefordert, die Endfassung ihrer Artikel dem Institut zur kostenlosen elektronischen Veröffentlichung zur Verfügung zu stellen. Die Autoren entscheiden selbst über den Zeitpunkt der digitalen Veröffentlichung, das NIH drängt jedoch seine Autoren, die Artikel so bald wie möglich nach Veröffentlichung durch einen Verlag elektronisch zur Verfügung zu stellen, auf jeden Fall aber noch innerhalb der ersten zwölft Monate.[29]

4. Die Zukunftsfähigkeit der Open-Access-Modelle

Welches der genannten Open-Access-Modelle wird sich eher durchsetzen, oder werden vielleicht beide Modelle nebeneinander existieren? Ich meine, dass die Selbstarchivierung am Institut die wesentlich besseren Überlebenschancen hat und das autorenfinanzierte Modell eher als Übergansmodell zu betrachten ist. Dies lässt sich aus den folgenden Gründen herleiten:
Zum Ersten gibt es Hinweise darauf, dass die Kosten für das Publizieren eines einzelnen Artikels deutlich über den 1.500,-$ liegen, die derzeit von der Public Library of Science gefordert werden. Um eine Zeitschrift kostendeckend zu betreiben, müsste möglicherweise sogar noch mehr als die derzeit von Springer geforderten 3.000,-$ je Artikel verlangt werden.[30] Aber selbst bei einem Höchstbetrag von 1.500,-$ je Artikel gibt es deutliche Anzeichen, dass das autorenfinanzierte Modell volkswirtschaftlich betrachtet keineswegs günstiger ist als die herkömmliche Praxis, eher sogar noch teurer. Eine Studie der Cornell University Library hat ergeben, dass sich eine durchgängig autorenfinanzierte Praxis jedenfalls für Cornell

[28] "However the Government has no present intention to mandate Research Council funded researchers to deposit a copy of their published material in institutional repositories". House of Commons, Government Response, a.a.O. Paragraph 44. Siehe aber auch: "…there may need to be additional investment by research councils to fund data facilities made available to support this objective. Institutional or thematic repositories should provide a useful environment for disseminating such information …", a.a.O. Paragraph 7.

[29] NIH, Notiz vom 3.2.05: Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-022.html.

[30] Siehe Donald W. King and Carol Tenopir: An evidence-based assessment of the 'author pays' model http://www.nature.com/nature/focus/accessdebate/26.html.

nicht lohnen würde. Ausgangspunkt der Studie war die Annahme, dass alle derzeit von der Cornell University abonnierten Zeitschriften auf ein autorenfinanziertes Open-Access-Modell umgestellt werden würden. Ferner wurde im Rahmen der Studie angenommen, dass die Universität ihre Forscher für die Publikation ihrer Artikel mit durchschnittlich je 1.500,-$ unterstützen müsste bei einer Gesamtzahl von etwas über 3.500 Artikeln je Jahr. Es wurde errechnet, dass dieses Modell die Cornell University rund 1,5 Mio. $ *mehr* kosten würde als jetzt für alle Zeitschriften-subskriptionen insgesamt ausgegeben wird.[31] Ob das autorenfinanzierte Modell volkswirtschaftlich gesehen wirklich günstiger ist als das traditionelle, ist also eher fraglich.

Zum Zweiten ändert sich, wie schon erwähnt, die Art des Wettbewerbs bei den Verlagen, was für manche Zeitschriften möglicherweise fatale Folgen hätte. Da unter der autorenfinanzierten Praxis nicht der Abonnent den Verlag für die Publikation bezahlen würde, sondern der Autor bzw. die Institution, die seine Forschung finanziert, entstünde ein Wettbewerb, der auf das beste Verhältnis zwischen dem Ranking der Zeitschrift und dem Publikationspreis, den der Autor bzw. sein Institut zahlen muss, hinausliefe. Setzte sich dieses Modell als das dominierende künftige System wissenschaftlicher Kommunikation durch, würde dies für Verlage bedeuten, dass für qualitativ hochwertige Zeitschriften sehr hohe Publikationspreise verlangt werden könnten, während weniger bedeutende Zeitschriften von der Einstellung bedroht wären, weil Autoren bzw. ihre Institute nicht bereit wären, für die relativ geringe Qualität der Zeitschriften die Preise zu zahlen, die erforderlich wären, um sie wirtschaftlich zu betreiben. Wenn von den Autoren bzw. ihren Instituten schon sehr hohe Publikationspreise abverlangt würden, dann würden diese auch erwarten, dass ihre Artikel in Zeitschriften mit hohen Impact Factor publiziert werden. Zeitschriften mit geringem oder gar keinem Impact Factor würden vom Markt verschwinden. Schlimmer noch: Selbst Zeitschriften mit hohem Impact Factor wären von der Einstellung bedroht, wenn es sich um Spezialzeitschriften auf einem kleinen Forschungsgebiet handelt, auf dem weltweit nur wenige Spezialisten arbeiten. Diese Zeitschriften haben naturgemäß nur kleine Auflagen und wenige Autoren je Jahrgang. Die Erlöse, die erforderlich wären, um die Zeitschrift wirtschaftlich zu betreiben, müssten von sehr wenigen Autoren erbracht werden, was schnell in fünfstellige Beträge gehen kann, die kein Autor und kein Institut mehr zu bezahlen bereit wären. Nach dem subskriptionsgestützten Modell würden sich die erforderlichen Erlöse immerhin auf deutlich mehr institutionelle Abonnenten verteilen.

Zum Dritten ist das autorenfinanzierte Modell nur für die STM-Bereiche geeignet. Es ist nicht damit zu rechnen, dass es sich sehr weit in die Geisteswissenschaften hinein ausdehnen wird. In den Naturwissenschaften pflegt allein schon der vielen

[31] Report of the CUL Task Force on Open Access Publishing Presented to the Cornell University Library Management Team August 9, 2004, http://dspace.library.cornell.edu/bitstream/1813/193/3/OATF_Report_8-9.pdf.

Instrumente und teuer ausgestatteten Labors wegen sehr viel mehr Geld zu fließen und es ist auch sehr viel geläufiger, dass Forschungseinrichtungen und sogar die Industrie gezielt einzelne Projekte mit sehr großen Summen unterstützen. All dies gilt für die Geisteswissenschaften nicht, so dass es relativ schwer sein wird, Sponsoren für ein autorenfinanziertes Modell zu finden.
Wie steht es demgegenüber mit der Selbstarchivierung am Institut? Auch das Selbstarchivierungsmodell ist volkswirtschaftlich betrachtet vermutlich eher ungünstig, betrachtet man nur die vielen Einzelinitiativen zum Aufbau digitaler Archive, die alle das Rad neu erfinden. Tatsache ist aber, dass überall auf der Welt massiv daran gearbeitet wird, digitale Archive aufzubauen. Und angesichts der Millionen von Dollars und Euros, die in diese Initiativen investiert wurden und werden, wird man sich sehr schwer tun, diese einfach wieder einzustellen. Sind die Archivserver aber erst einmal eingerichtet, wird sich das Modell schnell auf alle Wissenschaftsbereiche ausdehnen. Auch die Geisteswissenschaftler werden dann vermutlich die vorhandenen Infrastrukturen für ihre eigenen Inhalte nutzen. Anders als das autorenfinanzierte Modell hat das Selbstarchivierungsmodell daher die wesentlich besseren Überlebenschancen.
Auch ein Mangel an Kompatibilität zwischen den einzelnen Archiven ändert daran wenig. Zum einen wird viel Geld in die Entwicklung kompatibler Formate investiert, zum anderen ist abzusehen, dass die Wissenschaft sich in erster Linie auf fachgebietsorientierte Archive stützen wird. Es wird langfristig weniger darauf ankommen, ein „Superarchiv“ sämtlicher Dokumente aller Institute weltweit aufzubauen, sondern durch Konsortien und Zusammenschlüsse unter den Instituten für einzelne Disziplinen eigene Spezialarchive zu entwickeln. Die Forscher würden sich auf die Nutzung eines Archivs in ihrem Spezialgebiet beschränken, wie es in einigen Disziplinen bereits der Fall ist. Diese Spezialarchive müssten in sich natürlich konsistent sein, es wäre aber keine Kompatibilität zu anderen Archiven erforderlich.

5. Kommunikation in der Wissenschaft

Wie kann die wissenschaftliche Kommunikation grundsätzlich neu organisiert werden mit dem Ziel, die Schwierigkeiten des traditionellen Systems zu überwinden, ohne deren Vorteile aufgeben zu müssen?
Einvernehmen besteht in der Frage, dass jedes System wissenschaftlicher Kommunikation die folgenden fünf Kriterien erfüllen muss:

- Registration: Jedes Forschungsergebnis muss registriert und dem Urheber mit dem Datum der Entdeckung zugeordnet werden können.
- Certification: Jedes Forschungsergebnis muss auf seine Qualität hin geprüft werden.

- Awareness: Sobald ein Forschungsergebnis registriert und auf seine Qualität hin geprüft wurde, muss sichergestellt sein, dass alle relevanten Forscher sich dieses Forschungsergebnisses bewusst werden und bleiben.
- Archiving: Das Forschungsergebnis muss für künftige Generationen bewahrt werden.
- Rewarding: Schließlich erwartet der Forscher auch eine Anerkennung für die von ihm geleistete Arbeit.

Sowohl in dem traditionellen subskriptionsgestützten Modell als auch in dem autorenfinanzierten Modell wird die Funktion Registration durch den Verlag und die Funktion Archiving durch die Bibliothek vorgenommen. Für beide Modelle lässt sich die Wertschöpfungskette wie folgt darstellen (Abbildung 1):

Abbildung 1: Wertschöpfungskette im traditionellen subskriptionsgestützten Modell

Der Autor reicht seinen Artikel bei dem Verlag ein, der ihn einem Begutachtungsprozess unterzieht und anschließend in einer Zeitschrift veröffentlicht. Durch die Veröffentlichung wird das Forschungsergebnis registriert. Die Zeitschrift wird anschließend meist über Subscription Agents an die Universitätsbibliothek verkauft, die die Publikation archiviert und ihren Lesern zur Verfügung stellt. Die Gruppe der Leser ist weitgehend identisch mit der Gruppe der Autoren, nämlich die Wissenschaftler auf dem betreffenden Fachgebiet. Die Funktion „Certification“ wird zu einem Großteil durch den Begutachtungsprozess vorgenommen, findet aber auch nach Veröffentlichung durch die in der Regel dann eintretende Diskussion unter den Wissenschaftlern statt. Die Funktion „Awareness“ wird zum einen durch den Verlag mittels seiner Vertriebspartner sichergestellt, zum anderen aber auch durch die Diskussion in der Wissenschaftsgemeinschaft. Der Urheber erhält seine Anerkennung einerseits durch die Universität, andererseits aber auch durch die Reputation der Zeitschrift, in der er publiziert hat.

Möchte man dieses System wissenschaftlicher Kommunikation verbessern, dann liegt es zunächst nahe, es drastisch zu vereinfachen, indem alle zwischengeschalteten Schritte zwischen Autor und Leser weggelassen werden (Abbildung 2).

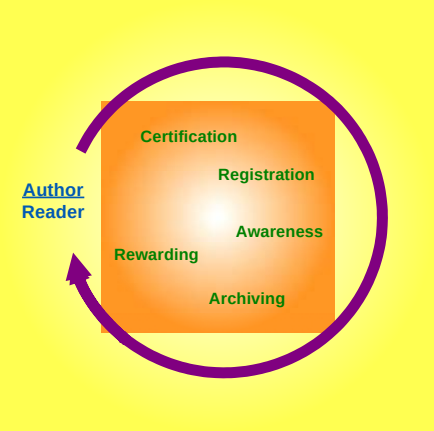


Abbildung 2: Vereinfachung der Wertschöpfungskette

Dieses Modell existiert und funktioniert auch tatsächlich, nahe liegender Weise allerdings nur in hoch spezialisierten Fachgebieten, auf denen es weltweit nur wenige Experten gibt, die sich alle gegenseitig kennen und miteinander vernetzt sind. Alle fünf Kernfunktionen können in diesen kleinen Wissenschaftsgemeinschaften erfüllt werden, selbst die Archivierung. Auf dem Gebiet der Hochenergiephysik sowie speziellen Bereichen der Mathematik steht z.B. der oben schon erwähnte ArXiv-Server zum Archivieren zur Verfügung. Auch in den Geisteswissenschaften gibt es Spezialgebiete mit wenigen Experten weltweit, die ohne jegliche Zwischenschritte bei ihrer Kommunikation auskommen.
Sobald aber das Fach breiter und die Zahl der beteiligten Forscher größer wird, kommt die wissenschaftliche Kommunikation ohne vermittelnde Zwischenschritte nicht aus. Ein ernsthaft diskutierter Vorschlag geht sogar von der Abschaffung der Bibliotheken innerhalb der Wertschöpfungskette aus.[32] Allerdings müsste die Archivierungsfunktion bei einem solchen System durch die Verlage übernommen werden. Die damit zu übernehmende Verantwortung und ökonomischen Risiken liegen aber nicht im Interesse der Verlage. Deutlich weiter verbreitet ist daher die Meinung, die Universitäten hätten auch selbst das nötige Know-how, einen

[32] Hans E. Roosendaal, P. A. Th. M. Geurts and P. E. van der Vet: Integration of Information for Research and Education: Changes in the Value Chain, in: Serials, vol. 15, no. 1, March 2002, p. 54, http://uksg.metapress.com/media/7HXDAF81WH0UWKF9EAAK/Contributions/9/C/A/1/9CA1XWFT1UNWERNH.pdf

Begutachtungsprozess zu steuern, so dass man also auch völlig ohne externe Verlage auskommen würde (Abbildung 3).

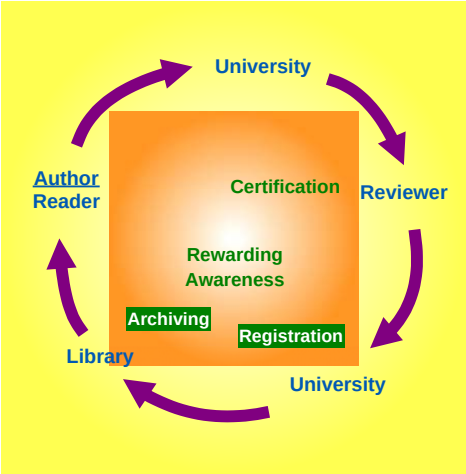


Abbildung 3: Begutachtungsprozess ohne externe Verlage

Die große Schwierigkeit bei diesem Modell besteht aber darin, dass die Autoren und ihre Universitäten dieselben Interessen haben, so dass ein neutraler und interessenunabhängiger Begutachtungsprozess nicht unbedingt gewährleistet ist. Die Frage, wer den Gutachter nach welchen Kriterien auswählt, ist für die Qualität des Begutachtungsprozesses von entscheidender Bedeutung, und selbst die größten Verfechter des selbstarchivierenden Open-Access-Modells legen den größten Wert auf ein strikt unabhängiges und qualitativ hochwertiges Begutachtungsverfahren. Die Organisation eines solchen unabhängigen Begutachtungsverfahrens wird auch für die Zukunft als eine der vornehmsten und wichtigsten Rollen für die Verlage betrachtet.[33]

Überhaupt ist es wenig realistisch anzunehmen, dass der eine oder andere Akteur in der Wertschöpfungskette völlig verschwinden wird. Und auch die Erfahrung lehrt, dass selbst bei revolutionären Veränderungen nur sehr selten ganze Gewerbe völlig untergehen. Mit sehr großer Wahrscheinlichkeit werden daher zwar die Akteure bleiben, allerdings mit jeweils anderen Aufgaben und Rollen.

Der Aufbau frei zugänglicher institutionaler Archivserver ist allein schon auf Grund der dort hinein investierten Energie und öffentlichen Mittel nicht aufzuhalten. Es ist

[33] Eher abwegig ist die ebenfalls vertretene Meinung, der Begutachtungsprozess ließe sich automatisieren, indem man einen Artikel durch die Anzahl der Zitate, die auf ihn verweisen, bewerten lässt. Die Internet-Suchmaschine „Google“ arbeitet z. B. nach diesem Prinzip. Um aber ein wirklich aussagefähiges Urteil zu erhalten, wird man kaum um einen von Personen durchgeführten Begutachtungsprozess herumkommen.

absehbar, dass in naher Zukunft jeder Wissenschaftler mindestens eine Kopie seiner Arbeiten in einem Institutsarchiv digital abzulegen verpflichtet sein wird. Dort stehen sie der Welt zwar kostenfrei zur Verfügung, aber es bleibt nach wie vor die Aufgabe der Verlage, für einen unabhängigen und qualitativ hochwertigen Begutachtungsprozess zu sorgen, wonach sie über die Universitätsbibliotheken an die Leser gelangen. Mit einer gewissen Wahrscheinlichkeit wird künftig die Wertschöpfungskette daher wie folgt aussehen (Abbildung 4):

Abbildung 4: Wertschöpfungskette mit hochwertigen Begutachtungsprozess durch die Verlage und frei zugänglicher institutionalem Archivserver

Das Copyright verbleibt dabei entweder beim Autor oder – wahrscheinlicher – bei der Institution, bei der angestellt ist. Nach diesem Modell wird es Aufgabe der Verlage sein, die in den vielen Institutsservern vorhandenen Informationen zu selektieren und einen Mehrwert dadurch zu schaffen, dass sie sie einem Begutachtungsprozess unterwerfen, miteinander verknüpfen, indexieren und der relevanten Zielgruppe unter einem Qualität bürgenden Markennamen zur Verfügung stellen. Verlage werden aber nicht mehr das Privileg für sich beanspruchen können, der Wissenschaft originale und bisher unveröffentlichte Forschungsliteratur bereitzustellen.
Dieses Szenario würde zu neuen Rollen der Akteure in einer geänderten Wertschöpfungskette führen. Universitäten und ihre Bibliotheken würden nach wie vor für die Archivierung der Inhalte sorgen, würden zusätzlich aber auch neue wissenschaftliche Arbeiten registrieren und auf ihren Servern zur Verfügung halten, damit auch etwas zur Funktion „Awareness“ beitragen. Verlage würden nach wie vor die Organisation des Begutachtungsprozesses übernehmen, zugleich würde es für sie aber auch den Aufbau neuer Geschäftsmodelle bedeuten: Aus den vielen frei zur Verfügung stehenden digitalen Inhalten der Universitätsarchive müsste der Verlag

einen Mehrwert schaffen durch Filtern, Selektieren, Begutachten, Indexieren und Aggregieren. Das Resultat wäre im Zweifelsfall nicht mehr als eine Sammlung von Links zu den Inhalten sehr unterschiedlicher Archivserver in der ganzen Welt. Der Mehrwert einer solchen Sammlung bestünde darin, dass der Verlag garantieren kann, dass die ausgewählten Informationen das umfassendste, aktuellste und einschlägigste Material zu einem bestimmten Themengebiet darstellt. Wenn ein derartiger Service einem Forscher eine Menge an Zeit und Recherchearbeit erspart, so wäre er (oder seine Bibliothek) sicherlich dafür auch zu zahlen bereit. Geschäftsmodelle dieser Art sind nicht grundsätzlich neu, sie sind neu nur auf diesem Gebiet. So verdienen z. B. jetzt schon Verlage wie die Thomson Corporation viel Geld mit der Aufbereitung und Indexierung von Patenten, die im Übrigen auf den Websites der jeweiligen Patentämter kostenlos zur Verfügung stehen.

Derartige Leistungen gehören zwar zu den Kernkompetenzen eines wissenschaftlichen Verlages, machen aber nur einen Teil des Verlagswesens aus, und es wäre sicherlich sehr unbefriedigend, würde man als wissenschaftlicher Verleger ausschließlich auf diese Leistungen beschränkt werden. Es spricht aber einiges dafür, dass in vielen Bereichen das Abonnementgeschäft noch bestehen bleiben wird, ähnlich wie die Erfindung des Fernsehens das Radio zwar zurückgedrängt, aber nicht gänzlich ersetzt hat. In den Bereichen, in denen das abonnementgestützte Modell gut funktioniert, gibt es schließlich für keine der beteiligten Parteien irgendeinen Grund, ein anderes System einzuführen. Warum sollten also auch in Zukunft nicht beide Modelle nebeneinander existieren können?

Wenn das so ist, dann würden wir uns in Zukunft einer durchaus größeren Bandbreite an Möglichkeiten wissenschaftlicher Kommunikation gegenübersehen. Universitäten übernehmen in bestimmten Bereichen Aufgaben, die ursprünglich Verlagsaufgaben waren, müssen dabei aber erhebliche Investitionen leisten, wobei nicht wirklich absehbar ist, dass dieser Weg langfristig volkswirtschaftlich günstiger ist. Verlage andererseits müssen die damit verbundenen Umsatzeinbußen durch den Aufbau neuer Geschäftsmodelle kompensieren. Insgesamt aber werden für beide Akteure innerhalb der Wertschöpfungskette die Aufgabenfelder eher reicher als ärmer, und für die Wissenschaftler die Möglichkeiten der Kommunikation eher vielfältiger.

Herausforderungen Wikipedia und Open Access – können Verlage etwas lernen von den Strategien angesichts Linux & Co. ?

Christoph Bläsi, Erlangen

Abstract

Lexikonverlagen erwächst in der freien Enzyklopädie Wikipedia zunehmend Konkurrenz – in der Gunst der Nutzer auf jeden Fall. Das community-basierte Entwicklungsmodell der Wikipedia umgeht auf dem Weg zwischen Autoren und Lesern die Wertschöpfungsstufe Verlag. Angesichts der Tatsache, dass auf dem Software-Markt mit Open-Source-Software schon seit einigen Jahren ein vergleichbares Phänomen eine wichtige Rolle spielt, untersucht dieses Paper, inwiefern Lexikonverlage davon lernen können, wie betroffene Softwareunternehmen sich in der neuen Situation einrichten.

Einleitung

Nutzer von Enzyklopädien[1] haben in der Wikipedia eine Alternative zu Offline-Lexika als Büchern oder auf Datenträgern sowie Web-Angeboten von Verlagen[2] – schon diesseits einer detaillierteren Bewertung und Abwägung von Produkteigenschaften, hat Wikipedia dabei einen ganz entscheidenden Wettbewerbsvorteil: Sie ist für die Nutzer kostenfrei. Es darf – Beteuerungen des im Wesentlichen unbeeindruckten Selbstbewusstseins zum Trotz[3] – vermutet werden, dass diese Tatsache im Lexikonbereich tätigen Verlagen Kopfzerbrechen bereitet.
In der Software-Branche macht der dem Wikipedia-Ansatz sehr ähnliche (ja, diesem als Idee zugrunde liegende) Open-Source-Software-Ansatz seit vielen Jahren den großen Software-Unternehmen mit kostenfreien und leistungsfähigen Produkten wie Linux, Mozilla und Apache Konkurrenz.

[1] Zur Klassifikation sachlexikographischer Werke vgl. Kapitel 2.1 „Enzyklopädien und Lexika – konstitutive und variante Eigenschaften“ in Bläsi 1998; ich werde „Lexikon“ und „Enzyklopädie“ im Folgenden synonym verwenden.

[2] Das vor wenigen Jahren selbst noch als Herausforderer angestammter Lexikonverlage aufgetretene Software-Unternehmen Microsoft werde ich – aufgrund seines Produktes „Encarta“ (Offline-.Enzyklopädie auf Datenträger) – im gegebenen Zusammenhang argumentativ wie einen (Lexikon-)Verlag behandeln.

[3] Vgl. z.B. Alexander Bob vom Brockhaus-Verlag nach Dambeck 2005a.

Ich habe es mir zur Aufgabe gestellt, zunächst zu untersuchen und zu klassifizieren, mit welchen Strategien von Open-Source-Software besonders bedrängte Software-Unternehmen auf diese Herausforderung reagieren. Dies als Folie nehmend werde ich dann systematisch Handlungsoptionen für Lexikonverlage angesichts der Wikipedia aufzeigen, das jedoch durch spezifische Aspekte anreichern.
Ein wissenschaftlicher Diskurs zur Wikipedia ist – oft aufbauend auf Argumentationen aus Marketing bzw. Public Relations der beiden Seiten – erst im Entstehen. Im Gegensatz dazu darf das Umfeld der Open-Source-Software, insbesondere auch deren ökonomische Auswirkungen betreffend, als gut erschlossen gelten (vgl. z.B. Brügge et al. 2004 oder Gehring / Lutterbeck 2004) – auf wichtige Erkenntnisse aus diesem Forschungsstrang werde ich im Folgenden näher eingehen.

Wikipedia

Die freie Enzyklopädie Wikipedia ist in den letzten Jahren auch in ihrer deutschen Ausgabe (im Web unter www.wikipedia.de) – und auf diese werde ich mich beziehen – zu einem bedeutenden Anbieter auf dem Markt für enzyklopädisches Wissen geworden – mit nach eigenen Aussagen über 265.000 Artikeln und um 500.000 Besuchern täglich. Das Konzept der Wikipedia beruht darauf, dass jeder Nutzer – nach Registrierung – mit einem Browser-basierten Web Content Management System auch als Autor bzw. Redakteur zu der Enzyklopädie beitragen kann; jeder Nutzer stimmt dabei mittels der „GNU-Lizenz für freie Dokumentation“[4] einer Vervielfältigung, Verbreitung und Veränderung seiner Hervorbringungen zu, verpflichtet etwaige Lizenznehmer aber seinerseits z.B. auch dazu, Abgeleitetes wieder unter die gleiche Lizenz zu stellen („Copyleft-Prinzip“).
Von der Wikipedia gibt es mittlerweile auch eine CD-ROM-/DVD-ROM-„Distribution“; thematische Printprodukte mit Wikipedia-Content sind für die zweite Jahreshälfte 2005 geplant. Wikimedia, die durch Spenden und die erwähnten Offline-Derivate finanzierte Dachorganisation der freien Informationsquellen, hat noch weitere Projekte in Arbeit, von denen im hier gegebenen lexikographischen Zusammenhang Wiktionary, ein mehrsprachiges Wörterbuch, sowie Wikiquote, eine Sammlung bekannter Zitate, von Bedeutung sein dürften.
Die argumentative Auseinandersetzung mit der Wikipedia aufgenommen hat von den Anbietern kostenpflichtiger Enzyklopädie-Angebote in letzter Zeit mehrfach und

[4] „Diese Lizenz gestattet die Vervielfältigung, Verbreitung und Veränderung des Werkes, auch zu kommerziellen Zwecken. Im Gegenzug verpflichtet sich der Lizenznehmer zur Einhaltung der Lizenzbedingungen. Diese sehen unter anderem die Pflicht zur Nennung des Autors bzw. der Autoren vor und verpflichten den Lizenznehmer dazu, abgeleitete Werke unter die selbe Lizenz zu stellen (Copyleft-Prinzip). Wer sich nicht an die Lizenzbedingungen hält, verliert damit automatisch die durch die Lizenz eingeräumten Rechte.“ (Wikipedia-Eintrag “GNU-Lizenz für freie Dokumentation”).

öffentlich der Brockhaus-Verlag. Von dessen Seite werden in erster Linie – dies hier nur ansatzweise, um kein einseitiges Bild der Wikipedia entstehen zu lassen – Bedenken die Verlässlichkeit und Ausgewogenheit betreffend vorgebracht („Einzelne Artikel sind zwar sehr gut, aber das Problem ist, ich weiß nicht, ist das nun gerade ein guter Artikel oder nicht. Kann ich mich darauf verlassen ?“[5]), sowie darauf hingewiesen, dass die Brockhaus-Enzyklopädie von professionellen Schreibern stammt (was das für Folgen für das Produkt haben könnte, darf der Leser in manchen Stellungnahmen selbst erschließen). Der Verlässlichkeits- und Ausgewogenheitsaspekt hat dabei insofern Bedeutung auch über den Einzelartikel hinaus, als „die Gesamtstruktur der Artikel [in der Wikipedia, C.B.] nicht den Proportionen einer Enzyklopädie [entspricht]. Dort [d.h., in der Brockhaus-Enzyklopädie, C.B.] sind die Themen so verteilt und die Artikel so gewichtet, dass sie der realen Bedeutung der Gegenstände [sic !] entsprechen. Bei ´Wikipedia´ sind Zahl und Umfang der Artikel hingegen vom Zufall oder vom aktuellen Interesse abhängig.“[6] Darüber hinaus wird einem wichtigen von der Wikipedia in eigener Sache in Anspruch genommenen Alleinstellungsmerkmal entgegengehalten: „Der Aktualisierungsbedarf von Lexika wird im Allgemeinen überschätzt.“[7] Grundlegender und mutmaßlich auf jeden Fall nicht interessengeleitet fragt zu diesem Themenkomplex Thomas Thiel: „Besteht die Befreiung des Wissens im Zeitalter seiner digitalen Modifizierbarkeit nicht vor allem in der Garantie seiner Qualität und weniger in seiner schieren Masse und schnellen Verfügbarkeit ? Wer schafft das Vertrauen in die jederzeit manipulierbaren Realitäten des Netzes, wenn nicht eine strenge Qualitätskontrolle ?“[8] Als Antwort auf Argumente in dieser Richtung verweist Wikipedia im Übrigen v.a. auf die lückenlose und für jeden einsehbare Änderungsgeschichte jedes Artikels, die es einem mündigen Leser ermöglichen sollte, sich positionsbasiert ein umfassendes und eigenes Bild des jeweiligen Gegenstandes zu machen – „im Herzen des Projekts steckt somit die Demokratisierung der Wahrheitsfindung“[9]. Die mit diesen Positionen und Gegenpositionen nur umrissene grundlegende Diskussion kann hier nicht vertieft werden.

Philosophisches

Sowohl der freien Enzyklopädie als auch der Open-Source-Software (sowie anderen Projekten des Arbeitens in virtuellen Gruppen unter Verzicht auf geistige Eigentumsrechte[10]) liegen nicht nur pragmatische Motive zugrunde. Es gibt vielmehr

[5] Alexander Bob vom Brockhaus-Verlag nach Dambeck 2005a.
[6] „Gedruckte Enzyklopädien haben Zukunft“ 2005.
[7] Alexander Bob vom Brockhaus-Verlag nach Dambeck 2005a.
[8] Thiel 2005.
[9] A.a.O.
[10] Folgende weitere Projekte können beispielhaft erwähnt werden: „[...] die NASA Clickworkers (ein Projekt, bei dem Freiwillige Krater auf dem Mars klassifizieren),

„philosophische“ Grundlagen der diese Projekte betreibenden Bewegungen; zu diesen gehören nicht zuletzt libertäre und kommunistische Wurzeln, die sich allerdings sehr heterogen äußern: „Die Selbstbindung der Bewegung [...] ist multidimensional: z.B. Gemeinwohlorientierung, Freiheit, utopische Weltentwürfe, Autoritätsferne, Gegnerschaft, Gefolgschaft, Sendungsbewusstsein [...] oder Neugier.“[11] In jüngerer Zeit kommt explizite Unterstützung für diese Bewegungen aber auch von der akademischen Informationsethik. Kuhlen stellt z.B. eingangs seines Buches „Informationsethik“ (Kuhlen 2004) Fragen, die die „Informationsindustrie“, wie er sie oft nennt, ohne Zweifel provozieren: „Wem gehört Wissen ? Darf Wissen überhaupt jemandem gehören, wenn dadurch andere von der Nutzung der aus Wissen abgeleiteten Informationsprodukte ausgeschlossen werden ? Sichert nur die private Verfügung über Wissen und Information deren Nutzung und Weiterentwicklung, stimmt also die These von der tragedy of the commons, nach der jedes öffentliche Gut, wenn es keine Zugriffsrestriktionen dafür gibt, tendenziell durch Übernutzung vernichtet wird ? Oder muss es einen unverzichtbaren Bereich der commons geben, in dem weite Bereiche von Wissen und Information der Öffentlichkeit allen gehören, gerade weil sich Wissen, anders als andere Güter, im Gebrauch nicht verbraucht ?“[12] Kuhlen erkennt in der Zusammenfassung seines Buches zwar die Notwendigkeit an, „dass [...] mit Informationsprodukten Geld verdient werden kann“[13] und sieht Probleme für Volkswirtschaften, „wenn keine von der Allgemeinheit akzeptierten Geschäfts- und Organisationsmodelle entwickelt werden, auf deren Grundlage investitionsintensive Mehrwertleistungen zur Erstellung und zum Vertrieb von Informationsprodukten zu einem befriedigenden return on investment und zu entsprechenden Gewinn ermöglichenden Einnahmen führen können“[14] – seine Hauptstoßrichtung ist allerdings erkennbar, dass „alle geistigen Werke, welcher medialen Form auch immer [...] Teil des menschlichen kulturellen Erbes [sind] und [...] in den Bereich der commons [gehören]“[15]. Schließlich sei – so sein gesellschaftspolitisches Argument – „die Chance für einen hohen Innovationsgrad der Wirtschaft,

Slashdot (ein Forum zum Kommentieren und Klassifizieren von Artikeln) und das Projekt Gutenberg (Scannen und Korrekturlesen von Büchern, deren Urheberrechte abgelaufen sind). [...] Gemeinsames Charakteristikum dieser Projekte ist die freiwillige kollektive Innovation unter weitgehendem Verzicht auf private geistige Eigentumsrechte. Ähnlich wie in weiten Bereichen der wissenschaftlichen Produktion tauschen die Mitglieder dieser virtuellen Gemeinschaften untereinander Beiträge aus, ändern oder verbessern sie, ohne Lizenzverträge abschließen zu müssen.“ (Osterloh et al. 2004). Dazu kommen entsprechende Ansätze im Bereich des Fiktionalen (Literatur im Internet etc.), auf die ich hier aber nicht eingehen kann.

[11] Weber 2005.

[12] Kuhlen 2004, S. 10.

[13] A.a.O., S. 380.

[14] A.a.O., S. 381.

[15] A.a.O., S. 380.

für einen hohen Inventionsgrad der Wissenschaft und einen hohen Demokratisierungs-/Transparenzgrad des politischen Systems“ umso größer, „je freizügiger der Umgang mir Wissen jeder Art ist“[16]. Kuhlen propagiert Anreize[17] „zum Schaffen neuer Geschäfts- und Organisationsmodelle“[18] nicht zuletzt von Seiten des Staates – diese könnten „sich nicht auf eine künstliche Verknappung der Güter Wissen und Information abstützen[19]“, sondern es könnte im Gegenteil damit gerechnet werden, „dass Märkte umso größer und wirtschaftliches Handeln mit Wissen und Information umso erfolgreicher sein werden, je offener, freizügiger und flexibler die Nutzung von Wissen und Information betrieben werden kann“[20].

Ein weiterer, mit Vorstehendem überraschenderweise (noch ?) überschneidendungsfreier aktueller Diskurs, in den die Wikipedia eingeordnet werden kann und auch wird, ist der um den v.a. von Trendforschern benutzen Begriff der „Schwarm-Intelligenz“ (vgl. dazu den Konferenztitel „Schwarm-Intelligenz – die Macht der smarten Mehrheit“), was v.a. im Zusammenhang mit Rheingold 2002 gesehen werden muss[21]. Für eine andere Zusammenschau solcher Phänomene vgl. Möller 2005.

Den Bogen von der Wikipedia zur Open-Access-Bewegung spannt – über die offensichtlichen strukturellen Ähnlichkeiten hinaus – explizit nicht zuletzt die Aussage von Jimmy Wales, dem Gründer der Wikipedia, sämtliche Lehrbücher, von der Schule bis zum Studium, werden – als Folge des Projektes Wikibooks – spätestens im Jahr 2040 als frei zugängliche Werke existieren: „Langfristig wird es sehr schwer für herkömmliche Verlage werden, mit frei lizenzierten Alternativen mitzuhalten“, einem offenen Projekt, an dem Dutzende Professoren mitarbeiten, könne ein Verlag kaum Paroli bieten[22].

[16] A.a.O., S. 362.

[17] Daneben erteilt er – das zur Abrundung des Bildes – sowohl der Verschärfung von Schutzrechten und Schutzmaßnahmen eine Absage als insbesondere auch der intensivierten Kriminalisierung von Verstößen (vgl. a.a.O., S. 380 / 381) sowie Geschäfts- und Organisationsmodellen, die „nicht dem normativen Verhalten und Bewusstsein der Nutzer der Dienste in den elektronischen Räumen entsprechen“ (a.a.O., S. 381).

[18] A.a.O., S. 381.

[19] Vgl. dazu auch Lawrence Lessig (in Lessig 2005), der die aktuell vonstatten gehende Verschiebung eines Gleichgewichts zugunsten der Rechteinhaber durch eine faktische Verwandlung von Immaterialgütern in quasi-materiale beklagt.

[20] Kuhlen (2004), S. 381.

[21] Vgl. Lehnartz 2005.

[22] Jimmy Wales nach Dambeck 2005b.

Analogie Wikipedia – Open-Source-Software

Wenn man für die folgende Argumentation nun – in dieser Richtung ! – explizit die Analogie von der Wikipedia zur Open-Source-Software etablieren will, geht man unhistorisch vor, denn die Wikipedia bezieht sich – nicht zuletzt in einer ihrer wesentlichen Bestimmungsgrößen, dem zugrunde liegenden Lizenzmodell – ganz explizit auf die vorgängige Open-Source-Software. Es ist sogar so, dass die bei der Wikipedia verwendete „GNU-Lizenz für freie Dokumentation" ursprünglich für begleitende Dokumentation von Open-Source-Software (nach der GNU General Public Licence – siehe Abschnitt >Open-Source-Software ...<) entwickelt wurde.
Nichtsdestotrotz kann diese Analogie phänomenologisch-abstrahiert folgendermaßen gefasst werden: Leute erbringen – typischerweise in ihrer Freizeit, nach einem nicht-hierarchischen „Basar-Modell", räumlich verteilt und für Mitarbeit grundsätzlich offen – intellektuelle Leistungen und wenden sich mit den entsprechenden Hervorbringungen an Abnehmer, die diese zwar einerseits fast uneingeschränkt nutzen dürfen[23], eigene mögliche Ableitungen davon aber den gleichen Bedingungen unterwerfen müssen. Diese Hervorbringungen – und das ist einer der wesentlichsten Punkte hier – treten auf dem Markt fast unweigerlich in Konkurrenz zu funktional vergleichbaren Produkten, die Unternehmen mit eigenen Angestellten und/oder mit über Honorare gebundenen freien Mitarbeitern entwickeln[24].

[23] Interessanterweise hat selbst die wichtige OSS-Lizenzverpflichtung, den Source-Code mit auszuliefern, eine Entsprechung in der GNU FDL für die Wikipedia: „If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy. Or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material."[23] Dabei gilt: "A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straight forwardly wit generic text editors […]". (GNU FDL o.J., Absätze 3 und 1).

[24] Eine Grenze der Analogie – das wird bei den für Verlage zu empfehlenden Strategien zu berücksichtigen sein – ist erreicht, wenn man bedenkt, dass „bei der OSS-Entwicklung [...] Software entwickelt wird, die teilweise in Bezug auf Stabilität und Qualität klassisch entwickelter Software überlegen ist" (Brügge et al. (2004), S. 54) und dem – im Abschnitt >Wikipedia< erwähnte – grundlegende Qualitätsvorbehalte gegenüber der Wikipedia im Vergleich zu kommerziellen Enzyklopädien gegenüberstellt. Allerdings: Selbst wenn diese Einschätzung Bestand hat, muss man der Frage nachgehen, inwieweit es Kunden gibt, die bereit oder darauf angewiesen sind, diese höhere Qualität dann auch zu honorieren. Wenn das nicht der Fall ist, bliebe in der Übertragung von den in der Literatur genannten Motiven für eine OSS-Einsatzentscheidung (vgl. Brügge et al. 2004, S. 115/116), nämlich monetären, strategischen (keine Exklusivrechte etc.) und operativen (also den Leistungsumfang und die Qualität betreffenden) für einen Wikipedia-Einsatz

Open-Source-Software (OSS), ihr community-basiertes Entwicklungsmodell und dessen Grenzen

Wie oben bereits erwähnt sind die der Entwicklungsarbeit zugrunde liegenden spezifischen Lizenzmodelle von OSS einerseits und Wikipedia andererseits ein Hauptansatzpunkt für die Analogie zwischen den beiden. Als Haupt-Bestimmungsgrößen von OSS werden zwar meist umfassender genannt „1. die Lizenz der Software, 2. nicht-kommerzielle Einstellung, 3. hoher Grad an Kollaboration bei der Programm Entwicklung und 4. die starke räumliche Verteilung der Entwickler“[25], es ist dabei aber eben nur der erste Punkt wirklich definitorisch, die anderen drei Punkte treffen auf viele OSS-Aktivitäten von Software-Unternehmen (siehe Abschnitt >Strategien der Softwareunternehmen ...<) nicht zu. Dabei gibt es bei OSS im Detail durchaus unterschiedliche konkrete Lizenzen[26], die allerdings alle ihrerseits den von der 1998 gegründeten Open Source Initiative (OSI) aufgestellten 10 („Meta“-) Bedingungen (z.B. Mitgabe des Quellcodes, Nicht-Diskriminierung von Personen und Gruppen, keine Einschränkung des Einsatzfeldes, Weitergabe auch der Lizenz, Technologieneutralität, etc.) genügen müssen. Das Recht, dass der „Quellcode [...] unbegrenzt gelesen, genutzt, modifiziert und distribuiert werden“ kann, räumen alle OSI-Lizenzen ein, insbesondere also auch die ursprüngliche und weitreichendste, die GNU General Public Licence GPL. Andere Lizenzen räumen – die ursprüngliche Idee letztlich einschränkend – darüber hinaus noch das Recht ein, dass OSS-Software „mit proprietärer Software verbunden und (re-)distribuiert werden“ darf (z.B. LGPL – GNU Lesser/Library General Public Licence), das Recht, dass „Modifikationen am OSS-lizenzierten Quellcode [...] im Distributionsfall proprietär bleiben“ können (z.B. BSD – Berkeley Software Distribution, z.B. im Fall des Web-Servers Apache) oder „spezielle Privilegien für den ursprünglichen Copyrighthalter über Modifikationen anderer“[27] (z.B. MPL – Mozilla Public Licence im Fall des Browsers Mozilla).

Die durch die Art der Lizenzierung von OSS ermöglichte (und, wie sich unten zeigen wird, auch eingeschränkte) Art des Arbeitens lässt sich softwaretechnologisch folgendermaßen beschreiben: Statt aktivitätsorientiertrer Methoden, bei denen der Entwicklungsprozess in Phasen eingeteilt ist und bei dem eine Phase erst beginnen kann, wenn die davor abgeschlossen ist (dieses Modell ist auch als Wasserfall-Modell bekannt), spielen zunächst entitätsorientierte Methoden eine größere Rolle; bei diesen werden Problemfelder definiert, die entweder offen (d.h. ungelöst) oder geschlossen (d.h. gelöst) sind, wobei geschlossene Problemfelder – aufgrund neuer Entwicklungen oder Anforderungen – wieder geöffnet werden können, es muss dann

---

prima facie nur der monetäre – wenn man von weltanschaulichen und imagebezogenen Motiven einmal absieht.

[25] Brügge et al. 2004, S. 19/20.

[26] Das Folgende vgl. a.a.O., S. 19-25.

[27] A.a.O., S. 23.

nur darauf geachtet werden, welche Problemfelder vom erneuten Öffnen bzw. geänderten Schließen mit betroffen sind. Da aber auch entitätsorientierte Prozesse typischerweise komplex sind und explizit geplant werden müssen, bietet sich in einem weiteren Schritt der Übergang zu agilen (entitätsorientierten) Prozessen an, bei denen – ohne Planung im traditionellen Sinne – in einem leichtgewichtigen Entwicklungsprozess Teillösungen immer sofort getestet und debugged werden. Grund für diesen zweiten Übergang ist, dass „mittlerweile [...] die Software-Entwicklergemeinde zu dem Schluss [kommt], dass die Erstellung von Software eher als kreativer Prozess zu verstehen ist, in dem Aktivitäten nichtlinear und oft parallel verlaufen." Daher „[deuten] neuere Forschungsergebnisse darauf hin [...], dass der Software-Prozess eher adaptiv sein muss. Damit ist gemeint, dass der Prozess selbst sich über die Zeit ändern muss, um sich an neue Ziele oder Lösungsanforderungen anzupassen."[28] Aus einer anderen Perspektive kann man „ein Open-Source-Projekt auch als einen genetischen, im Darwin´schen Sinne evolutionären Prozess auffassen. Das genetische Material ist der Quellcode. Der Entwickler nimmt Veränderungen vor (Mutationen) und die Community entscheidet, welche Mutationen Erfolg versprechen und weiterverfolgt werden."[29] Ein Haupt-Erfolgsfaktor richtig eingesetzter, erfolgreicher OSS-Entwicklungen ist in diesem Sinne – in einem als evolutionär zu beschreibenden Prozess – also nicht zuletzt „die große Anzahl von Entwicklern, die nicht nur mittels Tests, sondern auch mittels Inspektionen von Quellcode Defekte aufspüren und gegebenenfalls beheben. Die Offenlegung des Quellcodes gibt jedermann die Möglichkeit, Quellcode-Inspektionen nach eigens entwickelten Methoden auf verbreiteten Produkten durchzuführen."[30]
Wie oben angedeutet hat dieser typische Entwicklungsprozess[31] auch systematische Einschränkungen zur Folge, was die Art der mit diesem zu bewältigenden Projekte angeht: „Der community-basierte Open-Source-Entwicklungsprozess hat sich vorherrschend in der Konstruktion von Software-Systemen bewährt, 1. deren Funktion bekannt ist – z.B. in einer bereits existierenden Version – (Cloning) oder 2. die keinen komplizierten Entwurf haben, sondern einen Entwurf, der von Program

[28] A.a.O., S. 75.
[29] A.a.O., S. 85.
[30] A.a.O., S. 90.
[31] Dieser Entwicklungsprozess verzichtet typischerweise nicht nur faktisch auf Planung, sondern es gilt: „In keiner Open-Source-Lizenz wird verlangt, dass Entwurfsentscheidungen oder gar höhere Modelle [Analyse, Systementwurf, detaillierter Entwurf, C.B.] veröffentlicht werden. Der Entwickler wird lediglich gehalten, den Quellcode lesbar zu halten." (A.a.O., S. 72) Das führt zu dem ineffizienten Verfahren, dass Beiträger Reverse Engineering betreiben, um Entwurfsentscheidungen aus dem (kommentierten) Code in Form eines mentalen Modells wieder zu gewinnen, das sie dann wieder nur für das Forward Engineering in Form von Änderungen oder Ergänzungen verwenden, nicht aber explizieren oder gar weitergeben.[31] (vgl. a.a.O., S. 73)

mierern aus dem Quellcode abgeleitet werden kann, oder 3. deren Spezifikation ohne Zuhilfenahme aufwendiger Modellierungstechniken oder breitbandiger, synchroner Kommunikationswege, beispielsweise nur persönlich oder telefonisch, kommunizierbar ist. Mit anderen Worten, die Spezifikation muss intuitiv erfassbar, oder sehr leicht – auch asynchron – kommunizierbar sein [...]."[32], das heißt „dass das community-basierte Basar-Entwicklungsmodell von OSS sehr gut geeignet ist für inkrementelle Verbesserungen, während für umfangreiche nicht in Einzelaufgaben zerlegbare Neuentwicklungen der klassische proprietäre Ansatz („Kathetralen-Modell") besser geeignet erscheint. OSS, zumindest die in nicht-kommerziellen Communities im Basarstil entwickelte, tendiert eher zur Imitation der Funktionalitäten existierender Software [...]."[33] Außer diesen Einschränkungen, die auf die durch den Prozess nicht unterstützte Planung zurückzuführen sind, erreicht das community-basierte Basar-Entwicklungsmodell für Software dort Grenzen, wo 1. spezifische Domänen-Kompetenz erforderlich ist oder 2. die Gestaltung von Benutzeroberflächen betroffen ist, v.a. in dem Fall, dass Produkte nicht für die Unter-Zielgruppe gedacht sind, bei der Entwickler und Nutzer zusammenfallen und in der Funktionalität wichtiger ist als Ästhetik und Ergonomie. Brügge et al. schreiben zum ersten Punkt: „Die Programmierung der Funktionalitäten spezialisierter Anwendungssoftware, z.B. betriebswirtschaftlicher Software, erfordert entsprechende Fachkenntnisse. Entwickler werden selten über diese Kenntnisse und das Wissen um die erforderlichen Funktionalitäten der Software verfügen und gleichzeitig als Programmierer in der Lage sein, diese Anforderungen in Code umzusetzen. [...] die Entwicklung derartiger Software innerhalb der OSS-Community [ist] daher weniger wahrscheinlich als bei Software, deren Anwendungsbereich die Informationsverarbeitung selbst darstellt (Betriebssysteme, Netzwerksoftware)."[34] Zum zweiten Punkt – Bedienerfreundlichkeit – schreiben sie: „OSS, die einem Community-Entwicklungsprozess entstammt, wird typischerweise von hochqualifizierten Nutzern entwickelt. Soweit diese in erster Linie durch ihren eigenen Bedarf motiviert sind, wird auf eine einfache und leicht verständliche Bedienbarkeit zumeist weniger Wert gelegt als von Anbietern proprietärer Software."[35] Es muss allerdings an dieser Stelle erwähnt werden, dass das im Abschnitt >Strategien der Software-Unternehmen …< beschriebene Engagement von Unternehmen für OSS letztgenannte Beschränkungen zumindest ansatzweise heilen kann[36] – in keinem Fall jedoch natürlich die des systematisch nicht vorgesehenen Planungsprozesses.

---

[32] A.a.O., S. 85.

[33] A.a.O., S. 171.

[34] A.a.O., S. 166.

[35] A.a.O., S. 166.

[36] So kann man für den Bereich der Bedienerfreundlichkeit sagen, dass reife, populäre sowie unternehmensdominierte OSS-Projekte wie Linux, KDE oder OpenOffice diesen Makel weitgehend hinter sich gelassen haben, denn die beteiligten Unternehmen

Strategien der Softwareunternehmen, sich bei OSS zu engagieren

Ungeachtet der erwähnten Grenzen wird im community-basierten Software-Entwicklungsmodell vielfach hochwertige Software entwickelt, die potentiell Software von kommerziellen Software-Unternehmen direkt substituiert. Dadurch entstehen in einer Binnenbetrachtung in verschiedenen Software-Marktsegmenten neue Konkurrenzverhältnisse, die mittels marktüblicher Mechanismen (v.a. über Kosten- und Leistungsvergleiche) ausgetragen werden müssen. Darüber hinaus gibt es für Software-Unternehmen aber vielfältige weitere Möglichkeiten, mit der Herausforderung Open-Source-Software produktiv umzugehen. Die Gründe, sich als Software-Unternehmen ganz bewusst – also über evtl. mit OSS-Verpflichtungen verbundene Auftragsarbeiten[37] oder disziplinierende Effekte[38] hinaus – im OSS-Bereich zu engagieren, können dabei folgendermaßen klassifiziert werden:

Das Unternehmen möchte an OSS direkt mitverdienen, was v.a. durch den Verkauf von Distributionen von Open-Source-Software möglich ist; diese beinhalten oft außer dem offenem Programmcode selbst – lizenzrechtlich typischerweise ohne Probleme „bepreisbares" – proprietäres Material und – natürlich ebenfalls Kunden in Rechnung stellbare – nicht-digitale Leistungen um Marketing und Logistik.

Das Unternehmen[39] braucht für ein Projekt Entwicklungsressourcen, die es selbst nicht aufbringen kann oder will, und übergibt deswegen dieses Projekt an die Open-Source-Community – zu diesem Block gehören z.B. die Übergabe der Verantwortung für die Anpassung von Komponenten an veränderte Umgebungen (wie zum Beispiel im Falle von Gerätetreibern) oder die Übergabe der Verantwortung für die Weiterpflege eines nicht mehr rentablen, aber aus anderen Gründen weiterhin

„investieren entsprechend in die Verbesserung der Bedienbarkeit der Software und berücksichtigen dabei die Wünsche von Endnutzern." (A.a.O., S. 167).

[37] Vgl. a.a.O., S. 111/ 122.

[38] „Die bloße Möglichkeit, dass Unternehmensfremde den Quellcode inspizieren, kann einen disziplinierenden Effekt auf Programmierer haben (sauberer Code, klarere Struktur, bessere Dokumentation)." (A.a.O., S. 104/105).

[39] Um die in b) bis e) aufgeführten Ziele erreichen zu können, muss das Unternehmen tatsächlich Ressourcen in die Software selbst stecken – es gibt dabei drei Unterfälle, nämlich a) die spätere Freigabe des Quellcodes ursprünglich proprietärer Software, b) die Entwicklung von Software für die Freigabe des Quellcodes sowie c) die Betreuung eines bereits bestehenden OSS-Projektes. (Vgl. Leiteritz 2004, aber z.B. – mit Beispielen – auch Brügge et al. 2004, S. 112: „Dabei können drei Fälle unterschieden werden: ein Unternehmen kann zur Weiterentwicklung eines existierenden OSS-Programms beitragen, wie beispielweise IBM zu Linux; es kann eigene, vormals proprietäre Software als OSS freigeben (z.B. gab IBM Eclipse und Sun NetBeans frei); oder es kann ein neues OSS-Projekt initiieren.")

gewünschten Produktes, um einerseits Kosten zu sparen, andererseits aber Vertrauen zu erhalten bzw. zu schaffen.[40]

Das Unternehmen möchte mit seiner Lösung Standards setzen und in deswegen Preishürden vermeiden; „dies ist vor allem relevant bei Infrastruktursoftware und Softwaretools sowie allgemeiner dort, wo Netzeffekte vorliegen. Die Freigabe von OSS von sicherheitsrelevanter Software kann sogar als eine wichtige Methode angesehen werden, einen Standard für DRM oder Geheimhaltungsdienste von allen beteiligten Herstellern korrekt umsetzen zu lassen.“[41]

Das Unternehmen möchte durch wertvolle Beiträge eine Reputation als „good citizen“ (bei bestimmten Gruppen) erreichen oder diese weiter verbessern.

Das Unternehmen möchte seine Mitbewerber schwächen, indem es mit einem Produkt, das sich auf dem kommerziellen Markt nicht durchgesetzt hat, ein Stachel im Fleisch der Konkurrenz bleibt, dessen diese sich über Marktmechanismen (z.B. mittels Prizing) nur schwer erwehren kann[42].

Wenn man davon ausgeht, dass es – bei aller vorstellbaren „corporate policy“, die Community zu unterstützen, Monopole zu verhindern, etc. – insbesondere auch den Punkten a) – e) übergeordnetes Ziel eines Unternehmens ist, für seine Anteilseigner Gewinne zu erwirtschaften, erfüllt außer dem ersten keiner der oben angeführten Gründe dieses Ziel. Besser gesagt: Sie erfüllen es auf keinen Fall direkt, sondern – allenfalls – in Verbindung mit dem in diesem Zusammenhang überaus wichtigen „Meta-Ziel“ des so genannten Komplementverkaufs (in Marketing-Terminologie: Cross-/Up-Selling), bei dem es darum geht, durch die wie auch immer gestaltete Teilhabe an OSS Märkte zu erweitern und mit gewinnerwirtschaftenden Produkten aus dem eigenen Portfolio dann auch zu bedienen. Eine besondere Bedeutung haben dabei erfahrungsgemäß Dienstleistungen[43]: „Bei einem Dienstleistungs-Geschäftsmodell wird kein eigenes Produkt entwickelt, sondern es werden Dienstleistungen für existierende OSS-Produkte angeboten. Das OSS-Dienstleistungsmodell hat sich zum ´kleinsten gemeinsamen Nenner´ der OSS-Geschäftsmodelle entwickelt. Fast alle Geschäftsmodelle rund um OSS haben (auch) einen Dienstleistungsanteil. Unterschiedlich ist vor allem die Angebotstiefe und –breite: Sie kann vom einfachen E-Mail-Support bis zur kompletten Dienstleistungspalette reichen.“[44]

[40] Zu Letzterem vgl. Brügge et al. 2004, S. 113.

[41] A.a.O., S. 104.

[42] Vgl. zum Vorigen zum Teil a.a.O., S. 103-107.

[43] Generell gilt: „Solche Komplemente können Hardware sein [...], proprietäre Software [...] oder Software-Support [...].“ (Brügge et al. 2004, S. 107).

[44] Leiteritz 2004.

Der Analogieschluss: Strategien für Lexikonverlage

Wenn man im Analogieschluss die oben klassifizierten Handlungsoptionen für Softwareunternehmen angesichts Open-Source-Software auf (Lexikon-)Verlage angesichts der Wikipedia überträgt, kommt man zunächst v.a. auf die Optionen, lexikalische Substanzen freizugeben, freie Substanzen weiterzuentwickeln oder eine Community-Entwicklung von solchen Substanzen zu initiieren, um gewinnerwirtschaftende Komplementprodukte verkaufen zu können[45]. Dazu kommt der – bei der Darstellung der Strategien der Softwareunternehmen von mir nicht explizit gemachte, aber in der Beschreibung aktueller Software-Entwicklungsprozesse erwähnte – Aspekt, möglicherweise vom Community-Entwicklungsprozess für interne Strukturen und Prozesse (auch von Verlagen) zu lernen. Gemäß Fußnote 24, in der es um die relativen Qualitätsunterschiede zwischen OSS und proprietärer Software einerseits und der Wikipedia und kommerziellen Enzyklopädien andererseits geht, kommt aber – zumindest auf absehbare Zeit – ein entscheidender Aspekt hinzu, nämlich der, dass Verlage aufgrund von systematischen Schwächen des „Wiki-Ansatzes" die Chance haben, dem Markt zielgenau und systematisch „bessere" Produkte anzubieten bzw. offene Substanzen genau in diese Richtung hin weiterzuentwickeln und das auch so zu kommunizieren. Der Vollständigkeit halber muss unter die Handlungsoptionen noch das – von mir an der entsprechenden Stelle ebenfalls nicht ausgeführte, aber in der Literatur dargestellte[46] – „Mediatoren-Geschäftsmodell" gezählt werden; ein solches könnte z.B. durch den Betrieb eines Marktplatzes im Web verfolgt werden, der Entwickler, Nutzer, Dienstleister, etc. um die Wikipedia zusammenführt und sich über Transaktionen oder Werbung finanziert. Ebenfalls dem „Mediatoren-Geschäftsmodell" verhaftet wäre z.B. die – mit Letzterem erkennbar verwandte – Veröffentlichung von Literatur zur community-basierten Content-Entwicklung.

Wenn man nun die durch relativ direkte Analogieschlüsse erreichten und oben skizzierten Empfehlungen explizit in die Gegebenheiten der Verlagswelt zurücktransponiert, mit einigen spezifischen Gesichtspunkten anreichert und um Optionen, die sich eher nur systematisch ergeben, bereinigt, kommt man zu folgender Listung von Handlungsoptionen, die natürlich noch durch spezifische und detaillierte

[45] Ich sehe hier einmal vom offensichtlichen, aber relativ wenig Wertschöpfungspotential versprechenden Distributions-Fall ab, für den es nichtsdestotrotz gute Gründe geben muss, denn sowohl die im Abschnitt >Einleitung< erwähnte CD-ROM/DVD-ROM-„Distribution" als auch die dort erwähnten Buchprojekte werden von einem Verlag, Directmedia, betrieben. (Vgl. die Wikipedia-Einträge „Wikipedia: Wikipedia-Distribution" und „Wikipedia: Wiki Press").

[46] Vgl. Leiteritz 2004.

Recherchen verfeinert werden müssen, was die aktuellen Möglichkeiten eines Verlages und die tatsächlichen Chancen auf konkreten Teilmärkten betrifft[47,48]:
Aufgrund von mit der spezifischen Arbeitsweise an der Wikipedia (und vergleichbaren Projekten) zusammenhängenden Beschränkungen werden in der Sachlexikographie Marktsegmente sowohl „oberhalb" als auch „unterhalb" der Wikipedia durch solche Community-Projekte systematisch unbearbeitet bleiben – in der Komplexität „oberhalb" bei Produkten, die auf einen Top-Down-Planungs- und Managementprozess nicht verzichten können und in der thematischen Spezifität „unterhalb" bei Produkten, für die im „Wiki-Ansatz" nicht genügend freiwillige Mitarbeiter gefunden werden können; ein Beispiel für ersteres könnte eine vernetzte thematische Enzyklopädie sein, eines für letzteres sachlexikographische Werke in „spitzen" Segmenten des professionellen Fachpublizierens (außerhalb des Wissenschaftsbereiches).
Für wirklich innovative Produkte gibt es angesichts weiterer Beschränkungen community-basierter Entwicklungsprozesse, insbesondere der Tendenz zur (bloßen) Imitation von bestehenden Produkten, bei Verlagen potentiell bessere Chancen, und zwar hauptsächlich aufgrund von deren Möglichkeit, mit Kapitaleinsatz professionelle Marktforschung zu betreiben, sowie aufgrund anderweitig spezifischer Erfahrung und Ausbildung von Mitarbeitern, idealerweise auch dank professionellen Innovationsmanagements.
Für Kunden, denen die fachgerechte Ausführung im Detail wichtig ist, können Verlage mit der Kompetenz ihrer Mitarbeit und dem institutionellen Know-How ihrer Unternehmen mit potentiell sowohl inhaltlich als auch ergonomisch und ästhetisch hochwertigeren Produkten aufwarten; Beispiele hierfür sind Produkte mit durchgehendem sprachlichem Duktus[49] oder mit inhaltlich, wahrnehmungspsychologisch und medial begründetem (individuellem) Artikel-Layout.

---

[47] Eine ganz andere Möglichkeit, „diesseits" wie auch immer gearteter spezifischer Reaktionen auf Wikipedia Kuhlens Forderungen einzulösen, auch in proprietären Produkten niedergelegtes Wissen „offener" zugänglich zu machen, kann hier nicht näher behandelt werden. Es geht dabei im Wesentlichen darum, ohne Bezug auf Lizenzverträge auch für diese proprietären Produkte „neue Organisations- und Geschäftsmodelle" wie z.B. Provisions- und Creditingmodelle oder Auktionsmodelle zu entwickeln; Kuhlen handelt diese, für die Umsetzung seines Ansatzes so entscheidenden Modelle leider nur stichwortartig ab – auf 2 Seiten seines 444 Seiten umfassenden Werkes zur Informationsethik (vgl. Kuhlen 2004, S. 366/367).

[48] Ich gehe hierbei auf den Aspekt, dass es für Verlage einen u.U. unternehmensstrategisch wünschenswerten Reputationszuwachs bedeuten kann, durch Unterstützung der Wikipedia und verwandter Projekte als „good citizen" zu gelten, nicht näher ein.

[49] Die im Abschnitt >Wikipedia< erwähnte, von Brockhaus z.B. explizit in Anspruch genommene Verlässlichkeit und Ausgewogenheit einer redaktionell betreuten Enzyklopädie ordnen sich am ehesten hier und in a) ein.

Wissensprodukte sinnvoll ergänzende multimediale Elemente (Bilder, Tondokumente, Filme, Animationen, Interactivities) können typischerweise nicht im „Wiki-Ansatz“ erstellt werden, sondern müssen geschaffen und produziert oder lizenziert werden, wozu in beiden Fällen kraft Kapitalausstattung oder Konzernverknüpfung nur kommerzielle Verlage die Möglichkeit haben[50]. Existierende Multimedia-Enzyklopädien deutscher Lexikonverlage (hier ganz explizit einschließlich Microsoft mit „Encarta“) markieren, was so erreichbar ist.

Aufgrund der institutionellen Verknüpfung von Know-How über die abgedeckten Sach-Domänen und über die Kunden mit dem besseren Zugang zu computerlinguistischem und sprachtechnologischem Know-How können Verlage Produkte mit fach- und kundengerechten intelligenten Frontends schaffen. „Encarta“ und „Brockhaus multimedial“ zeigen z.B. mit Wissensnetzen über den Enzyklopädie-Inhalt, was hier möglich ist – Brockhaus möchte mit der nächsten Ausgabe des „Brockhaus multimedial“ (2006) hier offensichtlich sogar noch einen Schritt weiter gehen[51].

Die Gestaltung von Wissensmanagement- und Redaktionsprozessen kann nicht Gegenstand dieses Papers sein – es ist aber in Übertragung des OSS-Erfolgsfaktors Arbeitsorganisation möglicherweise wert zu überlegen, inwiefern im Einzelfall von den typischen Prozessen des „Wiki-Ansatzes“ (agil, evolutionär, entitätsorientiert statt aktivitätsorientiert etc.) für die Arbeit an Lexika auch in Verlagen gelernt werden kann[52].

Ergebnisse des „Wiki-Ansatzes“, also community-basierter Entwicklungsprozesse zu distribuieren ist für Verlage nur in Fällen ratsam, in denen entweder proprietäre Produktergänzungen (z.B. die Anzeigesoftware) einen so deutlichen Produktvorteil bringen, dass mit dem Gesamtprodukt ein einkömmlicher Preis erzielt werden kann, oder Wertschöpfungsschritte betroffen sind, an denen aus systematischen Gründen digital nicht weitergearbeitet werden kann, (z.B. herstellerische, marketingliche und logistische Aktivitäten in Richtung Offline-Produkte) oder durch das Produkt eine so

[50] Vgl. dazu allerdings „Das Wissen der Welt“ 2005: „Daneben sammelt Wikimedia Commons frei verfügbare Filme und Musikdateien. Mehr als 150.000 Dateien sind bisher zusammengekommen.“ Auch wenn die meisten wirklich attraktiven Medienelemente wahrscheinlich nicht frei verfügbar sein und Wikipedia deswegen auch nicht zur Verfügung stehen werden, heißt das, dass dieses Argument zumindest so kategorisch möglicherweise keinen Bestand haben kann.

[51] Bernd Kreißig von Brockhaus Duden Neue Medien in seinem Vortrag auf der AKEP-Jahrestagung am 14./15.6.2005 in Berlin.

[52] Der Aspekt, als systematische Bestimmungsgröße – ohne weitere Spezifizierung – über einen möglichst großen Mitarbeiterpool zu verfügen, dagegen ist – über die selbstverständliche Notwendigkeit angemessener Ressourcen hinaus – im Widerspruch zum erwähnten, für Verlage ratsamen Ansatz, mit qualifizierten Mitarbeitern in einem Top-Down-Planungs- und Managementprozess zu arbeiten, und kann von daher kein Orientierungspunkt für Verlage sein.

dringend erforderliche Vervollständigung der Produktpalette erreicht wird, dass die Nachteile (Verlust der exklusiven Nutzungsrechte, voraussichtlich eher bescheidene Marge) überwogen werden.
Auch die Weiterentwicklung vorhandener offener Substanzen empfiehlt sich für Verlage nur dann, wenn so eine so wichtige Vervollständigung der Produktpalette erreicht werden kann, dass die Bedingung der Lizenz, dass das Ergebnis der Arbeit an dieser Substanz wieder offen sein muss, nicht mehr ins Gewicht fällt.
Das wichtigste Kriterium für die Vervollständigung der Produktpalette mit offenen Substanzen (vgl. v.a. h)) ist in jedem Fall, ob der Verlag zu diesen Komplementprodukte absetzen kann, die Erfolg, insbesondere in Form eines Ergebnisbeitrags versprechen; das könnten thematisch abgeschlossene proprietäre Produkte zu bestimmten Themenbereiche sein, die bruchlos unter den für offene Substanzen verbreiteten Viewern betrachtet werden können (für in Sachen Contentqualität besonders sensibilisierte Kunden), oder auch proprietäre intelligente Frontends für offene Substanzen. Ebenfalls in diesen Block möglicher Komplementprodukte fallen Content-Dienstleistungen, die in größerem Ausmaß zeitlich konzentrierte und verlässliche Ressourcen erfordern, wie das z.B. beim Betrieb eines Call-Centers für telefonische Auskünfte im Umfeld bestimmten Contents der Fall ist.
Die Initiierung von Community-Projekten ist für Verlage (nur) in folgenden systematisch identifizierten Szenarien sinnvoll: Vervollständigung der Produktpalette für den ergebniseffektiven Komplementverkauf (s.o.), Schwächung kommerzieller Mitbewerber und möglicherweise – z.B. in den Bereichen Rechtschreibung oder Terminologie – Standardsetzung (s.o.); für die Freigabe einer ursprünglich proprietären Substanz könnte zusätzlich sprechen, dass ein „klassischer", reputationsträchtiger Titel nicht mehr kostendeckend weitergepflegt werden kann, in der public domain aber noch eine Zukunft mit positiven Effekten für den Verlag hat. Ein direkt aus dem Software-Bereich übertragbarer Grund könnte noch sein, dass für ein als wichtig eingestuftes Projekt Entwicklungsressourcen benötigt werden, die nur durch kollaborative Arbeit über die Unternehmensgrenzen hinaus zu mobilisieren sind – z.B. für eine lückenlose sachlexikographische Abdeckung aller deutschen Gemeinden; aber auch in diesem Fall ist natürlich entscheidend, dass das Ergebnis Komplementverkäufe des Verlages ermöglicht bzw. fördert.

Schluss

In Analogie zu den Strategien von Software-Unternehmen angesichts Open-Source-Software lassen sich zahlreiche plausible und zumindest im Einzelfall überprüfenswerte Strategie-Ansätze für Lexikonverlage angesichts der Wikipedia „konstruieren". Aus pragmatischer Verlagssicht bewertet dürften – so meine abschließende Einschätzung – davon in allererster Linie die beiden Handlungsoptionen eine Rolle spielen, sich erstens als Verlag auf solche Produkte zu konzentrieren, die aufgrund

des bei ihnen unabdingbaren Planungs- und Managementbedarfs das community-basierte Entwicklungsmodell systematisch überfordern, und zweitens auf die sachgerechte Perfektionierung in Form von Produkten zu setzen, die einer methodischen Qualitätssicherung sowie in größerem Umfang v.a. redaktionell-sprachlichen und herstellerischen Know-Hows bedürfen, wie es nur Verlagen in Form von qualifizierten Mitarbeitern und etablierten Workflows zu Gebote steht, und die genau aufgrund dieser Perfektion ihre Zielgruppen finden.

Literatur

[Bläsi 1998]: Christoph Bläsi: Artikel kleiner Lexika. Ein exemplarischer Vorstoß in die angewandte Metalexikographie sachlexikographischer Werke. Dissertation, Heidelberg 1998.

[Brügge et al. 2004]: Bernd Brügge, Dietmar Harhoff, Arnold Picot, Oliver Creighton, Marina Fiedler und Joachim Henkel: Open-Source-Software. Eine ökonomische und technische Analyse. Heidelberg 2004.

[Dambeck 2005a]: Holger Dambeck: „Wikipedia und Brockhaus haben sich lieb“, in: Spiegel Online, www.spiegel.de/netzwelt/netzkultur/0,1518,358610,00.html, 2.5.2005, abgerufen am 5.8.2005.

[Dambeck 2005b]: Holger Dambeck: „Zehn Dinge, die umsonst sein werden“, in: Spiegel Online (www.spiegel.de), 4.8.2005, abgerufen am 5.8.2005.

[„Das Wissen der Welt“ 2005]: o.V.: „Das Wissen der Welt“, FAZ, 8.8.2005, S. 17.

[„Gedruckte Enzyklopädien haben Zukunft“]: o.V.: „Gedruckte Enzyklopädien haben Zukunft“, in: börsenblatt, Onlinemagazin für den deutschen Buchhandel (www.boersenblatt.net), www.mvb-boersenblatt.de/sixcms/detail.php?id=93598, 5.8.2005, abgerufen am 5.8.2005.

[Gehring / Lutterbeck 2004]: Robert A. Gehring und Bernd Lutterbeck (Hg.): Open-Source-Jahrbuch 2004. Online, www.think-ahead.org, abgerufen am 27.7.2005.

[GNU FDL o.J.]: GNU Free Document Licence, www.gnu.org/licences/fdl.txt, abgerufen am 27.7.2005.

[Kuhlen 2004]: Rainer Kuhlen: Informationsethik. Konstanz 2004.

[Lehnartz 2005]: Sascha Lehnartz: „Schlauer schwärmen“, in: Die Zeit, 5.6.2005, S. 57.

[Leiteritz 2004]: Raphael Leiteritz: Open-Source-Geschäftsmodelle, in: Gehring / Lutterbeck 2004, keine Seitenzahlen.

[Lessig 2005]: Lawrence Lessig: Free Culture. How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity. New York 2005.

[Möller 2005]: Erik Möller: Die heimliche Medienrevolution. Wie Weblogs, Wikis und freie Software die Welt verändern. Hannover 2005.

[Osterloh et al. 2004]: Margit Osterloh, Sandra Rota und Bernhard Kuster: Open-Source-Softwareproduktion : Ein neues Innovationsmodell ?, in: Gehring / Lutterbeck 2004, keine Seitenzahlen.

[Rheingold 2002]: Howard Rheingold: Smart Mobs – The Next Social Revolution. Transforming Cultures and Communities in the Age of Instant Access. 2002.

[Thiel 2005]: Thomas Thiel: „Wir sind Sprengstoff !“, in: FAZ, 8.8.2005, S. 36.

[Weber 2004]: Karsten Weber: Philosophische Grundlagen und mögliche Entwicklungen der Open-Source- und Free-Software-Bewegung, in: Gehring / Lutterbeck 2004, keine Seitenzahlen.

Wikipedia-Einträge: “GNU-Lizenz für freie Dokumentation”, „Wikipedia: Wikipedia-Distribution“ und „Wikipedia: Wiki Press“, jeweils abgerufen am 29.7.2005.

Das Redaktionstandem als innovatives Kooperationsmodell zwischen Fachwissenschaftlern und Bibliothekaren am Beispiel des Open Access E-Journals Brains, Minds & Media

Cordula Nötzelmann, Sören Lorenz; Bielefeld

Abstract

Die verlagsunabhängige Etablierung eines neuen Open Access Journals erfordert neue redaktionelle Modelle. Ein Weg, neue redaktionelle Modelle zu erproben, ist die Etablierung von Redaktionstandems zwischen Fachwissenschaftlern und Bibliothekaren. Die Universität Bielefeld arbeitet aktiv am Ausbau innovativer Publikationsmodelle, u.a. auf Open Access Basis. Zu nennen ist hier etwa das Open Access E-Journal BRAINS, MINDS & MEDIA. Die Zeitschrift wird herausgegeben von Fachwissenschaftlern des Bielefelder Lehrstuhls für Neurobiologie und gefördert durch die Initiative *Digital Peer Publishing (DiPP NRW, MWF)* sowie durch die Universität Bielefeld. Neben klassischen Artikeln werden auch Materialien publiziert, z.B. Visualisierungen und Simulationen. Sowohl das Gutachterverfahren als auch der Publikationsprozess werden mithilfe eines elektronischen Systems abgewickelt, das durch das Hochschulbibliothekszentrum NRW (HBZ) als drittem Partner bereitgestellt und an die speziellen Erfordernisse des Journals angepasst wird. Ein neuartiges Redaktionstandem, gebildet aus Fachwissenschaftlern des Lehrstuhls für Neurobiologie und Mitarbeitern der Universitätsbibliothek Bielefeld, betreut die Zeitschrift vor Ort. Der Beitrag beschreibt eingehend die Struktur, die Aufgabenverteilung und die Erfahrungen des Redaktionstandems mit einem gemeinsamen Workflow, in dem beide Seiten die ihnen eigenen Expertenaufgaben übernehmen. Die Synergien aus dem Fachwissen der Wissenschaftler und der Informationsvermittlung durch die Bibliothek können helfen, nachhaltige Strukturen zu etablieren und so eine hohe und dauerhafte Präsenz des Journals zu erzeugen.

1. Einführung

BRAINS, MINDS & MEDIA ist ein neu gegründetes Open Access eJournal mit internationaler Ausrichtung für neue Medien in den Neuro- und Kognitionswissenschaften. Die Gründung des Journals wird gefördert im Rahmen des Projekts *Digital Peer Publishing (DiPP NRW;* http://www.dipp.nrw.de*)* sowohl vom Ministerium für Wissenschaft und Forschung des Landes NRW als auch von der Universität Bielefeld, die – wie z.B. in ihrer Resolution zur Unterstützung von Open Access-Aktivitäten vom 7. Juni 2005 niedergelegt (http://www.uni-bielefeld.de/ ub/wp/), im Sinne der „Berlin 3 Open Access"-Empfehlungen (http://www.eprints.org/ ber

lin3/outcomes.html) die Förderung von Open Access e-Journals an der Hochschule bereits seit einigen Jahren aktiv unterstützt.

BRAINS, MINDS & MEDIA (http://www.brains-minds-media.org) veröffentlicht nicht nur Artikel, sondern auch dazugehörige digitale Medien, wie Tutorien, interaktive Simulationen, dynamische Visualisierungen etc. Herausgegeben vom Lehrstuhl für Neurobiologie der Universität Bielefeld wird die redaktionelle Arbeit in enger Zusammenarbeit mir der Universitätsbibliothek Bielefeld durchgeführt. Dieses *Redaktionstandem* übernimmt Aufgaben, die traditionell von Verlagen übernommen werden.

Als reines Online-Journal konzipiert, erscheint BRAINS, MINDS & MEDIA im Allgemeinen nicht in der üblichen Heftstruktur, da angenommene Beiträge sofort veröffentlicht werden. Die Beiträge eines Jahres werden nachträglich zu einem Band zusammengefasst. Diese konsekutive Strategie beschleunigt nicht nur die einzelne Publikation sondern fördert auch den schnellen Austausch über aktuelle Ergebnisse. Dazu kann jeder veröffentlichte Beitrag von Lesern kommentiert werden. Eingereichte Kommentare werden nach einer zeitnahen Prüfung durch die Redaktion an die entsprechenden Artikel angehängt.

BRAINS, MINDS & MEDIA hat sich zum Ziel gesetzt, neue Publikationsformen in den Neuro- und Kognitionswissenschaften zu etablieren und neue Informationstechnologien dazu zu nutzen, das komplexe Thema Hirnforschung durch die Veröffentlichung interaktiver Materialien transparenter zu machen. Da der uneingeschränkte Zugang zu diesen Materialien einen wesentlichen Baustein der Zielsetzung von BRAINS, MINDS & MEDIA darstellt, spielt der Open Access Gedanke für das eJournal eine zentrale Rolle.

Open Access, der u.a. in der Berliner Erklärung vom Oktober 2003 geforderte schrankenlose Zugang zu wissenschaftlicher Information im Internet (http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html), gewinnt im Rahmen der Neuausrichtung im Wissenschaftlichen Publikationswesen immer weiter an Bedeutung. Die Vorteile alternativer Publikations- und Kommunikationsmöglichkeiten, die das Internet für Wissenschaft und Forschung bietet - und der damit verbundene Handlungsbedarf - setzen sich in Bibliotheks- und auch Wissenschaftskreisen immer mehr durch; dies vor allem in den sogenannten STM- *(Science, Technology, Medicine)* Fächern, deren Bedarf nach raschem Austausch wissenschaftlicher Erkenntnisse die traditionellen, langen Publikationswege entgegenstehen. Während im anglo-amerikanischen Bereich schon einige Open Access-Publisher Fuß fassen konnten (z.B. PLOS, BioMedCentral), werden in Deutschland verschiedene Initiativen konzeptioneller Art, aber auch konkrete Umsetzungsprojekte initiiert, die einen für den Nutzer kostenfreien Zugang unabhängig von Verlagsmonopolen fördern wollen. Einen Überblick über internationale und nationale Aktivitäten im Bereich Open Access bietet die Bielefelder Informationswebseite http://www.uni-bielefeld.de/ub/wp/.

Auch die Universität Bielefeld arbeitet aktiv am Ausbau innovativer Publikationsmodelle und hat in einer Resolution zur Unterstützung von Open Access vom 7. Juni

2005 den aktiven Ausbau von Open Access-Maßnahmen an der Hochschule nochmals besiegelt (http://www.uni-bielefeld.de/ub/wp/). Daher ist die Universität Bielefeld, insbesondere die Universitätsbibliothek ein wichtiger Partner für BRAINS, MINDS & MEDIA.
Zweiter essentieller Partner für Etablierung von BRAINS, MINDS & MEDIA ist das Hochschulbibliothekszentrum NRW (HBZ), an dem das DiPP-Projekt angesiedelt ist. Das Land NRW unterzeichnete als einziges Bundesland die Berliner Erklärung und setzt die Open Access-Idee im Rahmen des Projekts DiPP NRW derzeit durch die Gründung von mittlerweile zehn eJournals praktisch um. Das HBZ stellt allen am DiPP-Projekt beteiligten Zeitschriften ein elektronisches Publikationssystem und ein elektronisches Gutachterverfahren zur Verfügung, das von Mitarbeitern des DiPP-Projekts an die individuellen Bedürfnisse einzelner Journals angepasst wird.

2. Partnerstruktur und Rollenverteilung

Zentral für die redaktionelle Arbeit ist das Redaktionstandem aus Fachwissenschaftlern und Bibliothekaren der Universität Bielefeld, in dem beide Seiten die ihnen eigenen Expertenaufgaben übernehmen. Durch die Basisdienste des HBZ erweitert sich das Redaktionstandem aus Fachwissenschaftlern und Bibliothekaren zu einem Tridem (Abbildung 1). Die Struktur dieses Tridems verteilt die Lasten der redaktionellen Arbeiten auf drei Säulen und ermöglicht so die verlagsunabhängige Erstellung und Aufrechterhaltung des Journals aus dem Dauerbtrieb der beteiligten Institutionen Lehrstuhl, Bibliothek, DiPP (HBZ).
Das Redaktionstandem muss im Wesentlichen vier Aufgabenfelder abdecken: Koordination und Kommunikation, Erschließung, Technische Bearbeitung / Lektorat und Marketing.

a. Koordination und Kommunikation

Die Koordination und Kommunikation nach innen und außen liegt bei den Fachwissenschaftlern. Sie akquirieren und betreuen die Autoren der *scientific community* von der Einreichung des Papers über die Organisation des Peer-Review-Verfahrens bis zur Veröffentlichung. Sie verteilen die Aufgaben innerhalb des Tandems und koordinieren die Kommunikation zwischen dem *Editorial Board* und den Gutachtern.

b. Erschließung

Die Mitarbeiter der Universitätsbibliothek Bielefeld flankieren diese Aktivitäten durch bibliothekarisches Know-How: so geschieht die Erschließung der Forschungsbeiträge ebenso auf Bibliotheksseite wie deren aktive Vermittlung, etwa durch Platzierung in einschlägigen Datenbanken und Meldung an fachspezifische und allgemeine Online-Suchdienste.

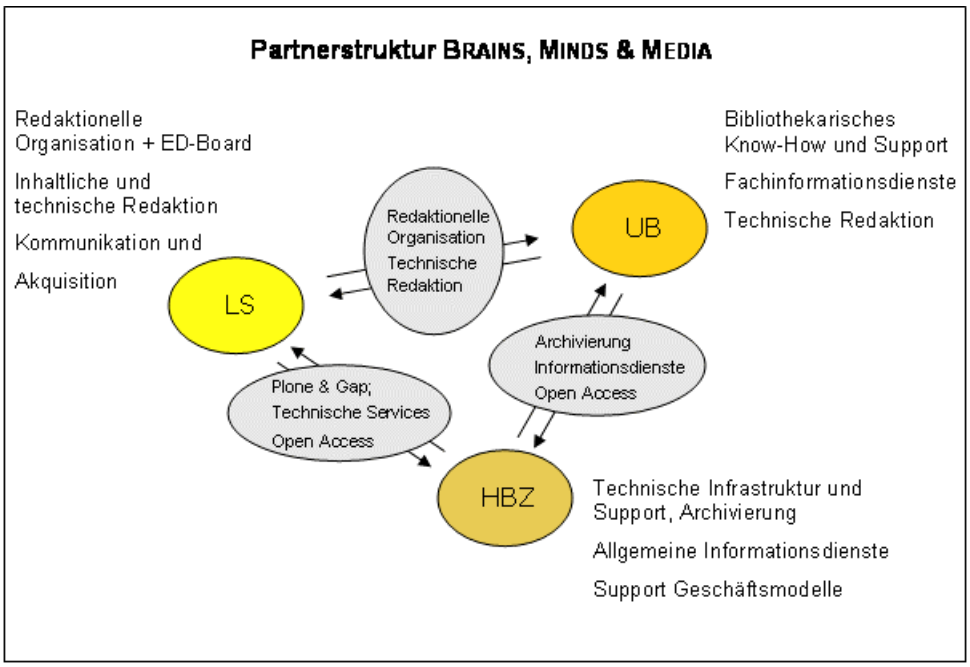


Abbildung 1: Partnerstruktur des Redaktionstandems zwischen dem Lehrstuhl für Neurobiologie (LS) und der Universitätsbibliothek (UB) sowie dem HBZ als drittem Partner.

c. Technische Bearbeitung / Lektorat

Beide Seiten des Redaktionstandems haben gemeinsam an den Kriterien für eine Formatvorlage gearbeitet, die eine reibungslose elektronische Weiterverarbeitung mit der Einbindung multimedialer Applikationen verbindet. Diese Weiterverarbeitung (Konvertierung) wird durch das vom HBZ bereitgestellte System automatisiert. Die automatische Konvertierung eines Artikels ist das technische Herzstück des Publikationssystems, dass durch das HBZ bereitgestellt wird (eine detaillierte Beschreibung findet sich in III. Technisches Basissystem). Dadurch reduziert sich die technische Bearbeitung der Artikel auf eine Prüfung der Formatkonsistenz eingereichter Artikel, auf einen Test etwaiger Materialien auf Lauffähigkeit sowie auf ein erstes Lektorat, unterstützt durch Rechtschreib- und Grammatikhilfen der einschlägigen Textprogramme. Eine weitere Prüfung wird durch die Gutachter und die Autoren erfolgen.

d. Marketing und Internetpräsenz

Über die eigentliche Redaktionsarbeit hinaus betreibt das Tandem Marketing für das Journal, die Fachwissenschaftler in der *scientific community*, die Bibliothek im BID-Bereich. In einschlägigen Mailinglisten und Datenbanken wird über neue Artikel

informiert, in der Fachpresse wird das Journal und seine Ziele in kurzen Beiträgen vorgestellt und auf Fachtagungen präsentiert. Durch eine leichte Auffindbarkeit über die gängigen Suchmaschinen und durch platzierte Links auf die Homepage (http://www.brains-minds-media.org) von anderen zentralen Internetseiten der *scientific community* wird die Auffindbarkeit des Journals ebenfalls erhöht. Die Internetpräsenz des Journal informiert detailliert über die Ziele und Ansprüche des Journals, stellt detaillierte Anleitungen zur Manuskripterstellung bereit und informiert ausführlich über die Lizenzbedingungen der *Digital Peer Publishing Licence* (DPPL; http://www.dipp.nrw.de/lizenzen/).

3. Technisches Basissystem

Die Abläufe der redaktionellen Arbeit werden maßgeblich durch die technische Infrastruktur bestimmt, auf die das Redaktionstandem zurückgreifen kann. Bevor der Workflow eingehend beschrieben wird, gibt dieser Abschnitt zunächst einen groben Überblick über die technischen Basissysteme:

Basierend auf dem Open Source Content Management System (CMS) *Plone* (http://plone.org) hat das HBZ ein individualisiertes Internetportal und Publikationssystem für jede Zeitschrift entwickelt. Erst dieser Dienst erlaubt dem Redaktionstandem überhaupt, das Journal mit minimalem Aufwand zu erstellen, da die Formatierung und die formattechnische Überprüfung der Artikel der zeitaufwändigste Arbeitschritts im Publikationsprozess darstellt. Ergänzt wird *Plone* durch ein *Fedora Repository* (http://www.fedora.info), in dem Artikel abgelegt und verwaltet werden können.

Neben dem Publikationssystem wird dem Tandem eine individualisierte Version des elektronischen Gutachtersystem German Academic Publishers (*GAPworks*; http://www.gapworks.de) zur Verfügung gestellt. Dieses System stellt einen differenzierten Workflow zwischen Autoren, Redaktion, Herausgebern und Gutachtern bereit und erleichtert damit die Abwicklung der Kommunikationsprozesse sowie die Terminplanung des Review-Verfahrens. Im Rahmen des DiPP-Projekts wurde eine Version des *GAPworks*-Systems entworfen, die einen einheitlichen Zugang über das *Plone*-System erlaubt, so dass die Nutzer den Eindruck haben, mit *einem* System zu arbeiten. Der einheitliche Zugang wird über die Internetseite des Journals ermöglicht (http://www.brains-minds-media.org).

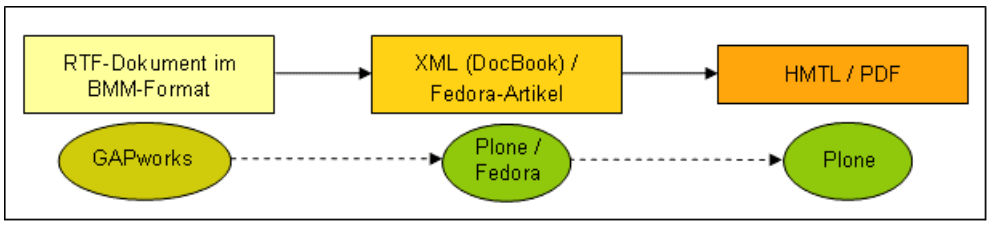


Abbildung 2: Konvertierungsschritte während der Publikation (oben) und die beteiligten Systeme (unten).

Im Zusammenhang der Publikation sind vor allem die Formatkonvertierungen wichtig: In *GAPworks* wird ein Artikel als rtf-Formatvorlage (von Autoren zu verwenden) hochgeladen, die bei Publikation in *Plone* importiert und nach XML (genauer: DocBook) konvertiert wird. Nach der Konvertierung liegen Artikel im *Fedora* Format vor und können in *Fedora Repositories* verwaltet werden. Bei dieser Konvertierung werden in *GAPworks* erstellte Metadaten mitgeliefert. Das XML-Dokument wird dann über individuelle Style-Sheets nach HTML konvertiert. Bei Bedarf kann eine Druckversion im PDF-Format erzeugt werden. Abbildung 2 illustriert die einzelnen Konvertierungsschritte.
Das BMM-Format ist die .dot-Formatvorlage des Journals, die von *Plone* erkannt und konvertiert wird. Eine konsequente (und verbindliche) Verwendung dieser Formatvorlage durch Autoren reduziert die redaktionelle Arbeit für das Tandem pro Artikel auf ein Minimum.

4. Workflow

Aus der Verlagsunabhängigkeit des Journals und den o.g. Rahmenbedingungen ergibt sich der spezifische Workflow der Zeitschrift. Grundsätzlich werden die Arbeitsprozesse in den o.g. zwei Systemen durchgeführt. Die Artikelbegutachtung erfolgt in *GAPworks*, die Veröffentlichung eines Artikels erfolgt in *Plone*. BRAINS, MINDS & MEDIA hat zum Ziel, ein zügiges Gutachterverfahren ohne qualitative Abstriche einzuführen. Artikel sollen innerhalb von höchstens drei Monaten zur Veröffentlichung gebracht werden. Während dieses Idealfalls spielt sich folgender Workflow ab:
Nach der Annahme des Artikels durch die Fachwissenschaftler am Lehrstuhl für Neurobiologe (LS) und der Entscheidung darüber, ob die Zielsetzungen des Journals durch den eingereichten Artikel abgedeckt sind, wird jeder eingereichte Artikel durch die Bibliothek (UB) auf Formatkonsistenz und -kompatibilität geprüft und ggf. an die Autoren zurückgegeben. Eventuell angehängte Materialien werden von der UB auf lauffähig getestet, d.h. ob sie zu öffnen und ausführbar sind. Falls für die Materialprüfung spezielle Software erforderlich ist, wird diese vom LS zur Verfügung gestellt.

Eine inhaltliche Beurteilung obliegt den Gutachtern. Grundsätzlich sind fast alle Formen von Materialien möglich, von Primärdaten bis hin zu elaborierten Anwendungen. Tabelle 1 gibt einen Überblick über mögliche bzw. zu erwartende Formate. Alle Materialien müssen von Autoren als Zip-Datei geliefert werden.
In *GAPworks* erstellt die UB eine Kurztitelaufnahme und vergibt ggf. Metadaten. Der eingereichte Artikel tritt in den *Review Cycle* (Abbildung 3) ein: die Redaktion leitet den Artikel über das *Editorial Board* an mindestens zwei Gutachter weiter (*Review-Request*), die in den folgenden sechs Wochen Zeit zur Begutachtung haben. Danach gelangt der Artikel, ggfs. mit Anmerkungen und Änderungswünschen der Gutachter, zurück an die Redaktion. Der LS leitet den Artikel für evtl. Änderungen an die Autoren, die innerhalb von zwei Wochen das redigierte Dokument an die Redaktion zurücksenden. Von dort aus erhalten die Gutachter den Artikel zur Prüfung. Sollten die Gutachter mit der redigierten Fassung einverstanden sein, liegt nach einer weiteren Woche die publikationsfähige Endfassung vor und der Autor erhält von der Redaktion eine Zusage über die Veröffentlichungsempfehlung der Gutachter. Andernfalls wiederholt sich der Zyklus.

Materialtyp	Format	Browser	Windows	Linux	Mac OS
Programm	Java-Applet	x			
	.jar (Javaprogramm)		x	x	X
	.exe (ausführbares Prog.)		X		
	.sh (ausführbares Prog.)			x	X
Datensätze	ASCII- Tabelle (für Import)		X	x	X
	SPSS		X		
	Statistica		X		
	MatLab		X		
Skripte	XML und Derivate	x			
	GENESIS		X	x	
	NEURON		X	x	
Medien	Gif-Animationen	x	X	x	X
	Flash-Animationen	x	X	x	X
	Filme (.mov, .mpeg, etc.)		X	x	X

Tabelle 1: Überblick der möglichen Materialtypen und Formaten sowie Plattformen (inkl. deren Derivate), unter denen diese Formate lauffähig sind. Browserfähige Formate, wie Java-Applets, Flashanimationen, mpeg-Movies etc. stellen die einfachsten Formate dar. Aber auch lokal zu installierenden Programme oder Datensätze und Skripte für bestimmte, weit verbreitete Statistikprogramme oder Simulationsumgebungen sind erwünscht, wie SPSS™, MatLab™ oder frei verfügbare

Simulationsprogramme, die in der Leserschaft verbreitet sind, z.B. GENESIS, NEURON etc. Die Tabelle erhebt keinen Anspruch auf Vollständigkeit, sondern stellt lediglich eine Beispielhafte Aufstellung zu erwartender Formate dar.

Der Review-Prozess ist damit abgeschlossen; es folgt der Publikationsprozess (Abbildung 3), der pro Artikel zwei Wochen umfassen soll: Der bisher in *GAPworks* vorliegende Artikel wird in das Publikationssystem *Plone* eingepflegt. Dies erfolgt durch einen automatischen Import aus *GAPworks* ins XML-Format und durch Formatkonvertierung anhand von Style-Sheets (siehe III. Technisches Basissystem).

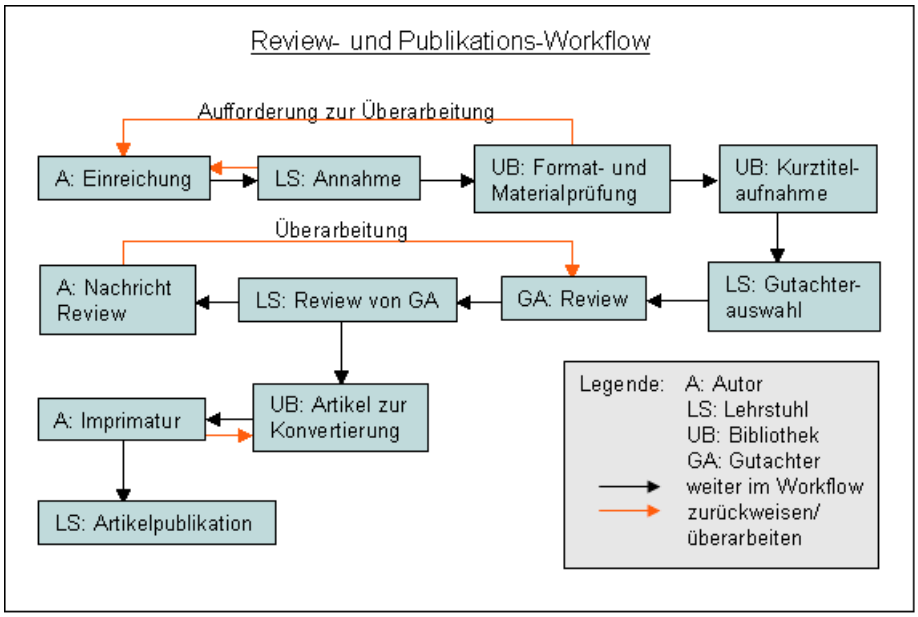


Abbildung 3: Vereinfachte Darstellung des beschriebenen Workflows für einen Artikel. Dargestellt sind Review-Prozess und Publikationsprozess.

Die UB kontrolliert diesen Import- und Formatierungsvorgang und ergänzt ggf. einzelne (Meta-)Daten. Die publikationsfähige Endversion wird durch den Autor freigegeben; zur Erteilung der Imprimatur hat der Autor eine Woche Zeit. Nach einem letzten Check durch die Fachwissenschaftler wird der Artikel umgehend auf der Homepage des Journals veröffentlicht.

V. Diskussion

In diesem Beitrag wurde das eJournal BRAINS, MINDS & MEDIA und seine spezifischen Merkmale vorgestellt und in den Kontext der aktuellen Open Access-Entwicklungen

eingebettet. Die Rollenaufteilung und der Workflow des lokalen Redaktionstandems wurden vor dem Hintergrund der Verlagsunabhängigkeit des Journals beleuchtet. Dabei stellt sich die Arbeitsteilung innerhalb des Redaktionstandems zusammenfassend folgendermaßen dar: Die Fachwissenschaftler akquirieren und betreuen die Autoren der *scientific community* von der Einreichung des Papers über die Organisation des Peer-Review-Verfahrens bis zur Veröffentlichung. Die Mitarbeiter der Universitätsbibliothek überprüfen einen jeden eingereichten Artikel auf Formatkonsistenz und -kompatibilität und flankieren diese Aktivitäten durch bibliothekarisches Know-How. Möglich macht diese Aufteilung das vom HBZ betreute Peer-Review- und Publikationssystem.
Neben der Kontinuität dieser in enger Kooperation zwischen den Fachwissenschaftlern und den Bibliotheksmitarbeitern geregelten Arbeitsabläufe gilt es, die Nachhaltigkeit des Journals auch nach Ablauf der Förderung zu sichern und ein tragfähiges Betriebsmodell zu entwickeln. Die für den Aufbau des Journals notwendige Infrastruktur und entsprechende Arbeitsläufe aufzubauen und zu etablieren war eine Aufgabe der Startphase von BRAINS, MINDS & MEDIA.

a. Betriebsmodell

Verlagsunabhängige Open Access-Publikationen sind zwar für den Nutzer kostenfrei, aber deshalb nicht kostenlos. Zur Finanzierung von Open Access-Publikationen finden z. Zt. mehrere non-profit-Geschäftsmodelle zur Deckung der entstehenden Kosten Anwendung – Gewinnmaximierung spielt beim Open Access keine Rolle. Langfristig durchzusetzen scheinen sich zwei Möglichkeiten der Finanzierung: ein verbreitetes Modell beinhaltet von den Autoren getragene Artikelbearbeitungsgebühren (*„author pays" model*, inzwischen auch in den Portfolios traditioneller Wissenschaftsverlage enthaltene Open Access-Option, z.B. Springer Open Choice (http://www.springeronline.com/), bisher jedoch oft verbunden mit prohibitiven Autorengebühren). In der *scientific communitiy* der Neurowissenschaften ist das *author pays*-Modell bereits gängige Praxis, nicht nur bei Open Access-Journalen (auch einige Printjournals erheben *page charges*). Voraussetzung für die Akzeptanz von Autorengebühren ist allerdings ein ausreichend hoher *Impact Factor* des Journals.
Eine weitere Finanzierungsmöglichkeit besteht in der Erhebung eines pauschalen Mitgliedsbeitrags für Institutionen, der die Autoren dieser Institution von den Autorengebühren befreit. Solche *institutional memberships*, wie z.B. bei BioMedCentral, eignen sich wahrscheinlich jedoch erst für etablierte, größere Open-Access-Anbieter.

b. Nachhaltigkeit

Die Förderung der Neugründung von BRAINS, MINDS & MEDIA durch das Land NRW sowie durch die Universität Bielefeld, die – im Sinne z.B. der Berlin 3 Open Access-Empfehlungen – günstige Rahmenbedingungen geschaffen hat, ermöglichte den

Aufbau einer technischen und redaktionellen Infrastruktur für das eJournal. Durch die Etablierung des Redaktionstandems sowie durch die Automatisierung wesentlicher redaktioneller Arbeitsschritte kann der zukünftige Arbeitsaufwand der beteiligten Institutionen und Personen relativ gering gehalten werden.
Eine wesentliche Voraussetzung für die Aufrechterhaltung dieses minimalen Aufwands ist jedoch die Mitarbeit der Autoren. Der Vorbereitung ihrer Manuskripte kommt durch die aus Zeitgründen eingeschränkten Möglichkeiten der redaktionellen Nachbearbeitung eine besondere Bedeutung zu, vor allem im Hinblick auf die zusätzlichen Materialien. Da dies jedoch bei bereits etablierten Online-Zeitschriften - ob kostenpflichtig oder Open Access - bereits gängige Praxis ist, ist dies für Autoren trotz des erhöhten Aufwands durch die Materialien eine gewohnte Aufgabe. Im Gegenzug ist die Veröffentlichung jedoch auch derzeit noch nicht mit Gebühren für Autoren verbunden.
Vor allem die Bereitschaft der Universitätsbibliothek Bielefeld, sich an der technischen Formatprüfung zu beteiligen hilft, den Betrieb der Zeitschrift neben den Kernaufgaben des Lehrstuhls zu ohne zusätzliche Mittel bewältigen. Es sei jedoch an dieser Stelle betont, dass ein neu gegründetes Journal mit kleinem oder gar keinem Budget nicht ohne ein großes Maß an Idealismus und Eigeninitiative der beteiligten Personen überlebensfähig ist.

Creative Commons-Lizenzen für Open Access-Dokumente

Jochen Brüning, Rainer Kuhlen; Konstanz

Abstract

Gegen die weltweit erkennbare Kommerzialisierung von Wissen und Information – in Deutschland abgesichert durch die Anpassung des deutschen Urheberrechts an EU-Vorgaben – formiert und artikuliert sich zunehmend Widerstand, insbesondere im Bildungs- und Wissenschaftsbereich. Open Access-Publikationen sind ein Weg, den in den Berliner[1], Göttinger[2] und Wiener[3] Erklärungen geforderten freien Zugang zu Wissen und Information insbesondere für Bildung und Wissenschaft zu ermöglichen. Um Autoren und Nutzern von Open Access-Publikationen Rechtssicherheit zu geben, ist eine Lizenzierung der Werke jeder medialen Art sinnvoll, die deren Nutzungsumfang eindeutig regelt. Eine dafür geeignete Lizenz muss verschiedene Kriterien erfüllen. Was die Creative Commons-Lizenzfamilie auszeichnet, welche Varianten und Freiheitsgrade den Autoren und Kulturschaffenden zur Verfügung stehen, wird ebenso wie deren Generierung / Anwendung erläutert.
Ebenfalls wird die Notwendigkeit einer (bislang noch fehlenden) digitalen Signatur von (CC-) lizenzierten Dokumenten angesprochen, durch die den gerade in der Wissenschaft wichtigen Anforderungen an Identität von Autoren, Authentizität der Werke und zeitliche Fixierung des Publikationsdatums entsprochen werden kann.

Sorgt das Urheberrecht noch für einen gerechten Interessen-ausgleich?

Auch wenn Stimmen laut werden[4], die geistiges Eigentum als ein in der „Wissens-gesellschaft“ überholtes, ja für die weitere positive gesellschaftliche und wirtschaft-liche Entwicklung schädliches Konzept ansehen, der bislang dominierende, offizielle und realistische/pragmatische Ansatz sieht einen gerechten und fairen Interessens-augleich zwischen den Autoren, Rechteinhabern (Verwertern) und Nutzern als einzig

[1] http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html
[2] http://www.urheberrechtsbuendnis.de/
[3] http://www.chaoscontrol.at/we.htm
[4] http://www.libresociety.org/library/libre.pl/Libre_Commons Noortje Marres (University of Amsterdam), Soenke Zehle und Geert Lovink auf der Incommunicado-Konferenz 2005 Amsterdam. Gegenteilige Position nimmt Péter Benjamin Tóth im Artikel „CREATIVE HUMBUG“ http://www.indicare.org/tiki-read_article.php?articleId=118 ein, indem er *Creative Commons* als kommunistisches Schreckensgespenst darstellt; vgl. Rainer Kuhlen „Wem gehört das Wissen“, http://www.inf-wiss.uni-konstanz.de/People/RK/Vortraege01-Web/publikationstext.pdf

mögliches Ziel an. Das Recht auf freien Zugang zu Wissen (einschließlich Information und Daten) wird unter anderem auch damit begründet, dass dieses nur einen momentan erreichten Zwischenstand darstellt, der aus dem entstanden ist, was von unseren Vorfahren bereits an Erkenntnissen erarbeitet wurde[5]. Entsprechend ist es nur konsequent, dass wir unser erarbeitetes Wissen restriktionsfrei gemäß *Open Access*-Prinzipien, der Gegenwart und den nachfolgenden Generationen zur Verfügung zu stellen. Ob man dieses Ziel auf der „*green* "- oder der „*golden"- road to Open Access*[6] anstrebt (gemeint ist damit das Publizieren *green:* auf Autoren- / Instituts- oder Fachgesellschaftsservern oder *golden:* in *Open Access*-Journalen): Auch das Publizieren nach dem *Open Access*-Modell bewegt sich nicht im rechtsfreien Raum. Die „Spielregeln" für den Interessensausgleich legt weiterhin das jeweils länderspezifische Urheberrecht fest. *Creative Commons* stellt sich nicht außerhalb des geltenden (Urheber-) Rechts. Sind jedoch, um bei dem Bild zu bleiben, die „Karten ungleich verteilt" – und viele Produzenten, unter ihnen Autoren und Kulturschaffende ebenso wie Nutzer und Konsumenten, empfinden dies so – dann mag dies der tiefere Grund für die eingangs erwähnte radikale Ablehnung bzw. zumindest für eine weitgehende Skepsis gegenüber der Nützlichkeit des Urheberrechts sein. Es stehen zwei sich nicht ausschließende Wege offen, das Ungleichgewicht auf weniger drastische Weise zu ändern:

- Zum einen wird versucht, auf die Gesetzgebung Einfluss zu nehmen, wie es im Aktionsbündnis „Urheberrecht für Bildung und Wissenschaft"
- http://www.urheberrechtsbuendnis.de, der Berliner, Göttinger und jüngst der Wiener Erklärung geschehen ist, die via Ausnahmeregelung (Schranken) /Änderungen /Ergänzungen[7] des Urheberrechts zugunsten eines freieren Zugriff für Wissenschaft und Bildung fordern.

[5] „If I have been able to see further it is because I have stood on the Shoulders of Giants". WIKIPEDIA: Isaac Newton zitiert Didacus Stella http://de.wikipedia.org/wiki/Auf_den_Schultern_von_Giganten

[6] Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. (2004) The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review 30. http://dx.doi.org/10.1016/j.serrev.2004.09.013
Shorter version: The green and the gold roads to Open Access. Nature Web Focus. http://www.nature.com/nature/focus/accessdebate/21.html

[7] Gerd Hansen schlägt vor, das Urheberrecht dahingehend zu erweitern, dass Autoren ein unabdingbares *Zweitveröffentlichungsrecht* zusteht, um wissenschaftliche Beiträge aus überwiegend öffentlich geförderter Lehr- oder Forschungstätigkeit auf eigenen Web-Seiten, Instituts- oder Archivservern zu nicht kommerziellen Zwecken nach OA-Prinzipien verfügbar machen zu können. Dies allerdings erst nach einer 6-monatigen Frist, in der den Verlagen das exklusive Veröffentlichungsrecht zusteht www.heise.denewsticker/meldungen/print59496. T. Pflüger und D. Ertmann gehen mit ihrem Vorschlag noch weiter, indem sie für Wissenschaftler im öffentlichen Dienst eine

- Zum anderen gilt es, durch kreative Anwendung des Urheberrechts, aber in seinem Rahmen bleibend dessen Potenzial für *Open Access*-Publikationen auszuschöpfen. Dafür ist eine Lizenzierung der zu publizierenden Werke notwendig, da ohne diese alle „Schutzbestimmungen" – oder vom Standpunkt der Nutzer aus gesehen, alle „Restriktionen" – des Urheberrechts automatisch, d.h., ohne jedes Zutun des Autors, völlig unabhängig von der medialen Art des Werks und der Art seiner Veröffentlichung, zum Tragen kommen. Dies gilt demnach auch uneingeschränkt für Netzpublikationen. Es reicht also nicht, ein Werk für Jedermann im Netz sichtbar – oder weiter gehend auch zum *download* – bereitzustellen, um Dritten legale Verwendungsmöglichkeiten und -rechte einzuräumen.

Rechtssicherheit für Autoren und Nutzer

Bis jedoch der erste Weg, nämlich die Änderung des Urheberrechts, zum Erfolg führt, sollten *Open Access*-Werke nur mit einer Lizenz versehen publiziert werden. Dabei hat die Lizenz bestimmte Kriterien zu erfüllen. Sie muss, um auch in Deutschland für Jedermann rechtsverbindlich zu sein – und das ist bei den wenigsten der über 30 allein für den *Content*-Bereich (der Softwarebereich ist hier nicht eingeschlossen) zur Verfügung stehenden Lizenzen der Fall –, an das nationale (hier deutsche) Urheberrecht angepasst und in deutscher Sprache formuliert sein. Um den unterschiedlichen Wünschen der Autoren bezüglich der an die Öffentlichkeit abzugebenden bzw. der beim Autor verbleibenden Autorenrechte zu entsprechen, sollte es darüber hinaus verschiedene Varianten der Lizenz geben.

- Zum einen muss einem juristischen Laien – und das dürfte der Großteil der angesprochenen Autoren und Nutzer sein – der Sachverhalt, also Rechte und Pflichten, die sich aus der Nutzung eines lizenzierten Werks ergeben, klar vermittelt werden.
- Zum anderen muss der Lizenztext juristisch einwandfrei formuliert sein, so dass im Falle von Rechtsstreitigkeiten – die ja durch die Lizenzierung gerade vermieden werden sollen, aber nicht auszuschließen sind – kein Auslegungsspielraum besteht.
- Zum dritten sollte der Lizenzinhalt (kodiert, also nicht notwendigerweise der Text selbst) auch in maschinell lesbarer Form vorliegen, um mit dem Werk verlinkt werden zu können. Damit können lizenzrelevante Parameter ma

Anbietungspflicht vorschlagen. Wissenschaftler hätten demnach das von ihnen zur Veröffentlichung vorgesehene Werk zuerst der Institution, der sie angehören, zur Veröffentlichung im institutseigenen Verlag oder auf dessen Server anzubieten. Dies verbunden mit dem Recht, das Werk auch kommerziellen Verlagen anbieten zu können http://www.ub.uni-konstanz.de/kops/volltexte/2004/1337/ .

schinell indexiert werden. Es wird dadurch ein gezieltes Suchen nach OA-Publikationen mit den gewünschten Freiheiten möglich.

- Nicht zuletzt sollten die Metadaten der Werke ebenfalls in maschinenlesbarer Form und DC-standardkonform erfasst und der Publikation beigefügt werden.

Hinsichtlich Flexibilität und Autonomie im Umgang mit den Autorenrechten sowie Rechtssicherheit für Autoren und Nutzer zeigen sich die *Creative Commons*-Lizenzen für *Open Access* als hervorragend geeignet. Dies ist sicherlich ein Grund für deren große Akzeptanz im Wissenschafts- sowie auch im kreativ-künstlerischen Bereich. Mehr als 17 Mio. (Stand Mitte 2005) Werke wurden bisher damit ausgezeichnet.

Anleitungen für den praktischen Einsatz der CC-Lizenzen finden sich zusammen mit allgemeinem Informationsmaterial auf dem ausgestellten Poster (http://www.inf-wiss.uni-konstanz.de/cc/cc_poster_150405_klein.pdf) und unter: http://www.inf-wiss.uni-konstanz.de/People/JB/#download.

Freiheitsgrade der Creative Commons-Lizenzen

Grundsätzlich ist jedem die Vervielfältigung und Verbreitung[8] CC-lizenzierter Werke erlaubt, hierin sind die Lizenzen völlig OA-konform. Hinzu kommt das Recht der öffentlichen Aufführung, sind doch die CC-Lizenzen gleichermaßen für Autoren wie Kulturschaffende konzipiert. Mit diesen Rechten ist die Verpflichtung verbunden, den Urheber (Autor, Komponist, Künstler …) zu nennen, einen Hinweis auf die Fundstelle (bei Netzpublikationen also einen URI, vorzugsweise als URN oder DOI, weniger als flüchtigen URL) anzugeben sowie einen Hinweis (i.A. einen Link) auf die Lizenz (-variante, dazu weiter unten) in einer der technischen Verbreitungsform des Werks angemessenen Weise[9] einzufügen. In der aktuellen Version der CC-Lizenzen kann anstelle des Urhebers oder zusätzlich dazu eine (Förder-) Institution oder ein (Gemeinschafts-) Projekt angegeben werden. Die letztere Möglichkeit trägt der in elektronischen Räumen zunehmenden Praxis des kollaborativen Arbeitens, z.B. in Wikis, Rechnung. Der Lizenzgeber hat die Wahl, weitere, ihm durch das Urheberrecht garantierte Rechte an die Öffentlichkeit zurückzugeben. Dazu gehören das Recht der kommerziellen Verwertung und die Möglichkeit, neben der wörtlichen Weitergabe auch Derivate (abgeleitete Werke, Bearbeitungen durch Dritte) und

[8] Hierzu gehört auch die Vervielfältigung und Verbreitung in einem vom Original abweichenden Format. So darf ein vom Autor im digitalen Format (z.B. eine Netzpublikation) veröffentlichtes Werk durchaus auch in analoger Form, also gedruckt, verbreitet werden. Übersetzungen in andere Sprachen hingegen sind per se nicht erlaubt. An einer eigenen Lizenzvariante für Übersetzungen wird gearbeitet.

[9] Sichtbarer Hinweis bei Druckmedien, eingebettete Metadaten bei digitalen Formaten, ID-Tags bei MP3-kodierter Musik, …

deren Verbreitung zu gestatten, wobei für den letzteren Fall dies mit den Auflagen verbunden werden kann, die so entstandenen Werke nur unter den Lizenzbestimmungen der Originalarbeit zu verbreiten und / oder dabei – dies im Widerspruch zur ursprünglichen Forderung – die Namensnennung des Urhebers der Originalarbeit zu unterlassen. Mit den genannten Lizenzgrundelementen Namensnennung, kommerzielle Verwertung, Verbreitung von Derivaten ggf. mit der Verpflichtung, die Lizenz des Originals beizubehalten, lassen sich sechs Lizenzvarianten formulieren. Diese Lizenzierung ist unwiderruflich, d.h. einmal per Lizenz abgetretene Rechte können nachträglich nicht mehr eingefordert werden. Hingegen ist eine Relizenzierung mit geringeren Rechtsvorbehalten durchaus möglich.

Lizenzgenerierung und -format

Über ein intuitiv zu bedienendes Web-Interface kann unter der Web-Adresse http://www.creativecommons.org/license mit wenigen Mausklicks eine den Wünschen des Urhebers gerecht werdende Lizenzvariante (siehe vorangehender Absatz) generiert werden. Dazu gehört auch die Wahlmöglichkeit für die der Lizenz zu Grunde liegende Juristikation, wobei über interne Abfrage der Web-Browser-Einstellung die jeweilige Ländervariante[10] vorgeschlagen wird. In einem Arbeitsgang können auf derselben Web-Seite zusätzlich Metadaten zum Werk erfasst werden, die dann zusammen mit der Lizenzinformation im „Dreierpack“ dem Autor zur Integration in sein Werk zur Verfügung gestellt werden. Dieser „Dreierpack“ selbst ist abhängig vom Format der zu lizenzierenden Arbeit entweder ein zur Auszeichnung von Web-Seiten und XML-/ html- Dokumenten geeignetes html-Code-Fragment oder eine XMP-Datei, geeignet zur Auszeichnung von pdf-/ Photoshop-/… Dokumenten. Für die Auszeichnung von Audiodateien im MP3-Format steht eine spezielle Schnittstelle (http://ccmixter.org/) zur Verfügung. Im Werk sind nach der Einbindung URLs auf (1.) einen allgemein verständlichen Text, der die gewählten Verpflichtungen und Rechte der Lizenzvariante erläutert, (2.) die juristische Formulierung der Lizenzbestimmungen und (3.) die maschinenlesbaren Metadaten enthalten. Die CC-Lizenz wird so integraler Bestandteil des Werks, auch ohne dass der Lizenztext selber darin enthalten ist[11]. Durch die direkte Einbettung der Metadaten in das Werk hingegen eröffnen sich – das *Semantic Web* antizipierend – Möglichkeiten für die maschinelle Verarbeitung, die weit über einfache Suchfunktionen hinausgehen. Wird das Werk anschließend elektronisch signiert (siehe folgender Abschnitt), kann weder Werk

[10] Derzeit liegen die CC-Lizenzen in 21 an das nationale Urheberrecht angepasste Fassungen vor, an weiteren Anpassungen wird im Rahmen des *iCommons*-Projekts gearbeitet. Diese Anpassung ist ein weiteres Merkmal, das die CC-Lizenzen von anderen *content*-orientierten Lizenzen abhebt.

[11] Auch hierin unterscheiden sich die CC-Lizenzen von andern *Open Content/Source*-Lizenzen.

noch Lizenz geändert werden, ohne dass dies durch eine als ungültig gekennzeichnete Signatur sichtbar wird.
Um den Anforderungen spezieller Klassen / Genres (Wissenschaft, Lehre, Musikszene, Dritte Welt, …) gerecht zu werden, wurde neben der hier erläuterten allgemeinen (*generic*) CC-Lizenz eine Reihe speziell darauf abgestimmter Lizenzen entwickelt, die ebenfalls über die oben genannte Web-Seite generiert werden können.

Urheberschaft, Werkauthentizität und Publikationsdatum

Es zeigt sich jedoch, dass selbst im Umgang mit lizenzierten *Open Access*-Dokumenten die folgenden, im wissenschaftlichen Umfeld besonders relevanten Fragen nicht oder nur unzulänglich mit der notwendigen Sicherheit beantwortet werden:

- Wer ist der Autor / Urheber eines Werks (Authentizität)?
- Ist das vorliegende Werk unversehrt (Integrität)?
- Wann wurde das Werk öffentlich gemacht, bzw. welche Version des Werks liegt vor?

Antworten auf diese Fragen könnten vorzugsweise[12] durch eine digitale Signatur des Werks gegeben werden. Durch die Signatur wird ein eindeutiger Bezug des signierten Werks zum Signierenden (in der Regel dem Autor, ersatzweise einer Registrierungsautorität) hergestellt. Die Signatur enthält einen vom Werk abgeleiteten verschlüsselten hashcode, der jede Veränderung am Werk nach dessen Signierung zuverlässig anzeigt. Bei der Signierung wird zudem ein Zeitstempel generiert und in die Signatur eingebunden. Signaturen nach dem X.509-Standard sind – liegt ein entsprechend fortgeschrittenes oder qualifiziertes Zertifikat zu Grunde und entspricht die verwendete Signiereinrichtung den Vorgaben – signaturgesetzkonform und damit rechtsverbindlich.

[12] Der vorgeschlagene Einsatz von DRM-Systemen für diesen Zweck http://www.indicare.org/tiki-read_article.php?articleId=92 bis articleID=95, =98 und =106 bis =108 ist u.E. nicht mit den *Open Access*-Prinzipien vereinbar bzw. konform. Der ebenfalls vorgeschlagenen Verwendung digitaler Wasserzeichen für diese Zwecke fehlt die rechtliche Relevanz.

Zusammenfassung

Um Rechtssicherheit zu erlangen, und somit Autoren wie Nutzern einen sorgenfreien Umgang mit den OA-publizierten Werken zu ermöglichen, zeigt es sich, dass *Open Access*-Publikationen einer den genannten Ansprüchen genügenden Lizenzierung bedürfen, und die so ausgezeichneten Werke digital zu signieren sind.

Kooperationsmodelle für Open Access eJournals in der Publikationsinitiative DiPP NRW

Wolfram Horstmann, Köln

Abstract

Die neuen Möglichkeiten des digitalen Informationsmanagements in der Wissenschaft und die Forderungen des offenen Zugangs zu Wissen (Open Access) verändern das Publikationswesen im Zeitschriftensektor. Neue Betriebs- und Geschäftsmodelle für Online-Zeitschriften traditioneller Machart werden entwickelt („goldener" Weg zu Open Access), Autoren veröffentlichen ihre Artikel aus traditionellen Zeitschriften parallel in Online-Archiven („grüner" Weg) und gänzlich neue Publikationsformen entstehen. Die Publikationsinitiative Digital Peer Publishing NRW fokussiert auf die Entwicklung neuer Publikationsformen. Sie wurde vom Ministerium für Wissenschaft und Forschung NRW ins Leben gerufen, um neuen, innovativ ausgerichteten eJournals in ihrer Entstehung, bei ihrem Aufbau und in ihrem dauerhaften Betrieb unter Open Access Bedingungen zu unterstützen. In der Stimulationsphase wurden seit dem Frühjahr 2004 acht eJournals in einem breiten disziplinären Spektrum von Geisteswissenschaften über Technik bis zu Naturwissenschaften auf- und ausgebaut (s.a. http://www.dipp.nrw.de). Parallel wurde im Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) eine leistungsfähige Infrastruktur für eJournals entwickelt, die nicht nur technische, sondern auch geeignete organisatorische und rechtliche Rahmenbedingungen bietet. Die Dienstleistungen können nach dem individuellen Anforderungsprofil der eJournals maßgeschneidert werden. Zunehmend nutzen auch weitere, neue und bestehende eJournals die Angebote der Initiative.
Die komplexen und anspruchsvollen Prozesse bei der Gestaltung eines eJournals in der redaktionellen Arbeit, der qualitätssichernden Begutachtung (z.B. „Peer-Review"), der bibliographischen Erschließung und der technischen Realisierung der eJournals wird von einem hochgradig differenzierten Netzwerk von Wissenschaftlern, Bibliothekaren und IT-Spezialisten aus unterschiedlichsten Disziplinen getragen. Abhängig von den in der jeweiligen Fachkultur zu findenden Qualifikationsprofilen der Akteure ist die konkrete Realisierung der Kooperationen bei einem einzelnen eJournal sehr individuell organisiert: in einigen Fällen werden inhaltliche, technische, bibliografische und administrative Aufgaben in Personalunion durchgeführt, in anderen ist eine strikte personelle Trennung im redaktionellen Workflow verteilt über verschiedene Hochschuleinrichtungen realisiert. Die praktische Arbeit der eJournals in der DiPP Initiative zeigt, welche Kooperationsmodelle sich als geeignet erweisen.

Entwicklungen im Publikationswesen

Neue Publikationsformen entstehen fortlaufend, da sich wissenschaftliche Disziplinen ständig ausdifferenzieren und neue ‚Organe' hervorbringen. Hinzu kommen die

neuen technischen Möglichkeiten des elektronischen Publizierens im Internet. Hiermit verändern sich auch die Akteure, die am elektronischen Publikationsprozess beteiligt sind und ihre Rollen: Die herkömmliche Trennung von inhaltlicher Redaktion und technischer Produktion von Druck und Satz wird mehr und mehr aufgelöst, da die technischen Prozesse *en passant* von Autoren und Redakteuren getragen werden. Schließlich entfällt durch die vereinfachten Veröffentlichungsformen, die das Internet bietet, die zwingende Notwendigkeit des aufwändig produzierten Druckerzeugnisses. Dies gilt besonders für Zeitschriften, deren bedeutendste semantische Einheit ein einzelner Artikel ist, der sich auch auf dem lokalen Drucker einfach auf Papier bringen lässt, so dass die ästhetische Überlegenheit der vorgefertigten Druckform – etwa im Vergleich zum kompletten Buch – vernachlässigbarer wird. Die schnellere und effizientere Verbreitung (Brody & Harnad 2004) und letztendlich die Möglichkeiten zur Integration interaktiver Medien können schließlich zu einer Unabdingbarkeit der elektronischen Publikation führen.

Eine neue ökonomische, aber auch ethische Qualität bekommt die elektronische Publikation wissenschaftlicher Ergebnisse durch die Möglichkeit, jedem Menschen von jedem Ort auf der Welt über das Internet einen offenen Zugang zum Wissen zu verschaffen. Unter dem Schlagwort „Open Access“ arbeiten seit einigen Jahren Initiativen vor allem aus Bereichen der Wissenschaft, aber auch aus dem Bibliothekswesen, die diese neue Qualität in der Wissenschaft etablieren wollen. Die bekanntesten Initiativen sind die sind die „Budapester“ (Open Society Institute 2002), die „Bethesda“ (Suber 2003), und „Berliner“ (Max-Planck-Gesellschaft 2003). Förderinstitutionen, etwa die DFG (Deutsche Forschungsgemeinschaft 2005a) oder das NIH (National Institute of Health 2005a) beginnen den offenen Zugang als Mandat zu verstehen und in die Richtlinien zur Veröffentlichung der Ergebnisse aus den von ihnen (oft aus Steuermitteln) finanzierten Projekten zu verankern. Selbst die großen Verlage erlauben in großem Rahmen die parallele, offene Veröffentlichung von Artikeln ihrer Zeitschriften durch die Autoren (Sherpa 2005) und es entstehen Modelle, in denen Autoren (bzw. ihre Geldgeber, meist ebenfalls Steuergelder) die Publikationskosten tragen, damit die Artikel letztendlich kostenfrei im Internet erscheinen können (z.B. BioMed Central Ltd 2005, PLoS 2005a).

Hier verzweigt sich die Open Access Bewegung auch in mindestens zwei separate, aber langfristig komplementäre Bereiche – einen so genannten „grünen Weg“ , der auf eine offene Nebenverwertung ansonsten lizenzkostenpflichtiger, traditioneller Zeitschriftenartikel abzielt und einen so genannten „goldenen Weg“ , der auf einen lizenzkostenpflichtigen Zugang zum Wissen von vorne herein verzichtet und stattdessen alternative Betriebs- und Geschäftsmodelle etabliert (Harnad et al. 2004). Diese beiden Wege besetzen grundsätzlich verschiedene Arbeitsgebiete: Der grüne Weg behandelt die *Bereitstellung* des offenen Zugangs bereits produzierter und anderweitig veröffentlichter Ergebnisse, der goldene Weg behandelt die direkte *Publikation* unter den Voraussetzungen des offenen Zugangs. Die beiden Wege werden häufig irreführend als Alternativen dargestellt, verfolgen aber beide das

gleiche Ziel und sind langfristig sogar aufeinander angewiesen. Um die Kooperationsmodelle in der Initiative „Digital Peer Publishing“ im Lichte der aktuellen Open Access Diskussion zu verstehen, seien diese Hintergründe eingehender beleuchtet.

Schattierungen von Open Access

Der grüne Weg heißt „grün“, weil die Verlage, die einen Artikel veröffentlichen, den Autoren rechtlich „grünes Licht“ zur Parallelveröffentlichung in freien Dokumentarchiven im Internet (z.B. arXiv 2005, PubMed 2005, Cogprints 2005) geben. Da der grüne Weg somit auf das bereits etablierte Publikationssystem aufsetzt, kann er quasi ohne Übergangsphase rezeptartig verschrieben werden und somit unmittelbar auf breiter Basis seine Wirkung entfalten. Daher wird der Anspruch, hundert Prozent der wissenschaftlichen Ergebnisse im offenen Zugang zu veröffentlichen, von Meinungsführern als nicht unrealistisch betrachtet (Harnad 2005). Dieses Szenario eines offenen Zugangs zu hundert Prozent aller wissenschaftlichen Ergebnisse über die freien Dokumentarchive des „grünen Wegs“ wirft die Frage auf, wie das Publikationswesen langfristig organisiert wird, da ja nach wie vor für technische Infrastruktur, allgemeine Redaktion, Lektorat, Formatierung, Grafik, Rechte- und Lizenzmanagement, Koordination etc. Ressourcen benötigt werden? Wenn alle Publikationen frei in Dokumentarchiven zur Verfügung stehen, warum sollten Institute, Bibliotheken oder Einzelpersonen langfristig ihre Zeitschriftenabonnements behalten, wenn sie sonst sparen, wo sie nur können? Ohne die Einnahmen aus diesen so genannten Subskriptionsgebühren, wird aber wird den Zeitschriften konventioneller Verlage und Fachgesellschaften die Geschäftsgrundlage entzogen. Wer soll die Prozesse bezahlen, die typischerweise durch die Subskriptionsgebühren refinanziert werden?

Neue Modelle für Zeitschriften

Ein zu Ende gedachtes Szenario des „grünen Weges“ verdeutlicht, warum parallel zum „grünen Weg“ , der unmittelbar eine kritische Masse an Open Access Publikationen erzeugt, der „goldene Weg“ verfolgt wird: hier werden zukünftige Betriebs- und Geschäftsmodelle für Zeitschriften entwickelt und erprobt, die die Aufbereitung und Primärpublikation von Wissen für den offenen Zugang ermöglichen. Auf dem goldenen Weg werden wiederum mindestens zwei Strategien verfolgt, die ebenfalls zum Teil komplementär sind. Die erste Strategie kopiert im Prinzip die traditionelle wissenschaftliche Zeitschrift (online), kehrt aber das Geschäftsmodell um: an Stelle von Subskriptionsgebühren werden Autorengebühren verlangt. Die zweite Strategie entwickelt und erprobt neue Publikationsformen, die sich nicht direkt im traditionellen Publikationswesen abbilden lassen (s.u.).

Bei der ersten Strategie, die in den meisten Fällen gemeint ist, wenn der „goldene Weg“ genannt wird, ist neben BioMed Central (s.o.) wohl die Public Library of

Science (PLoS s.o.) das bekannteste Beispiel einer verlagsähnlichen Organisationsform. PLoS hat es innerhalb kurzer Zeit geschafft, Open Access Zeitschriften mit hoher Reputation aufzubauen (PLoS 2005b). Aber auch einzelne Zeitschriften, wie das „New Journal of Physics“ (Deutsche Physikalische Gesellschaft & Institute of Physics 2005) zeigen die Erfolgschancen dieses Modells. Die meisten dieser Ansätze sind nicht Profit orientiert. Doch selbst die großen kommerziellen Verlage wie Springer bieten an, Artikel gegen Autorengebühren online frei zur Verfügung zu stellen (Springer Science and Business Media 2005). Andere starten Open Access Journals, etwa das „Molecular Systems Biology“, das „Nature“ in Zusammenarbeit mit der „European Molecular Biology Organisation“ (EMBO) herausgibt (Nature Publishing Group & EMBO 2005). Autorengebühren werden in der Regel nicht von den Autoren privat getragen, sondern häufig aus den Fördermitteln bestritten. Die Geldgeber, gerade solche, die sich aus Steuergeldern speisen, bieten spezifische Fördermöglichkeiten, die zur Kostendeckung von Publikationen verwendet werden können. Ein Beispiel ist die Publikationspauschale der Deutschen Forschungsgemeinschaft DFG (Deutsche Forschungsgemeinschaft 2005b). Dieses Szenario des goldenen Weges zu Ende gedacht, könnte also letztendlich zu einer Verschiebung von eingesetzten Steuergeldern kommen, weg von Subskriptionsgebühren hin zu Autorengebühren.

Bei der Höhe der erhobenen Autorengebühren ist eine große Vielfalt zu beobachten. Letztgenanntes „Molecular Systems Biology“ etwa erhebt pro veröffentlichtem Artikel 3000 $ (ebenso Springer mit ihrem „Open Choice“ Modell), PloS 1500 $, BioMedCentral zwischen 400 und 1600 $ je nach Zeitschrift und das „Journal of Atmospheric Chemistry and Physics“ (European Geosciences Union 2005) zwischen 15 und 45 € pro Seite. Oft werden Vergünstigungen für weniger zahlungsstarke Autoren oder für Mitgliedschaften angeboten. Die Druckausgaben sind meist kostenpflichtig. Es gibt aber auch viele Open Access Zeitschriften, die keine Autorengebühren erheben, etwa „Documenta Mathematica“ (Rehmann 2005). In den Geisteswissenschaften ist es gänzlich unüblich, die Autoren zu belasten, so dass man hier vergeblich nach solchen Modellen suchen wird.

Die große Vielfalt der Autorengebühren-Modelle macht bereits deutlich, dass hier eigentlich eine Vielzahl ganz unterschiedlicher Publikationsstrategien vorliegt, die nicht einheitlich als „goldener Weg“ zu Open Access bezeichnet werden kann. Wollte man diese Vielfalt ordnen, müsste man eine Fülle von Ordnungskriterien berücksichtigen, etwa die disziplinäre Publikationskultur, die nationale Förderpolitik, ob Profitorientierung vorliegt, Quersubventionierung gemacht wird oder in welcher Form Mitgliedsbeiträge von Fachgesellschaften, Fördergelder oder öffentliche Ressourcen hinzugezogen werden. Ferner muss berücksichtigt werden, wie die Ressourcen eingesetzt werden: wie groß sind die Anforderungen und Budgets für die verschiedenen Arbeitsbereiche, etwa für Aufgaben im Betrieb wie Autoren- und Gutachterbetreuung, Ausgabenplanung, Formatierung, Lektorat oder für strategische Maßnahmen wie Distribution, Marketing oder technische Entwicklung?

Erfolgsfaktoren

Gibt es denn einen Zusammenhang zwischen den Gebühren, die verlangt werden und dem Erfolg einer Zeitschrift? Auch diese Frage lässt sich nicht eindeutig beantworten. Generell finden sich in bei Analysen der Maßzahlen für Zeitschriften, wie der „Journal Impact Factor“ (Garfield 1994), dass viele der Zeitschriften mit hohem JIF oder Verlage mit generell hoher Reputation auch hohe Autorengebühren nehmen. Aber wie der kausale Zusammenhang zwischen Erfolg und Autorengebühren letztendlich aussieht, ließe sich wohl nur in der Einzelfallbetrachtung klären[1].

Ob es überdies sachdienlich ist, allein den JIF als Bewertungsmassstab für eine Zeitschrift heranzuziehen, sei dahingestellt (vgl. Seglen 1997): Er stellt zwar faktisch zur Zeit *die* entscheidende, quantitative Messlatte für Zeitschriften im Sinne des traditionellen Publikationswesens dar, aber es gibt durchaus auch andere Aspekte, die berücksichtigt werden müssen: Zum einen bildet der JIF ausschließlich das traditionelle Publikationssystem ab, indem die Häufigkeit von Zitaten aller Artikel einer Zeitschrift analysiert werden. Dies liefert zwar zutreffende und empirisch belegbare Daten, sagt jedoch beispielsweise wenig über den Einfluss von Veröffentlichungen einzelner Artikel im Internet aus. Eine validierte und weithin akzeptierte Metrik zur Beurteilung der Nutzung wissenschaftlicher Ergebnisse im Internet steht noch aus (vgl. aber Björneborn & Ingwersen 2001, 2004). Zum anderen sind es eben zunehmend nicht nur die traditionellen „Artikel“, die als wissenschaftliches Ergebnis in Betracht kommen: Datensätze, Bilddaten, Modelle, interaktive Computerprogramme, Simulationen, und dynamische Textformen wie Diskussionen oder kollaborative Enzyklopädien beleben das traditionelle Monopol von statischem Text und Bild. Die Entwicklung von Begutachtungs- und Bewertungssystemen für solche Ergebnisse ist noch in den Kinderschuhen und es ist nicht unangemessen, kritisch zu hinterfragen, wie ein ähnlich potentes Qualitätssicherungsinstrument wie das „Peer-Review“ für solche Veröffentlichungen aussehen kann. Eben dies sind spezifische Fragestellungen, die die Zukunft des Publizierens bestimmen können.

Die Relativierung der konventionell verwendeten Erfolgsfaktoren durch neue Typen und Nutzungsformen von wissenschaftlichen Ergebnissen im Internet und verdeutlichen, dass der „goldene Weg“ zu Open Access sehr vielschichtig und eng mit grundsätzlichen strukturellen und funktionellen Aspekten des Publikationswesens –

[1] In jedem Fall können Autorengebühren nur dann durchgesetzt werden, wenn eine ausreichend hohe Reputation vorliegt, um die notwendige Einreichungsrate zu erzielen. Zeitschriften in der Start-Phase haben es schwer, vom Start weg Autorengebühren zu verlangen, da diese wertvolle Autoren von der Einreichung abhalten könnten. Sie unterliegen immer einer Art „Prestige-Paradox“ (Crow & Goldstein 2003): ohne Autoren keine Reputation und ohne Reputation keine Autoren. Immerhin hat PLoS als Beispiel einer Zeitschrift, die direkt mit Autorengebühren gestartet ist, allein seit der offiziellen Gründung im Oktober 2000 drei Jahre geplant und einen enormen Marketing-Aufwand betrieben, um die erste Ausgabe im Oktober 2003 zu veröffentlichen.

allgemeiner ausgedrückt des digitalen Informationsmanagements in der Wissenschaft – verknüpft ist, die unter anderem das Format einer Publikation an sich in Frage stellen.
Neben dem Pfad, dem goldenen Weg, auf dem in erster Linie das traditionelle Publikationssystem unter Veränderung des Geschäftsmodells kopiert wird, gibt es einen weiteren Pfad, der sich an spezifischen Fragestellungen des digitalen Informationsmanagements in der Wissenschaft orientiert und dabei Open Access anstrebt. Während der grüne Weg und der „traditionelle" goldene Weg einen sich am bestehenden Publikationssystem orientieren und auf dieser Basis einen programmatischen, zum Teil vorhersagbaren Charakter entwickeln, ist der „explorative" Weg im Entwicklungs- und Innovationssektor des digitalen Informationsmanagements in der Wissenschaft anzusiedeln[2]. Ob auf diesem „explorativen" Weg die Trennung zwischen „grün" und „gold" dauerhaft aussagekräftig bleibt, ist eine offene Frage.

Digital Peer Publishing

Die Initiative Digital Peer Publishing befindet sich auf dem zuletzt beschriebenen „explorativen" Weg. Sie strebt an, in der praktischen Arbeit mit Zeitschriften, Aufschluss über die Fragen des digitalen Informationsmanagements in der Wissenschaft unter Open Access Bedingungen zu erhalten, besonders an einer einzelnen Hochschule (s. Abbildung 1). Die Entwicklung neuer Kooperationsmodelle ist auf der strukturellen Ebene eine übergeordnete Frage der DiPP-Initiative: Wie können die ohnehin schon stark vom Wissenschaftssystem, d.h. von Autoren und Herausgebern getragenen Prozesse beim Publizieren innerhalb des Wissenschaftssystems systematisch verteilt werden, um eine möglichst schnelle, vielfältige und qualitativ hochwertige Informationslandschaft im offenen Zugang zu unterstützen?

[2] Da die im Entwicklungs- und Innovationssektor vorzufindende Vielfalt an Ansätzen oberflächlich gesehen einem „weißen Rauschen" ähnelt, könnte dieser Weg zu Open Access auch als „weißer Weg" bezeichnet werden. Aber diese weitere Differenzierung würde der Übersichtlichkeit der Open Access Bewegung auch nicht zuträglich sein. Schließlich ist es das Ziel dieser Diskussion, die Polarisierung in der Open Access Diskussion aufzuheben – und nicht den programmatischen Charakter der Bewegung aufzuweichen.

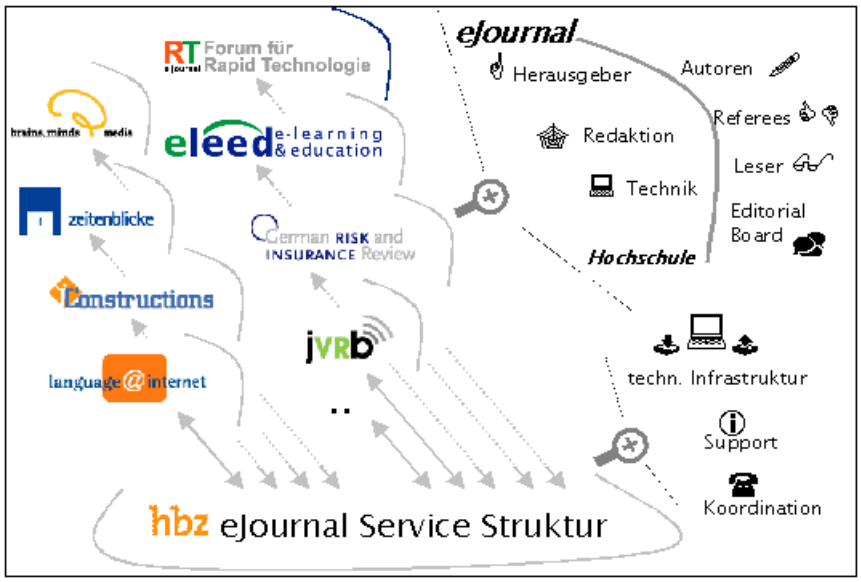


Abbildung 1: Grundsätzliche Organisation in DiPP. Die eJournals (Logos links oben), sind verankert an einer Hochschule (oben rechts) und regeln die Kommunikation mit der „Außenwelt“, also den Autoren und Gutachtern und Mitherausgebern. Die eJournals greifen dabei zurück auf die eJournal Service-Struktur von DiPP (unten).

Mindestens drei Grundmotivationen der Zeitschriften sind hier zu berücksichtigen:
Das Entstehen neuer Zeitschriften in der wissenschaftlichen Differenzierung: Die fortschreitende Spezialisierung wissenschaftlicher Fachbereiche ist eine offenkundige Entwicklung. Die neu entstehenden Bereiche finden sich nur zum Teil im traditionellen Publikationswesen wieder, so dass immer neue wissenschaftliche Fachgemeinschaften, neue Publikationsorgane hervorbringen, während andere verschwinden. Diese Dynamik wird durch die Veränderungen des digitalen Informationsmanagements in der Wissenschaft verstärkt. Fast alle Zeitschriften in der DiPP-Initiative haben thematisch differenzierende Ausprägungen und treten nicht mit dem Anspruch auf, andere Publikationsformen zu verdrängen, sondern einem spezifischen Ansatz eine Stimme zu verleihen. Als Beispiel sei hier die an der Fernuniversität Hagen angesiedelte Zeitschrift „eLearning and Education“ genannt, die die Konsolidierungstendenz der neuen Möglichkeiten des Lernens mit neuen Medien durch ein eigenes Forum unterstreicht (Krinke et al. 2005). Wie kann das Entstehen neuer, thematisch motivierter Zeitschriften also unter den veränderten Bedingungen erfolgen? Die DiPP-Initiative stellt hierzu einen Rahmen her, der es erlaubt über eine einzelne Zeitschrift hinaus, rechtliche, technische und organisatorische Rahmenbedingungen zu entwickeln, zu erproben und zu verankern. So ergibt sich die dezentral-zentral aufgebaute Struktur im Kooperationsmodell der DiPP Initiative: verallgemeinerbare Aspekte werden von DiPP im Hochschulbiblio

thekszentrum des Landes NRW in Köln („hbz“) analysiert und erprobt und ggf. allen Partnern angeboten (s.a. Abbildung 2).

Die Verankerung neuer Publikationsformen: Die Veränderungen des digitalen Informationsmanagements in der Wissenschaft bedingen langfristige und tief greifende Umstrukturierungen in der Arbeitsverteilung zwischen den am Publikationsprozess beteiligten Akteuren. Während Autoren früher auf Schreibmaschinen oder sogar handschriftlich verfasste Manuskripte bei den Redaktionen, den Verlegern oder Druckern abgaben, werden heute zum Teil druckfertige Manuskripte eingereicht, die – nach Begutachtung – direkt im Internet veröffentlicht werden können. Besonders den naturwissenschaftlich, technisch und medizinisch orientierten Bereichen des Publikationswesens, den so genannten STM-Bereich (STM steht für „Science, Technology and Medicine“), werden immer mehr Formatvorlagen oder Autorenwerkzeuge angeboten, die diese Verlagerung von traditionell im verlegerischen Bereich angesiedelten Arbeitsprozessen auf die Autorenseite verdeutlichen. Die technische Plattform in der DiPP-Initiative unterstützt diese Entwicklungen durch eine praxisnahe und am Nutzer orientierte Unterstützung der Redaktionen mit den Entwicklungen im hbz. In den Geisteswissenschaften ist dieser Prozess nicht so weit fortgeschritten wie im STM-Bereich. Da hier der Text das eigentliche Forschungsinstrument ist – nicht etwa das Labor – spielt der gedruckte Text und die Ästhetik des Druckerzeugnisses eine weitaus größere Rolle. Aber auch in den Geisteswissenschaften lässt sich eine lebhafte Entwicklung digitaler Publikationsformen entwickeln, wie es etwa am Beispiel der Zeitenblicke (Gersmann et al. 2005) aus der Geschichtswissenschaft der Universität Köln auch in der DiPP Initiative deutlich wird.

Die Erprobung neuer Publikationsformen: Die Veränderungen des digitalen Informationsmanagements in der Wissenschaft bringen jedoch auch viele Publikationsformen hervor, die weit über den traditionellen Zeitschriftenartikel hinausgehen. So sind Text und Bild in vielen Fachgebieten eigentlich nur noch als statisches Dokumentationsformat der eigentlich viel dynamischeren wissenschaftlichen Arbeit von Bedeutung. Die wissenschaftliche Bearbeitung einer Fragestellung liefert häufig Ergebnisse, die sich als Text und Bild nur in sehr beschränkter Form darstellen lassen, besonders im STM-Bereich: Aufwändige experimentelle Prozeduren liefern komplexe und große Datensätze, die mit Computerprogrammen analysiert werden, und erzeugen Modelle, Animationen und Simulationen. Alle diese einzelnen Teilergebnisse – Experimente, Datensätze, Computerprogramme, Animationen, Simulationen etc. – müssen aber nach wissenschaftstheoretischen Grundsätzen nachvollziehbar und reproduzierbar sein und entsprechend publiziert werden. Nur ein Bruchteil dieser Schritte ist allerdings tatsächlich mit Text und Bild darstellbar. Daher bieten immer mehr Zeitschriften über das WWW Zusatzmaterialien an und große Förderorganisationen wie das NIH nehmen entsprechende Verpflichtungen in ihre Förderbedingungen mit auf (National Institute of Health 2003). In der DiPP Initiative werden innovative Publikationsstrategien eben solcher Ergebnisse ernst genommen:

Die Zeitschrift „Brains, Minds and Media“ von der Universität Bielefeld (Egelhaaf et al. 2005) hat die Veröffentlichung von Zusatzmaterialien als Charakteristikum ihrer Publikationsstrategie an erste Stelle gesetzt. Auf der Seite der technischen Plattform werden im hbz die entsprechenden Voraussetzungen für die Publikation geschaffen.

Übergreifende Angebote von DiPP im „hbz“

Wie bereits in der Beschreibung der Grundmotivationen angedeutet, wird der Publikationsprozess wird von verschiedenen Akteuren getragen (s. Tabelle 1). Die einzelnen Prozesse können durch die Nutzung der vorhandenen DiPP-Infrastruktur effektiv gestaltet werden. Die Arbeiten werden dabei zwischen verschiedenen Akteuren verteilt. Hier sind viele Möglichkeiten denkbar. In DiPP wird davon ausgegangen, dass es einen vornehmlich inhaltlichen Anteil gibt, der von einer Redaktion geleistet wird und einen vornehmlich nicht-inhaltlichen (organisatorisch, rechtlich, technischen) Anteil, der von DiPP unterstützt wird (s.a. Abbildung 1). Auf der organisatorischen Seite bietet DiPP eine allgemeine Begleitung und Beratung an. Es finden regelmäßig Workshops statt, auf denen die verschiedenen Aspekte der Arbeit an eJournals beleuchtet werden. Auf der rechtlichen Seite werden im Rahmen der Digital Peer Publishing Lizenz, Modelle zur Open Access Veröffentlichung der Artikel angeboten, bei denen Autoren den Zeitschriften die Rechte zur elektronischen Veröffentlichung erteilen, aber andere Rechte beim Autor verbleiben. Auf der technischen Seite hilft DiPP bei der selbstständigen Erstellung und Redaktion von Inhalten. Einige Beispiele für unterstütze Prozesse sind die Produktion und Veröffentlichung von Informationen für die WWW-Seiten, Formatvorlagenentwicklung und Konvertierung, Integrierte URN-Vergabe für Artikelpublikationen oder Nachweis und Indexierung von Artikeln (mehr in Tabelle 1).

Prozess	Beispielaufgaben	fachlich	neutral	DiPP
Einreichung	Formale / technische Prüfung	o	o	BS
Annahme und Zuweisung	Rückmeldung an Autoren	o	o	BS
Vorbegutachtung	Formale / technische Prüfung	o	o	BS
Inhaltliche Begutachtung	Auswahl, Anfragen Gutachter	x	*	BS
redaktionelle Bearbeitung	fachliche und formale Kontrolle	*	o	BS
Formatierungen	Artikelkonvertierung	o	o	PS
Metadatenerfassung	Titelaufnahme	*	x	BS/PS
Bibliografische Erschließung	Suchmaschinen, Datenbanken	*	x	BS/PS
Redaktionelle Inhalte	Schreiben, Formatieren von News	x	*	PS
Veröffentlichung	Ausgabenplanung, Freischaltung	x	*	PS
Kontakte	Betreuung v. Autoren u. Gutachtern	x	*	PB
Rechte-Clearing	Einholen von Nachdruckerlaubnis	o	o	PB
„Marketing“	Verbreitung, Publikation, Vorträge	x	*	PB
Mehrwertdienste	Kongressberichte	o	o	PB
Finanzierungsmaßnamen	Abrechnung, Mittelbewirtschaftung	o	o	PB
Verwaltung	Verträge	o	o	PB
Pflege WWW	Anpassung/Erweiterung der Inhalte	o	o	PB
Formatvorlagen für Autoren	Spezifikation, Erweiterung	x	o	PB
Technischer Support	Unterstützung der Redaktionen	*	*	x
Software-Pflege	Ergänzung der Plattform	*	*	x
Software-Entwicklung	Erweiterung der Plattform	*	*	x
Lizenz-Management	Digital Peer Publishing Lizenz	*	*	x
Koordination	Kooperationsverträge	*	*	x

Tabelle 1: Beispiele für Arbeitsprozesse bei der Zeitschriftenpublikation (im laufenden Betrieb – die Aufgaben beim Aufbau sind der Übersichtlichkeit halber nicht berücksichtigt) und mögliche Arbeitsverteilungen. Die Spalte „fachlich“ bezieht sich auf die Beteiligung von Fachwissenschaftlern (Professoren oder in der fachwissenschaftlichen Einrichtung angesiedelte Mitarbeiter). Die Spalte „neutral“ bezieht sich auf die Beteiligung von anderen Einrichtungen an einer Hochschule, etwa Bibliotheken oder Medienzentren. Die Spalte DiPP bezieht sich auf die Beteiligung von DiPP, also in erster Linie auf übergreifende Aufgaben und die Nutzung der DiPP-Plattform. Bei Prozessen, die von Akteuren außerhalb der Redaktionen durchgeführt werden (Autoren, Gutachter, Mitherausgeber) sind nur die Arbeitsanteile gemeint, die Aufwand innerhalb der Redaktionen verursachen (z.B. die Betreuung, Prüfung etc.). [x] = Mitarbeit erforderlich; [*] = Mitarbeit möglich; [o] = Aufteilung variabel; BS = Begutachtungssystem/Support, PS = Publikationssystem/ Support, PB = Persönliche Bearbeitung

Das Begutachtungssystem der DiPP-Plattform beinhaltet ferner unterstützende Funktionen für die selbstständige Vor-Verarbeitung von Artikeln. Der Übergang vom Begutachtungssystem in das Publikationssystem ist weitestgehend automatisiert. Beispiele für unterstütze Prozesse sind hier die Einreichung durch Autoren, die Annahme und Zuweisung an Gutachter, die formale / technische Prüfung oder die inhaltliche Begutachtung.
Soweit möglich, wurde bei der Entwicklung auf bestehenden Systemen aufgesetzt (z.B. GAPworks, Fedora, Plone, s.a. Horstmann et al. 2005). Die Modularität des Ansatzes ergibt eine flexible Entwicklungsumgebung, die es erlaubt, neue Anforderungen der Zeitschriften zu berücksichtigen und anderen Partnern in der Initiative zur Verfügung zu stellen (s.a. Abbildung 2).

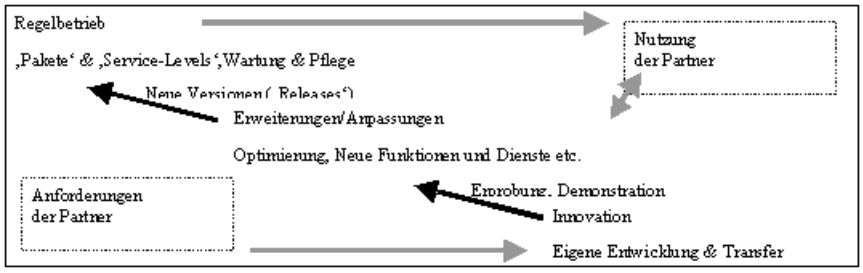


Abbildung 2: Entwicklungsmodell in DiPP. Die Anforderungen der Partner werden nach Möglichkeit in der DiPP Infrastruktur als Erweiterung berücksichtigt und im Testbetrieb entwickelt und erprobt. Sie fließen ggf. in den Regelbetrieb des „Live-Systems" ein und können allgemein genutzt werden.

Redaktionsmodelle

Die entscheidenden Prozesse der Publikation sind lokal an den Hochschulen angesiedelt. Die meisten Redaktionsmodelle setzen auf das traditionelle System aus Herausgebern, Gutachtern und Redakteuren auf, die in einer fachwissenschaftlichen Einrichtung der leitenden Herausgeberin oder des leitenden Herausgebers verankert sind, und beziehen Bibliotheken ein, die bei der Erschließung und Verbreitung und bei technischen Aufgaben helfen. Die Aufgabenverteilung ist hierbei hochgradig unterschiedlich. Viele Arbeitsprozesse können sowohl in der fachwissenschaftlichen Einrichtung als auch in der Bibliothek betreut werden (s.a. Tabelle 1). Die genaue Aufteilung hängt von den jeweils vorhandenen Ressourcen und von der Qualifikation des Personals ab. Häufig ist ein wissenschaftlicher Mitarbeiter des Professors für die „operativen" Anteile zuständig. Je nach Qualifikation und Auslastung werden so alle wesentlichen Prozesse vom wissenschaftlichen Mitarbeiter betreut oder eben durch die Bibliothek getragen. Es existieren aber auch Modelle, in denen gar kein weiterer fachwissenschaftlicher Mitarbeiter eingebunden ist und alle inhaltlichen Prozesse – strikt getrennt von fachunabhängigen Prozessen – nur von der Herausgeberin oder dem Herausgeber getragen werden. Hier sind die Redakteure in der Bibliothek

angesiedelt und arbeiten aus einer fachunabhängigen Perspektive den Fachwissenschaftlern zu, die sich ansonsten selbst organisieren. Besonders an Fachhochschulen, die typischerweise keinen ausgeprägten Mittelbau besitzen, ist dieses Modell zu finden. Wiederum andere beziehen weitere übergeordnete Stellen, wie hochschulweit arbeitende Dienstleistung, Verwertungs- und Koordinierungseinrichtungen mit ein. Die rationelle Aufteilung, die durch die Auslagerung von verallgemeinerbaren Prozessen , wie Rechte und Technik zu DiPP im hbz angefangen wurde, kann so innerhalb der Hochschule fortgesetzt werden, indem hochschulspezifische Anteile, etwa die Medienproduktion, die Verwertung oder die Verwaltung in hochschulweit arbeitenden Strukturen verankert werden. Dies können Bibliotheken, integrierte Medienzentren oder Informationsdienstleister sein.

Schluss

Ein Patentrezept für den Betrieb von eJournals kann aufgrund der sehr unterschiedlichen Anforderungen und Voraussetzungen in den einzelnen Teil-Initiativen nicht verschrieben werden. Alle habe sich auf ihre Art und Weise bewährt, wie es die Ergebnisse der Zeitschriften es widerspiegeln. Nach nur einem guten Jahr sind die Zeitschriften an der Öffentlichkeit und arbeiten auf Basis der DiPP-Infrastruktur im hbz. Die Erfahrung auf dem wissenschaftlichen Zeitschriftensektor lehrt, dass es einige Jahre dauert bis eine Zeitschrift sich etabliert (Crow & Goldstein 2003). Und die Umwälzungen im digitalen Informationsmanagement sind in vollem Gange: Die Hochschulen richten sich mit ihrem Einrichtungen neu aus und Verlage fangen an, auf die durch die Open Access Bewegung verursachten Veränderungen im Publikationswesen auf dem „grünen“ und „goldenen Weg“ zu reagieren, so dass auch die Verschmelzung und der Transfer von innovationsorientierten Initiativen wie DiPP (auf dem „weißen Weg“, s.o.) und den an konventionellen Rahmenbedingungen ausgerichteten Verlagen in den nächsten Jahren zu erwarten ist.

Literatur

arXiv.org (2005) e-Print archive. http://arxiv.org/ (Stand 15.08.2005)

BioMed Central Ltd (2005) The Open Access Publisher. http://www.biomedcentral.com/ (Stand 15.08.2005)

Björneborn, L., Ingwersen, P. (2001) Perspectives of Webometrics. Scientometrics Vol. 50, No. 1 65-82. http://www.db.dk/lb/ (Stand 15.08.2005)

Björneborn, L. , Ingwersen, P. (2004) Toward a Basic Framework for Webometrics. Journal of the American Society for Information Science and Technology, 55(14):1216–1227. http://www.db.dk/lb/ (Stand 15.08.2005)

Brody, T. and Harnad, S. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. D-Lib Magazine 10(6). http://www.dlib.org/dlib/june04/harnad/06harnad.html (Stand 15.08.2005)

Cogprints (2005) Cognitive Sciences ePrint Archive. http://cogprints.org/ (Stand 15.08.2005)

Crow, R., Goldstein, H. (2003) Guide to Business Planning for Launching a New Open Access Journal (Edition 2). Open society Institute. http://www.soros.org/openaccess/oajguides/

Deutsche Forschungsgemeinschaft (2005b) Publikationsstrategien im Wandel? (pp. 58 -60) Wiley VCH. http://www.dfg.de/dfg_im_profil/zahlen_und_fakten

Deutsche Forschungsgemeinschaft, Fournier, J. (2005a) Wege zum Wissen. Aktionsfelder zur Förderung des Open Access durch die DFG. http://www.dfg.de/dfg_im_profil/zahlen_und_fakten (Stand 15.08.2005)

Deutsche Physikalische Gesellschaft & Institute of Physics (2005) New Journal of Physics. http://www.iop.org/EJ/njp (Stand 15.08.2005)

Egelhaaf et.al (2005) Brains, Minds & Media. http://www.brains-minds-media.org

European Geosciences Union (2005) Atmospheric Chemistry and Physics. http://www.copernicus.org/EGU/acp/ (Stand 15.08.2005)

Garfield, E. (1994) The Impact Factor. Current Contents (25):3-7. http://scientific.thomson.com/knowtrend/essays/journalcitationreports/impact-factor/ (Stand 15.08.2005)

Gersmann et al. (2005) Zeitenblicke. http://www.zeitenblicke.de (Stand 15.08.2005)

Harnad, S. (2005) Fast-Forward on the Green Road to Open Access: The Case Against Mixing Up Green and Gold. Ariadne 43.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Yves, G., Charles, O., Stamerjohanns, H. and Hilf, E. (2004) The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials review 30(4). http://dx.doi.org/10.1016/j.serrev.2004.09.013 (Stand 15.08.2005)

Horstmann W., Reimer, P. Schirrwagen, J. (2005) Multi-level eJournal support structures in the initiative "Digital Peer Publishing NRW". Joint Workshop on Electronic Publishing Organised by Delos. Lund, Sweden. http://www.dipp.nrw.de/publikationen/horstmann_lund_full.pdf/download (Stand 15.08.2005)

Krinke et al. (2005) eLearning and Education. http://eleed.campussource.de

Max-Planck-Gesellschaft (2003) Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html (Stand 15.08.2005)

National Institute of Health (2003) NIH Data Sharing Policy. NIH GUIDE 2003 http://grants.nih.gov/grants/policy/data_sharing/ (Stand 15.08.2005)

National Institute of Health (2005a) Policy on Enhancing Public Access to Archived Publications Resulting From

National Institute of Health (2005b) PubMed Central: A free archive of life sciences journals. http://www.pubmedcentral.nih.gov/ (Stand 15.08.2005)

Nature Publishing Group & EMBO (2005) Molecular Systems Biology. http://www.nature.com/msb/about/oa.html (Stand 15.08.2005)

NIH-Funded Research. Federal Register Vol. 70, No. 26. http://www.nih.gov/about/publicaccess/ (Stand 15.08.2005)

Open Society Institute (2002) Budapest Open Access Initiative. http://www.soros.org/openaccess/read.shtml (Stand 15.08.2005)

PLoS (2005a) Public Library of Science. http://www.plos.org/ (Stand 15.08.2005)

PLoS (2005b) The First Impact Factor for PLoS Biology—13.9. http://www.plos.org/news/announce_pbioif.html (Stand 15.08.2005)

Rehmann, U. (2005) Documenta Mathematica. http://www.math.uni-bielefeld.de/documenta/ (Stand 15.08.2005)

Seglen, P.O. (1997) Why the impact factor of journals should not be used for evaluating research. BMJ ;314:497. http://bmj.bmjjournals.com/cgi/content/full/314/7079/497 (Stand 15.08.2005)

Sherpa (2005) Publisher copyright policies & self-archiving. http://www.sherpa.ac.uk/romeo.php (Stand 15.08.2005)

Springer Science and Business Media (2005) Springer Open Choice™. http://www.springeronline.com/sgw/cda/frontpage/0,11855,1-40359-0-0-0,00.html (Stand 15.08.2005)

Suber, P. (2003) Bethesda Statement on Open Access Publishing. http://www.earlham.edu/~peters/fos/bethesda.htm (Stand 15.08.2005)

Erfahrungen mit dem Open-Access-Journal “eleed (e-learning and education)”

Jens Krinke, Martin Roos; Hagen

Abstract

Im Rahmen der Open Access Initiative „Digital Peer Publishing NRW” gibt die FernUniversität in Hagen in Kooperation mit der Initiative CampusSource seit 2005 das eJournal „eleed (e-learning and education)“ heraus. eleed ist ein elektronisches Journal, bei dem alle Veröffentlichungen unter eine Open Access Lizenz gestellt werden. Ein Redaktionstridem, bestehend aus Fachwissenschaftlern sowie Vertretern der Universitätsbibliothek und CampusSource akquiriert und bewertet wissenschaftliche Aufsätze, Projektberichte und Buchbesprechungen. Dieser Artikel beschreibt, welche Erfahrungen mit diesem Journal gemacht wurden.

Einführung

Im Rahmen der Open Access Initiative „Digital Peer Publishing NRW“[1] (DiPP) gibt die FernUniversität in Hagen in Kooperation mit der Initiative CampusSource seit 2005 das eJournal „eleed (e-learning and education)“[2] heraus. Das zweisprachige, international ausgerichtete Journal soll eine Plattform für neue wissenschaftliche Forschungsergebnisse aus dem weit gefächerten Gebiet des e-learning sein. Alle Aufsätze, Projektberichte und Buchbesprechungen sind im Internet unter einer Open Access Lizenz frei zugänglich. Ein Begutachtungsprozess durch internationale Fachleute stellt die hohe wissenschaftliche Qualität der Beiträge sicher. Gewahrt bleibt die Aktualität, weil jeder Beitrag nach abgeschlossener Begutachtung unmittelbar veröffentlicht wird. Neben wissenschaftlichen Beiträgen ergänzen Projektberichte, die über den praktischen Einsatz von e-learning in Unternehmen und Bildungseinrichtungen berichten, sowie Besprechungen aktueller Literatur das eJournal.

Die Kombination aus traditioneller Journal-Form und der Verbreitung über das Internet stellt besondere Anforderungen, bietet aber andererseits interessante Möglichkeiten. Die begutachteten Beiträge stehen in klassischen Formaten zur Verfügung und müssen für die Veröffentlichung in der vom Hochschulbibliothekszentrum NRW (HBZ) zur Verfügung gestellten Software-Plattform automatisiert aufbereitet werden.

[1] http://www.dipp.nrw.de/

[2] http://eleed.campussource.de/

Dazu wurde ein XML-Format gewählt, in dem die Artikel in der Endfassung nachhaltig gespeichert werden. Aus diesem Format werden Online- und Druckfassung in einer einheitlichen Form generiert. Zusätzlich enthält es Metadaten, Referenzen etc., die in der Präsentationsplattform eingebunden werden. eleed ist ein Kooperationsprojekt der FernUniversität in Hagen (Fachbereich Elektrotechnik und Informationstechnik sowie der Universitätsbibliothek) und der Landesinitiative CampusSource. Die Universitätsbibliothek ist einerseits für die Erstellung der Metadaten verantwortlich, andererseits betreut sie die Beitragskategorie „Literatur“ eigenständig. Darüber hinaus baut sie aus den Literaturangaben der Beiträge eine Referenzdatenbank auf, die durch Verlinkung zu weiteren Diensten zu einem Portal ausgebaut wird.

Hintergrund

Die FernUniversität in Hagen besteht seit nunmehr 30 Jahren und ist die erste Adresse in Deutschland im Bereich Fernstudium, Distance Learning und e-learning. Von den einfachen technischen Bedingungen zu Beginn haben sich die Möglichkeiten der Lehre in Bezug auf das Fernstudium gravierend geändert. So spiegeln die Aufbereitungen der Materialien, die im Fernstudium eingesetzt werden, den technischen Fortschritt der Medien selbst wider. Mit den heutigen Möglichkeiten des Internet können insbesondere studienbegleitende Lehr- und Lernmaterialien schnell und einfach bereitgestellt und erschlossen werden. Durch den Einsatz elektronischer Medien und nicht zuletzt der Online-Medien wurde der Informationsaustausch in den letzten Jahren rasant beschleunigt. Die FernUniversität in Hagen nimmt an diesem Prozess nicht nur teil, sondern sieht sich in einer aktiven Vorreiterrolle.

In diesem Umfeld wurde der Bedarf an einer qualitativ hochwertigen Austauschmöglichkeit für Wissenschaftler im Bereich des e-learnings erkannt. Die klassische Form des wissenschaftlichen Austauschs findet in den gedruckten Journalen statt. Diese Form ist etabliert und wird seit Jahrhunderten erfolgreich praktiziert. Mittlerweile gewähren namhafte Verlage und Institutionen einen Online-Zugang auf ihre Print-Journale. Diese sind jedoch in der Regel ein elektronisches Abbild der gedruckten Zeitschrift.

Da jedoch das e-learning selbst sehr schnell auf die technischen Änderungen im Informationswesen reagieren muss, lag es nahe, ein adäquates Medium für den wissenschaftlichen Austausch einzusetzen. Dieses Medium ist das eJournal, also die elektronische Zeitschrift. Hiermit ist nicht gemeint, ein elektronisches Abbild einer gedruckten Zeitschrift zu publizieren, sondern die Möglichkeiten, die die Internet-Technologie bietet, gezielt für den wissenschaftlichen Informationsaustausch einzusetzen. Eine Möglichkeit ist die der schnellen Veröffentlichung. Sobald ein Beitrag von der Redaktion zur Veröffentlichung freigegeben ist, ist er auch tatsächlich (online) zugänglich. Weder ein Redaktionsschluss, noch Versand- oder Drucktermine schieben den Veröffentlichungstermin hinaus. Damit reagiert ein eJournal zeitlich

wesentlich schneller und flexibler als es im Print möglich ist. Somit bildet das eJournal die Schnelligkeit des Internets bzw. des elektronischen Informationsaustausches direkt ab.
Diese Umstände haben die FernUniversität in Hagen bewogen ein eigenes eJournal unter dem Titel „e-learning and education“ kurz „eleed“ herauszugeben. Einerseits setzt die FernUniversität in Hagen e-learning aktiv ein, andererseits ist sie auf Grund ihrer Ausrichtung her aktiv an der Fortentwicklung von e-learning-Methoden beteiligt. Aus den oben genannten Gründen wird die Zeitschrift als eJournal unter Open Access publiziert.

Open Access

Eine Neuerung im Bereich des Publizierens wird mit dem Begriff Open Access umschrieben. In Bezug auf die „Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen“ (Berlin Declaration 2003) aus dem Herbst 2003 wird auf die Problematik der Kostensteigerung der wissenschaftlichen Zeitschriften im STM-Bereich (Science, Technology, Medicine) hingewiesen. Im engen Rahmen der Hochschulbudgets und damit der Universitätsbibliotheken ist eine umfassende Literaturversorgung immer schwieriger zu bewerkstelligen. Notwendige Kosteneinsparungen machen Zeitschriftenabbestellungen unabdingbar. Damit werden jedoch Wissenschaftler von der Literaturversorgung abgeschnitten, was zur Einschränkungen bei Forschung und Lehre führt. Eine Möglichkeit, der Kostenexplosion in der Literaturversorgung entgegen zu wirken, ist das Publizieren unter Open Access. Hierbei fallen beim Endnutzer keinerlei Kosten an, die Inhalte sind zur Benutzung frei. Damit entfallen auch keine Kosten auf die Universitätsbibliotheken, die ja für die Literaturversorgung und –beschaffung zuständig sind.
Open Access hat inzwischen einen mehr als deutlich messbaren Einfluss. So stellen Harnad und Brody (2004) fest, dass Open Access Artikel inzwischen deutlich häufiger zitiert werden als Artikel, die nicht elektronisch frei zugänglich sind. Open Access bedeutet heutzutage allerdings meistens, dass die Autoren ihre Artikel frei zugänglich ins Netz stellen – sie also selbst die Artikel (die primär in traditionellen Publikationsformen erscheinen) elektronisch und frei zugänglich archivieren (Pinfield 2004). Bei dieser Art der individuellen Publikation im Netz sind Aspekte der Nachhaltigkeit des offenen Zugangs und der Verfügbarkeit nicht sicher gestellt. Komplette Open Access Journale vermeiden solche Probleme, gehen aber weit darüber hinaus und bieten ihre Inhalte primär elektronisch als eJournal an.

Vorteile elektronischer Journale

Weiterere Vorteile des eJournals sind die direkte Zuordnung von Diskussionsforen zu Artikeln sowie die Einbindung multimedialer Elemente. Tondokumente, Filme, Animationen, Simulationen etc. lassen sich problemlos einbinden und bieten so einen qualitativ neuen Zugang zu Forschungs-, Lehr- und Lerninhalten. Darüber hinaus sind weitere Eigenschaften realisierbar, die Internet-Technologien abbilden. Künftige

Entwicklungen können ebenfalls auf ihre Tauglichkeit geprüft und dann eingesetzt werden.
Dies alles zeigt, dass ein wissenschaftliches Publikationsorgan bezüglich des e-learning die gleichen oder ähnliche Technologien einsetzen muss wie das e-learning selbst, wenn es den aktuellen Stand und auch den Fortschritt des e-learnings in wissenschaftlicher Sicht begleiten will. Hierdurch lassen sich neue Technologien nicht nur anhand von Prototypen demonstrieren, sondern können im Alltagsbetrieb auf ihre Praxistauglichkeit getestet werden. Aus diesen Gründen ist nicht nur der Bedarf an einem eJournal für das e-learning vorhanden, vielmehr ist eine zwingende Notwendigkeit gegeben, das wissenschaftliche Publizieren hier mittels eines eJournals zu befördern.

Strukturen des eleed-Journals

Im Gegensatz zu anderen Open Access Journalen hat eleed mehrere Besonderheiten. Da ist zuallererst das Redaktionsteam zu nennen, das bei eleed aus einem Tridem gebildet wird: Den Fachwissenschaftlern sowie Vertretern der Initiative CampusSource und der Universitätsbibliothek. Außerdem versucht eleed zugleich ein wissenschaftliches Fachjournal und ein Publikumsjournal zu sein, indem sowohl begutachtete Fachbeiträge, als auch Projektberichte und Literaturbesprechungen veröffentlicht werden. Diese beiden Besonderheiten sind miteinander verknüpft und werden im Folgenden erläutert.
Das eJournal eleed beinhaltet drei Beitragskategorien. Erstens werden wissenschaftliche Beiträge unter der Kategorie „e-learning Beiträge“ veröffentlicht. Ein Begutachtungsprozess durch internationale Experten und Expertinnen sichert die notwendige wissenschaftliche Qualität. Die Autoren und externen Gutachter werden durch die Fachwissenschaftler betreut. Zweitens sind im eJournal Projektberichte vertreten. Es werden Zustandsbeschreibungen, Einsatz von e-learning, Softwarewerkzeuge im Alltagsbetrieb usw. aufgeführt. CampusSource ist auf Grund seines Netzwerkes an Entwicklern hier in der Lage, Beiträge dieser Kategorie zu akquirieren und zu bewerten. Drittens wird aktuelle Literatur zum Thema e-learning vorgestellt. Diese Kategorie wird von der Universitätsbibliothek betreut. Sie steht im Kontakt mit Autoren, Verlagen und Rezensenten. Die Kategorien entsprechen somit der Ausrichtung des Redaktionstridems. Das eJournal erreicht damit eine größere inhaltliche Bandbreite und ermöglicht eine umfassendere Übersicht über das breite Themenspektrum des e-learning. Hiermit wird ein weiteres Ansinnen deutlich: eleed will nicht nur ein Publikationsorgan des wissenschaftlichen Veröffentlichens sein, sondern darüber hinaus die weit gefächerte Welt des e-learning mittels Beiträgen vermitteln.
Die Redaktion findet sich zu regelmäßigen Sitzungen zusammen. Die eingegangen Beiträge werden vorgestellt und diskutiert. Für die wissenschaftlichen Beiträge werden Gutachter vorgeschlagen und ausgewählt. Der Stand der im Begutachtungsprozess befindlichen Beiträge wird dargelegt, etwaige Rückmeldungen an

die Autoren und Gutachter besprochen sowie Terminsetzungen festgelegt. Dieser Teil bindet naturgemäß einen großen Anteil der Redaktionsarbeit.
Die im Rahmen des DiPP-Projekts gestarteten eJournals weisen unterschiedliche Konzepte auf. Im eJournal eleed ist die Universitätsbibliothek Hagen mit einer eigenen Rubrik vertreten, nämlich der Auswahl und Vorstellung aktueller Literatur. Dies ist insofern beachtenswert, als dann üblicherweise die Hauptfunktion der Bibliotheken in der Bestandserschließung und hier in der Erschließung der Journalbeiträge gesehen wird. Damit eröffnet sich für die Universitätsbibliothek eine vollkommen neue Möglichkeit. War ihre Rolle bislang eher passiv zu sehen, nämlich in der Aufbereitung bzw. Erschließung der eingereichten Beiträge, so ist nun die Rolle der Universitätsbibliothek wesentlich aktiver. Die Universitätsbibliothek gestaltet mit ihren Beiträgen, also den Besprechungen und Rezensionen, das eJournal direkt mit. Somit werden auch neue Anforderungen an die Universitätsbibliothek gestellt: Sie muss den Kontakt zu Autoren, Verlagen und Rezensenten pflegen. Durch diese neuen Aufgaben ist die Universitätsbibliothek ein vollwertiges Mitglied des Redaktionstridems.

Qualität des eleed-Journals

Die Qualität der wissenschaftlichen Beiträge muss nicht nur gewahrt sein, vielmehr muss darauf geachtet werden, dass für ein neues wissenschaftliches Journal ein dauerhaft hochwertiger Qualitätssicherungsprozess implementiert wird. Nur dadurch lassen sich langfristig namhafte Autoren und Gutachter für das eJournal gewinnen. Üblicherweise dauert es fünf oder mehr Jahre, bis sich ein neues Journal erfolgreich auf dem Markt etabliert hat. Der Begutachtungsprozess eines einzelnen Artikels kann sich – wie bei Print-Journalen – über einen Zeitraum von mehr als sechs Monaten hinziehen. Dies widerspricht zwar dem Gedanken des schnellen Publizierens, was ja gerade ein eJournal, also ein Online-Medium auszeichnet, dieser Prozessschritt ist aber für die wissenschaftliche Qualitätssicherung unabdingbar.
Ebenfalls problematisch ist die Akquise neuer Beiträge. Vor allem namhafte Wissenschaftler legen sehr großen Wert darauf, in anerkannten Journalen mit entsprechendem Ansehen zu veröffentlichen, da dies ihr eigenes Renommee steigert. Umgekehrt steigern solche Journale selbst ihr Renommee durch Beiträge namhafter Autoren. Dieser Kreis muss durchbrochen werden, wenn ein neues (e-)Journal erfolgreich sein will. Daraus folgt, dass das Vertrauen des Autors in das neue eJournal vorhanden sein und gestärkt werden muss. Dies ist in der Regel dann der Fall, wenn das Herausgebergremium mit namhaften Wissenschaftlern besetzt ist und damit die Seriosität des eJournals gewährleistet wird. Somit nimmt die Autorenbetreuung einen zentralen Stellenwert in der Redaktionsarbeit ein. Nur wenn dauerhaft eine hohe Qualität des Journals besteht, kann eine Marktpräsenz gesichert werden. Die Notwendigkeit ergibt sich auch aus der Tatsache, dass etablierte Datenbanken, die Artikel einzelner Journale aufnehmen, von diesen aber eine Marktpräsenz von mindestens ein bis zwei Jahren als Bedingung für die Aufnahme von Artikeln in ihre Daten

bank voraussetzen. Diese Bedingung ist nachvollziehbar, denn nur auf diese Weise können die Datenbankanbieter die Qualität ihrer eigenen Produkte, nämlich die Datenbanken selbst, gewährleisten. Im konkreten Fall bedeutet dies, dass das eJournal eleed mindestens zwei Jahre mit hochkarätigen Beiträgen namhafter Autoren regelmäßig erscheinen muss, um bei weiteren Informationsanbietern als wahrnehmungswürdig eingestuft zu werden. Hieran wird der hohe Aufwand erkennbar, der für die Autorenbetreuung aufgewendet werden muss. Eine größere Anzahl an Beiträgen anzuwerben, ist in der Regel das geringere Problem. Die hochwertigen Beiträge einzufordern bzw. auszuwählen, diese zudem noch einen kritischen Begutachtungsprozess durchlaufen zu lassen, lässt die Annahmequote für Artikel deutlich sinken. Hingegen greift eleed auf das Netzwerk von CampusSource und der Fachwissenschaftler im Redaktionstridem zurück, welches europa- bzw. weltweit Personen und Institute im Bereich des e-learning umfasst. Dies ist eine notwendige, wenn auch keine hinreichende Bedingung für eine dauerhaft erfolgreiche Akquise. Insofern profitiert eleed in nicht unerheblichem Maße von dem CampusSource-Netzwerk.

Ferner, wenn eleed Renommee gewinnen will, müssen die wissenschaftlichen Beiträge Erstveröffentlichungen sein. Die Digital Peer Publishing Lizenz (DPPL), die in NRW eigens für die DiPP-Initiative entwickelt worden ist, belässt die Rechte für die Nutzung in körperlicher Form, insbesondere die Rechte zur Verbreitung in Druckform oder auf Trägermedien, beim Autoren (Metzger, Jaeger 2004). Er kann daher nach der Veröffentlichung in eleed seinen Artikel einem Verlag zur gedruckten Veröffentlichung überlassen. Insofern kommt die DPPL dem Autor entgegen. Die Erstveröffentlichung ist notwendig, um das Leserinteresse auf eleed zu lenken. Eine Zweitveröffentlichung ist naturgemäß weniger interessant, weil ja der Artikel bereits an anderer Stelle publiziert wurde. Der geneigte Leser wird daher kaum die Anstrengung aufwenden, den Artikel ein zweites Mal in einer anderen Publikation zu konsumieren. Auch der Autor selbst wird in der Regel eine Erstveröffentlichung in einem anerkannten Journal publizieren wollen, denn hier wird ihm von Seiten der Leser mehr Aufmerksamkeit entgegengebracht, als dies bei einer unbekannten Zeitschrift der Fall sein wird.

Die hier genannten Punkte betreffen auch ein klassisches Journal, welches neu publiziert wird. Im vorliegenden Fall jedoch handelt es sich um ein Open Access Journal. Oftmals wird hier die Befürchtung (das Vorurteil?) geäußert, dass das, was nichts kostet, auch nichts wert sein kann. Dies ist analog zur IT-Branche zu sehen, in der anfangs nur wenige Open Source Softwareprodukte aus der Marktnische heraus eine etablierte Stellung erreicht haben. Dies ist in erster Linie – und hier ist durchaus eine Parallele zu Open Access zu erkennen – auf Qualität, Nutzen und Preisdifferenz zu kommerziellen Produkten zurückzuführen. Gerade gegenüber teuren Produkten ist es für Open Source wie Open Access leichter, Marktterrain zu gewinnen und sich als dauerhafte Alternative zu klassischen Produkten zu etablieren. Insofern haben Open Access Journale durchaus die Chance, eine signifikante Marktposition zu erreichen.

Infrastruktur

Die Journale der DiPP-Initiative werden vom Hochschulbibliothekszentrum in Köln (HBZ) in technischer Sicht betreut. Hier werden den Mitgliedern Softwareprodukte für den Begutachtungs- und Publikationsprozess zur Verfügung gestellt sowie die Datenhaltung übernommen. Die zur Verfügung gestellten Systeme dienen der Unterstützung des Begutachtungsprozesses und der Veröffentlichung der Beiträge (Horstmann et al 2005). Im Folgenden werden die Erfahrungen mit den Systemen dargestellt.

Das GAPworks-System

In der ersten Planung sollte den Journalen der DiPP-Initiative ein Zugang zu einem kommerziellen Produkt zur Unterstützung des Begutachtungsprozesses ermöglicht werden. Ein solches Produkt wie etwa ScholarOne’s Manuscript Central[3] wird sehr erfolgreich von Verlagen und Organisationen zur Durchführung des Begutachtungsverfahrens und zur Unterstützung der Herausgeber eingesetzt. Stattdessen wurde aber entschieden, sich an dem GAP-Projekt[4] (German Academic Publishers) zu beteiligen. In diesem Projekt wird das GAPworks-System entwickelt, das dem Verwalten von Publikationsprozessen dient. Leider ist das diesem System zu Grunde liegende Modell ein schwergewichtiges Workflow-Modell aus der Sicht eines Verlags. Es hat sich herausgestellt, dass für die Herausgabe eines einzigen eJournals wie eleed der Einsatz dieses Systems nicht sinnvoll ist: Das GAPworks-System erfordert nicht nur einen sehr hohen Aufwand, sondern ist insbesondere für potentielle Autoren abschreckend. Für eleed wurde daher entschieden, auf die Benutzung von GAPworks zu verzichten und stattdessen den Begutachtungsprozess manuell mit Office-Systemen (Email und Tabellen) durchzuführen.

Das Publikationssystem

Vom HBZ wurde im Rahmen der DiPP-Initiative ein webbasiertes Workflow-System entwickelt, welches einerseits die Darstellung der zu veröffentlichenden Beiträge erzeugt und andererseits die nachhaltige Präsentation der Beiträge übernimmt.
Die von den Herausgebern zur Veröffentlichung freigegebenen Artikel liegen im RTF-Format vor. Das System des HBZ übernimmt daraufhin den ersten Schritt der Aufbereitung und erzeugt aus dem RTF-Dokument ein XML-Dokument. In einem zweiten Schritt wird aus dem XML-Dokument ein HTML-Dokument erzeugt, das in das Publikationssystem übernommen wird. Die Herausgeber stellen dann aus den Dokumenten eine Ausgabe des eJournals zusammen, die freigeschaltet wird. Das Publikationssystem versieht die veröffentlichten Beiträge mit URNs (Unified Resource Name) und bietet sie potentiellen Lesern an.

[3] http://www.scholarone.com/
[4] http://www.gap-portal.de/

Dieses Vorgehen stellt besondere Anforderungen an das Redaktionstridem als Herausgeber. Insbesondere ist die Konvertierung der RTF-Dokumente in XML-Dokumente problematisch. Trotz Autorenrichtlinien sind die eingereichten RTF-Dokumente sehr unterschiedlich und eine automatische Konvertierung produziert nur in Ausnahmen die gewünschten Ergebnisse. Daher muss das Redaktionstridem die RTF-Dokumente immer wieder so ändern, dass die Konvertierung in XML (und später HTML) das erwünschte Ergebnis erzielt. Diese redaktionellen Arbeiten stellen einen nicht unerheblichen Aufwand dar.

Geschäftsmodelle

Im Rahmen der DiPP Initiative des Landes Nordrhein Westfalen wird das eJournal eleed in den Jahren 2004 und 2005 gefördert. Die Förderung dient einerseits dem Aufbau des Journals und andererseits der Finanzierung des Publikationsprozesses. Im Gegensatz zu Print-Journalen müssen bei einem Open-Access eJournal die Herausgeber selbst die Verlagstätigkeiten übernehmen. Dazu gehören insbesondere das Redigieren und die Aufbereitung der eingereichten Beiträge und die Betreuung und Benutzung der technischen Infrastruktur.
Da die Förderung jedoch ausläuft und es der FernUniversität nicht möglich ist, die entstehenden Kosten zu übernehmen, werden unterschiedliche Geschäftsmodelle diskutiert. Typische Geschäftsmodelle für Open Access Journale (Crow, Goldstein 2003) sind im deutschen Markt zumindest im Bereich e-learning nicht anwendbar. Insbesondere die Unterschiede des Verlags- und Urheberrechts verhindern eine Übertragbarkeit von amerikanischen Geschäftsmodellen. Eine Möglichkeit ist das Modell Pay-by-Author, bei dem der Autor bzw. seine Institution die Kosten für die Veröffentlichung trägt oder auch dass das eJournal kostenpflichtige Zusatzdienste anbietet. Pay-by-Author ist für ein neu eingeführtes Journal wie eleed nicht nutzbar, da die Autoren dann auf Journale ohne dieses Modell ausweichen würden. Ein solches Modell setzt ein entsprechendes Renommee voraus. Die Diskussion ist hier noch nicht abgeschlossen, die Zukunft wird zeigen, welche Geschäftsmodelle auf dem Markt tragfähig sind.
Vorerst ist eine weitere Förderung durch öffentliche Mittel unumgänglich. Diese Kosten werden aber durch die Einsparungen auf Seiten der Bibliotheken und Instituten mehr als aufgewogen. Insgesamt muss es eine Verlagerung des Einsatzes öffentlicher Gelder geben: Statt den Zugriff auf Forschungsergebnisse zu finanzieren, sollte deren Veröffentlichung gefördert werden.

Erschließung der Beiträge

Natürlich befasst sich die Universitätsbibliothek auch mit der Erschließung der Beiträge. Hier wird ein aufwändiges Verfahren benutzt. Die ursprüngliche Idee, eine

reine Fachklassifikation für e-learning zu verwenden, wurde verworfen. Es stellte sich nämlich heraus, dass zwar für dieses Thema Fachklassifikationen existieren, diese jedoch in der Regel einen rein akademischen Charakter besitzen, d.h., dass es praktisch keine größere Anwendergruppe für diese Klassifikationen gibt, also etwa Datenbankanbieter, Fachgesellschaften oder dergleichen. Aus diesem Grund wird eine Kombination aus Universalklassifikation (Dewey Decimal Classification, DDC) und einer Fachklassifikation (Physical and Astronomical Classification Scheme, PACS) verwendet. Letztere wird zwar hauptsächlich in der Physik und Astronomie eingesetzt, wie der Name auch sagt, sie besitzt aber darüber hinaus eine detaillierte Klassifikation aus den Bereichen Computer Science und Information Technology. Da das eJournal eleed gerade aus diesen Bereichen viele Beiträge aufweist, ist die Verwendung dieser Klassifikation sinnvoll. Zudem werden, soweit sinnvoll, mehrere Classification Codes angegeben. Darüber hinaus werden von den Autoren, Gutachtern und der Universitätsbibliothek freie Schlagworte (free terms) vergeben. Dieser große Aufwand ist ein typisches Merkmal für die hohe Qualität des eJournals.

Über diese Tätigkeit hinaus plant die Universitätsbibliothek aus den Beiträgen weitere Angebote zu entwickeln. Sowohl die wissenschaftlichen Beiträge als auch die Projektberichte enthalten Angaben zu weiterführender Literatur bzw. Literaturverzeichnisse. Diese Angaben beinhalten die aktuelle Literatur zum Thema e-learning. Formal sind die verschiedensten Medien vertreten: vom klassischen Buch und Zeitschriftenaufsatz (Print) über elektronische Medien wie CD-ROM bis hin zu Online-Quellen, die sowohl eJournals umfassen können als auch „einfache“ URLs. Aus diesen Literaturangaben wird eine Datenbank (citation index, Referenzdatenbank) erstellt. Dies ist insofern von Bedeutung, als dass es kaum Datenbanken gibt, die sich nur mit dem Thema e-learning befassen. Auch hier zeigt sich, dass mit dieser Datenbank eine Marktlücke besetzt werden kann.

Die Pflege solcher Datenbanken ist sehr aufwändig, wenn gewisse Qualitätskriterien berücksichtigen werden sollen. Von bibliothekarischer bzw. dokumentarischer Seite ist natürlich die Erschließung der Beiträge zu nennen. Erst hierdurch wird eine sinnvolle Recherche ermöglicht, und die Erschließung ist ein Qualitätsmerkmal der Datenbank. In der Regel werden Beiträge manuell erschlossen, was sehr personalintensiv und damit teuer ist. Hier stehen neue Überlegungen an, wie diese Kosten reduziert werden können ohne dass die Qualität der Erschließung merklich vermindert wird. Der Kerngedanke an dieser Stelle ist die Übernahme der classification codes bzw. free terms des Beitrags, der die Literaturangaben enthält. Anders formuliert: Die Literaturangaben erhalten dieselbe Klassifikation wie der Originalbeitrag. Dies ist nahe liegend, weil Beitrag und Literatur sich sicherlich mit einem ähnlichen Themenkomplex befassen. Da der Beitrag selbst, wie oben beschrieben, sehr aufwändig erschlossen wird, ist eine gewisse Qualität in der Erschließung somit auch bei den Literaturangaben gewährleistet.

Unabhängig davon, ob nun dieses Verfahren oder ein anderes zur Erschließung der Literaturstellen gewählt wird, ist der Aufbau der Datenbank sinnvoll. Zum einen

befasst sich die FernUniversität in Hagen von ihrem Selbstverständnis her mit e-learning bzw. distance learning. Zum anderen hat die Universitätsbibliothek der FernUniversität in Hagen die Deutsche Fernstudiendokumentation aus Tübingen (DFSD) übernommen. Diese enthält retrospektiv die Literatur zum Thema distance learning. Die oben genannte Datenbank schließt hier inhaltlich an und sammelt die aktuelle Literatur. Sinnvollerweise wird die Produktion dieser Datenbank dann auch von der Universitätsbibliothek durchgeführt, damit die Datenbestände, die zwar unterschiedlicher Herkunft sind, in einer Institution beheimatet sind. Hier ist natürlich geplant, dass beide Datenbanken, also die Deutsche Fernstudiendokumentation und die aus eleed gewonnene Literaturdatenbank, strukturell, sofern sinnvoll, aufeinander abgestimmt werden. Damit stehen dann in der Zielvorstellung einheitliche Recherchemöglichkeiten für beide Datenbanken zur Verfügung. Es ist noch abzuklären, inwieweit sich diese Zielvorstellung realisieren lässt.

Erfahrungen

Trotz der hier vorgestellten umfangreichen Planungen und Vorüberlegungen ist eleed ein knappes Jahr nach Projektbeginn mit der ersten Ausgabe online gegangen. Aus jeder der vorgestellten Rubriken konnten mehrere aktuelle Beiträge namhafter Autoren gewonnen werden. Insgesamt sind in der ersten Ausgabe 13 Beiträge veröffentlicht worden. Diese zeigen bereits die Vielfältigkeit des eJournals.
Nicht nur der Start verlief positiv, auch konnten die Zugriffszahlen überzeugen. Eine unabhängige Evaluation der Zugriffszahlen auf die erste Ausgabe ergab, dass allein in den ersten vier Monaten über 190.000 Zugriffe gezählt werden konnten. Hier zeigt sich die erfolgreiche Einbettung und Verbreitung eleeds über das internationale CampusSource-Netzwerk. Dies ist umso interessanter, als dass die begleitenden Marketingmaßnahmen erst im zweiten Quartal 2005 angelaufen sind. Hierunter ist nicht die Werbung im herkömmlichen Sinn, sondern die Verankerung des eJournals in (Internet-)Suchmaschinen bzw. Datenbanken und Newslettern gemeint. Während die Verankerung bei den Suchmaschinen schnell durchgeführt werden konnte, ist dies, wie oben erwähnt, bei den klassischen Datenbankanbietern nicht der Fall. Hier wird es noch ein bis zwei Jahre dauern, um von diesen aufgenommen zu werden. Es ist dann davon auszugehen, dass dadurch eine deutliche Steigerung der Zugriffszahlen erreicht wird.

Bewertung

Die Einführung von eleed hat die besonderen Anforderungen an die Herausgabe von elektronischen Open-Access-Journalen herausgestellt. Insbesondere sind hier zu nennen:

- Noch haben es Open-Access-Journale schwer, sich gegenüber Print-Journalen zu behaupten. Insbesondere müssen Open-Access-Journale von Anfang an eine hohe Qualität aufweisen, um sich ein entsprechendes Renommee aufzubauen. Dass dies gelingen kann, zeigt die hohe Qualität der Beiträge in eleed.
- Es ist abzusehen, dass Open-Access-Journale guter Qualität in Zukunft ein hohes Renommee erwerben können, wenn man die Entwicklung der Zitierhäufigkeit betrachtet.
- Für wissenschaftliche Veröffentlichungen spielt der zeitliche Vorsprung des elektronischen zum traditionellen Publizieren noch eine geringe Rolle. Die größten Verzögerungen entstehen durch den aufwändigen Begutachtungsprozess, der schon mal mehrere Monate in Anspruch nehmen kann. Allerdings wird erwogen, auch hier neue Wege zu gehen, um den Begutachtungsprozess zu straffen bzw. die Möglichkeit der Publikation einer Vorfassung zu eröffnen.
- Ein eJournal kann aber die Zeit zwischen der Annahme der endgültigen Version eines Beitrags und der Veröffentlichung minimieren. Einerseits fällt keine Verzögerung durch Satz, Druck und Verteilung an und andererseits können andere Ausgabe-Formate gewählt werden. So wird eleed in Zukunft akzeptierte Beiträge sofort veröffentlichen und Beiträge erst im Nachhinein zu Ausgaben zusammenfassen, was die Verzögerungszeit minimiert und die Aktualität der Beiträge sichert.
- Zur Unterstützung des Begutachtungs- und Publikationsprozesses sowie zur Produktion und Bereitstellung des Journals ist der Einsatz qualitativ hochwertiger Systeme, die ständig gepflegt werden müssen, notwendig. Leider sind in diesem Bereich zurzeit so gut wie keine Open-Source-Produkte verfügbar.
- Die durch die Herausgabe eines eJournals erzeugten Kosten sind nicht unerheblich. Daher müssen tragende Geschäftsmodelle für Open-Access-Journale gefunden werden. Bis diese vorhanden sind, ist eine öffentliche Förderung unumgänglich.
- Sowohl für die Leser, als auch für Institutionen wie Bibliotheken werden etablierte Open Access Journale zu erheblichen Einsparungen führen, die die Kosten der Förderung bei weitem übertreffen.

Der Aufbau eines neuen eJournals wie eleed zeigt, dass das Publikationswesen im wissenschaftlichen Bereich sich einem starken Wandel unterzieht. Sowohl die technische Publikationsform (print oder elektronisch) als auch das Herausgebermodell (Open Access) unterscheidet sich grundlegend von dem, was vor noch nicht allzu langer Zeit im Verlagswesen Standard war. Es ist daher nicht verwunderlich, dass sich erst allmählich die neuen Möglichkeiten zeigen. Diese werden vom Redaktionstridem sehr aufmerksam verfolgt. eleed will ja nicht nur passiv über diese Möglich

keiten berichten, sondern sie vielmehr aktiv ausprobieren und gestalten. Das Redaktionstridem ist optimistisch, diesen Prozess erfolgreich begleiten zu können.

Literatur

Berlin Declaration (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, Berlin, Max-Planck-Gesellschaft http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.

Crow, Raym and Howard Goldstein (2003). *Guides to Business Planning for Launching a New Open Access Journal*, Edition 2, Open Society Institute, New York, Juli 2003 http://www.soros.org/openaccess/oajguides/html/business_planning.htm.

Harnad, Stevan und Tim Brody (2004). *Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals,* D-Lib Magazine 10(6), Juni 2004.

Horstmann, Wolfram and Peter Reimer, Jochen Schirrwagen (2004). *Multi-level eJournal support structures in the initiative „Digital Peer Publishing NRW*“, Joint Workshop on Electronic Publishing, Lund, Sweden, 14.-15. April 2005 http://www.dipp.nrw.de/publikationen/horstmann_lund_full.pdf/download

Metzger, Axel und Till Jaeger (2004). *Digital Peer Publishing Lizenz*, Köln, Ministerium für Wissenschaft und Forschung des Landes Nordrhein-Westfalen http://www.dipp.nrw.de/lizenzen/dppl/.

Pinfield, Stephen (2004). *Self-achiving publications*, in: *International Yearbook of Library and Information Management 2004-2005*, London, Facet, S. 118-145 http://eprints.nottingham.ac.uk/archive/00000142/01/IYLIM04.PDF.

Optimierungspotenziale bei der praktischen Umsetzung von Open Access

Christian Woll, Hürth

Abstract

Seit der Unterzeichnung der Berlin Declaration durch die führenden deutschen und einige internationale Wissenschaftsorganisationen am 22. Oktober 2003 hat sich das Thema „Open Access“ auch in Deutschland zu einem „Dauerbrenner“ in der bibliothekarischen und informationswissenschaftlichen Fachdiskussion entwickelt. Doch inwieweit erweisen sich die beiden maßgeblichen Open Access-Strategien, „Self-Archiving“ und „Open Access-Zeitschriften, auch als praxistauglich? Welche Schwachstellen sind bei den bisherigen Angeboten auszumachen und mit welchen generellen Akzeptanzproblemen haben sie noch zu kämpfen? Welche Faktoren sind letztlich für den Erfolg von Open Access ausschlaggebend? Diesen Fragen widmet sich der erste Teil des Beitrages. Darauf aufbauend wird dann im zweiten Teil ein Strategiekonzept zur Akzeptanzsteigerung von Open Access als Publikationsform in Wissenschaft und Forschung entwickelt.

1 Einleitung

„Open Access“ wird häufig mit „kostenfrei zugänglich“ gleich gesetzt. Zwar ist die Kostenfreiheit eine wesentliche Komponente von Open Access, sie ist als alleiniges Kriterium aber nicht hinreichend, wenn man die für die Open Access-Bewegung[1] maßgeblichen BBB-Erklärungen (Budapest Open Access Initiative[2], Bethesda Statement on Open Access Publishing[3], Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[4]) betrachtet. Denn neben den „Preisbarrieren“ (Subskriptions-, Pay-per-view- und Lizenzgebühren) sollen auch die meisten „Zugangsbarrieren“ (“permission barriers“) zu Gunsten der Nutzung durch die Allgemeinheit wegfallen.[5] In der Berlin Declaration[6] heißt es diesbezüglich:

[1] Zur Entstehung und zur Philosophie der Open Access-Bewegung vgl. beispielsweise Mruck et al. 2004 oder Woll 2005, S. 27-29.

[2] http://www.soros.org/openaccess/read.shtml (Zugriff: 29.07.2005)

[3] http://www.earlham.edu/~peters/fos/bethesda.htm (Zugriff: 29.07.2005)

[4] http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html (Zugriff: 29.07.2005)

[5] Vgl. hierzu Peter Subers „Open Access Overview“, abrufbar unter http://www.earlham.edu/~peters/fos/overview.htm (Zugriff: 29.07.2005)

[6] Der Wortlaut des Bethesda Statements ist hier nahezu identisch.

"The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship […] as well as the right to make small numbers of printed copies for their personal use."[7]

Darüber hinaus ergibt sich aus der Berlin Declaration (und in großen Teilen auch aus dem Bethesda Statement) als weitere Anforderung an Open Access-Publikationen, dass diese in mindestens einem Online-Archiv abgelegt werden müssen, welches den OAI-Standard einhält und bemüht ist, die Langzeitarchivierung sicherzustellen.

Inhaltlich geht es den Protagonisten der Open Access-Bewegung in erster Linie um den freien Zugang zu wissenschaftlicher Zeitschriftenliteratur. Hintergrund hierfür ist die sog. Zeitschriftenkrise, die an dieser Stelle nicht weiter behandelt wird.[8]

Letztlich sollen aber auch alle anderen Arten wissenschaftlicher Texte berücksichtigt werden, welche Autoren publizieren, ohne dafür einen finanziellen Gegenwert zu erhalten. In den FAQs[9] der BOAI werden beispielhaft wissenschaftliche Monografien, Tagungsbände, Diplomarbeiten und Dissertationen, von staatlichen Stellen veröffentlichte wissenschaftliche Schriftenreihen, Gesetzestexte und juristische Kommentare genannt. Die Berlin Declaration hat den Open Access-Gedanken zudem auf digitalisiertes Kulturgut in Archiven, Bibliotheken und Museen ausgeweitet. Dies ist auf die Beteiligung der Initiative ECHO (European Cultural Heritage Online)[10] zurückzuführen.

2 Status quo von Open Access

Nachdem der Terminus Open Access definiert worden ist, soll nun in diesem Kapitel eine Analyse des Status quo der beiden maßgeblichen Open Access-Strategien, dem Open Access Publishing mittels Open Access-Zeitschriften sowie dem sog. Self-Archiving vorgenommen werden.

Dabei sind die unterschiedlichen Rezeptions- und Publikationsgewohnheiten der verschiedenen Wissenschaftsdisziplinen zu berücksichtigen, die teilweise erheblich differieren können. Einige Beispiele hierfür lassen sich der aktuellen DFG-Studie „Publikationsstrategien im Wandel?"[11] (DFG 2005a) entnehmen:

[7] http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html (Zugriff: 29.07.2005)

[8] Stattdessen sei auf einige Arbeiten verwiesen, die sich eingehend mit dieser Thematik befasst haben, wie zum Beispiel Andermann/Degkwitz 2004, Meier 2002 und Woll 2005.

[9] http://www.earlham.edu/~peters/fos/boaifaq.htm (Zugriff: 29.07.2005)

[10] http://echo.mpiwg-berlin.mpg.de/home (Zugriff: 29.07.2005)

[11] Im Rahmen der Studie wurde von Oktober bis November 2004 unter mehr als 1.000 Wissenschaftlern, die im Zeitraum von 2002 bis 2004 in unterschiedlichen Programmen

- Während in den Lebens- und Naturwissenschaften Zeitschriftenaufsätze die mit Abstand am häufigsten genutzte Informationsquelle darstellen, spielen in den Ingenieurwissenschaften auch Beiträge in Proceedings bzw. Tagungsbänden (83,9 %) und in den Geistes- und Sozialwissenschaften Beiträge in Sammelbänden (74,8 %) und Monografien (70,9 %) zur Beschaffung aktueller Informationen im Fachgebiet eine wichtige Rolle (vgl. DFG 2005a, 22). Zu berücksichtigen ist hierbei jedoch die große Streuung innerhalb der Gruppe der Geistes- und Sozialwissenschaftler: „Beiträge in Sammelbänden [..] werden von rund 88 % der Geisteswissenschaftler, doch nur von 58 % der Sozial- und Verhaltenswissenschaftler häufig oder sehr häufig rezipiert. Ähnliches ist für Monografien zu konstatieren, über die sich 85 % der Geisteswissenschaftler, doch nur 53 % der Sozial- und Verhaltenswissenschaftler relativ häufig informieren." (DFG 2005a, 23)
- Im Hinblick auf die für die Publikation eigener Forschungsergebnisse präferierten Form zeigt sich zwar, dass mit Ausnahme der Ingenieurwissenschaften[12] durchweg Zeitschriftenaufsätze an erster Stelle stehen (vgl. zu diesem Abschnitt DFG 2005a, 24f.). Doch ist die Streuung innerhalb einzelner Wissenschaftsbereiche zwischen Vertretern unterschiedlicher Disziplinen teilweise erheblich, zum Beispiel innerhalb der Naturwissenschaften, wo die Vertreter aus den Fachgebieten Physik und Chemie mit rund 27 bzw. 25 Aufsätzen oberhalb, Geowissenschaftler und Mathematiker mit 14 bzw. 13 Veröffentlichungen deutlich unterhalb des Durchschnitts liegen. Darüber hinaus lässt sich bei den Geistes- und Sozialwissenschaftlern eine gegenüber den anderen Wissenschaftsbereichen deutlich erhöhte Nutzung von Sammelbänden und Monografien für die Publikation ihrer Forschungsergebnisse erkennen.

2.1 Open Access-Zeitschriften

Nimmt man die BBB-Definitionen als Maßstab, so müssen Open Access-Zeitschriften die folgenden drei Kriterien erfüllen:

1. Sie praktizieren wie die konventionellen Zeitschriften ein *Peer-Review.*
2. Die angenommenen Beiträge werden *ohne zeitlichen Verzug*, das heißt sofort mit ihrer Veröffentlichung, *kostenfrei* online zugänglich gemacht.
3. Das Copyright wird so gehandhabt, dass Interessierten, unter der Voraussetzung der korrekten Zitierung, bestimmte *Nutzungsrechte*[13] eingeräumt werden.

---

der DFG gefördert wurden, von der Gesellschaft für Empirische Studien bR in Kassel eine Befragung durchgeführt.

[12] Hier nehmen die Beiträge in Proceedings bzw. Tagungsbänden die Spitzenposition ein.

[13] Der Umfang der für die Allgemeinheit vorgesehen Rechtseinräumungen geht insbesondere im Bethesda Statement und der Berlin Declaration sehr weit. So wird hier beispielsweise das Anfertigen von "derivative works" (Bearbeitungen) erlaubt. Da dies

Der derzeit größte Anbieter von Open Access-Zeitschriften ist BioMed Central (BMC)[14], ein profitorientiertes britisches Verlagshaus mit einem Portfolio von über 170 Titeln, darunter 137 reine Open Access-Journale sowie weitere 5 Titel, bei denen zumindest die Forschungsartikel frei zugänglich sind (Stand: 02. August 2005).[15]
Der zweite renommiertere Anbieter von Open Access-Zeitschriften, ebenfalls im biomedizinischen Bereich angesiedelt, ist die Non-Profit-Organisation Public Library of Science (PLoS)[16]. Neben den beiden bereits etablierten Journalen, *PLoS Biology* und *PLoS Medicine*, werden seit 2005 drei weitere Titel herausgegeben: *PLoS Computational Biology* (erste Ausgabe Juni 2005), *PLoS Genetics* (erste Ausgabe Juli 2005) und *PLoS Pathogens* (erste Ausgabe soll am 30. September 2005 erscheinen).
In Deutschland sind im Bereich der Open Access-Zeitschriften vor allem die beiden Projekte German Medical Science (GMS)[17] und Digital Peer Publishing (DiPP) NRW[18] zu nennen.

2.1.1 Bekanntheit und Nutzung durch Wissenschaftler

Die Ergebnisse der bereits erwähnten DFG-Studie zeigen, dass nur etwa 38 % der befragten Wissenschaftler Open Access-Zeitschriften bekannt sind (vgl. DFG 2005a, 41).[19] Dem entsprechend dürftig fällt bislang auch die Nutzung von Open Access-Zeitschriften als Publikationsplattform aus (vgl. ebd.): Von den 381 Wissenschaftlern, die angegeben hatten, Open Access Zeitschriften in ihrem jeweiligen Fachgebiet zu kennen, gaben nur weitere 122 (11,9 %) an, in den zurückliegenden fünf Jahren zumindest einen Aufsatz in einer Open Access-Zeitschrift publiziert zu haben. Geht man von der Gesamtzahl der Befragten (1028) aus, ist dies ein Prozentsatz von gerade mal knapp einem Prozent. Somit kann auch nicht verwundern, dass der Durchschnittswert der in diesem Zeitraum in Open Access-Zeitschriften veröffent

jedoch ebenso wie die Frage einer möglichen kommerziellen Nutzung, die von keiner der BBB-Definitionen explizit ausgeschlossen wird, ein sensibler Punkt ist, erscheint es mir sinnvoll, dieses Kriterium nicht ganz so streng auszulegen. Eine mögliche Ausgangsbasis wäre zum Beispiel die gemeinsame Schnittmenge der BBB-Definitionen.

[14] http://www.biomedcentral.com/ (Zugriff: 02.08.2005)

[15] Eine alphabetische Übersicht aller BMC-Titel kann über http://www.biomedcentral.com/browse/journals/ (Zugriff: 02.08.2005) abgerufen werden.

[16] http://www.plos.org/ (Zugriff: 02.08.2005)

[17] http://www.egms.de/de/ (Zugriff: 02.08.2005)

[18] http://www.dipp.nrw.de/ (Zugriff: 02.08.2005)

[19] Am bekanntesten ist dieser Open Access-Kanal dabei mit 47,6 % unter den Lebenswissenschaftlern, der geringste Bekanntheitsgrad besteht innerhalb der Ingenieurwissenschaften mit einem Wert von nur 24,2 %.

lichten Beiträge unter eins liegt (er reicht von 0,2 bei den Geistes- und Sozialwissenschaftlern bis zu 0,8 bei den Naturwissenschaftlern).
Der Anteil der Artikel, die in Open Access-Zeitschriften innerhalb der letzten fünf Jahre veröffentlicht wurden liegt zwischen 1,6 % (Geistes- und Sozialwissenschaften) und 5,4 % (Ingenieurwissenschaften) (vgl. DFG 2005b, 13, Tabelle 14c).

2.1.2 Erfolgskritische Faktoren[20]

2.1.2.1 Rechtliche Rahmenbedingungen

Im Bereich der Open Access-Zeitschriften sind keine nennenswerten urheberrechtlichen Probleme zu verzeichnen: Die Beiträge werden kostenfrei im Internet zugänglich gemacht, unabhängig davon, ob das Copyright[21] beim Autor verbleibt oder ob es an die Zeitschrift übertragen wird. Im Detail geht es dann darum, die richtige Balance zwischen der kompletten Freigabe aller Rechte ("public domain") einerseits und dem Festhalten an allen Rechten ("all rights reserved") andererseits zu finden. Die hierfür notwendigen Rahmenbedingungen sind mit der Bereitstellung verschiedener Lizenzen der Kategorie "some rights reserved" (z.B. den Creative Commons-Lizenzen[22] oder auch den Digital Peer Publishing Lizenzen[23]) bereits geschaffen worden.

2.1.2.2 IT-Infrastruktur

Der technische Aufwand für den Aufbau von Open Access-Zeitschriften kann individuell sehr unterschiedlich sein und richtet sich neben den bereits gegebenen Voraussetzungen bezüglich der IT-Infrastruktur in erster Linie nach dem Umfang der angebotenen Features (vgl. Björk 2004):

- Speichermechanismus für die Dokumente (statische Webseiten versus Datenbank)
- Formate (HTML, PDF, XML etc.)
- Grafiken, multimediale Elemente
- Indexierungs- und Linkingsysteme
- Alerting- und Personalisierungsmöglichkeiten
- Diskussionslisten
- Nutzungs- und Zitationsstatistiken

---

[20] Die Auswahl der erfolgskritischen Faktoren orientiert sich an einer Studie von Björk (2004), in welcher die Relevanz verschiedener „ Barrieren" für die beiden maßgeblichen Open Access-Strategien Open Access-Zeitschriften und Self-Archiving eruiert wurde.

[21] Das deutsche Urheberrecht ist im Gegensatz zum Copyright als Persönlichkeitsrecht an den Autor gebunden. Der Urheber hat jedoch gemäß § 31 UrhG die Möglichkeit einfache oder ausschließliche Nutzungsrechte einzuräumen.

[22] http://creativecommons.org/ (Zugriff: 02.08.2005)

[23] http://www.dipp.nrw.de/lizenzen/ (Zugriff: 02.08.2005)

- Sicherheits-Back-ups, Spiegelseiten etc.

Erhebliche Kosten lassen sich auf lange Sicht durch die gemeinsame Nutzung von Ressourcen einsparen, z.B. durch das Teilen von Softwareanwendungen, durch kollaboratives Webhosting oder auch durch das Zurückgreifen auf ein gemeinsames Workflow-System, wie es beispielsweise im Projekt German Academic Publishers (GAP)[24] entwickelt worden ist.

2.1.2.3 Geschäftsmodelle

Am weitesteten verbreitet sind derzeit Geschäftsmodelle, welche die konventionellen Finanzströme umkehren: Es müssen keine Subskriptionsgebühren gezahlt werden, sondern die Autoren bzw. deren Institutionen zahlen eine „Publikationsgebühr" für die sofortige freie Zugänglichmachung ihrer Beiträge. Als Prototypen für dieses Geschäftsmodell können die beiden bereits erwähnten Verlage BMC und PLoS angesehen werden:

BMC erhebt von den Autoren sog. Artikelbearbeitungsgebühren ("article processing charges"), die bei den meisten Zeitschriften 590 US$ betragen[25]. Alternativ können diese Gebühren von den Institutionen der Autoren durch eine Mitgliedschaftsjahres-pauschale abgegolten werden, deren Höhe von der Gesamtanzahl der Forscher und Fakultätsmitglieder in den entsprechenden Fakultäten, z.B. Biologie und Medizin, abhängig ist (die Spannbreite reicht derzeit von 1725 US$ für sehr kleine Institute bis zu 8625 US$ für sehr große Institute)[26]. Auch die Autoren selbst können zu einer Reduzierung der Gebühr beitragen, indem sie ein formatiertes Manuskript ein-reichen.[27] Ganz erlassen wird die Gebühr in Fällen unzumutbarer Härte, z.B. für Autoren aus Entwicklungsländern.[28]

PLoS erhebt zur Kostendeckung von den Autoren eine Publikationsgebühr in Höhe von 1.500 US$ pro akzeptiertem Artikel.[29] Durch verschiedene Arten institutioneller Mitgliedschaften können Rabattierungen von 10 % ("Active Member"; Jahresbeitrag: 2.000 $) bis 75 % ("Championing Member"; Jahresbeitrag: 100.000 $) erreicht werden.[30] Wie bei BMC wird auf die Gebühr verzichtet, wenn ein Wissenschaftler nicht über die notwendigen Forschungsgelder verfügt.[31]

---

[24] http://www.ubka.uni-karlsruhe.de/gap-c/index_de.html (Zugriff: 02.08.2005)

[25] Die Spannbreite bei den übrigen Zeitschriften reicht von US$395 (*International Journal of Behavioral Nutrition and Physical Activity*) bis zu US$1610 (*BMC Biology*, *BMC Medicine*, *Genome Biology, Journal of Biology*). Der Durchschnittswert dieser 19 Titel liegt bei 1020 US$. (http://www.biomedcentral.com/info/authors/apcfaq, Zugriff: 10.08.2005)

[26] http://www.biomedcentral.com/info/about/instmembership (Zugriff: 10.08.2005)

[27] http://www.biomedcentral.com/info/authors/apcfaq (Zugriff: 10.08.2005)

[28] http://www.biomedcentral.com/info/authors/apcfaq (Zugriff: 10.08.2005)

[29] http://www.plos.org/faq.html (Zugriff: 10.08.2005)

[30] http://www.plos.org/support/instmembership.html (Zugriff: 10.08.2005)

[31] http://www.plos.org/faq.html (Zugriff: 10.08.2005)

Dass dieses Geschäftsmodell häufig pauschal als "authors-pays model" tituliert wird, obwohl die Gebühren vielfach von den Institutionen oder Fördereinrichtungen übernommen werden, ist als äußerst erfolgskritisch einzustufen, denn verschiedene Umfragen[32] zeigen, dass (bislang) nur etwa 50 % der Autoren überhaupt bereit wären, für eine Veröffentlichung in einer Open Access-Zeitschrift zu zahlen, und hiervon wiederum nur 15 - 20 % mehr als 500 $ und nur etwa 5 % mehr als 1.000 $ zahlen würden.

Weiterhin finden verschiedene Formen hybrider Geschäftsmodelle in der Praxis Anwendung. Die beiden am häufigsten anzutreffenden Varianten sind:

- „Teilweiser" Open Access: Bei diesem Geschäftsmodell werden die primären Forschungsartikel gegen eine bestimmte „Autorengebühr" frei zugänglich angeboten. Für alle anderen Inhalte (zum Beispiel Editorial, Review-Artikel, Kommentare, Rezensionen) werden weiterhin Subskriptionsgebühren erhoben. Beispiel: *Breast Cancer Research* von BMC (Gebühr: 1345 US$)[33].
- „Optionaler" Open Access: Hier wird das traditionelle Subskriptionsmodell dahingehend modifiziert, dass Artikel, für welche die Autoren eine bestimmte Gebühr zahlen, zusätzlich frei zugänglich im Internet angeboten werden. Beispiele: Springer Open Choice (Gebühr: 3.000 US$)[34]; Oxford Open von Oxford University Press (Gebühr: 1500 £ bzw. 800 £, falls die Institution eine laufende Online-Subskription aufweist)[35].

Als dritte Gruppe sind Geschäftsmodelle zu nennen, die vollständig auf die Erhebung von „Autorengebühren" verzichten und die benötigten Finanzmittel aus verschiedenen anderen Quellen beziehen:

- "Dual mode" Open Access: Die Zeitschrift erscheint nach wie vor als Printausgabe auf Subskriptionsbasis, zusätzlich wird jedoch unmittelbar nach der Veröffentlichung eine komplette Online-Version frei zugänglich angeboten. Beispiele: *Documenta Mathematica*[36]; *Journal of Postgraduate Medicine*[37].
- Subventionierung: Die Finanzierung erfolgt über (öffentliche oder private) Mittel, z.B. direkt aus Zuschussfonds oder indirekt durch die Institution, welche das Personal bezahlt und die Infrastruktur bereitstellt. Beispiel: *D-Lib Magazine* (Finanzierung aus Mitteln von DARPA und NSF).

---

[32] Vgl. Cozzarelli et al. 2004; Rowlands et al. 2004: 28

[33] http://breast-cancer-research.com/info/faq/apcfaq.asp?txt_faq=howmuch (Zugriff: 14.08.2005)

[34] http://www.springeronline.com/sgw/cda/frontpage/0,11855,1-40359-12-115393-0,00.html (Zugriff: 14.08.2005)

[35] http://www.oxfordjournals.org/oxfordopen/about (Zugriff: 14.08.2005)

[36] http://www.math.uiuc.edu/documenta/ (Zugriff: 14.08.2005)

[37] http://www.jpgmonline.com/currentissue.asp (Zugriff: 14.08.2005)

Weitere Einnahmequelle sind Mehrwertdienstleistungen wie z.B. Printing on demand-Services, Werbung, Sponsoring oder das Eintreiben von Spenden.
Die Frage, welches Geschäftsmodell am besten geeignet ist, lässt sich nicht pauschal beantworten, da der Erfolg von verschiedenen, teilweise (noch) nicht genau kalkulierbaren Parametern abhängig ist: Ein entscheidender Faktor ist in jedem Fall die grundsätzliche Frage, ob eine Gewinnerzielung im Vordergrund steht oder ob es nur um eine (annähernde) Kostendeckung (Non-Profitbereich) geht und welcher Qualitätsmaßstab jeweils zu Grunde gelegt wird. Ein weiterer wichtiger Aspekt, der berücksichtigt werden muss, sind die in Kapitel 2 kurz angerissenen (teilweise gravierenden) disziplinären Unterschiede, die dazu führen können, dass Geschäftsmodelle in manchen Fächern oder Nischen gut funktionieren und in anderen weniger.

2.1.2.4 Indexierungsservices und -standards

Björk (2004) schreibt den kommerziellen Indexierungsdiensten (sein spezieller Fokus liegt dabei auf den Zitationsindices[38] des ISI) eine wichtige Doppelfunktion zu:
"First, they [indexing services] help in attracting occasional readers who may not even be aware of the journal's existence. Secondly, the fact that a journal can claim being 'indexed in' lends prestige to the journal and thus helps in attracting more and better submission."
Er betrachtet es daher als großes Manko, dass Open Access-Zeitschriften bislang kaum in kommerziellen Indexierungsdiensten nachgewiesen sind.

2.1.2.5 Akademisches Reward-System

Im Rahmen der DFG-Studie wurde auch die Relevanz bestimmter Kriterien für die Auswahl einer Zeitschrift zur Veröffentlichung eigener Forschungsergebnisse untersucht (vgl. zu diesem Abschnitt DFG 2005a, 25-28). Das Renommee der Zeitschrift wird dabei von allen vier Wissenschaftsbereichen als wichtiges bzw. sehr wichtiges Auswahlkriterium angesehen.[39] Eine noch größere Bedeutung wird (mit Ausnahme der Geistes- und Sozialwissenschaften) jedoch dem internationalen Verbreitungsgrad der Zeitschrift beigemessen (Durchschnittswert: 92,2 %). Im Vergleich dazu wird der Impact Faktor insgesamt gesehen deutlich seltener als wichtig eingestuft (Durchschnittswert: 61,7 %). Zu berücksichtigen ist dabei aber der signifikante Unterschied zwischen den Geistes- und Sozialwissenschaften (42,7 %) einerseits und den Lebenswissenschaften (83,3 %) andererseits.

[38] Science Citation Index, Social Sciences Citation Index, Arts and Humanities Citation Index

[39] Den geringsten Wert weisen die Geistes- und Sozialwissenschaften mit 85,1 %, den höchsten die Naturwissenschaften mit 93,7 % auf.

2.1.2.6 Marketing und kritische Masse

Insgesamt machen Open Access-Zeitschriften erst einen Anteil von knapp 8 %[40] des wissenschaftlichen Zeitschriftenmarktes aus. Björk (2004) führt dies unter anderem auf ein gänzlich fehlendes oder unzureichendes Marketing zurück. Vor allem gehe es darum, das Prestige des Journals zu erhöhen, wobei folgendes zu beachten sei:
"First, the reputation of the editor and the constitution of the editorial board are important. Secondly, attracting enough papers from leading academics early on is important. This can again lead to a positive chain reaction of citations in other articles and eventually (in the long term) inclusion in the SCI."

2.2 Self-Archiving (elektronische Archive)

Self-Archiving meint die durch einen Wissenschaftler oder dessen Institution selbst vorgenommene digitale Speicherung seiner Fachbeiträge[41] in geeigneten „elektronischen Archiven" (Repositorien) , wobei drei Arten zu unterscheiden sind:[42]

1. Individuelle Repositorien: Selbstarchivierung durch den Autor auf seiner eigenen Website;
2. Fachliche (zentrale, disziplinäre) Repositorien: Archivierung der Beiträge erfolgt auf fachbezogenen Servern, wobei Veröffentlichungen aus verschiedenen Einrichtungen gebündelt werden (Beispiel: arXiv[43]);
3. Institutionelle Repositorien (lokale Publikationsserver): Archivierung der Beiträge erfolgt auf dem Server der Forschungseinrichtung, des Institutes, der Fakultät oder der Bibliothek. Vor allem in Deutschland handelt es sich dabei vielfach noch um (mehr oder weniger) reine Hochschulschriftenserver.

Das Hauptaugenmerk der nachfolgenden Analyse gilt den institutionellen Repositorien.

2.2.1 Bekanntheit und Nutzung durch Wissenschaftler

Laut der DFG-Studie wurde die Möglichkeit, auf konventionelle Weise publizierte Zeitschriftenaufsätze nochmals für den entgeltfreien Zugriff im Internet zugänglich zu machen, innerhalb der letzten fünf Jahre von 20,1 % der Naturwissenschaftler und

[40] Dabei wurde von einer Gesamtzahl von ca. 21.000 Peer-Reviewed wissenschaftlichen Journals (Quelle: Homerton College Library Online Resource Guide, http://www.homerton.cam.ac.uk/libguide_resources_26jan05.pdf, Zugriff: 14.08.2005) sowie rund 1.670 Peer-Reviewed Open Access Journals (Quelle: Directory of Open Access Journals, http://www.doaj.org/, Zugriff: 14.08.2005) ausgegangen.

[41] Dies können sowohl die endgültigen, redigierten Versionen der Artikel sein, die ein Peer-Review durchlaufen haben (Postprints), aber auch Preprints, also die noch nicht begutachtete Manuskriptfassung.

[42] Vgl. http://www.isn-oldenburg.de/publications/11argumente.html (Zugriff: 31. Juli 2005)

[43] http://arxiv.org/ (Zugriff: 14.08.2005)

von 17,6 % der Ingenieurwissenschaftler[44] genutzt (DFG 2005b, 15, Tabelle 16c). Die Lebenswissenschaftler fallen mit 12,3 % etwas, die Geistes- und Sozialwissenschaftler mit 5,9 % deutlich ab (ebd.), was sich bei Letzteren vor allem auf die im Vergleich zu den drei anderen Wissenschaftsbereichen deutlich geringere Relevanz der Zeitschrift als Publikationsmedium zurückführen lässt. Im Hinblick auf Preprint-Archive ergibt sich folgendes Bild: Diese sind zwar 49 %[45] der Naturwissenschaftler, aber nur 20,3 % der Ingenieurwissenschaftler und sogar nur 13,9 % der Geistes- und Sozialwissenschaftler für ihr eigenes Fach bekannt (DFG 2005b, 16, Tabelle 18).

2.2.2 Erfolgskritische Faktoren

2.2.2.1 Rechtliche Rahmenbedingungen

Die notwendigen rechtlichen Rahmenbedingungen für das Self-Archiving sind vielfach bereits gegeben: So gestatten laut der SHERPA/RoMEO-Liste[46] 71 % der 118 erfassten Zeitschriftenverlage zumindest eine Form des Self-Archving, 49 % erlauben sowohl das Self-Archiving von Preprints als auch der Postprints. Problematisch ist jedoch, dass dies unter den Wissenschaftlern bislang wenig bekannt ist[47] und nur ein geringes Interesse an urheberrechtlichen Fragen zu bestehen scheint (vgl. Rowlands et al. 2004, 14).

2.2.2.2 IT-Infrastruktur

Soweit bereits die notwendig Basisausstattung (insbesondere ein geeigneter Server) vorhanden ist, können die Kosten für den Aufbau eines elektronischen Archivs relativ gering gehalten werden, zumal inzwischen eine Reihe von Open Source Softwareprodukten hierfür zur Verfügung steht (z.B. Eprints, DSpace, Fedora, MyCoRe, OPUS). Einzukalkulieren sind in jedem Fall Personalkosten für die Pflege und die Administration des elektronischen Archivs.

2.2.2.3 Geschäftsmodelle

Da die institutionellen Repositorien von der jeweiligen Einrichtung getragen werden, besteht hier kein konkreter Handlungsbedarf, Geschäftsmodelle zu entwickeln.

---

[44] Diese stellen darüber hinaus auch 26 % ihrer Proceedings- und Tagungsbeiträge kostenlos im Internet bereit.

[45] Diese vergleichsweise hohe Zahl relativiert sich, wenn man die in diesem Wissenschaftsbereich bereits früh ausgeprägte „Preprint-Kultur“ berücksichtigt.

[46] http://www.sherpa.ac.uk/romeo.php (Zugriff: 14.08.2005)

[47] So kennen laut einer aktuellen Umfrage von Swan/Brown (2005, 7) nur 10 % der Autoren die SHERPA/RoMEO-Liste und in einigen Fällen bestehen nach wie vor urheberrechtliche Bedenken, vor allem aufgrund des sog. „Ingelfinger-Gesetzes“, welches besagt, dass nur solche Artikel angenommen werden, die noch in keiner Form publiziert worden sind.

2.2.2.4 Indexierungsservices und -standards

Damit die Inhalte der elektronischen Archive überhaupt von einer größeren Öffentlichkeit wahrgenommen werden können, müssen sie zumindest über allgemeine Suchmaschinen wie Google[48] auffindbar sein. Darüber hinaus sollten die Repositorien mit OAI-Schnittstellen ausgestattet und die Metadaten nach Dublin Core beschrieben sein, so dass sie mit speziellen Suchwerkzeugen wie z.B. OAIster[49], CiteSeer[50] oder Citebase[51] unter einer einheitlichen Suchoberfläche durchsucht werden können.

2.2.2.5 Akademisches Reward-System

Da die in institutionellen Archiven befindlichen Dokumente in der Regel kein Peer-Review durchlaufen haben und hier oftmals verschiedenste Materialien unterschiedlicher Qualität eingestellt werden, stellt sich die Frage, wie dieses mangelhafte „professionelle Prestige" kompensiert werden kann. Beispielsweise könnte durch eine inhaltliche Differenzierung des Angebotes die Transparenz deutlich erhöht werden (vgl. hierzu den Vorschlag von Töwe/Piguet 2005, 178f.) Darüber hinaus sollte geprüft werden, inwieweit Verfahren zur Qualitätsprüfung realisierbar sind. Zumindest sollte vor der Einspeisung der Beiträge in den Server eine Prüfung durch Hochschullehrer, Institutsmitarbeiter oder durch die Fachreferenten der Hochschulbibliothek vorgenommen werden.

2.2.2.6 Marketing und kritische Masse

Die kritische Masse, ab der sich institutionelle Dokumentenserver als ernsthafte Alternative zu kommerziellen Zeitschriftenverlagen etablieren könnten, ist bei Weitem noch nicht erreicht, wie Ware (2004) für 45 institutionelle Repositorien aus verschiedenen Nationen und Woll (2005, 49ff.) exemplarisch für die Dokumentenserver der Universitäten in Nordrhein-Westfalen festgestellt haben. Töwe/Piguet (2005) sehen den größten Handlungsbedarf zunächst „bei der aktiven Information der Autoren und der Fakultäten sowie bei der Akquisition von Postprints". Hierzu sollte die Bandbreite der verschiedenen kommunikationspolitischen Instrumente so weit wie möglich ausgeschöpft werden.[52]

[48] http://www.google.de/ (Zugriff: 14.08.2005)

[49] http://oaister.umdl.umich.edu/o/oaister/ (Zugriff: 14.08.2005)

[50] http://citeseer.ist.psu.edu/ (Zugriff: 14.08.2005)

[51] http://www.citebase.org/cgi-bin/search (Zugriff: 14.08.2005)

[52] Zum Einsatz kommunikationspolitischer Instrumente vgl. im Detail Gattuso 2004, S. 64-70.

3 Strategiekonzept

Ein (nicht nur) im Wissenschaftssektor häufig anzutreffendes Phänomen ist, dass Probleme zwar identifiziert, für deren Lösung aber keine übergreifende koordinierte Strategie entwickelt wird. Daher soll als Ergebnis dieses Beitrages abschließend ein Strategiekonzept vorgestellt werden, welches als konkrete Handlungsanweisung für die im Zusammenhang mit der Optimierung von Open Access-Angeboten vorzunehmenden Maßnahmen gedacht ist. Diese Maßnahmen sind auf der Basis der Analyseergebnisse aus Kapitel 2 in nachfolgender Übersicht zusammengestellt worden (Tabelle 1):

Handlungsfeld	Einzelmaßnahmen
Gezieltes Marketing, Öffentlichkeitsarbeit, Werbung	Akquirierung hochrangiger Wissenschaftler
	Transparenz hinsichtlich der Qualität einzelner Beiträge in elektronischen Archiven schaffen
	Vorzüge des elektronischen Publizierens im Allgemeinen und von Open Access im Speziellen aufzeigen
Geschäftsmodelle	Nachhaltigkeit sichern auf der Basis von Business Plänen und Vollkostenrechnungen
Autorenbetreuung („Advocacy“)	Allgemeine Beratung, Telefon- und Email-Support
	Spezielle Kurse und Schulungsangebote
	Entwicklung und Bereitstellung von Authoring Tools zur Minimierung des zeitlichen Aufwands für die Autoren (z.B. Richtlinien und Hinweise für die Autoren, Konvertierungswerkzeuge, Metadatenwerkzeuge, Dokumentvorlagen)
	Schaffung von Mehrwertdienstleistungen (z.B. Publikationslisten, bibliografische Auswertungen, Nutzungsstatistiken, Print on demand-Service)
Wissenschaftliches Gratifikations- und Bewertungssystem	Entwicklung und Etablierung alternativer Instrumente zur Leistungsbewertung und Evaluation (z.B. „bereinigte“ Downloadstatistiken, Faculty of 1000[53]) bzw. Modifikation des Impact Faktors (vgl. die Empfehlungen der AWMF zur Verwendung des IF)[54]
	Angemessene Berücksichtigung von Open Access-Publikationen bei Promotions-, Einstellungs- und Berufungsverfahren sowie bei der Vergabe von Drittmitteln oder Fördergeldern
	Etablierung von Verfahren zur Qualitätsprüfung (z.B. Herausgebergremium) auch für Dokumentenserver

[53] http://www.facultyof1000.com/start.asp (Zugriff: 14.08.2005)

[54] http://www.uni-duesseldorf.de/WWW/AWMF/awmf-fr2.htm (Zugriff: 14.08.2005)

Technik	Wahrung von Integrität und Authentizität der Dokumente
	Sicherstellung der Langzeitverfügbarkeit der Dokumente
	Workflow-Management
	Umsetzung der DINI-Empfehlungen für elektronisches Publizieren an Hochschulen[55] sowie der Richtlinien des DINI-Zertifikates für Dokumenten- und Publikationsserver[56]
	Vernetzung der Open Access-Angebote mit internationalen Angeboten (z.B. internationale Sichtbarkeit über Wissenschaftsportale)
Urheberrechtliche Rahmenbedingungen	Transparenz hinsichtlich Urheberrecht/Copyright schaffen (Hinweis auf die RoMEO/Sherpa-Liste, ggf. spezielle Aufbereitung dieser Liste für die Wissenschaftler des Instituts)
	Konkrete Empfehlungen für Autoren der Institution für das Abschließen von Verlagsverträgen vorlegen
Hochschul-/wissenschaftspolitische Ebene	Offizielles Bekenntnis der einzelnen Hochschule/Wissenschaftsinstitution zu Open Access im Sinne der "Berlin 3 Open Access"-Empfehlung[57], z.B. durch die Verabschiedung einer entsprechenden Resolution[58]
	Anschluss neuer Projekte an bereits bestehende Initiativen und Netzwerke wie DiPP, GAP oder GMS, Erschließung möglicher Synergieeffekte (ggf. in Form eines Kompetenznetzwerkes ähnlich wie für den Bereich der Langzeitarchivierung kopal[59])

Tabelle 1: Maßnahmen zur Optimierung von Open Access-Angeboten

Diese lange Liste von Einzelmaßnahmen lässt sich nicht sofort vollständig umsetzen. Neben der Relevanz einer Maßnahme sind auch der Zeitaufwand und die anfallenden Kosten zu berücksichtigen. Daher wird für die Realisierung ein „Vier-Stufen-Plan“ vorgeschlagen, welcher diese drei Faktoren in Relation zueinander setzt (Tabelle 2):

[55] http://www.dini.de/documents/DINI-EPUB-Empfehlungen-2002-03-10.pdf (Zugriff: 14.08.2005)

[56] http://www.dini.de/dini/documents/DINI-Zertifikat-2003-10-08.pdf (Zugriff: 14.08.2005)

[57] http://www.eprints.org/berlin3/outcomes.html (Zugriff: 14.08.2005)

[58] Siehe hierzu die Pressemitteilung „Universität Bielefeld als Vorreiter für ’Open Access’ an deutschen Hochschulen“ vom 07.06.2005, http://bis.uni-bielefeld.de/infomanager/SilverStream/Pages/Pressemitteilungen_Detail_Web.html?query=PRESSEMITTEILUNGEN.ID+%3D+28428 (Zugriff: 14.08.2005)

[59] http://kopal.langzeitarchivierung.de/ (Zugriff: 14.08.2005

Einzelmaßnahme	P[1)]	FZ[2)]	FK[3)]	Stufe[4,5)]
Vorzüge des elektronischen Publizierens im Allgemeinen und von Open Access im Speziellen aufzeigen	1	1	1	1
Transparenz hinsichtlich Urheberrecht/Copyright schaffen	1	1	1	1
Akquirierung hochrangiger Wissenschaftler	1	2	2	1
Angemessene Berücksichtigung von Open Access-Publikationen bei Promotions-, Einstellungs- und Berufungsverfahren sowie bei der Vergabe von Drittmitteln oder Fördergeldern	1	2	2	1
Nachhaltigkeit des Geschäftsmodells auf der Basis von Business Plänen und Vollkostenrechnungen sichern	1	3	2	2
Allgemeine Beratung, Telefon- und Email-Support	2	1	1	2
Spezielle Kurse und Schulungsangebote	2	1	1	2
Konkrete Empfehlungen für Autoren der Institution für das Abschließen von Verlagsverträgen vorlegen	2	2	1	2
Workflow-Management	2	2	2	2
Entwicklung und Bereitstellung von Authoring Tools zur Minimierung des zeitlichen Aufwands für die Autoren	2	2	2	2
Transparenz hinsichtlich der Qualität einzelner Beiträge schaffen	2	2	2	2
Anschluss neuer Projekte an bereits bestehende Initiativen und Netzwerke wie DiPP, GAP oder GMS, Erschließung möglicher Synergieeffekte	2	2	2	2
Etablierung von Verfahren zur Qualitätsprüfung auch für Dokumentenserver	2	3	1	3
Umsetzung der DINI-Empfehlungen für elektronisches Publizieren an Hochschulen sowie der Richtlinien des DINI-Zertifikates für Dokumenten- und Publikationsserver	2	3	2	3
Entwicklung und Etablierung alternativer Instrumente zur Leistungsbewertung und Evaluation bzw. Modifikation des Impact Faktors	2	3	2	3
Wahrung von Integrität und Authentizität der Dokumente	2	3	3	4
Sicherstellung der Langzeitverfügbarkeit der Dokumente	2	3	3	4
Schaffung von Mehrwertdienstleistungen	3	2	2	4
Offizielles Bekenntnis der einzelnen Hochschule/Wissenschaftsinstitution zu Open Access im Sinne der "Berlin 3 Open Access"-Empfehlung	3	2	2	4
Vernetzung der Open Access-Angebote mit internationalen Angeboten (z.B. internationale Sichtbarkeit über Wissenschaftsportale)	3	3	3	4

Tabelle 2: „Vier-Stufen-Plan“

Erläuterungen zu Tabelle 2:

1) Priorität

1 = hohe Priorität

2 = mittlere Priorität

3 = wünschenswert

2) Faktor Zeit

1 = unmittelbar bzw. kurzfristig realisierbar

2 = mittelfristig realisierbar

3 = nur langfristig realisierbar

3) Faktor Kosten

1 = gegen null bis gering

2 = mittel

3 = hoch

4) Stufe 1: Maßnahmen mit hoher Priorität, die kurz- bis mittelfristig realisiert werden können und mit geringem bis mittlerem finanziellen Aufwand verbunden sind;
Stufe 2: Maßnahmen mit hoher Priorität, die sich aber nur langfristig realisieren lassen und/oder mit größerem finanziellen Aufwand verbunden sind, sowie Maßnahmen mit mittlerer Priorität, die kurz- bis mittelfristig und mit einem geringen bis mittleren finanziellen Aufwand realisiert werden können;
Stufe 3: Maßnahmen mit mittlerer Priorität, die sich jedoch nur langfristig realisieren lassen und/oder mit größerem finanziellem Aufwand verbunden sind;
Stufe 4: Maßnahmen, die wünschenswert, aber nicht vordringlich sind sowie Maßnahmen, die zwar eine mittlere oder sogar hohe Priorität haben, sich aber nur langfristig und mit hohem finanziellem Aufwand realisieren lassen.

5) Die Zuordnung zu den vier Stufen kann Tabelle 3 entnommen werden.

Stufe	Priorität	Faktor Zeit	Faktor Kosten
1	1	1	1
1	1	1	2
1	1	2	1
1	1	2	2
2	1	1	3
2	1	3	1
2	1	2	3
2	1	3	2
2	2	1	1
2	2	1	2
2	2	2	1
2	2	2	2
3	2	1	3
3	2	3	1
3	2	2	3

3	2	3	2
4	1	3	3
4	2	3	3
4	3	1	1
4	3	1	2
4	3	2	1
4	3	1	3
4	3	2	2
4	3	3	1
4	3	2	3
4	3	3	2
4	3	3	3

Tabelle 3: Matrix zum „Vier-Stufen-Plan“ (Tabelle 2)

Literatur

Andermann, H., Degkwitz, A. (2004): Neue Ansätze in der wissenschaftlichen Informationsversorgung: eine Überblick über Initiativen und Unternehmungen auf dem Gebiet des elektronischen Publizierens. In: Bibliothek. Forschung und Praxis 28 (1): 35-59

Björk, B.-C. (2004): Open access to scientific publications - an analysis of the barriers to change. In: Information Research 9 (2) <http://informationr.net/ir/9-2/paper170.html> (Stand 14.08.2005)

Cozzarelli, N. R., Fultion, K. R., Sullenberger, D. M. (2004): Results of a PNAS Author Survey on an Open Access Option for Publication. Proceedings of the National Academy of Sciences 101 (5): 1111 < http://www.pnas.org/cgi/content/full/101/5/1111> (Stand 14.08.2005)

DFG (2005a): Publikationsstrategien im Wandel? Ergebnisse einer Umfrage zum Rezeptions- und Publikationsverhalten unter besonderer Berücksichtigung von Open Access. Weinheim: Wiley <http://www.dfg.de/dfg_im_profil/zahlen_und_fakten/statistisches_berichtswesen/open_access/download/oa_ber_dt.pdf> (Stand 14.08.2005)

DFG (2005b): Publikationsstrategien im Wandel? Ergebnisse einer Umfrage zum Rezeptions- und Publikationsverhalten unter besonderer Berücksichtigung von Open Access: Tabellenband. Weinheim: Wiley <http://www.dfg.de/dfg_im_profil/zahlen_und_fakten/statistisches_berichtswesen/open_access/download/oa_tabband.pdf> (Stand 14.08.2005)

Gattuso, M. (2004): Verbesserung der Akzeptanz und Nutzung von Hochschulschriftenservern, dargestellt am Beispiel des Online Publikationsver

bundes Stuttgart, Diplom-Arbeit. Stuttgart: Hochschule der Medien, Fachbereich Information und Kommunikation <http://elib.uni-stuttgart.de/opus/volltexte/2005/2200/pdf/opus0205.pdf> (Stand 14.08.2005)

Meier, M. (2002): Returning Science to the Scientists: der Umbruch im STM-Zeitschriftenmarkt unter Einfluss des Electronic Publishing. München: peniope

Mruck, K., Gradmann, S., Mey, G. (2004): Open Access: Wissenschaft als öffentliches Gut. Forum Qualitative Sozialforschung 5 (2) <http://www.qualitative-research.net/fqs-texte/2-04/2-04mrucketal-d.htm> (Stand 14.08.2005)

Rowlands, I., Nicholas, D., Huntingdon, P. (2004): Scholarly Communication in the digital environments: what do authors want? Findings of an international survey of author opinion: project report <http://ciber.soi.city.ac.uk/ciber-pa-report.pdf>

Swan, A., Brown, S. (2005): Open access self-archiving: An author study. Technical Report, Joint Information Systems Committee (JISC) <http://www.keyperspectives.co.uk/OpenAccessArchive/2005_Open_Access_Report.pdf> (Stand 14.08.2005)

Töwe, M., Piguet, A. (2005): Konzeptstudie E-Archiving. Zürich: Konsortium der Schweizer Hochschulbibliotheken <http://e-collection.ethbib.ethz.ch/ecol-pool/bericht/bericht_412.pdf> (Stand 14.08.2005)

Ware, M. (2004): Pathfinder Research on Web-based Repositories: Final Report <http://www.palsgroup.org.uk/palsweb/palsweb.nsf/0/8c43ce800a9c67cd80256e370051e88a/$FILE/PALS%20report%20on%20Institutional%20Repositories.pdf> (Stand 14.08.2005)

Woll, C. (2005): Wissenschaftliches Publizieren im digitalen Zeitalter und die Rolle der Bibliotheken (= Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft ; Bd. 46). Köln: Fachhochschule Köln <http://www.fbi.fh-koeln.de/institut/papers/kabi/volltexte/Band046.pdf> (Stand 14.08.2005)

DataMining – Verfahren und Anwendungen

Generierung ontologischer Konzepte und Relationen durch Text Mining-Verfahren: Methoden und Bewertung

Christian Wolff, Regensburg

Abstract

Dieser Kurzbeitrag beschreibt zunächst den Prozess des ontology engineering und geht auf die verschiedenen Verfahren zur Erzeugung und Erweiterung von Ontologien durch Verfahren des Text Mining ein. Dabei wird die aktuelle Forschungssituation skizziert und Probleme der Bewertung von Ontologien werden diskutiert.

1 Einleitung

Die Wissensakquisition stellt einen Flaschenhals bei der Erzeugung und Pflege von Ontologien dar. Anwendungen für das Semantic Web (vgl. Berners-Lee 1998), insbesondere im Bereich des Wissensmanagements und der Informationsaufbereitung erfordern hochspezialisierte Wissensstrukturen, bei denen nur bedingt auf vorgefertigte Ontologien zurückgegriffen werden kann. Verfahren des Text Mining (vgl. Mehler & Wolff 2005), die aus großen und un- oder semi-strukturierten Dokumentbeständen für den Aufbau einer Ontologie relevante Konzepte und die zwischen ihnen bestehenden Relationen identifizieren können, haben deshalb im Bereich des *ontology engineering* an Bedeutung gewonnen. Text Mining fußt dabei auf der Annahme, dass mit halb- oder vollautomatischen Verfahren aus großen un- oder semi-strukturierten Textbeständen relevante Konzepte für den Aufbau von Ontologien identifiziert und extrahiert werden können und es auch möglich ist, Beziehungen zwischen Konzepten zu erkennen und – im Sinne des Typensystems einer Ontologie – auch zu klassifizieren. Wenn mit ihrer Hilfe in der Regel zwar die vollautomatische Generierung vollständiger Ontologien (noch) nicht möglich ist, so können sie doch den Modellierungsaufwand erheblich reduzieren.

2 Der Prozess des Ontologieaufbaus

Gruber 1993 präzisiert den Ontologiebegriff in seiner weithin akzeptierten Definition wie folgt:[1]

[1] Die nachfolgenden ausführungen beziehen sich auf den Ontologiebegriff, wie er in Künstlicher Intelligenz-Forschung und Informatik Verbreitung gefunden hat.

"An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what “exists” is exactly that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse." (Gruber 1993: 199)

Beim Aufbau von Ontologienist zunächst zwischen der formalen Grundlegung der ontologischen Struktur und der tatsächlichen Erzeugung einer Ontologie für ein bestimmtes Wissensgebiet zu unterscheiden. Das erstere Problem besteht in der Definition geeigneter grundlegender Relationstypen und Primitive, aus denen jede konkrete Ontologie aufgebaut ist (sog. top-level ontology):

"The term formal ontology has its origin in philosophy but here we use it in a special sense to designate a research area in theoretical computer science which is aimed at the systematic elaboration of formalized axiomatic theories of forms and modes of being, and at the development of formal specification tools and methods to support the modeling of the complex structures of the world.“ (Heller & Herre 2003: 2).

Auf dieser Ebene kommen vor allem mengentheoretische und logische Methoden zum Einsatz (top-down-Ansatz). Eine typische Grundstruktur einer Ontologiekann in diesem Sinne aus

- einer Hierarchie wesentlicher ontologischer Kategorien,
- einem hierarchischen System elementarer Relationen
- einem Axiomensystem für die Operationalisierung von Ontologien und
- einer Basiseinteilung in Mengen (sets), Klassen (classes) und Urelemente

bestehen, wobei Mengen und Klassen meta-mathematische Superstrukturen sind, d. h. letztlich „Arbeitsmittel“ für die Beschreibung der Konstrukte in einer Ontologie (nach Heller & Herre 2003).

Bei der konkreteren Fragestellung nach der Erzeugung einer Ontologie für ein *bestimmtes Wissensgebiet* auf der Basis einer großen Textkollektion, wie sie das Text Mining zu lösen versucht, ist dagegen ein bottom-up- oder bootstrapping-Ansatz üblich. Der Metapher des Schürfens folgend wird versucht, aus der Textmenge ein Extrakt in Form ausgewählter Begriffe und Beziehungen zwischen Begriffen zu erzeugen, das das Ausgangsmaterial für eine neu zu definierende Ontologie bildet oder dazu dient, bestehende ontologische Strukturen zu erweitern und ergänzen.

Im ontology engineering hat man es sich zur Aufgabe gemacht, diesen Prozess zu beschreiben und zu systematisieren:

"Ontology engineering has as its goal effective support of ontology development throughout its life cycle – design, evaluation, maintenance, deployment, mapping, integration, sharing, and reuse." (Gruninger & Lee 2002: 40).

Im Mittelpunkt dieses zyklisches Prozesses steht dabei die Erarbeitung und Verfeinerung ontologischer Konzepte, die anschließend durch Gebrauch, Validierung und Entdeckung neuer Konzepte modifiziert werden kann. Dies zeigt die folgende Abbildung 1:

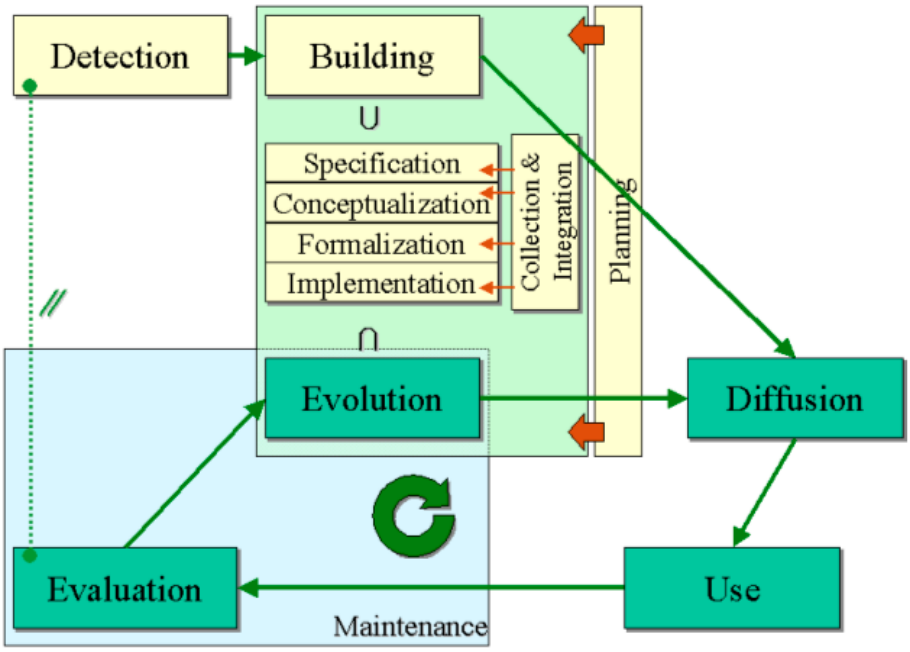


Abbildung 1: Der zyklische Prozess der Ontologieerzeugung (Gandon 2002: 26, Abbildung 7)

3 Text Mining-Verfahren für das Ontology Learning

Text Mining ist ein noch junges Arbeitsfeld, das aufbauend auf der Verfügbarkeit sehr großer digitaler Textmengen automatische Analyseverfahren entwickelt. Es speist sich sowohl aus dem etwas älteren Gebiet des Data Mining und den dort entwickelten Verfahren des maschinellen Lernens (KDD – Knowledge Discovery in Data, vgl. Fayyad et al. 1996), als auch aus Ansätzen der Künstlichen Intelligenz-Forschung und deren traditionellen Arbeitsgebiet des Knowledge Engineering, wie sie schon seit den 60er Jahren vor allem für das Gebiet der Expertensysteme entwickelt wurde. Ein dritter Ansatz sind linguistische Ansätze, die sprachliches Wissen in den Analyseprozess integrieren.[2]

Verfahren aus dem Bereich des maschinellen Lernens verwenden dabei für die Konzeptextraktion und –relationierung statistische Verfahren der Textanalyse auf. Zu ihnen gehören

- statistische Klassifikationsverfahren für Texte, bei denen Texte zunächst einem bestehenden Klassifikationssystem zugeordnet werden und diese Zu

[2] Einen Überblick zum Text Mining geben Mehler & Wolff 2005, Berry 2003 und Weiss et al. 2004.

ordnung dann auf dem Wege der Begriffsextraktion für die Erweiterung von Ontologien genutzt werden kann,

- Clusteranalysen, bei denen Texte ohne vorliegende Klassifikation zunächst gruppiert und dann einer Begriffsanalyse unterzogen werden,
- die Berechnung von statistisch signifikanten Assoziationen zwischen Begriffen (statistische Kollokationsanalyse, vgl. Heyer et al. 2001),
- vergleichende Corpusanalyse, die auf der Basis von Frequenzvergleichen relevante Konzepte eines Fachgebietes entdecken helfen (vgl. Böhm et al. 2002) sowie
- neuere Analyse-Verfahren wie Support Vector Machines (und Latent Semantic Indexing (Landauer & Dumais 1997), die durch Dimensionsreduktion die Begriffsvielfalt in großen Textkollektionen eingrenzen und damit wesentliche Konzepte einer Ontologie identifizieren helfen können.

Bei den linguistischen Analysemethoden spielen neben der einfachen Annotation eines Textes mit Kategorieninformation (Part-of-Speech Tagging, vgl. Brants 2000) auch Verfahren der syntaktischen und semantischen Analyse eine Rolle: Ein nahe liegender Ansatz ist es dabei, textpragmatisch typische syntaktische Strukturen der Begriffsdefinition zu beschreiben und auf dieser Basis eine syntaktische Analyse von Texten vorzunehmen, die in diesen Strukturen relevante ontologische Konzepte und Relationen entdeckt (vgl. Hearst 1998).

Seit größere ontologische Datenbanken wie WordNet (Fellbaum 1998) oder GermaNet (Kunze et al. 2004) elektronisch verfügbar geworden sind, können Verfahren für Ontologieaufbau und –erweiterung auch sie in den Analyseprozess integrieren. Soweit sich für neu identifizierte Konzepte eine Relation zu einem bereits in einer solchen Referenzontologie enthaltenen Konzept herstellen läßt, ist die Einordnung des neuen Konzeptes in eine Ontologie möglich; der Schwerpunkt liegt hier vor allem auf der Erweiterung bestehender Ontologien.

Allen Verfahren ist gemeinsam, dass sie beim derzeitigen Stand der Technik nicht in der Lage sind, vollautomatisch neue Ontologien auf dem Wege des bootstrapping zu generieren. Die Ontologieerstellung ist daher in der Regel auf die intellektuelle Überarbeitung bei Begriffsauswahl und Typisierung von Begriffsrelationen angewiesen. Hier kommt unterstützenden Werkzeugen, insbesondere Ontologieeditoren besondere Bedeutung zu. Werkzeuge wie Protégé erlauben die Überarbeitung ontologischer Konzepte in graphischen Editoren, die z. T. die Ontologiedarstellung auch durch geeignete Verfahren der Informationsvisualisierung unterstützen.

Gleichzeitig ist anzunehmen, dass nicht ein einzelnes Verfahren oder ein einzelner Algorithmus sich als optimale Lösung für einen allgemeinen Prozess der Ontologieerstellung herausstellen wird. Insofern liegt ein großes Potential in der Kombination von Verfahren und Wissensquellen, z. B. bei der Koppelung von statistischer Kollokationsanalyse und sprachlichem Wissen und der damit verbundenen Integration unterschiedlicher Verfahren in Software-Frameworks, die den

gesamten Ontologiebearbeitungsprozess (automatisch wie intellektuell) unterstützen (Bloehdorn et al. 2005, die ein solches Framework vorstellen).

4 Bewertung von Ontologien

Ein wesentliches Problem des ontology engineering besteht im Bereich der Bewertung und Evaluation von Ontologien: In der Regel ist eine optimale Vergleichsgrundlage für die Qualität einer Ontologie nicht gegeben. Gleichzeitig erscheinen traditionelle Effektivitätsmaße, wie sie im Information Retrieval verwendet werden (recall, precision, F-Maß), für die Übertragung auf die Bewertung von Ontologieverfahren weniger geeignet, da sie jeweils eine benutzerbezogene Bewertung der Ergebnisqualität voraussetzen, die für abstrakte Metadatenstrukturen, wie sie in Ontologien auftreten, nur schwer umsetzbar erscheint.
Qualitätskriterien für Ontologien müssen daher auf unterschiedlichen Ebenen angesiedelt sein: Mit Hinblick auf die modellierte Domäne spielen Vollständigkeit ebenso wie Korrektheit der Begriffe und der Relationierung (korrekte Zuweisung von Relationentypen wie Teil-Ganzes, Ober-/Unterbegriff etc.) eine Rolle. In struktureller Hinsicht zählen mittlerer Verzweigungsgrad der Begriffshierarchie, Hierarchietiefe und Ausgewogenheit des Begriffsstruktur eine Rolle – hier können bewährte Parameter zur Beschreibung von Datenstrukturen wie Bäumen und Graphen herangezogen werden (vgl. Botafogo et al. 1992). Schließlich kann eine Ontologie auch in quantitativer Hinsicht beschrieben und Bewertet werden: Gesamtzahl der Konzepte, mittlere Anzahl zugeordneter Konzeptvarianten (verwandte Begriffe, Teilsynonyme etc.) können hier ebenso betrachtet werden wie der Umfang der Ontologie im Verhältnis zur Begriffswelt der Ausgangstexte.

5 Fazit

Die Vielfalt der in den letzten zehn Jahren entwickelten Text Mining-Verfahren kann mittlerweile einen wesentlichen Beitrag zur Erstellung von Ontologien leisten. Zwar ist das Ziel einer vollautomatischen Ontologieerzeugung und –wartung noch weit entfern (und möglicherweise auch nicht zu erreichen), der Gesamtaufwand für das ontology engineering läßt sich aber durch Einsatz von Text Mining-Verfahren deutlich verringern.
Neben der Weiterentwicklung und Kombination von Text Mining-Algorithmen kommt dabei vor allem der Integration in Wissensmanagement-Prozesse und der Standardisierung eine große Rolle vor. Hier ist durch die Bemühungen des World Wide Web Consortium, auf der Basis der Metasprache XML (eXtensible Markup Language) Beschreibungsformate für das Semantic Web und Ontologien zu entwickeln (OWL - Web Ontology Language, vgl. McGuiness & van Harmelen 2003), für den deskriptiven Aspekt des ontology engineering eine wichtige Grundlage geschaffen

worden. In softwaretechnischer Hinsicht zeichnen sich mit der Initiative einer „Unstructured Information Management Architecture“ (UIMA), die einheitliches Interface für Textanalysen definiert, erste Bemühungen um Standardisierung ab.

Literatur

Aussenac-Gilles N.; Biebow B., Szulman S. (2000). „Revisiting Ontology Design: a Method Based on Corpus Analysis.“ In: Proc. EKAW'2000, Juan-les-Pins, 172-188

Berners-Lee, Tim (1988ff). „Semantic Web Roadmap.” World Wide Web Consortium, Design Issue Drafts, http://www.w3.org/DesignIssues/Semantic.html [letzter Zugriff September 2005].

Berry, M. W. (2003). Survey of text mining. New York: Springer.

Bloehdorn, St. et al. (2005). “An Ontology-based Framework for Text Mining.” In: LDV-Forum 20(1) (2005), 87-112

Boehm, K.; Heyer, G.; Quasthoff, U.; Wolff, Ch. (2002). “Topic Map Generation using Text Mining.” In: J.UCS (Journal of Universal Computer Science) 8(6) (2002), 623-633.

Botafogo, R. A.; Rivlin, E.; Shneiderman, B. (1992). “Structural analysis of hypertexts: Identifying hierarchies and useful metrics.” In: ACM Transactions on Information Systems, 10(2), 142–180.

Brants, T. (2000). “TnT – a statistical part-of-speech tagger.” In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000), Seattle, WA.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996a). “The kdd process for extracting useful knowledge from volumes of data.” In: Communications of the ACM, 39(11), 27–34.

Fellbaum, Ch. (Ed.) (1998). WordNet: An Electronic Lexical Database. MIT Press.

Gandon, F. (2002). Ontology Engineering: A Survey and a Return on Experience. Institut de Recherche en Informatique et Automatique (Inria), Research Report N° 4396, März 2002.

Gandon, Fabien (2002). Ontology Engineering: a survey and a return on experience. Institut de Recherche en Informatique et Automatique (Inria), Research Report, N° 4396, März 2002, http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/research/RR4396/, http://www.inria.fr/rrrt/rr-4396.html

Gruber, Tom R. (1993). “A translation approach to portable ontologies.” In: Knowledge Acquisition, 5(2) (1993), 199-220, 1993 [online: ftp://ftp.ksl.stanford.edu/pub/KSL_Reports/KSL-92-71.ps.gz, Zugriff September 2005]

Gruninger, M.; Lee, J.; Ontology Application and Design". In: Communications of the ACM 45(2) (2002), 39-41.

Hearst, M. (1998). Automated discovery of wordnet relations. In Fellbaum 1998.

Heller, Barbara; Herre, Heinrich (2003). Formal Ontology and Principles of GOL. Onto-Med Report Nr. 1, Juli 2003. Institute for Medical Informatics, Statistics and Epidemiology / Institute for Informatics. Leipzig University, http://www.onto-med.de/en/publications/scientific-reports/om-report-no1.pdf [letzter Zugriff September 2005]

Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, Th.; Wolff, Ch. (2001). "Learning Relations using Collocations." In: Maedche, A.; Staab, St.; Nedellec, C.; Hovy, E. (Hrsg.). Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, 19-24.

IBM (2005). Unstructured Information Management Architecture SDK. A Java SDK that supports the implementation, composition, and deployment of applications working with unstructured information. IBM Corp.; http://www.alphaworks.ibm.com/tech/uima [letzter Zugriff September 2005]

Joachims, T. (1998). "Text categorization with support vector machines: learning with many relevant features". In Proceedings of the Tenth European Conference on Machine Learning (ECML 1998), 137–142.

Kunze, C.; Lemnitzer, L.; Wagner, A. (edd.) (2004). Anwendungen des deutschen Wortnetzes in Theorie und Praxis. Beiträge des Germanet-Workshops, Tübingen, Oktober 2003. LDV-Forum (19) (1/2) (2004).

Landauer, T. K. & Dumais, S. T. (1997). "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological Review, 104(2), 211–240.

Maedche, A.; Staab, St. (2001). "Ontology Learning fort he Semantic Web." In: IEEE Intelligent Systems 16, 2 (2001), 72-79.

Maedche, A.; Staab, St.; „Ontology Learning for the Semantic Web"; IEEE Intelligent Systems 16, 2 (2001), 72-79.

McGuinness, Deborah L.; van Harmelen, Frank (2003). OWL Web Ontology Language Overview. World Wide Web Consortium Recommendation, Dezember 2003, http://www.w3.org/TR/owl-features/.

Mehler, Alexander; Wolff, Christian (2005). Einleitung: Perspoektiven und Positionen des Text Mining." In: LDV-Forum 20(1) (2005), 1-18 [Themenheft des LDV-Forum zum Text Mining].

Noy, N. F. ; Sintek, M. ; Decker, St. ; Crubézy, M. ; Fergerson, R. W. ; Musen, M. (2001). "Creating Semantic Web Contents with Protégé-2000" In: IEEE Intelligent Systems 16, 2 (2001), 60-71.

Weiss, S. M.; Indurkhya, N.; Zhang, T.; & Damerau, F. J. (2004). Text Mining. Methods for Analyzing Unstructured Information. New York: Springer.

Zhou, L.; Booker, Qu. E.; Zhang, Dongsong; “ROD – Toward Rapid Development for Underdeveloped Ontologies”; Proc. 35th Annual Hawaiian Int’l Conf. On System Sciences (HICSS-35 ’02), Vol. 4.

Entwicklung eines sehr flexiblen Internet-Basierten Datenbanksystems für die Umweltforschung

Reiner Krause, Jena

Abstract

In diesem Papier wird das zentrale Datenbanksystem für Forschungsdaten des Max-Planck-Instituts für Biogeochemie in Jena vorgestellt. Es wird für Daten und Metadaten des Instituts und von Projekten, an denen das Institut beteiligt ist, verwendet. Externe Datenquellen können verlinkt werden, sodass die Datenbank auch als Datenkatalog für dezentral gespeicherte Daten eingesetzt werden kann. Die Datenbank zeichnet sich durch eine hohe Flexibilität aus, die durch einen objektorientierten Ansatz für die Datenhaltung erreicht wird. Hierdurch können unterschiedlichste Daten in integrierter Weise in einem System dokumentiert werden, und das System einfach an neue Datentypen angepasst werden. In diesem Papier wird die Datenstruktur, Möglichkeiten der Versionierung und Vergabe von Benutzerrechten und die hauptsächlich genutzten Fenster der Benutzeroberfläche vorgestellt.

1. Einführung

Das Max-Planck-Institut für Biogeochemie untersucht langfristige Wechselwirkungen zwischen der Bio-, Atmo- und Geosphäre, sowie den Ozeanen, um zu einem besseren Verständnis von Klimaänderungen beizutragen. Das Institut ist an internationalen Projekten beteiligt, die mit unterschiedlichen Arten von Daten zu tun haben: Treibhausgas-Messungen, Parameter zu Pflanzen, Tieren, Mikroben und Böden, Satellitenbilder und Ergebnisse von Modellläufen.

Gegenwärtig werden Daten in unterschiedlichen Formaten in Dateien und Datenbanken auf verschiedenen Systemen gespeichert. Die Dokumentation von Daten folgt individuellen Systematiken. Die Entwicklung der Forschungsdatenbank, die hier beschrieben werden soll, wurde auf den Weg gebracht, um Daten auf einfache Weise allen interessierten Wissenschaftlern verfügbar zu machen und die Standardisierung der Dokumentation zu unterstützen, die auch denen nachvollziehbar sein soll, die nicht die Daten produziert haben. Das soll auch die Verpflichtung erfüllen, im Zweifelsfall die Nachprüfung von Schlüssen zu ermöglichen, die mit Hilfe der Daten gezogen wurden.

Weil wir an vielen Projekten beteiligt sind, in denen wir mit anderen Europäischen Instituten zusammenarbeiten, hält die Datenbank auch Daten und deren Beschreibungen vor, die von solchen Instituten kommen. Damit dient sie auch als Projektdatenbank.

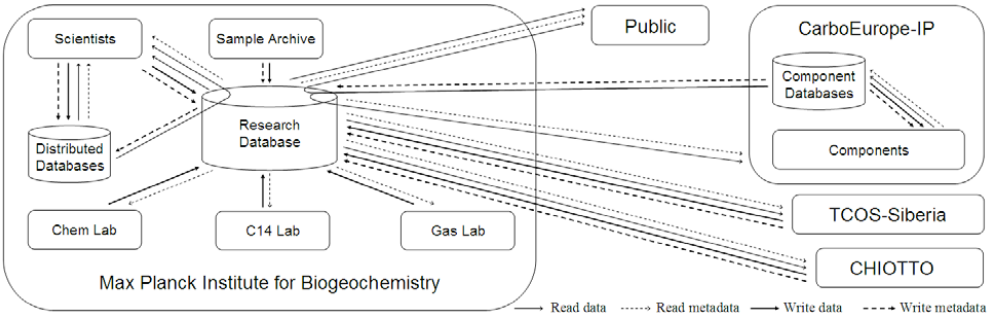


Abbildung 1: Verbindungen

Abbildung 1 zeigt unterschiedliche Verbindungen der Forschungsdatenbank, die geplant und teilweise bereits verwirklicht sind. Die Datenbank wird Daten von Wissenschaftlern und Serviceabteilungen des Instituts enthalten. Einige verteilte Datenbanken im Institut werden weiterhin Teile der Daten halten, die aber in der zentralen Forschungsdatenbank dokumentiert werden sollen. Labore erhalten Informationen zu Proben und Analyse-Anfragen von der Datenbank und liefern Analyse-Ergebnisse an die Datenbank zurück.
Das europäische Klimaforschungsprojekt CarboEurope-IP besteht aus Teilprojekten, den Komponenten. Unsere Forschungsdatenbank dient als zentraler Datenkatalog, der den Inhalt der Komponentendatenbanken beschreibt. Einige Daten werden jedoch auch für dieses Projekt direkt in der Forschungsdatenbank vorgehalten, so wie es für ein paar andere Projekte wie TCOS-Siberia und CHIOTTO der Fall ist. Die Öffentlichkeit kann freigegebene Inhalte lesen und dezentral gehaltene Daten über die Dokumentation in der zentralen Forschungsdatenbank finden.

2. Entwurfs-Prinzipien

2.1 Integration verschiedener Datentypen

Alle Daten und Metadaten sollen in einem System integriert verwaltet werden. Das wird dadurch erreicht, dass alle Informationen als Beschreibungen behandelt werden, unabhängig davon, ob es sich um gemessene Werte handelt, oder Beschreibungen von Forschungsgeländen, Projekten oder Messmethoden, und unabhängig von den Datenformaten und deren Struktur. Einerlei ob wir an Referenzen zu anderen Datenquellen im Internet oder Texte, Zahlen, Datums- und Zeitwerte oder komplette Dateien denken, ob wir es mit einzelnen Werten oder Datenserien zu tun haben – alles soll einheitlich behandelt werden. Darüber hinaus soll alles miteinander verknüpft werden können, soweit es Sinn macht. Alle anderen Prinzipien wie Versionierung und Zugriffsbeschränkungen können auf alle Arten von Beschreibungen angewandt werden.

2.2 Versionierung

Sobald aus wissenschaftlichen Daten Schlüsse gezogen wurden oder weitere Daten berechnet wurden, müssen sie in ihrer ursprünglichen Version archiviert werden, damit die wissenschaftlichen Ergebnisse nachvollziehbar sind. Auf der anderen Seite muss es immer möglich sein, Daten zu korrigieren. Das Datenbanksystem unterstützt die Eingabe neuer Versionen bei Beibehaltung der alten Versionen und ihre Kennzeichnung als ungültig. Benutzer sehen standardmäßig die aktuellen Versionen, können aber auch zu alten Versionen wechseln. Bei der Eingabe werden Daten normalerweise als vorläufig gekennzeichnet. Dieser Status erlaubt noch das Überschreiben und Löschen von Daten. Die Versionshaltung wird erst aktiviert, wenn Daten als geprüft gekennzeichnet werden. Dann muss bei Datenänderungen ein Grund oder die Art der Änderung dokumentiert werden. Diese Information kann zusammen mit dem Änderungsdatum und dem Benutzernamen von allen eingesehen werden, die auf die jeweiligen Versionen zugreifen.

2.3 Zugriffsrechte

Die Projekte, die in der Datenbank Daten speichern, werden aus öffentlichen Haushalten finanziert. Auch wenn daher die meisten Daten öffentlich zugänglich sein werden, ist der Zugriff zunächst beschränkt. Der Grund liegt darin, dass die Dateneigner als erste von Publikationen neuer Einsichten profitieren wollen und weil sie die Daten sorgfältig prüfen und mit Qualitätskennzeichen versehen wollen, bevor sie anderen den Zugriff gestatten, die nicht so vertraut mit den Herausforderungen und Unsicherheiten sind, mit denen es die jeweiligen Wissenschaftler zu tun haben.
Im Datenbanksystem kann der Lesezugriff zu Daten zunächst auf ausgewählte Benutzergruppen beschränkt werden, um sie erst später öffentlich zugänglich zu machen. Die Möglichkeit, Daten zu ändern und Rechte zu ändern, können ebenfalls Gruppen zugeordnet werden, die typischerweise an dem Projekt beteiligt sind. Benutzerrechte können für jede Angabe individuell eingestellt werden. Ein Weitergabe-Mechanismus macht die explizite Zuweisung für die meisten Fälle jedoch unnötig.
Nicht nur die interessierte Öffentlichkeit ist überall auf unserem Planeten bzw. in Europa zu finden, sondern auch unsere Projektpartner. Daher kann auch die Forschungsdatenbank über das Internet erreicht werden. Die Browser-Technologie macht es nebenbei einfach, andere Online-Informationen zu referenzieren, sofern sie nicht direkt in der Datenbank gespeichert werden sollen.

2.4 Verteilte Pflege der Datenbank

Soweit wie möglich, sollen Änderungen in der Forschungsdatenbank von denen ausgeführt werden, die die Daten produziert haben oder Projekte leiten. Hierdurch kann der Kreis der Anwender und Anwendungen wachsen ohne dass eine zentrale Verwaltung wesentlich mitwachsen muss. Außerdem dokumentieren hierdurch

diejenigen die Daten und legen Zugriffsrechte darauf fest, die mit den Daten vertraut sind.

2.5 Standardisierung

Trotz der Verschiedenartigkeit der Daten und beteiligten Gruppen soll die Datenbank ein einheitliches Datenarchiv sein, das es nachnutzenden Projekten erlaubt, Daten unterschiedlicher Projekte zusammenzustellen. Das erfordert ein wenig Standardisierung:

- Gleiche Arten von Beschreibungen sollten denselben Namen im ganzen Datenbanksystem haben. Das gilt für Metadatenfelder genauso wie für Variablen und Maßeinheiten.
- Derselbe Schlagwortkatalog sollte benutzt werden, soweit er anwendbar ist.
- Ein Minimum an Informationen sollte bereitgestellt werden, um eine zuverlässige Suche zu ermöglichen.

Alle diese Aspekte unterstützen eine einfache Orientierung in der wachsenden Datenbasis. Um von Bemühungen zu profitieren, die bereits in anderem Zusammenhang unternommen wurden, und um für zukünftigen Datenaustausch mit anderen Systemen vorbereitet zu sein, die zur Zeit noch nicht berücksichtigt werden, orientieren wir uns an Standards wie ISO 19115 und Dublin Core. Aber daneben haben wir auch mit Bedürfnissen zu tun, die spezifisch für die von uns unterstützten Projekte sind. Dies hat zur Folge, dass wir Felder und Schlagwörter vorsehen müssen, die nicht in allgemeinen Standards zu finden sind.

2.6 Projektspezifische Anforderungen

Dadurch dass unterschiedliche Gruppen unser Datenbanksystem benutzen, kommen auch verschiedene Anforderungen auf das System zu. Während zum Beispiel die Mitglieder eines Projekts alle Proben sehen wollen, die einem Forschungsgelände entnommen wurden, ist ein anders Projekt überhaupt nicht an diesen Proben interessiert, wenn sie die Beschreibung des Forschungsgeländes vor sich haben. Die beste Lösung für individuelle Anforderungen ist nicht, unterschiedliche Arten der Namensgebung und Beschreibung zu erlauben, sondern den Benutzern zu ermöglichen, ihre eigenen Profile zu definieren. Ein Profil zeigt den gesamten Datenbankinhalt in einem gemeinsamen Stil, selbst wenn die Informationen von verschiedenen Projekten stammen. Benutzer können zwischen unterschiedlichen Profilen wechseln und hierdurch das Datenbanksystem ihrem gegenwärtigen Bedarf anpassen.

2.7 Datenstruktur

Der elementare Datensatz in dieser Datenbank ist sehr einfach: Er repräsentiert ein Objekt einer Klasse, wie z.B. ein Projekt, eine Luftprobe, eine Methodenanwendung, eine Variable oder einen Namen. Werte-Datensätze, wie eine CO2-Konzentration oder ein Projektname, haben einen Wert aber geben nicht an, welche Probe oder

welches Projekt sie spezifizieren. Datensätze ohne Wert (z.B. ein Forschungsgelände) repräsentieren lediglich ein Objekt einer Klasse ohne weitere Information (nicht einmal der Name des Forschungsgeländes ist Teil dieses Datensatzes). Ein elementarer Datensatz hat also keinen echten Informationsgehalt. Dieser kommt erst durch Verknüpfungen von Datensätzen zustande.
Verknüpfungen verbinden zwei Datensätze miteinander, wie ein Forschungsgelände mit seinem Namen oder mit einer dort entnommenen Luftprobe, und die Luftprobe mit ihrer CO2-Konzentration. Verknüpfungen können unterschiedliche Bedeutungen haben: Eine Person kann mit einem Projekt als Leiter oder Kontaktperson verbunden sein, ein Platz kann ein Flughafen oder der Probenentnahmeort eines Fluges sein. Ergebnisse von Aktivitäten werden durch Dreifach-Verknüpfungen zwischen Aktivität, Quelle und Ergebnis beschrieben (Siehe die Verknüpfung 7 in Abbildung 2, die die Analyse mit der Luftprobe und der CO2-Konzentration verbindet.

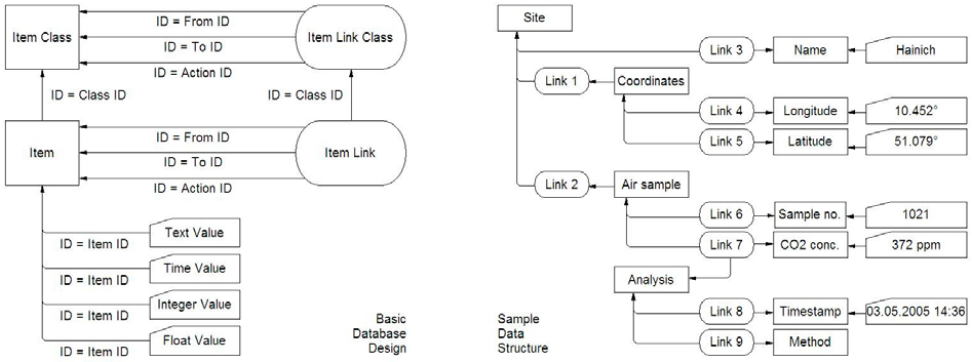


Abbildung 2: Datenstruktur

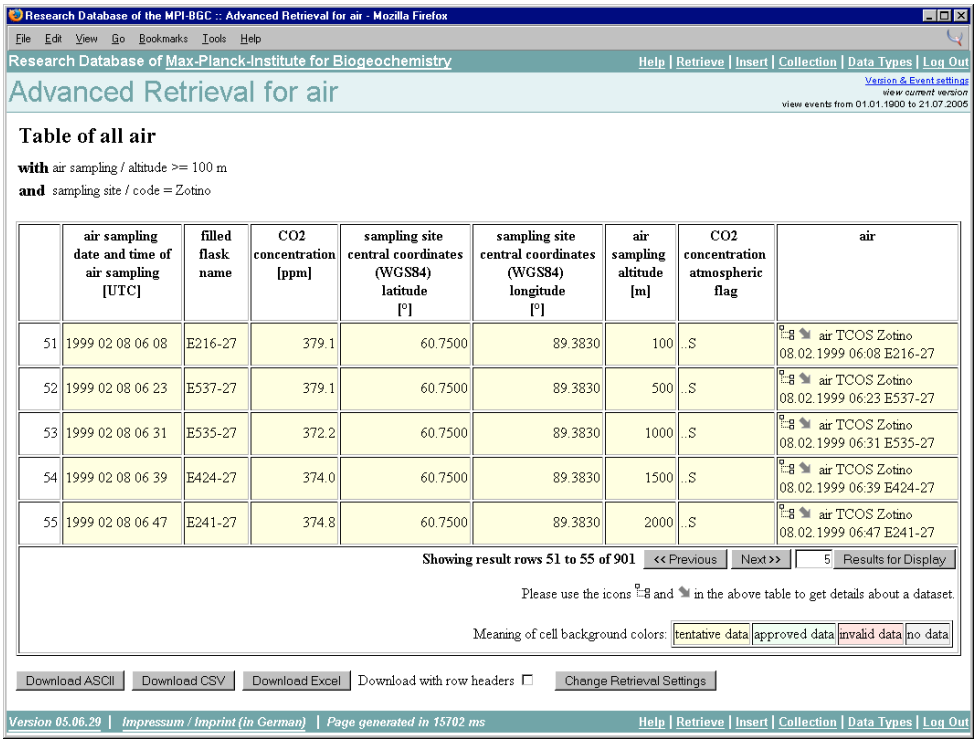
Der Hauptgrund für diese zergliederte Struktur ist ihre Flexibilität. Man braucht keine Tabellen zu ändern oder neu anzulegen, wenn neue Beschreibungsarten benutzt werden sollen. Fast alles kann durch ein Netz von Objekten beschrieben werden. Neue Datentypen können ohne Software-Entwicklung gespeichert werden. Programmierung ist nur nötig, wenn weitere Möglichkeiten für neue oder alte Beschreibungsarten benötigt werden. Auf der anderen Seite erfordert dieses Datenbank-Design sorgfältige Erwägungen bei der Software-Entwicklung, was das Antwortverhalten des Systems betrifft.
Um die Benutzung von Standards sicherzustellen, müssen alle Klassen definiert werden, bevor entsprechende Daten eingegeben werden können. Die Definitionsebene ist ein eigener Teil im Datenbank-Modell, dessen zwei zentrale Tabellen (Item Class und Link Class) in der vereinfachenden Abbildung 2 (links) zu finden sind. Neben der Beschreibung, welche Objekte miteinander verknüpft werden können, definiert die Definitionsebene auch das Erscheinungsbild der Daten und die Felder in der Eingabemaske pro Klasse. Neue Beschreibungsfelder und Variablen werden

lediglich in der Definitionsebene eingetragen und können sofort verwendet werden, da alle Formulare sich nach diesen Definitionen richten.
Die Definitionsebene kann wie die Daten über Internet-Seiten editiert werden. Weil diese Änderungen koordiniert vorgenommen werden sollen, hat jedoch nur eine kleine Benutzergruppe das Recht dazu

3. Benutzerschnittstelle

3. 1 Daten lesen

Es gibt vier Arten von Internetseiten zum Ansehen von Daten. Das Suchfenster (Retrieval Form) ermöglicht es Daten aufzulisten und herunterzuladen, die eingegebenen Suchkriterien genügen. Es gibt eine sehr einfache Variante, aber auch eine komplexere, in der Suchkriterien, Spalten für das Suchergebnis und Formate für das Herunterladen sehr flexibel eingestellt werden können. Solche Einstellungen können für die wiederholte Nutzung durch Gruppen von Wissenschaftlern gespeichert werden.



	air sampling date and time of air sampling [UTC]	filled flask name	CO2 concentration [ppm]	sampling site central coordinates (WGS84) latitude [°]	sampling site central coordinates (WGS84) longitude [°]	air sampling altitude [m]	CO2 concentration atmospheric flag	air
51	1999 02 08 06 08	E216-27	379.1	60.7500	89.3830	100	..S	air TCOS Zotino 08.02.1999 06:08 E216-27
52	1999 02 08 06 23	E537-27	379.1	60.7500	89.3830	500	..S	air TCOS Zotino 08.02.1999 06:23 E537-27
53	1999 02 08 06 31	E535-27	372.2	60.7500	89.3830	1000	..S	air TCOS Zotino 08.02.1999 06:31 E535-27
54	1999 02 08 06 39	E424-27	374.0	60.7500	89.3830	1500	..S	air TCOS Zotino 08.02.1999 06:39 E424-27
55	1999 02 08 06 47	E241-27	374.8	60.7500	89.3830	2000	..S	air TCOS Zotino 08.02.1999 06:47 E241-27

Abbildung 3: Suchfenster mit Suchergebnis

Das Datensatzfenster (Display Form) zeigt einen elementaren Datensatz und alle direkt verknüpften elementaren Datensätze und entspricht hiermit am ehesten der Vorstellung von einem Fenster zur Darstellung von Datensätzen im herkömmlichen Sinne. Es bietet Knöpfe und Links für die Dateneingabe und zum Pflegen der Benutzerrechte, sofern man den Editiermodus eingestellt hat.

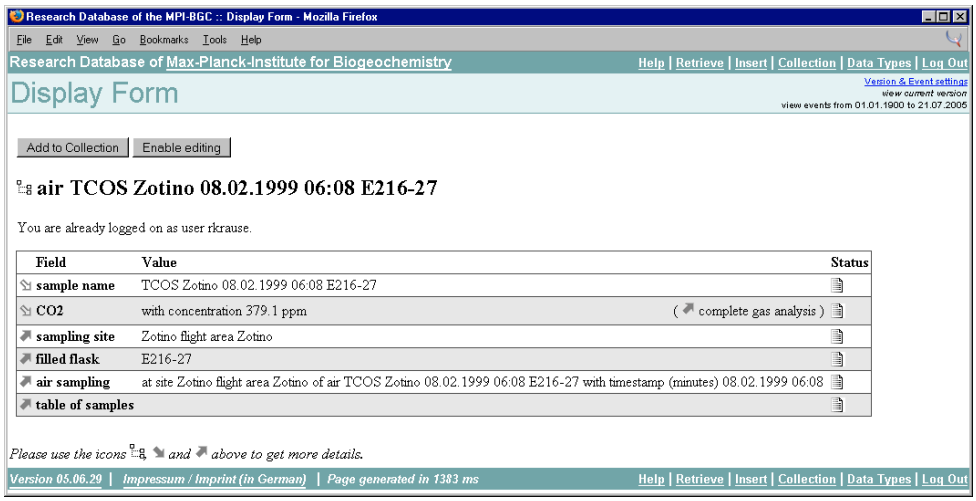


Abbildung 4: Datensatzfenster

Das Hierarchiefenster (Tree View) zeigt zusätzlich indirekt verknüpfte elementare Datensätze in rekursiver Manier, wobei einstellbar ist, welchen Arten von Verknüpfungen gefolgt werden soll.

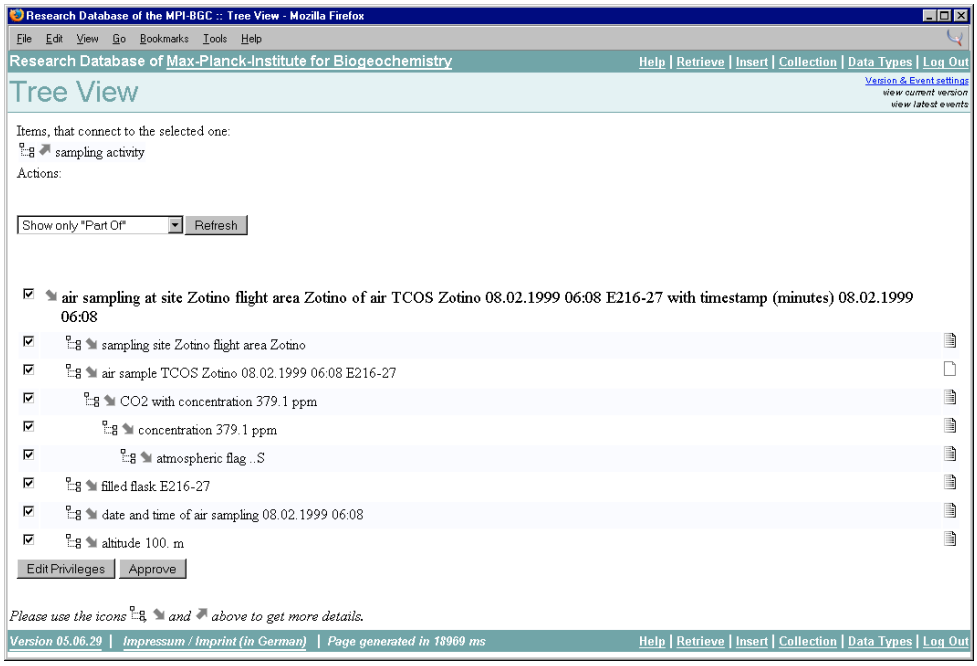


Abbildung 5: Hierarchiefenster

Während die vorher genannten Fenster fast alle Objektarten darstellen können, benötigen lediglich Zeitserien ein spezielles Serienfenster (Series Display), das es ermöglicht, solche Daten für einen gewählten Zeitbereich anzuzeigen und herunterzuladen mit der Option, mehrere Serien unter Verwendung einer gemeinsamen Zeitspalte zu mischen

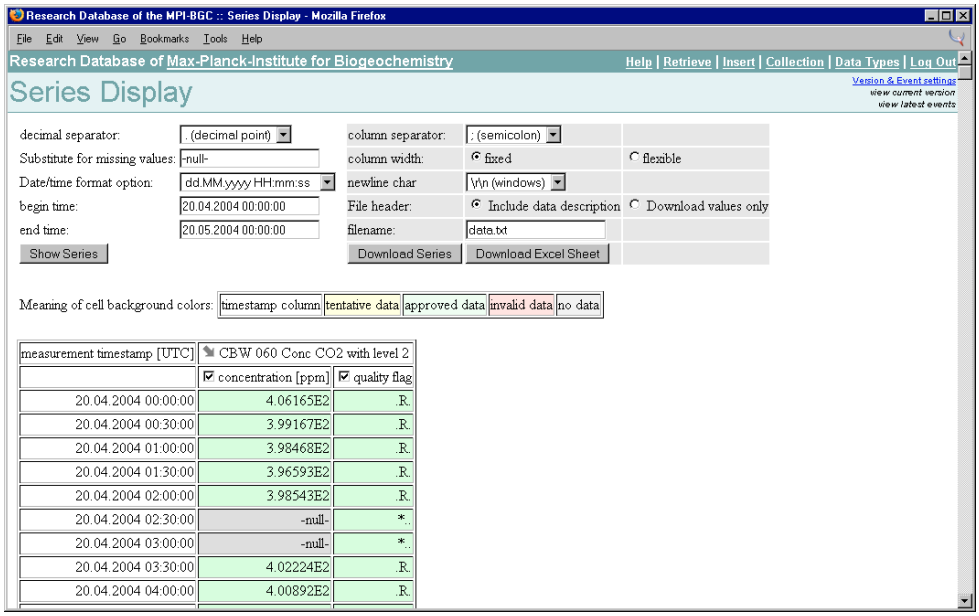


Abbildung 6: Serienfenster

3.2 Daten eingeben

Das Eingabefenster (Input Form) wird für das Hinzufügen und Ändern von Daten genutzt. Es stellt Felder für die gewählte Klasse und – abhängig von Einstellungen in der Definitionsebene – für direkt und indirekt verknüpfte elementare Datensätze zur Verfügung. Felder können je nach Einstellung Auswahllisten und Texteingabefelder sein.

Während eine manuelle Eingabe für kleine Änderungen sinnvoll ist, ist das Hochladen von größeren Datenmengen effektiver, besonders wenn dieselben Dateiformate wiederholt genutzt werden. Mit Hochladen ist in diesem Zusammenhang das Analysieren und Zerlegen des Dateiinhalts in elementare Datensätze gemeint, die dann in der Datenbank angelegt werden.

Beschreibungen, wie Dateien zu analysieren sind und welche Datensätze angelegt werden sollen, können für die wiederholte Anwendung gespeichert werden. Es können sehr unterschiedliche Arten von Dateien verarbeitet werden.

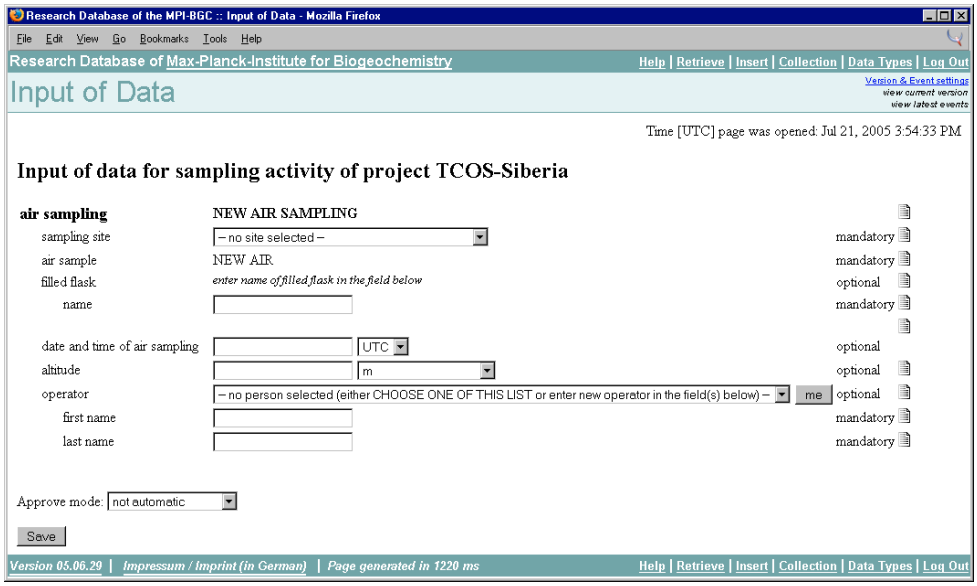


Abbildung 7: Eingabefenster

Ein zweiter Typ des Hochladens von Dateien speichert komplette Dateien im Datenbanksystem für den Zweck ihrer Dokumentation und des späteren Herunterladens. Für diese Art der Archivierung wird allerdings keine Suche angeboten, die auf Dateiinhalten basieren würde.

4. Rück- und Ausblick

Bevor mit der Entwicklung eines eigenen Datenbanksystems begonnen wurde, hatten wir nach brauchbaren vorhandenen Systemen Ausschau gehalten. Wir fanden zwar Systeme, die ähnliche Anforderungen erfüllten, aber es fehlten immer wesentliche Teile. So gibt es reine Metadaten-Systeme, die nicht die Datenspeicherung selbst integrieren, Systeme mit gleicher Flexibiltät wie unseres, aber ohne Internetanbindung und ein System mit fast demselben Ansatz, aber ohne Versionierung.
Wir standen natürlich vor der Frage, ob man ein eigenes System entwickeln oder aber ein vorhandenes anpassen sollte. Letztlich haben wir uns für eine Eigenentwicklung entschieden, um den Bedürfnissen des Instituts möglichst gerecht werden zu können. Bei dieser Entscheidung spielten aber auch andere Gesichtspunkte eine Rolle, wie die mangelnde Dokumentation mancher vorhandener Systeme.

Zu diesem Zeitpunkt wage ich noch kein Urteil darüber, ob das wirklich der beste Weg war. Auf jeden Fall war die Entwicklung aber weitaus aufwändiger als zu Anfang angenommen.
Anwendererfahrungen liegen zwar schon vor, aber zum Zeitpunkt der Erstellung dieses Textes (Mitte Juli 2005) nur in sehr begrenztem Umfang. Auf der Konferenz „Knowledge eXtended“ werde ich sicherlich auf mehr Erfahrungen zurückgreifen können, da gerade jetzt der Umfang der Anwendungen zunimmt.

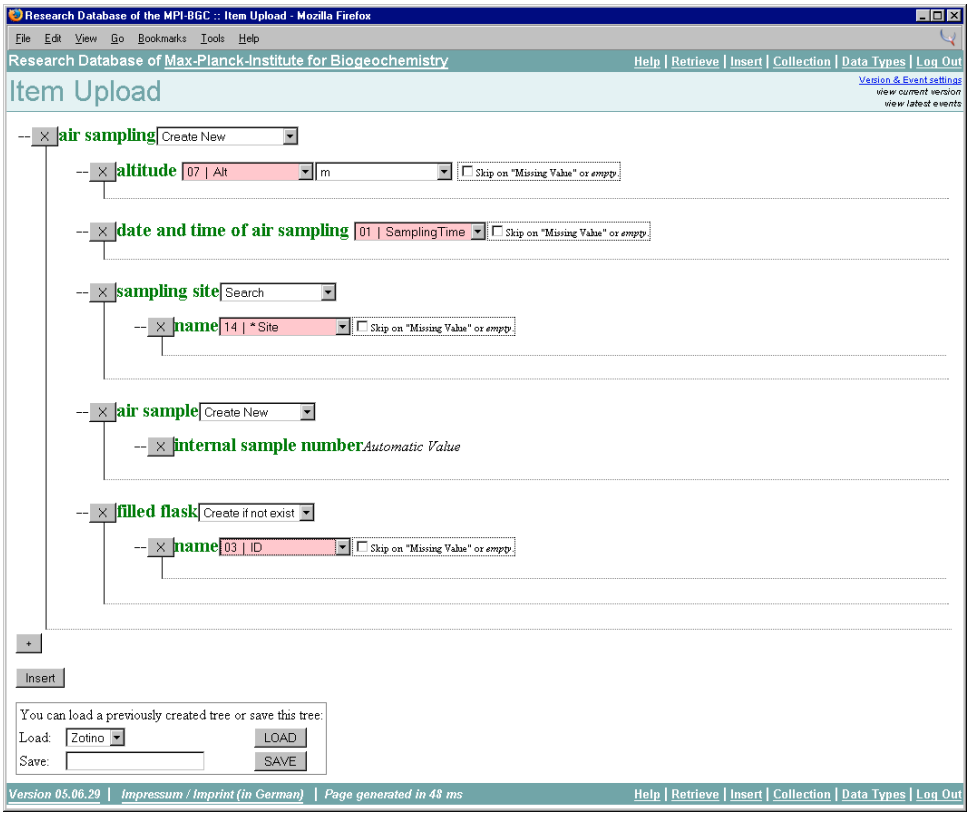


Abbildung 8: Hochladefenster (Feldnamen der Datenbank grün, der Datei rosa)

Von der Tendenz her wird das System gut angenommen, es kommen aber immer wieder Wünsche nach einer Vereinfachung der Eingabe und dem Abfragen von Daten. Das geht z.B. in die Richtung der oben angesprochenen Profile oder das Kopieren und anschließende Ändern vorher eingegebener Daten. Außerdem ist das System bereits jetzt zu langsam; eine gründliche Prüfung, wie die Geschwindigkeit gesteigert werden kann, ist daher notwendig. Bei vermehrter Nutzung der Datenbank wird die weitere Entwicklung immer stärker von Benutzerwünschen geprägt werden. Insofern bin ich gespannt, wie der weitere Weg aussehen wird.

Bibliometric Mining: Mehrwert aus Analyse und Retrieval

H. Peter Ohly, Bonn

Abstract

Bibliometrie ist herkömmlich als deskriptive statistische Analyse von wissenschaftlichen Strukturen und Prozessen anzusehen. Die Daten hierzu resultieren aus Informations- und Verwaltungshandeln. Mit der Anforderung, Qualitätsurteile zu fällen oder bisher unerkannte Strukturen und Informationen zu finden, kommt der Bibliometrie in Kombination mit komplexen Information-Retrieval-Verfahren die Rolle zu, explorativ und entscheidungsunterstützend zu wirken. Insofern hat sie inzwischen wichtige Merkmale des Data Mining, wobei die Analyse von Texten und Internetmaterial als zusätzliche Herausforderung anzusehen ist. Im Sinne eines evaluativen Forschungsansatzes sind bei beratender Bibliometrie auch Inferenzverfahren und Navigationsverfahren hinzuzuziehen.

Einleitung

Bibliometrie ist eine relativ junge Forschungsrichtung[1], die wesentlich von der gründlichen Arbeit der wissenschaftlichen Bibliotheken profitierte und inzwischen durch die weltweiten Zugänglichkeit von Datenbanken nicht mehr wegzudenken ist. Bibliometrie misst indirekt, indem sie Schlussfolgerungen aus formalen Zählungen von Literaturaufkommen zieht, ohne nähere Information über die Inhalte oder Entstehungsbedingungen einholen zu müssen. Von daher hat sie eine gewisse Faszination, denn sie ermöglicht, 'objektiv' den Wissensentstehungsprozess zu vergleichen, zu bewerten und zu lenken. Die damit untersuchten Wissenschaftler und Wissenschaften sind aber insofern auch ausgeklammert, als sie weder bei der Erhebung der Daten noch ihrer Interpretation gefragt werden müssen. Unter dem Aspekt zunehmend größerer Datenmengen und neuer Informationstechnologien soll im Folgenden den Fragen nachgegangen werden, ob nicht eine Verwandtschaft zu wissensbasierten Technologien, wie dem Data Mining, besteht und ob diese nicht wiederum zu erweiterten Möglichkeiten und Vorsichtsmaßnahmen anregt.

Die Metapher des Mining

Herkömmliche Analyse, so auch die bibliometrische Analyse, geht nach einem festen Forschungsplan vor. Zunächst erfolgt eine theoretische Aufarbeitung zu einem -

[1] Als einer der Begründer gilt de Solla Price etwa mit 'Little Science, Big Science' (1963).

meist aus praktischen Problemen und Interessen heraus eingegrenzten - Gebiet. Hieraus leitet sich eine konkrete Hypothese ab, die dann nach geeignetem Erhebungs- und Analyse-Design bestätigt oder widerlegt werden kann. Abschließend erfolgt eine Interpretation der Analyseergebnisse innerhalb des theoretischen und praktischen Kontextes.
Anders ist es beim "Mining", was sich aus dem Schürfen nach Metall, vorzugsweise Edelmetallen, herleitet. Hier gibt es eine nur vage Vorstellung darüber, welche Information in einer Datenbasis zu erwarten ist. Erst das systematische und gleichzeitig zielorientierte Suchen führt zu Erkenntnissen, die von selbst so nicht offen zu Tage getreten wären. Mining ist damit als ein exploratives Vorgehen anzusehen, bei dem heuristische Verfahren Anwendung finden mit dem Ziel, aus einer Menge von ungeordneten Informationen, einige übergeordnete Erkenntnisse oder besonders wertvolle Details herauszufiltern[2]. Statistische und logische Kriterien finden hier nur insofern Anwendung, als sie dazu dienen, den Suchraum immer stärker einzugrenzen und sicherzustellen, dass man sich noch auf einem erfolgversprechenden Weg befindet.

Knowledge Discovery

Meist werden aber noch weitreichendere Erwartungen an das Mining gerichtet. So wird 'Data Mining' – auch mit 'Knowledge Discovery in Databases' (KDD) gleichgesetzt - in den KD Nuggets wie folgt definiert: "Data Mining is the process of finding new and potentially useful knowledge from data"[3]. Oder nach Frawley et al. ist Data Mining: "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (1992)[4], wobei dies sich auf große Datenmengen oder Datenbanken bezieht[5]. Die gefundene Information ist also einerseits empirisch begründet, andererseits aber neu und wichtig. Dies ist zwar letztlich für jede wissenschaftliche (neue) Erkenntnis zu fordern, hier aber ist der Ansatz von vorneherein darauf gerichtet, die Spreu vom Weizen zu trennen und nur außergewöhnliche Ergebnisse zu finden. Dieses Verfahren ist damit wissenschaftlich der Induktion, also der Generierung von Fakten und Hypothesen zuzurechnen. Knowledge Discovery hat aber neben der Anwendung statistischer Verfahren auf Datenbanken noch

[2] Beispiel hierfür konnen die selbstorganisierenden Karten von Kohonen sein (1996).
[3] < http://www.kdnuggets.com/faq/data-mining.html >
[4] Weitere Definitionen finden sich in < http://www.orbiscope.net/en/data_mining.html >.
[5] Auch wenn oft synonym verwendet, muss genau genommen zwischen Daten, Informationen und Wissen unterschieden werden. Daten sind physische Zeichen, die aber interpretierbar sind, Informationen sind nützlich bei einer Problemlösung und Wissen ist genereller in ein Wissensgebäude integrierbar. Meist dürfte es sich im Kontext von Data Mining um Informationen handeln, die eben aus Daten gewonnen werden. Sicher sind solche Informationen wertvoller, wenn sie generalisierbar sind.

weitere Aspekte: Computerbasierte Entscheidungsunterstützung, maschinelle Lernverfahren und Visualisierung (Schmidt-Thieme 2002). KDD zielt also auf problementscheidende Suchergebnisse und ist bei der Auswertung von großen Datenmengen auf maschinelle Verbesserungen der Suchstrategie und anschauliche Darstellung angewiesen. Die Ausweitung auf möglichst viele relevante Basisdaten ergibt sich aus dem Ziel der möglichst treffenden Entscheidung, andererseits wird gerade damit eine Vielschichtigkeit der Ausgangslage erzeugt, die zu unzulässigen Verallgemeinerungen führen mag.

Text Mining

Text (Data) Mining versteht sich als Erweiterung des Data Mining, indem es sich auf alphanumerische Daten bezieht und damit die semantische Komponente explizit anspricht. Als Anwendungen werden genannt: Fragenbeantwortung, Informationsextraktion, Thesauruserstellung, Inhaltszusammenfassungen, Kategorisierungen und Textgruppierungen (Hearst 1999). Entweder handelt es sich um die Suche nach interessanten (oft synthetisch gewonnenen) Inhalten, den 'Nuggets', oder um globale Übersichten, die beide eine automatische Sprachverarbeitung erfordern. Durch die nicht-numerische Form der Daten sind graphische Darstellungen schwieriger zu erzeugen, da quantitative Merkmale nur abgeleitet werden können. Wie beim Data Mining besteht die Zielsetzung in mehr als nur optimalem Retrieval von vorhandenen (Text-)Daten. Vielmehr sind verstreute Informationen im Kontext zu bewerten und zusammenzuführen.
Unter Betonung der praktischen Verwertbarkeit des in Datenbanken Gefundenen kann auch von Information Mining gesprochen werden, dann aber ggf. wieder unter Einbezug von quantitativen Daten. Kruse und Borgelt führen aus: 'Information mining is the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in heterogeneous information sources' (2003).
Die Verdichtung und Selektion von Information aus dem WWW beinhaltet Prinzipien des Text und Information Mining. Web Mining geht aber noch darüber hinaus, da hier sehr viele unterschiedliche Datentypen vorkommen, die Datenbasis enorm groß ist und Informationen über Seitenverlinkung und Zugriffe einbezogen werden können (Hearst 1997; Ghani et al. 2000). Web Mining kann damit unter verschiedenen Gesichtspunkten betrieben werden: Analyse des Inhaltes (einer Web-Seite oder eines Suchergebnisses), der Strukturen und der generellen oder problembezogen Nutzung (Schmidt-Thieme 2002).

Bibliometrie als indirekte Erhebung

Die Bibliometrie als Vermessung von wissenschaftlichen Texten, wobei Szientometrie und Informetrie[6] eingeschlossen sein sollen, ist als wissenschaftliche Disziplin der explorativen Beschreibung und Testung von Hypothesen zuzurechnen. Es werden genau definierte Textdaten quantifiziert (z.B. durch Zählung von Wortformen) und miteinander verglichen, korreliert oder in Schaubildern oder Koeffizienten verdichtet. Es sollte ein Bezug zu dem gegeben sein, was statistisch betrachtet als "normal" anzusehen ist. Das Besondere der Bibliometrie liegt in der Datenherkunft, da sie sich auf Informationen bezieht, die in anderem Kontext generiert wurden. Die Produktion von Büchern oder Artikeln geschieht mit der Absicht, einen bestimmten Inhalt anderen Wissenschaftlern zu vermitteln. Ihre Interpretation als Gradmesser für wissenschaftliche Produktivität, geistige Verwandtschaft oder Akzeptanz der Nutzer ist nicht zwingend gegeben. Grundsätzlich ist von einer Diskrepanz zwischen Datenintention und Analyseintention auszugehen, wodurch die Validität der Interpretation in Frage gestellt wird. Es gibt eine Vielzahl von Verzerrungen durch die ursprünglich andere Verwendungsbestimmung, die auch bei Gesamtbetrachtungen sich nur bedingt aufheben[7]. Die Resultate bleiben lediglich Indikatoren, die je nach Granularität vor dem Hintergrund der Entstehung der Daten problematisiert werden müssen.

Bibliometric Mining

Mit der Verwendung als Verfahren zur Bewertung und Steuerung von Wissenschaft und Informationsflüssen ist die Bibliometrie von der Tendenz her eher als ein Mining- und Discovery-Verfahren anzusehen. Auch sie arbeitet mit großen Datenmengen, die statistisch-explorativ erschlossen werden, um Übersichten oder Highlights zu ausgewählten Wissenschaftsgebieten zu bekommen. Thematische Karten und Autorennetzwerke entsprechen dem Ziel einer verdichteten Darstellung, die als Unterstützung bei Entscheidungen im Wissenschaftsmanagement dienen können. Ranking-, Co-Citation- und Impactfaktor-gestütztes Vorgehen erlauben eine indirekte Bewertung der Quellen und die Einschränkung des Suchraums. Mit dem Ranking von Ergebnissen, der Listung von 'ähnlichen' Quellen oder der Erstellung von semantischen Karten zu gezielten Suchen kann eine Unterstützung auch im

[6] In einem engeren Verständnis ist Bibliometrie die Analyse von Veröffentlichungen, Szientometrie die von Wissenschaftsvorgängen und Info(r)metrie die von Informationsvorgängen, was oft zusammenfällt. Üblicherweise werden Datenbanken und Bibliotheksbestände analysiert, was inzwischen nahtlos in die Cybermetrie (Internetometrie, Webometrie) überleitet.

[7] Relativierende Betrachtungen nehmen u.a. Kraft (1998), Fröhlich (1999) und Stock (2000) vor.

Retrieval gegeben werden. Ein prominenter Vertreter ist die Google-Search, wo von vorneherein der Suchraum auf häufig verlinkte Seiten im Internet beschränkt wird[8]. Andere Suchverfahren gewichten extrahierte Wörter nach ihrem Vorkommen in bestimmten Seitenabschnitten und im Verhältnis zum Gesamtaufkommen im betrachteten Text oder in der Datenbank[9].

Synthetisch-deduktive[10] Verfahren, wie Text-Summarizing oder -Clustering über mehrere Dokumente oder Autorengewichtungen gemäß ihrem Kooperationskontext finden dagegen eher selten im Retrieval oder in bibliometrischen Analysen Anwendung. Unter der Voraussetzung von theoretischer Herleitung und Offenlegung der Operationalisierung wäre allerdings bei logischer Verarbeitung nicht nur ein bibliometrisches Mining[11], sondern auch eine mehrwert-orientierte Analyse denkbar: Analyseeinheit ist nicht mehr das einzelne Dokument sondern eine Information in ihrem Kontext, so wie bei Netzwerkanalysen Akteure in Bezug auf ihren Kontext bewertet und verglichen werden[12]. Hilfen bieten hierfür die bibliographischen Beschreibungsschemata der Datenbankanbieter und im WWW die Metainformationen und das Semantic Web[13]. Als Ergänzung der Suche nach hervorgehobenen Einzelfällen und der Darstellung grober Strukturen wäre auch die Herausarbeitung von dynamischen Entwicklungen zu wünschen, worin generelle Trends sowie markante Einzelentwicklungen aufgezeigt werden. Cited Half-Life und Immediacy Index[14] sind hierbei nur eine sehr reduzierte Information zu einer dynamischen Betrachtung.

[8] Der sog. Page-Rank von Google, geht davon aus, dass eine WWW-Seite umso wichtiger ist, je mehr Seiten auf sie verlinken. (Brin/Page 1998)

[9] siehe etwa: Rosenbaum 1997

[10] Gemeint sind Verfahren, die aus Einzelfällen generalisieren (etwa 'lernen') wie auch aus bekanntem Wissen Schlussfolgerungen für einzelne Fälle ziehen.

[11] 'Bibliomining' wird dagegen enger verwendet im Sinne von Data Mining für Bibliotheken (Nicholson 2003).

[12] Beispiel hierfür kann etwa TétraFusion sein, wo eine Vielzahl von Verfahren und Wissen verwendet werden, um zu verdichteten Darstellungen von Web- und konventionellen Dokumenten zu kommen (Crimmins 1999) oder das FINGRID-Projekt, wo Simulation mit Informationsextraktion kombiniert wird, um Markttrends aus News-Reports zu gewinnen (Ahmad et al. 2005).

[13] Siehe etwa: Studer 2003 und Chiravegna/Chapman 2005.

[14] Immediacy Index(Unmittelbarkeitsfaktor) ist ein Maß für die Zitierschnelligkeit: Wieviele Artikel einer Zeitschrift innerhalb des Erscheinungsjahres zitiert worden sind. Cited Half-Life (Zitierte Halbwertszeit) ist ein Maß für den zeitlichen Wertverlust der Inhalte: Publikationsjahre, nach denen die Zitierungsrate auf die Hälfte ihres anfänglichen Wertes abgesunken ist.
Siehe dazu: JCR Glossary < http://jcrweb.com/www/help/hjcrgls2.htm >

Resumee

Bibliometrie und ihre Anwendung im Information Retrieval sind bereits von ihrem Anspruch her eine Art Information Mining, welches Data Mining und Text Mining umfasst. Dabei hat die Bibliometrie in Rechnung zu ziehen, dass sie zunehmend mehr und größere Datenbestände zur Verfügung hat und dies bei Ihren Aussagen – etwa durch Duplizitätskontrollen und Gewichtungen – berücksichtigen muss. Besondere Beachtung finden sollte auch der Einbezug uneinheitlicher, unstandardisierter und nicht-konventioneller Datenformen. Gerade bei Internetanalysen muss eine Heterogenitätsbehandlung erfolgen. Im Falle einer qualitativen Aussage ist unbedingt der jeweilige Kontext in Rechnung zu ziehen. Gerade bei wertenden Schlussfolgerungen ist die quantitative Bedeutung des Untersuchten genauer zu hinterfragen. Werden Übersichten gebraucht, so sind Visualisierungen der Ergebnisse und Indexbildungen oft geeigneter als Zahlenreihen, um hieraus relative Beurteilungen treffen zu können. Um Analyse- oder Suchmotivation und Datenwahl gegeneinander abzustimmen, ist es günstig, einen benutzergesteuerten Datenselektionszugang zu haben, der die Granularität und Betrachtungsdimension gestuft anpassen kann.

Literatur

Ahmad, Khurshid; Gillam, Lee; Cheng, David: Textual and Quantitative Analysis: Towards a new, e-mediated Social Science. Paper given at the First International Conference on e-Social Science, 22-24 June 2005, Manchester, UK. < http://www.ncess.ac.uk/events/conference/programme/papers /ncess2005_paper_K_Ahmad.pdf > (Stand 20.9.2005)

Brin, Sergey; Page, Lawrence: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Computer Networks and ISDN Systems30(1),1998, pp. 107-117. < http://www-db.stanford.edu/pub/papers/google.pdf > (Stand 20.9.2005)

Ciravegna, Fabio; Chapman, Sam: Mining the Semantic Web: Requirements for Machine Learning. Dagstuhl workshop on Learning for the Semantic Web , 13-18 February 2005, Dagstuhl, Germany. < http://www.smi.ucd.ie/Dagstuhl-MLSW/proceedings/ciravegna-chapman.pdf > (Stand 20.9.2005)

Crimmins, Francis; Smeaton, Alan F.; Dkaki, Taoufiq; Mothe, Josiane: TétraFusion: Information Discovery on the Internet. In: IEEE Intelligent Systems, vol. 14 no. 6, 1999, pp. 55-62, November/December

Frawley, William J.; Piatetsky-Shapiro, Gregory; Matheus, Christopher J.: Knowledge Discovery in Databases: An Overview. AI Magazine 13(3), 1992, pp. 213-228. < http://www.aaai.org/Library/Magazine/Vol13/13-03/Papers/AIMag13-03-005.pdf > (Stand 20.9.2005)

Fröhlich, Gerhard: Das Messen des leicht Messbaren. Output-Indikatoren, Impact-Maße: Artefakte der Szientometrie? In: Jörg Becker, Wolf Göhring: Kommunikation statt Markt – Zu einer alternativen Theorie der Informationsgesellschaft (GMD Report 61 (1999)). St. Augustin: Gesellschaft für Mathematik und Datenverarbeitung – Forschungszentrum Informationstechnik GmbH. S. 27-38. < http://www.gmd.de/publications/report/0061/Text.pdf > (Stand 20.9.2005)

Ghani, Rayid; Jones, Rosie; Mladenic, Dunja; Nigam, Kamal; Slattery, Sean: Data Mining on Symbolic Knowledge Extracted from the Web. In: Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining, pp. 29-36, Boston, MA, August 2000. < http://www.accenture.com/xdoc/en/services/technology/publications/7.pdf > (Stand 20.9.2005)

Hearst, Marti: Distinguishing between Web data mining and information access. Position statement for Web Data Mining KDD 1997. < http://www.sims.berkeley.edu/~hearst/talks/data-mining-panel/ > (Stand 20.9.2005)

Hearst, Marti: Untangling Text Data Mining. In: Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 1999, pp. 20-26. < http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html > (Stand 20.9.2005)

Solla Price, Derek J. de: Little Science, Big Science. New York: Columbia University Press, 1963 (dt. Übersetzung 1974 im Suhrkamp Verl.)

Kohonen, Teuvo: Self-Organizing Maps. Berlin: Springer, 1995.

Kraft, Rolf-Peter: Meßbarkeit von Forschungsqualität? Der Zeitschriften-Impact-Faktor und die Zitieranalyse. In: Krebsforschung heute, Berichte aus dem Deutschen Krebsforschungszentrum 1998, Darmstadt 1998, S. 264-269. < http://www.dkfz-heidelberg.de/zbi/infos/messbark.htm > (Stand 20.9.2005)

Kruse, Rudolf; Borgelt, Christian: Information mining. In: Int. Journal of Approximate Reasoning 32 (2-3), 2003, pp. 63-65. < http://fuzzy.cs.uni-magdeburg.de/~borgelt/papers/ijar.pdf > (Stand 20.9.2005)

Nicholson, Scott: Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research works. Journal of the American Society for Information Science and Technology 54 (12). 2003, pp. 1081-1090.

Rosenbaum, Jana: Die Leistungsfähigkeit ausgewählter Suchmaschinen im World Wide Web hinsichtlich ihrer Retrievalmethoden unter besonderer Berücksichtigung des Rankings von Dokumenten, Magisterarbeit, Bibliothekswiss., HU-Berlin,1997.

Schmidt-Thieme, Lars: Webmining. Vorlesungsunterlagen vom 15.10.2002. < http://www.informatik.uni-freiburg.de/cgnm/lehre/wm-02w/webmining-1.pdf > (Stand 20.9.2005)

Stock, Wolfgang G.: Was ist eine Publikation? Zum Problem der Einheitenbildung in der Wissenschaftsforschung. In: Klaus Fuchs-Kittowski et al.: Wissenschaft und Digitale Bibliothek. Wissenschaftsforschung Jahrbuch 1998. Berlin: Gesellschaft für Wissenschaftsforschung 2000. < http://www.wissenschaftsforschung.de/JB98_239-282.pdf > (Stand 20.9.2005)

Studer, Rudi; Hotho, Andreas; Stumme, Gerd; Volz, Raphael: Semantic Web - State of the Art and Future Directions. In: KI Künstliche Intelligenz 2003 (3), S. 5-9.

Integration of Innovation: Linking the Innovation Systems of Japan and China

Han Shucheng, Harayama Yuko; Sendai-shi (Japan)

Abstract

The increasing geographical inter-linkages of technologies, markets, firms, capital, and production factors foster the integration of innovation between countries. This paper, specifically focusing on the innovation systems of Japan and China, aims to reveal the dynamic and nature of innovation integrations between the two countries. Beginning with a theory review on the openness of national innovation systems and their cross-border linkage, this article proposes a new analytical framework that suits our study purposes. Five different key activities are identified according to the nature of innovation process. They are regulation and policy, education, R&D, production, and market. The framework reflects that the innovation processes are implemented by activities and interactions created by organizations and institutions within the innovation system.

The main part of the article applies the proposed framework to empirically analyze the integration of innovation between Japan and China. Employing plenty of data and facts, this study draws a comprehensive comparison of advantages of each innovation system, analyses the motivations, patterns, and effects of the innovation linkages.

The study concludes that the innovation systems of Japan and China are getting increasingly open to each other. The innovation linkage is intensified and broadened as time goes on, from trade of capital goods to production transfer, then to R&D collaboration. The key reason for the linkage is that there exist numerous complementary innovation resources between Japan and China. With institutional and technical change, the complementary innovation resources become more and more valuable to each other. The third finding of our study is that through the dynamic linkages, both countries' innovation activities are expanded; their innovation infrastructures are enhanced; national innovation capabilities are improved. All these can be attribute to mutual learning and reciprocal exchange between two sides.

1. Introduction

Technical change is a fundamental force in shaping the patterns of transformation of the economy (Dosi et al., 1988). Technological innovations occupy a key position both for the competitive advantage of private sector firms as for the establishment of a self-enforcing process of economic growth and prosperity (koopmann and Münnich, 1999). In such an economy driven by on-going innovation, the institutional setting will have a major impact upon how economic agents behave and as well upon the conduct and performance of the technological system as a whole.

In the late 1980s, the concept of national system of innovation (NSI) has first appeared in the industrial innovation literature to reveal the interrelationship between technological development and the institutional embeddedness of innovative organizations (Freeman, 1988; Lundvall, 1992; Nelson, 1993). With the OECD as an early proponent of the concept, it has entered the vocabulary of national and international policymakers in the industrialized world remarkably quickly and is now also gradually spilling into policy making circles in developing countries. There has been a rapidly growing literature on this topic (OECD, 1999; Amable.. et al., 1997; Correa, 1998; Foray, 1994; Freeman, 1997; Lundvall,1998; Nelson, 1988,1992,1993; Niosi, 1991,1995; Saviotti,1996). Much of this literature (especially Hu, 1992; Porter, 1990; Patel, 1995) insists on the central importance of national systems, e.g. national institutions and networks, as well as nation-specific socio-cultural environments. However, a number of scholars (notably Ohmae, 1990) have argued that "globalization" has greatly diminished or even eliminated the importance of the nation-state, pointing to the internationalization of markets, technologies, production and other corporate activities, with an extreme example of Europeanization of public policies.

In this article, we take the NSI-concept as the analytical starting point, though our main purpose is to explore those features within a national system of innovation that have been reconfigured and restructured along transnational lines. That is, how does a country's NSI link to another country's? In particularly, our paper will empirically examine the dynamics and changing logics of organization and institutions with reference to the innovation systems of Japan and China.

The environment surrounding the innovation system of Japan has been changing tremendously since the end of last century. The changing environment suggests that industrial composition of the Japanese economy is shifting and has to shift towards science-based industries (SBI), in which the development is pursued by means of innovations based on science (Hiroyuki Odagiri, 2003). Kaiser (2004) pointed out that an emerging science-based industry makes new demands on the institutional environment of a NSI. Moreover, in Japan, the population is expected to age and decrease at a rate no other country has yet experienced. This trend challenges the technological progress and makes it of vital importance to create an efficient national innovation system (Akira Goto). In this context, Japan requires new institutional arrangements to incorporate and exploit outside innovation resources and capabilities to achieve greatest innovation performance.

In the similar vein, with China's reform and openness since 1980s', great changes have taken place in this country's innovation system. The ratio of R&D spending to GDP increased from 0.64% in 1997 to 1.23% in 2002, though it is still much lower than Japan's 3.35%. China's large-scale science and technology development plans and projects are dependent upon indigenous research and technological advancement as well as foreign investment, research, and technology.

There are several reasons why we choose the innovation systems of Japan and China to empirically explore the changing logics of national innovation system at

cross-border linkage. First, in recent years, business interaction between Japan and China has been increasingly intensified. Given their close culture, race, as well as their geographical proximity, we believe, the dynamic of Tran-boundary innovation activities a valuable issue. Second, Japan is a highly developed country and China is a rapidly developing country. Though most scholars focus international innovation integrity on the developed world, our interest is in the relationship of two different worlds. Thirdly, the political tension between Japan and China intrigue us to look into how the political factors have an impact on innovation system.

2. Conceptual framework and analytical model

According to Lundvall (2002, p. 215) the main background of innovation system concept should be found in the needs of policy makers and students of innovation. Researchers realized that innovation stemmed not only from university and technical research and development, but also from such other sources as production engineers, customers, and marketing, etc. The “innovation system” concept, introduced in Lundvall (1985) to integrate these broader contributions, then still without the adjective “national” added to it.

Although various scholars defined the national system of innovation in different ways (see an overview, Niosi, 2002, p. 292), there is something important in common. Galli and Teubal (1997) defined national system of innovation as “the set of organization, institutions and linkages for the generation, diffusion, and application of scientific and technological knowledge operation in a specific country”.

Besides their emphasis on the “national”, studies about the NSI consider that innovation processes not merely proceed within the national sphere (Lundvall, 1992). There are many arguments for enriching the innovation system approach by the international dimension (Hotz-Hart, 2002; Bunnell and Coe, 2001). So in this context, we can say that institutions is not only “national institution”, organizations is not only “national organization”, as well as linkages are not confined to national linkages.

It is clear that systems of innovation do not operate in isolation from each other. National systems of innovation are open systems that relate to domestic and international environments. The degrees and types of their openness differ from one country to another (Noisi, 1994). Under pressure of a globalizing economy it is becoming increasingly apparent that “close”, local learning relationships do not suffice to note sustain innovation but need to be complemented by extra-local knowledge flows (Asheim, 2002). Rather, international science, innovation and diffusion networks turn nationally based systems of innovation into open systems (Galli and Teubal, 1997). Thus a pertinent analysis of the systems of innovation approach needs to integrate two systemic dimensions: home and abroad.

These limitations have stimulate research efforts to draw on the NSI approach to analyze technical change in developing countries as well as to formalize the NSI

concept and carry out system-level analysis. A good case in point that is meant to capture the structure and performance of an NSI is the conceptual framework proposed by Liu and White (2001). This framework is built on five different activities of innovation processes. These activities are research, implementation (manufacturing), "end-use (customers of the product or process output)", "linkage", and "education"(Liu and White, 2001, p. 1094).

Based on the above theory debate, a general analytical model is formulated (Fig. 1). The innovation processes are implemented by activities and interactions created by organizations and institutions within the innovation systems of home and host countries. Regulation, education, R&D, production, and market are the five main innovation activity fields. Between them, there exist certain linkage and knowledge flow. Government plays an important role in the cross-border linkage process. In most developing countries, especially Pacific Asian countries, the host country's government not only serves as an intervene by employing macro-economy policy, but also as a participant of many technology transfer.

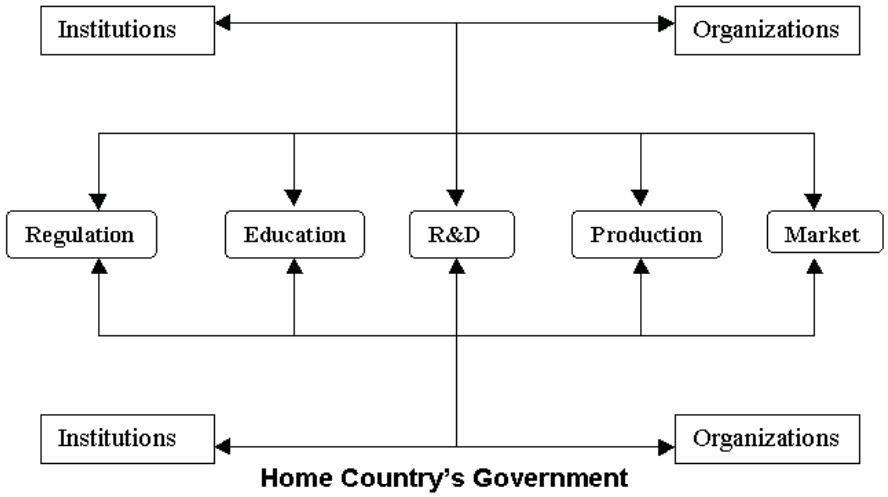


Figure1: an analytical model to study cross-border linkage of innovation systems

3. The cross-border linkage of innovation systems of Japan and China: some empirical evidence

3.1 market

It seems obvious that most of the innovation did not come directly from R&D, but rather from other sources like production engineers, customers, marketing, etc (lundvall et al., 2002). Among key factors that contribute to technological innovation

are government's market protection (Amsden, 1989; Hobday, 1995; Lall and teubal, 1998); environmental conditions of the product/market segment; favorable product market conditions (Bell and Pavitt, 1992); collaboration or strategic alliance with external technological agencies, such as customers, suppliers, and foreign technological sources (Kim and Lee, 2002). Many studies show the importance of user-producer linkages and the disadvantage of underdeveloped, small market, and unsophisticated demanding users to innovation (Freeman, 1990; Vernon, 1960; Lundvall, 1998; Porter, 1990). The size of the market is a key incentive for firms to invest in technological innovation. Innovation is costly. When a given cost of innovation is spread over more units of output, it will be easier to convince firms to place emphasis on the technological innovation.

China's huge market and its increasing upgrade have great significance to innovation systems of Japan and China and enhanced their linkages. For Japan, first, the attraction of huge market potential and market competition stimulate Japanese firms to keep active innovation. Secondly, through selling products to hundreds of millions of new customer in China, Japanese firms can exploit their technology and productivity fully and the profit from the huge market ensured fund for heavy investment in R&D.

Meanwhile, Japanese firms also export assembly lines, mainly through equipment purchase, technology licensing or OEMs. Usually, the assembly lines are not most advanced (the equipment is somewhat out of date and the technology is mature). By this strategy, Japanese companies can transfer their old industries and obtain profitable payment without losing their technology and innovation advantages.

However, Chinese market's openness is not without condition. Government takes measures to develop the national industry. "Bargain market for technology" is a strategy that is used to acquire technology from foreign countries including Japan. Through importing capital goods and technology licensing, plus technology learning, China can improve his national innovation capability.

A very good case is China's color TV (CTV) industry. After China opened up in the late 1970s, there was explosive growth in the consumer CTV market. Imported CTVs (mainly Japanese) exploited this opportunity and captured most of the market share with their higher-quality products and greater product variety.

To support national industrial development, Chinese central government initiated to support firms that imported assembly line. In this context, China imported the first CTV assembly line from Japan. The Chinese government took a flexible and pragmatic attitude towards the development of China's CTV industry. On the one hand, government protected local market through high tariffs. Government protection helped domestic firms maintain high-profit local market that contributed greatly to investment in learning and innovation. On the other hand, government liberalized domestic market step by step to increase the competition pressure on local firms. With the reduction in tariffs, by the end of 1999, exposure to Japanese imports and outside techniques had the following effects. First, foreign suppliers of technology

and assembly lines played an important role in helping to establish the industrial base of the Chinese CTV industry through licenses and parts. Second, the presence of foreign companies in the Chinese market provided competitive incentives, models and styles to be emulated by Chinese firms. Sino-Japan joint venture is an important path for Japanese firms to enter Chinese market. Table 1 shows major CTV firms in China during 1990s.

Name of company	Date of entry	Most recent reported production capacity	Ownership
Chonghon	1979	12,000,000	State-owned
Konka	1994	7,000,000	Former Hong Kong joint venture
Panda	1982	4,000,000	State-owned
TCL	1992	36,000,000	State-owned
Hisense	1983	13,000,000	State-owned
Haier	1997	1,000,000	State-owned
West Lake	1982	1,000,000	State-owned
Hitachi	1981	800,000	50%-owned joint venture (Japan)
Sanyo	1992	12,000,000	50%-owned joint venture (Japan)
Skyworth	1990	2,000,000	Private firm, Hong kong-controlled
Philips	1992	800,000	51%-owned joint venture (Netherlands)
Samsung	1994	800,000	50%-owned joint venture (Korea)
Sony	1996	3,000,000	50%-owned joint venture (Japan)
Matsushita	1996	500,000	50%-owned joint venture (Japan)
Sharp	1996	1,000,000	50%-owned joint venture (Japan)
Toshiba	1996	1,000,000	50%-owned joint venture (Japan)

Table 1: Major CTV firms in China during 1990s
Source: White and Linder (2002) and Xie (2001)

In 1971, the first domestic CTV product appeared in China. After the door opened to the outside world in the year of 1978, Chinese CTV industry, in the very beginning with the help of Japanese technology, imported capital goods, components and parts, achieved significant growth, as shown in table 1. Since 1990, China's CTV industry's output has been ranked first across countries in the world.

Based on the active learning process and imported technologies, China CTV firms have accumulated necessary technological capability to proceed in product and process innovation. Today, most of Chinese CTV firms have the design capability and engineering process capability. A number of innovative products appeared recently, including flat-screen CTV, 3-D CTD, multisystem CTV, WEB CTV, satellite CTV, PDP CTV, digital CTV, and high-definition TV (HDTV). So, Chinese CTV industry has benefited most from technological learning from Japan and had

considerable success in catching up with Japan. Table2 reflects a significant change of Japanese firm's market share in China's CTV market.

1980-1990		1996		1997		2000	
Leading brands	Market share	Leading brands	Market share	Leading brands	Market share	Leading brands	Market share
Foreign brands mainly from Japan such as Panasonic, Sony, Hitachi, NEC, Sanyo, etc.	Foreign brands account for about 80% of the total.	Changhong	20.5	Changhong	25.0	Changhong	18.7
		Panasonic	13.3	Konka	15.1	Konka	13.0
		Konka	12.2	TCL	9.5	TCL	8.3
		Beijing	7.1	Panasonic	6.7	Hisense	8.3
		Sony	5.5	Goldenstar	4.5	Skyworth	5.9
		Panda	4.6	Skyworth	4.4	Haier	5.1
		Toshiba	4.2	Panda	3.9	Goldenstar	6.7

Table 2: Leading brands in China's CTV market (%)
Sources: Sino-Market Research (Oct. 31, 2000)

3.2 production

Global production network disruptively changed geography of innovation system and creates new opportunities for international knowledge diffusion (Ernst, 2000). Japan and China are closely involved in current global production network, enhancing the linkages between innovation system of Japan and China.

Against the background of intense international competition, Japanese companies are attempting to select optimum bases for their activities in an effort to achieve advantage. In order to do so, Japan shifts oversea production bases for mass-produced core products to secure competitiveness mainly in terms of cost and facilitating entry into markets of countries where bases are located.

Kojima (1973) explained that Japanese firms invested abroad because of changes in macroeconomic conditions in Japan, which made it impossible for firms to continue producing at home. The change was also been reflected in the idea of the flying geese, which can be used to analyze the production integration between Japan and China. The flying geese model pointed out that as Japanese wages increased and the Yen appreciated, production facilities were relocated from Japan, first to the four NIEs, then to the second-tier ASEAN economies, and to China. Later, as wage costs rose in the dragons, more and more investment fluxed into China. This improved Lall's prediction that foreign corporation's new investment would concentrate on other locations particularly in countries such as China that can offer low wages, huge domestic market and a larger base of skills. Table 3 and table 4 show this change clearly.

Year	Japan's investment in 9 East Asia countries	Japan's investment in China	Share of Japan's investment in China to Japan's investment in 9 East Asian countries
1990	1015.8	511	5.0%
2002	6356	2152	33.9%

Table 3: Japan's investment in 9 East Asian countries and China's share
Sources: Ministry of Finance, Japan, 2003

Year	Number of firms transferred to East Asia	Number of firms transferred to China	Share of firms transferred to China to firms transferred to East Asian
1990	2862	150	5.2%
2000	6919	1712	24.7%

Table 4: Number of firms transferred to East Asia and China's Share
Sources: Ministry of Finance, Japan, 2001

It is an important form of exploiting the innovation generated at home to install direct investment productive facilities in host country and produce *in loco* new products and processes (Archibugi and Pietrobelli, 2002). Shifting production base from Japan to China benefited Japan's innovation system at least in the following several aspect. First, just as Kojima (1973) argued that FDI originating in Japan was in line with the exploitation of the host country's comparative advantages, production in China made Japan fully use of China's low wage but skilled labors and reduce the cost of new product and make it more competitive. Second, moving old industries to China can make Japan concentrate energies to the innovation of new product and industries. Thirdly, production in China make it easier for Japanese firm's entry to Chinese market. Fourth, production in China can reap the maximum return from the Value Chain, for example, by control over the key industry technology and components.

What is the impact of the production integration on the development of China's innovation system? This issue involves theoretical debates on the role of FDI and multinational corporations (e.g., Cantwell, 1994; Dunning, 1998). The dominant position has been that innovation, in contrast to most other stages of the value chain, is highly immobile: it remains tied to specific locations, despite a rapid geographic dispersion of markets, finance and production (e.g., Archibugi and Michie, 1997). Knowledge and innovation thus do not easily migrate across borders: they do not automatically follow, once production moves. Another proposition qualifies this argument and contending that cross-order production network facilitates knowledge diffusion, increasing the variety of international knowledge linkage. This creates new opportunities and challenges for the development of innovation system. Production

integration in the linkages of innovation systems poses a fundamental dilemma. An increased mobility of production resources and capabilities may enhance the knowledge diffusion and profit creation that are beneficial to innovation system. However, to guarantee this outcome needs proper institutional arrangement in each country's innovation systems. Network integration may equally well erode a country's resource of competition advantage and deter the promotion of national innovation capability.

A case in point is China's automobile industry. There have ever been two distinct arguments about Japan's production of automobile in China. Some argue that manufacturing and related support services of Japan's automobile corporation in China give ride to considerable learning and innovation. It includes for instance trial production, tooling and equipment, benchmarking of productivity, testing, process adaptation, product customization, and supply chain coordination. In contrast to this argument, many scholars prove that Sino-Japan automobile joint venture, including other automobile FDI in China, cannot lead to technological learning in product design. Even worse, it brings about a "crowding-out effect " to domestic R&D.

Here we do not want to make judgment about which one of the two opposite perception is right. In our opinion, it is better to understand this issue in a dynamic way.

In 1990s, or even earlier, because of Japan's excessive protection of its technology, and because of Japan's inharmonious policy with China's automobile industry policy, Japan missed a lot of opportunity to cooperate with Chinese partner to exploit its technology and new product. This made the US and European automobile companies take a lead in production and marketing in China. A type example is the investment failure of Toyota Corporation (one of the largest Japanese automotive corporations) in China.

In the same time, there was some weakness in China's innovation system that leads to the poor technology learning from Japanese companies. Kelly (2004) found that FDI in the automotive sector did not strongly contribute to improving Chinese technological capabilities because little knowledge was transferred along with the production. He attributed this to the weakness and inconsistency of Chinese policy in acquiring technology from foreign companies.

With China's accession into WTO and rapid expansion of automobile market, Japanese automotive companies take active measures to change the past strategy and increase their investment in China. Japanese companies in China outsource not only manufacturing, but also a variety of high-end, knowledge-intensive activities; such cross-functional, knowledge-intensive support services that are intrinsically linked with production can help considerable learning and innovation. Rather, Japanese companies take steps to integrate the functions of production, R&D and marketing and looking for Chinese long-term strategy partners.

Also, China realized that it has failed to develop its own technological capabilities and remains almost completely totally reliant on Japan and other foreign joint venture

partners for advanced automotive technologies. To reverse this situation, China has taken measures to improve the learning efficiency and self-innovate capability.

3.3 R&D

Industrial R&D

Innovation activities are increasingly shared between different national innovation systems in recent years. Beginning with the latest wave of globalization and accelerating in the 1990s, a growing number of multinational firms have begun to explore opportunities to expand and outsource R&D work to the developing world. There are numerous factors driving the increasing cross-border linkage of corporation R&D. First, innovation, especially in IT and biotechnology, is costly and risky. R&D cooperation is a best strategy to share the cost and uncertainty. Second, the way technological advances are realized in the information and communications industries, which rely as much on manufacturing technology as on services. Thirdly overseas R&D is also made easier by the enhanced mobility of both human beings and financial capital. Finally, an important driver facilitating cross-border linkage of R&D is the move toward normalization of international trade through the WTO.

Japan's R&D first emerged on the Chinese Mainland in the mid-1990s. Since then, this trend has evolved much the same way the global R&D trend has unfolded in other parts of the world. Especially after China's accession into WTO, more and more Japanese companies chose to conduct R&D activities in China. They show a tendency to locate important departments and research personnel within expatriate R8D subsidiaries in China.

What can Japan's innovation system reap from its deploy of R&D activities in China? First, Overseas R&D units support the marketing and production activities by developing products tailored to the need and taste of Chinese market, usually by modifying the products developed originally in the headquarters laboratory in Japan, and by providing technical service to the dealer and user. For instance, with the expansion of Chinese automobile market and with more and more Japanese automobile manufacturing in China, Japanese automotive companies have began to conduct R&D in China and R&D investment increases rapidly. In 2004, Japan's Toyota Motor opened its two auto R&D centers in Shanghai and Guangzhou. In the beginning of 2005, Dongfeng Mortor Corporation, one of China's automobile giants, have set up the country's first joint-venture auto R&D center with Nissan of Japan.

Second, China is turning into a global R&D center, in addition to a sales center (Miao Xu, CEO of Dongfeng Mortor Corporation). Accumulated knowledge and skills, as well as experienced suppliers make a broad technological base in China. Such a broad technological base insures that, whatever direction the path of R&D may take, the results are likely to be of value to the Japanese innovation. What is more, Chinese universities have raised a great many of high-quality students, plus a large pool of researchers and engineers as well as an influx of returnees from abroad. Therefore Japanese R&D activities in China can capitalize on the supply of R&D

resources, scientists and engineers, the supply of scientific and technological knowledge that may be gained through proximity to the universities, the public and private research institutions, or even the laboratories of other firms in China. Japan benefits from the advantage of economies of scale and scope through innovating in China.

Third, Chinese government provide financial and other incentives to attract foreign R&D investment in China, for example tax reduction and government procurement. Japan can take advantage of these favorable policies to be a member of the local innovation community.

On the other hand, there are numerous ways in which China's innovation system benefits form Japanese R&D investment. In addition to direct technology transfers through Sino-Japan contract or venture agreements, R&D investments from Japan are having considerable spillover effects on Chinese innovation.

First, Japanese-funded R&D centers and other R&D activities in China, which focus mainly on the key areas of applied research and industry technology, are helping China to bridge innovation gap between the realization of new advances in basic research and the implementation of these ideas in market. This contributes increasingly to China's long-term S&T development goals, one of which is to acquire from foreign investors the modern innovation concepts and technology development skills to fill the well-recognized "Valley of Death" in China.

Second, through partnerships and collaborations with industry researchers from Japan, China gained the applied research and technology development capabilities. These capabilities, along with the advantage of a sound and highly skilled scientific base, are sure to aid, and possibly accelerate, China's efforts to advance its technological modernization. Actually, Chinese government already has put in place many of the policy, institutional, and legal foundations China needs to better exploit technology and know-how from Japan and other foreign countries.

Japan's R&D investments also are aiding China to expand, enhance, and disseminate scientific knowledge throughout the country. The emergence of Japanese high-tech industry and R&D centers in China allows many more Chinese to improve their skills without having to go to Japan. What is more, Chinese employees working for Japanese R&D in China have access to modern research and development practices and processes, including innovative management techniques and other aspects of high-tech industry development that previously were accessible only to students and workers able to interact with high-tech companies in Japan.

Chinese S&T community, too, is benefiting by more fully participating in world scientific conferences, large-scale research projects, and other professional activities. Japanese R&D centers play a role in fostering these interactions by sending Chinese employees to visit corporate offices abroad for training, by sponsoring interns and research fellows, and by hosting visiting foreign scientists, engineers, technicians, and others at their centers in China.

Japanese R&D investment in China also contributes to improvements to the China's technological infrastructure. For instance, Japanese Export-Import Bank supplemented China expenditures on high-tech infrastructure.

Many Chinese companies enjoy R&D relationships with Japanese MNCs. These companies, through a combination of strategic alliances with Japanese companies (including on R&D), talented and comparatively low-cost labor, low-priced national brands, broad government support, and, increasingly, their own high-tech R&D efforts are beginning to change the image of the trade in China's label. As these enterprises' growing competitiveness suggests, China is reaping the benefits of these interactions, which have helped industry become China's leading contributor to R&D funding.

Academic R&D

Academic R&D is conducted mainly by universities and professional research institutes. Unlike industrial R&D, Academic R&D has an indirectly, time-cost feature for the promotion of a country's innovation capability. However, it is crucial for national innovation system in a long run.

Academic R&D collaborations between Japan and China are initiated and advanced by the two country's governments in 1970s. Japan and China signed an official agreement of science and technology cooperation in 1980. Since then, horizontal collective researches began to develop rapidly between the two countries. The S&T cooperation mainly includes collaboration in large-scale science and research projects, researcher communication, and other relevant coordinate activities. An important outcome of academic R&D collaborations is published articles coauthored by researchers. Japan has a close relationship with China's internationally co-authored articles (see tables 5 and 6).

Rank	Country	Number of Papers
1	US	1499
2	Japan	882
3	Germany	414
4	England	294
5	Australia	213

Table 5: China's internationally coauthored articles (first author is Chinese) in 2002
Sources: Chinese science and technology papers statistic (2002)

Rank	Country	Number of Papers
1	US	1382
2	Japan	723
3	Germany	371
4	England	331
5	Australia	176

Table 6: China's internationally coauthored articles (first author is foreigner) in 2002
Sources: Chinese science and technology papers statistic (2002)

Industry-academy cooperation

In recent years, a massive movement towards innovation cooperation has been launched in Japan and China among enterprises, universities, and public research institutes. Part of this industry-academy cooperation is crossing the border between Japan and China.

For example, in 2003, Japanese Sumitomo Corporation signed a cooperation agreement with Shanghai Communication University. The two sides will develop comprehensive cooperation in technology transfer, contract research, and researcher communication. Trough it, Sumitomo Corporation aims at reaping from the research capability that Shanghai Communication University owns in biotechnology and nano-technology. They also plan to fund joint venture in China to commercialize some promising R&D results.

This is not the solo example. Japanese Daikin Air Conditioner Corporation cooperates with Tsinghua University to research and develop air conditioner technology. An R&D center was built in Tsinghua University, which is the first time for Japanese air conditioner companies to locate an R&D center overseas. Each of the two sides sent 10 researchers to conduct R&D on key technologies in new generation of air conditioner.

3.4 Education

Education relates to the potential of future S&T human resources development (Dahlman, 1994; Nasierowski ad Arcelus, 1999). Thus, students exchange, as the main way of education cooperation between different countries, represents flows of technical and scientific personnel, especially at graduate level. Knowledge and technology move from one nation to the other, together with students bearing them.

Historically, student exchange has been one of the most important mechanisms for transfer of technology between Japan and China. Usually, there are more Chinese students move to study in Japan, in public or private way, for the purposes of researching and learning. After graduation, many of them chose to stay in and work for Japan, which enhances Japan's S&T base. In this sense, it is kind of "brain drain" for China to some extent.

However, in the opposite direction, more and more students awarded advanced degree or having worked in Japan for several years are going back to China, thereby promoting Chinese scientific and technical capabilities and reinforcing its innovation system.

3.5 Regulation and public innovation policy

Since the mid 1980s, market and product regulatory activities have gradually move to organizations of economic integration at multilateral level (Kaiser and Prange, 2003). Especially with China's access to WTO, Japan and China have made progress in establishing the framework for market liberalization and market access. Japanese companies and their high-tech products are getting more and more easier to enter into Chinese market. Meanwhile, public technology and innovation policies are reconfigured towards beyond the border between Japan and China. China faced growing portions of R&D-financing from Japan, and Japanese companies began to take part in Chinese public technology development programs.

4. Conclusion: to link the innovation systems of Japan and China effectively

"Systems of innovation are increasingly complex and intertwined" (Hotz-Hart, 2002). Regarding the issue how the innovation systems of Japan and China may effectively be linked, and how the linkages, in turn, have impact on the two country's innovation system, this paper bases on Liu and White's study to proposes an analytical framework for offering some new ideas. Five kinds of activities in national innovation system are identified according to the nature of innovation process. In each kind of activities, the interplay dynamics of the innovation systems of Japan and China is examined, including organizational and institutional changes as well as their interactions. We believe that the approach developed here could be applied to the study of cross-border linkages between any two or more country, not only Japan and China.

First, our study found that the innovation systems of Japan and China are getting increasingly open to each other. There are two peaks for the integration. One is China's "Open the Door" in the 1980s, and the other is China's entry into WTO. What is more, the depth of the linkage is intensified as time goes on, from trade of capital good to production transfer, then to R&D collaboration.

We also found that the key reason for the linkage is that there exist numerous complementary innovation resources between Japan and China. With institutional and technical change, the complementary innovation resources become more and more valuable to each other. This leads to the interdependence of the innovation systems of Japan and China. All these suggest the possibility and necessity of comprehensive cooperation between the two sides on S&T subjects.

The third finding of our study is that through the dynamic linkages, both countries' innovation activities are expanded; their innovation infrastructures are enhanced; national innovation capabilities are improved. All these can be attribute to mutual learning and reciprocal exchange between two sides.

Finally, It seems that China in fact dominants the linkage and can reap more from the linkage because most of the activities related to innovation linkage are conducted in China rather than in Japan. In the linkage, China usually plays a "host" actor, which can decide who is a popular guest and who is not. If favorable institutional and organizational change can be realized, China can catch up successfully by active learning.

However, there are still some politically unharmonious problems and old grudge between the two nations, and some ideological conflicts mean that this antagonistic political relationship is unlikely to change fundamentally in a short term. Nevertheless, the private sectors of the innovation systems on both sides have been influencing each other and this will definitely speed up academic and industrial cooperation. Pursuing higher economic growth will force policymakers on both sides to look ahead and accelerate the opening of the public and private sector to increase to active linkage, thus enhancing the efficiency and effectiveness of the innovation systems on both sides.

Main References

Archibugi and Michie, 1997. "Technological Globalisation or National Systems of Innovation", Future. Vol.29, pp.121-137.

Dosi, G, 1988. The Nature of the Innovative Process, in: G Dosi et al. eds, Technical Change and Economic Theory. Pinter, London.

Freeman, Christopher., 1988. Japan: a new national system of innovation, in: G Dosi et al. eds, Technical Change and Economic Theory. Pinter, London.

Goto, Akira., 2000, Japan's National Innovation System: Current status and Problems, Oxford Review of Economic Policy, Vol.16, No.2.

Galli, Riccardo. and Teubal, Morris., 1997. Paradigmatic Shifts in National Innovation Systems, in C. Edquist (ed), Sysrem of Innovation. Technologies, Institutions, and Organizations. Pinter: London, pp.342-370.

Hobday, M. (1995) Innovation in East Asia: the Challenge to Japan, Edward Elgar, Aldershot.

Kaiser, Robert., Prange, Heiko., 2003. The Reconfiguration of National Innovation Systems in OECD Countries. Paper prepared for international conference "Innovation in Europe: Dynamics, Institutions, and Values", Roskilde/Denmark, 8-9 May 2003. Koopmann, George. and Münnich, Felix., 1999. National and Inter

national Developments in Technology: Trends, Patterns and Implications for Policy, HWWA Discussion Paper 76/1999, Hamburg.

Lundvall, B., 1992. Introduction, in Lundvall, B. eds, National Systems of Innovation. Toward a Theoty of Innovation and Interactive Learning. Pinter: London, pp. 1-19.

Liu, X. and White, S., 2001. Comparing innovation systems: a framework and application to China's transitional context. Research Policy 30, (2001) 1091-1114.

Niosi, Jorge., 2002. National systems of innovations are "X-efficient" (and x-effective): Why some are slow learners. Research Policy 31 (2002) 291-302.

Nelson, R., 1993 (ed), 1993, National Innovation System. A comparative Analysis. Oxford University Press, pp. 3-21.

Odagiri, Hiroyuki., 2003. Japan's National Innovation System: Its Evolution and Current Situation, Seminar on Innovation Systems in Asian Economies, 4-5 September 2003.

Xie, W. and WU, G., 2003. Differences between learning processes in small tigers and large dragons. Research Policy, 32, 1463-1479.

Makro- und Mikro-Mining am Beispiel von Webserver Logfiles

Philipp Mayr, Christian Nançoz; Bonn

Abstract

Webserver Logfiles sind eine hochinteressante Informationsquelle zur Untersuchung der Zugänglichkeit, Sichtbarkeit und Verlinkung von beliebigen Webinhalten. Dieser Beitrag stellt zwei neuere Ansätze der Logfile Analyse bzw. des Web Mining vor (Makro-Mining & Mikro-Mining). Der weitverbreiteten Methode der Makro-Analyse, die hauptsächlich allgemeine Zugriffszahlen aggregiert (z. B. Anzahl der Downloads eines Dokuments), wird die bislang weniger bekannte Methode der Mikro-Analyse gegenübergestellt. Die Mikro-Analyse konzentriert sich auf schmale Segmente des Logfiles, die bis auf Transaktionen einzelner User zurückgehen. Beide Analysemethoden werden anhand eines Beispiels erklärt. Weiterhin wird versucht neue Einsatzbereiche der beiden Web-Mining Verfahren zu identifizieren und Formen der kombinierten Nutzung der beiden Methoden zu skizzieren.

1. Einleitung

Als Folge des Paradigmenwechsels im wissenschaftlichen Publikationswesen werden immer häufiger Forderungen nach Bewertungskriterien deutlich, die den spezifischen Eigenschaften der neuen Publikationstypen Rechnung tragen. Die zunehmende Anzahl und wissenschaftliche Bedeutung von Online-Publikationen z.B. aus dem Open Access Bereich, macht es daher notwendig, innovative webbasierte S&T-Indikatoren zu entwickeln. Diese Web-Indikatoren sollen helfen, die neuen Publikationen einzuordnen und ggf. zu bewerten. Web-Indikatoren werden laut WISER charakterisiert als:

„A policy relevant measure that quantifies aspects of the creation, dissemination and application of science and technology in so far as they are represented on the internet or the World Wide Web." (vgl. dazu WISER Projekt[1])

Das Fehlen von robusten webbasierten Messmethoden, beispielsweise zur Impactbestimmung wissenschaftlicher Literatur, wirkt sich für alle Akteure, die in den wissenschaftlichen Prozess involviert sind (Wissenschaftler, Bibliothekare, Gutachter, …), als sehr störend aus. Das Hauptproblem bestellt darin, dass bislang noch keine tragfähigen Indikatoren in Sicht sind, die die Einschränkungen aber auch Potenziale des neuen Publikationsmediums Internet genügend berücksichtigen.

[1] WISER steht für Web Indicators for Science, Technology & Innovation Research, siehe http://www.wiserweb.org

Die Webometrie (Björneborn & Ingwersen 2001, Thelwall, Vaughan & Björneborn 2003), die wie die Zitationsanalyse (vgl. Bibliometrie) auf quantitativen Methoden basiert, verspricht, hier ansatzweise Hilfestellung, aber noch keine Lösungen anzubieten. Neben der Analyse der Hyperlinkstrukturen im Web, bieten sich zur Analyse auch die Nutzungsinformationen an, die auf Webservern durch die Protokollierung der Onlinebenutzung entstehen. Dieses Analyseverfahren nennt sich Logfile Analyse.
Die Logfile Analyse, z.T. auch Web-Mining genannt, soll im Mittelpunkt des folgenden Beitrags stehen.

2. Logfile Analyse

Die Logfile Analyse ist als Verfahren zur nichtreaktiven Nutzungsmessung von Webnutzern allgemein anerkannt und erfreut sich trotz einiger systematischer Einschränkungen (Nicholas et al. 1999) großer Beliebtheit. Beispiele für Anwendungsgebiete der Logfile Analyse sind neben der weit verbreiteten Websiteanalyse bzw. -statistik auch das User Modelling oder die quantitative Untersuchung des Informationsverhalten auf Websites, sowie die Nutzung von Suchmaschinen (Thelwall, Vaughan & Björneborn 2003). Logfile Analysen liefern darüber hinaus wertvolle Informationen zur Zugänglichkeit (accessibility), Sichtbarkeit (visibility) und Verlinkung (interlinking) von Webinhalten eines spezifischen Webservers (Thelwall 2001, Mayr 2004a).
„With the web being such a universally popular medium, accounting forever more of people's information seeking behaviour, and with every move a person makes on the web being routinely monitored, web logs offer a treasure trove of data. This data is breathtaking in its sheer volume, detail and potential ... Unfortunately the logs turn out to be good on volume and (certain) detail but bad at precision and attribution." (Nicolas et al. 1999)
Dem verwandten Konzept Web-Mining liegt die Annahme zugrunde, dass die im Web verfügbaren Daten (Webdaten) ausreichend strukturiert sind, um mit Algorithmen nach Mustern zu suchen. Unter Web-Mining werden allgemein Data Mining-Techniken[2] verstanden, die zum automatischen Auffinden und Extrahieren von Informationen aus Webdokumenten und -services dienen. Siehe dazu die Definition von Kosala & Blockeel:
„The Web mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing." (Kosala & Blockeel 2000)

[2] Data Mining wird allgemein als explorative Verdichtung von Daten verstanden, um neue Erkenntnisse aus den Daten zu gewinnen.

Das Verfahren der automatischen Zitationsextraktion und -analyse, dass von Lawrence, Giles und Bollocker beschrieben und im Citeseer-System eingesetzt wird, (Lawrence, Giles & Bollocker 1999) kann beispielsweise als Web-Mining Verfahren bezeichnet werden.
Sowohl in der Webometrie (Thelwall 2001, Thelwall, Vaughan & Björneborn 2005) als auch in anderen informationswissenschaftlichen Disziplinen haben sich Logfiles z. B. ansatzweise zur Indikatorenbildung (Brody & Harnad 2005, Mayr 2004b) und Messung des Nutzerverhaltens (Nicholas et al. 1999, Koch, Golub & Ardö 2004) bewährt.
Im nachfolgenden Kapitel soll sehr knapp auf die Spezifika der Webserver Logfiles eingegangen werden.

2. Webserver Logfiles

Webserver Logfiles sind aufgrund ihrer Struktur, Größe und Verfügbarkeit exzellente Datenquellen für Untersuchungen des Benutzungs- und Kommunikationsverhaltens im Web. Die Webserver Logfiles stellen große Informationsmengen (i.d.R. ohne Unterbrechung) über alle Zugriffe auf einen bestimmten Webserver bereit. Die Qualität der enthaltenen Daten ist ein wichtiger Aspekt der Logfile Analyse, der in vielen Standardanalysen leider wenig Beachtung findet. Es ist aus vielen Untersuchungen bekannt, dass einige Störfaktoren die Ergebnisse der Analyse stark verfälschen können und die erste Reinigungsphase (“Cleaning”) ein außerordentlich wichtiger Schritt des Web-Mining darstellt.
Einer der bekanntesten Störfaktoren in Logdaten öffentlicher Webserver sind die Einträge der Suchroboter, die sogenannten “Spider”[3]. Diese Programme generieren im Logfile Einträge, die keine Hinweise über das Verhalten des User geben und müssen vor der Analyse unbedingt entfernt werden. Teilweise bietet schon die Analyse-Software diese Funktionalität an, indem sie die bekanntesten Such-Roboter innerhalb der Logfiles verfolgen und herausfiltern. Die Anwesenheit von unbekannten Robotern kann mit letzter Gewissheit nicht ausgeschlossen werden. Weitere Faktoren sind noch schwieriger zu beseitigen: u.a. die breite Verwendung der Rück-Navigation und die Optimierung durch die Zwischenablage der neusten Seiten inmsogenannten “Cache Memory” auf der User Seite. Dazu gehören auch die “proxy caching servers”, die Bestandteil der Website Architektur sind, und die Geschwindigkeit der Website erhöhen sollen. Alle Abfragen der Benutzer, die zwischengelagert werden, kommen nicht im Logfile vor und können daher auch nicht untersucht werden. Laut Gutzman (Gutzman 1999) verursachen die “Spider” und Such-Roboter, die von Suchmaschinen wie Google verwendet werden, ein Drittel der Einträge der

[3] Ein Programm, das automatisch Webseiten herunter lädt und der Suchmaschine zuführt. (Quelle Google: http://www.google.ch/intl/de/ads/glossary.html)

Logfiles (siehe dazu auch Nicholas & Huntington 2003). Neben diesen Störfaktoren für die Logfile Analyse existieren noch weitere, wie z.B. die immer mehr für das E-Business verwendeten dynamisch gebildeten Webseiten. Wegen der großen Informationsmenge, der zugrundliegenden Ungenauigkeit und der Schwierigkeit Störfaktoren zu identifizieren, würden Logfiles einen idealen Einsatzbereich für die unscharfe Logik darstellen (Nançoz 2004).
Diese Aspekte führen zur Behauptung, dass Logfiles per se nicht vollständig sind und ein verfälschtes Bild des Verhalten des Websitebesuchers abgeben können.
Neben den zahlreichen Einschränkungen der Webserver Logfiles existieren eine Reihe von Potenzialen, die hier aber nur ansatzweise aufgezählt werden können:

- Vergleichsweise einfach zugängliche Datenquelle zur Untersuchung einer umfangreichen und heterogenen Nutzergruppe.
- Zeitnahe und nahezu vollständige Analyse des Informationsaufnahme-verhaltens einzelner Websitebesucher oder Gruppen.
- Hinweise zur Optimierung, Evaluation und Adaption von Webinhalten und -services.
- Potenzial zur Messung von Web Impact (vgl. Brody & Harnad 2005) und zur Entwicklung weiterer webbasierter Indikatoren.

Im folgenden Abschnitt wird erläutert inwiefern Webserver Logfiles zur Makro- bzw. Mikro-Analyse eingesetzt werden können und ob eine kombinierte Analyse der beiden methodisch unterschiedlichen Verfahren denkbar und sinnvoll ist.

3. Makro- und Mikro-Mining

In diesem Kapitel werden zwei neuere und bereits publizierte Methoden vorgestellt mit denen Webserver Logfiles analysiert werden können. Die erste Untersuchungs-methode lässt sich als Makro-Mining-Methode beschreiben. Makro-Mining von Webserver Logfiles meint hier die Extraktion von Daten, die sich zu einfachen Maßen wie Anzahl der Visits oder Downloads zusammenfassen lassen. Neben diesen einfachen Standardmaßen wird in Kapitel 3.1 eine erweiterte Makro-Methode be-schrieben, die auf der Unterscheidung verschiedener Einstiegszugriffe einer Website basiert.
Nicholas & Huntington haben 2003 erstmals eine Methode vorgestellt, die als Mikro-Mining-Methode bezeichnet wird.
“The aim of the study is to show how microanalysis can enhance current log analyses techniques. In particular the paper seeks to demonstrate three potential ‘micro’ techniques

1. the construction of a subgroup of users for which we can feel confident in regard to their geographical origin;
2. the analysis of a subgroup of users whose Internet Protocol (IP) addresses are more likely to reflect the use of individuals, and the same individuals;
3. the tracking and reporting of the use made by individuals rather than groups."

(Nicholas & Huntington 2003)

Eine Idee dieses Beitrags ist es die Potenziale der beiden unterschiedlichen Analyseformen zu kombinieren und die Stärken der jeweiligen Methode zu nutzen, um sie künftig zu aussagekräftigeren Kennzahlen zu verdichten bzw. spezifischere Analysen auf ihnen aufzubauen. Des weiteren soll motiviert werden, warum es für bestimmte Analysen notwendig ist, über die Standard Logfile-Auswertungen hinaus eine erweiterte Logfile Analyse durchzuführen. Der Beitrag beschränkt sich hier auf die Vorstellung der beiden Verfahren und skizziert lediglich mögliche Einsatzszenarien. Die Kennzahlenbildung bzw. empirische Prüfung der Analysen steht nicht im Mittelpunkt dieses Papers.

3.1 Makro-Mining Ansatz

Der erste Ansatz betrachtet Logfile-Einträge aus einer Makro-Perspektive. Die unten aufgezählten Makro-Analysen können weitgehend als Standardverfahren angesehen werden und werden in vielen Logfile Analyseprogrammen angeboten. Ihre Kennzahlen (z.B. Hit, View, Visit) sind zwar verbreitet, geben aber lediglich ein sehr grobes und eingeschränktes Bild des eigentlichen Onlineverhaltens. Typische Makro-Mining Auswertungen beantworten z.B. folgende Fragen:

- Wie viele Zugriffe erhalten bestimmte Bereiche bzw. Entitäten (Directory, Page) einer Website?
- Welche sind die wichtigsten Einstiegsseiten einer Website? (siehe Abbildung 1)
- Über welche Suchmaschinen bzw. Suchbegriffe finden die Nutzer ein bestimmtes Webseitenangebot?

Die aus den Logdaten extrahierten Informationen liefern bei dieser Methode Hinweise auf einer relativ abstrakten Makroebene und sind i.d.R. nur hilfreich, wenn Vergleichswerte (Benchmarks) z. B. Werte vom vorherigen Quartal vorliegen.

Top Entry Pages			
	Page	% of Total	Visits
1	http://www.ib.hu-berlin.de/	5.35%	229
2	http://www.ib.hu-berlin.de/~ hab/amd/Start.html	3.74%	160
3	http://www.ib.hu-berlin.de/~ mh/gedv/ascii.htm	2.26%	97
4	http://www.ib.hu-berlin.de/~ mh/css/css2/fonts.html	1.96%	84
5	http://www.ib.hu-berlin.de/~ mh/projekte/metaopac/	1.75%	75
6	http://www.ib.hu-berlin.de/ jaw/Html/studwohn.html	1.26%	54
7	http://www.ib.hu-berlin.de/~ hab/christine/gaudi1.html	1.16%	50
8	http://www.ib.hu-berlin.de/~ pbruhn/russgus.htm	1.14%	49
9	http://www.ib.hu-berlin.de/~ wumsta/rehm8.html	1.14%	49
10	http://www.ib.hu-berlin.de/~ wumsta/rehm4.html	1%	43

Abbildung 1: Beispiel einer typischen Makro-Analyse. Die wichtigsten Einstiegsseiten (Top Entry Pages) einer Website gemessen an der absoluten Anzahl der Besuche (Visits).

Die Abbildung 1 zeigt einen Ausschnitt einer Liste mit Einstiegs-Webseiten einer Website. Die Webseiten sind nach der Häufigkeit der Websitebesuche (visits) geordnet, die auf den jeweiligen Seiten begonnen haben. Die Seite mit dem Rang 1 http://www.ib.hu-berlin.de ist für den Untersuchungszeitraum die wichtigste Einstiegsseite (229 visits, d.h. 5.35% aller visits, haben auf dieser Seite begonnen).

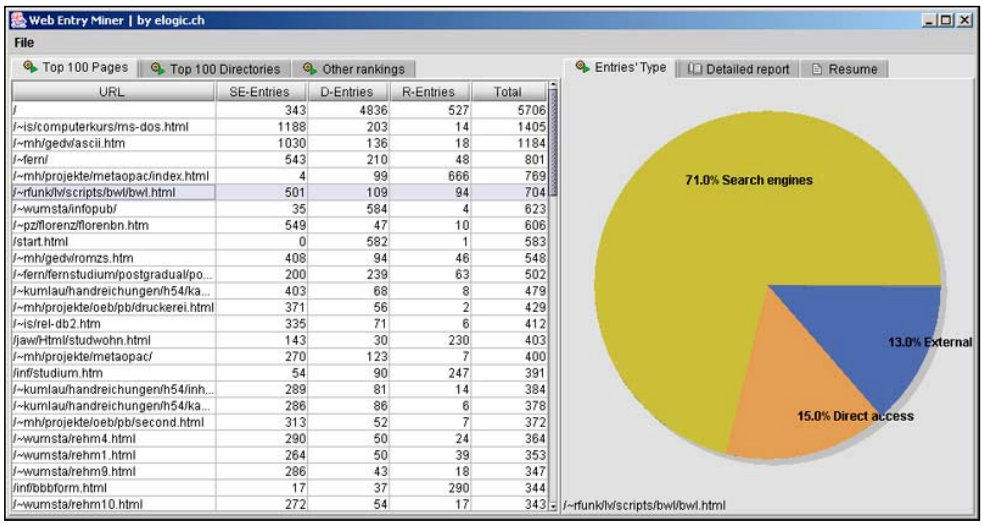


Abbildung 2: Screenshot einer Makro-Analyse (Web Entry Analyse) mit dem Prototypen Web Entry Miner[4]

[4] Web Entry Miner, siehe http://www.ib.hu-berlin.de/~mayr/wem/

Abbildung 2 zeigt die Analyse und Visualisierung von Logdaten über eine neuere Makro-Mining-Methode (Mayr 2004b, siehe Abbildung 2). Die Web Entry Analyse basiert auf der eindeutigen Unterscheidung von Einstiegszugriffen (Web Entries) aller Besucher einer Website, die sich auf die drei Einstiegsarten Suchmaschinen, Backlinks[5] oder direkte Zugriffe zurückführen lassen. Diese drei sehr unterschiedlichen Zugriffs- bzw. Einstiegsarten, die sich zu drei Anteilswerten (Suchmaschinen-Einstieg, Backlink-Einstieg und direkter Einstieg) pro Einstiegsseite aggregieren lassen (siehe linker Tabellenbereich in Abbildung 2), geben detaillierte Einsichten über die Bedeutung und Qualität einer konkreten Seite als Einstiegsseite innerhalb der Website aus. Die in Abbildung 2 ausgewählte Seite (siehe Spalte URL in Abbildung 2) erhält z.B. 71% der Einstiegszugriffe über die Einstiegsart Suchmaschine. Die übrigen Einstiegszugriffe verteilen sich auf die beiden anderen Einstiegsarten. Aus dieser Auswertung lässt sich schließen, dass die entsprechende Seite zum Zeitpunkt der Analyse sehr gut bei einer oder mehreren Suchmaschinen positioniert ist. Eine tiefere Analyse der Zusammensetzung des Zugangsverhalten kann diese Makrosicht aber vorerst nicht liefern. An dieser Stelle wird es notwendig das nachfolgend beschriebene Mikro-Mining an die Analyse anzuschließen.

3.2 Mikro-Mining Ansatz

Als ein Anwendungsbeispiel der Mikro-Analyse wird das detaillierte Online-Verhalten (User Tracking) einzelner akademischer Benutzer nachvollzogen. Die grundsätzliche Idee der Mikro-Analyse (Nicholas & Huntington 2003) besteht darin, eine Untergruppe von Besuchern einer Webseite zu analysieren und sich an Benutzer-Sessions zu orientieren. Innerhalb dieser Sessions wird die Abfolge der gesichteten Seiten extrahiert, um daraufhin Benutzer-Tendenzen zu identifizieren. Die Untergruppe kann so ausgewählt werden, dass sie repräsentativ ist oder die Merkmale einer größeren identifizierbaren Gruppe trägt. Zum Beispiel wählt Gutzman eine Untergruppe, (eine akademische Benutzergruppe), die über ihre geographische Zuordnung zuverlässig extrahiert werden kann (Gutzman 1999). Die Mikro-Analyse erlaubt gewisse Tendenzen im Verhalten der Untergruppe zu identifizieren, indem der Fokus auf einzelne Benutzer gesetzt wird. Es wird weiterhin versucht die Benutzer-Sessions, die von Störfaktoren sehr wahrscheinlich beeinflusst wurden, auszuschließen.

[5] Hyperlinks, die sich außerhalb der analysierten Website befinden und auf eine Seite der analysierten Website verweisen, werden als Backlinks bezeichnet.

IP Adresse	Webseite	Referrer	Browser	Zeit
18 Oktober 2003				
128.xxx.xxx.xxx	/	"http://www.google.com/search?q=disinfo"	"Mozilla/3.01 [de] (Win16; I)"	01:03:52
128.xxx.xxx.xxx	/content.htm	"http://www.disinfojournal.net"	"Mozilla/3.01 [de] (Win16; I)"	01:11:01
128.xxx.xxx.xxx	/issue1_1.htm	"http://www.disinfojournal.net/content.htm"	"Mozilla/3.01 [de] (Win16; I)"	01:11:01
128.xxx.xxx.xxx	/free.htm	"http://www.disinfojournal.net/issue1_1.htm"	"Mozilla/3.01 [de] (Win16; I)"	05:23:34
128.xxx.xxx.xxx	/hilights.htm	"http://www.disinfojournal.net/free.htm"	"Mozilla/3.01 [de] (Win16; I)"	02:21.56
128.xxx.xxx.xxx	/authors.htm	"http://www.disinfojournal.net/hilights.htm"	"Mozilla/3.01 [de] (Win16; I)"	12:54:05
128.xxx.xxx.xxx	/about-us.htm	"http://www.disinfojournal.net/authors.htm"	"Mozilla/3.01 [de] (Win16; I)"	03:34:41
128.xxx.xxx.xxx	/index.html	"http://www.disinfojournal.net/about-us.htm"	"Mozilla/3.01 [de] (Win16; I)"	00:30:31
19 Oktober 2003				
128.xxx.xxx.xxx	/hilights.htm		"Mozilla/3.01 [de] (Win16; I)"	24:09.36
128.xxx.xxx.xxx	/authors.htm	"http://www.disinfojournal.net/hilights.htm"	"Mozilla/3.01 [de] (Win16; I)"	03:44:02

Abbildung 3: Beispiel mehrerer Sessions eines anonymisierten Users (User Tracking Protokoll)

Das User Tracking erfolgt durch die Identifikation der einzelnen Sessions des Benutzers. Durch die Einbeziehung sämtlicher in Logfiles enthaltenen Informationen wird es möglich den Weg einer IP-Adresse durch ein Logfile zu verfolgen. Folgende Logfile-Felder werden dazu benötigt: IP-Adresse, Referrer[6] und verwendeter Browser-Typ. Die Zugriffszeit ermöglicht die Zugriffsdauer zu berechnen und mehrere Sessions logisch zu unterscheiden, indem ein gewisser Time-out definiert wird. Das Beispiel der Abbildung 3 illustriert die Sessions eines Users auf der disinfojournal.net Website. Die Benutzung des gleichen Browsers weist darauf hin, dass sich höchstwahrscheinlich ein einziger User hinter dieser IP-Adresse verbirgt. Ein anderer Hinweis, dass es sich bei den Sessions um den gleichen User handelt, ist die gemeinsame Thematik die sich aus den angesehenen Seiten und vielleicht sogar aus der kurzen Zeitspanne zwischen den Sessions ableitet. In unserem Beispiel war der User offensichtlich auf der Suche nach allgemeinen Informationen über die online Zeitschrift „Disinfojournal“. Einen Tag später, am 19. Oktober, hat der User direkt auf die „Hilights“ Seite zugegriffen oder hat die Seite vielleicht in seinen Favoriten gespeichert und hat wiederum nach den gleichen Informationen gesucht. Die genauere Uhrzeit und vor allem die Dauer der Sessions geben zudem Erläuterungen über die Verweildauer von bestimmten Benutzern auf der Website.
Durch diese feine und bewusst eingeschränkte Analyse kann die Navigationsmethodik des einzelnen Benutzers nachvollzogen werden. Damit wird auch der Inhalt der Logfiles mit allen Details ausgenutzt und der Verlust großer Informationsmengen vermieden. Im Gegensatz zur Makro-Analyse sind diese Resultate und Aussagen viel genauer und zuverlässiger, beschränken sich aber auf eine viel kleinere Benutzergruppe.

[6] URL der Webseite von der der Besucher die Seite aufgerufen hat.

Auf der Basis der komplementären Eigenschaften beider Analysen stellt das nächste Kapitel ihre sinnvolle Kombination anhand verschiedener Szenarios vor.

4. Einsatzszenario – kombinierte Analysen

Zum Abschluss dieses Beitrags wird zur Verdeutlichung der Möglichkeiten der beiden beschriebenen Analysemethoden ein potenzielles Einsatzszenario einer kombinierten Analyse skizziert. Das folgende Beispiel, bei dem Makro- und Mikro-Analyse ineinander greifen, basiert auf folgenden abgestuften Analyseschritten:
Schritt 1: Zunächst wird ein Logfileausschnitt über die in Kap. 3.1 beschriebene Makro-Methode Web Entry Mining (Mayr 2004b) analysiert. Nach der Analyse steht eine typische Makrosicht (vgl. Abbildung 2) zur Verfügung, die einzelne Webseiten und ihre Zugänglichkeit über die drei Zugangsarten (Suchmaschinen, Backlinks, Direkte Zugriffe) darstellt. Diese Makrosicht lässt sich über den folgenden Schritt weiter vertiefen.
Schritt 2: Der Betrachter will sich im zweiten Schritt die genaue Zusammensetzung der Einstiege z.B. über Suchmaschinen einer beliebigen Webseite ansehen. Dazu startet er ein sogenanntes „Drill down“[7], indem er sich die genaue Zusammensetzung der Gesamtzahl der Suchmaschinen-Einstiege anzeigen lässt. Beispielsweise gehen die 71% der Suchmaschinen-Einstiege einer Seite auf mehrere Suchmaschinen und verschiedene Suchbegriffe zurück. Über die Drill down-Analyse werden z.B. alle Suchmaschinen sichtbar, die zu den Suchmaschinen-Einstiegen der Webseite geführt haben. Über einen weiteren Drill down-Schritt werden z.B. die Suchbegriffe einer konkreten Suchmaschine sichtbar. An dieser Stelle würde die Makro-Analyse enden und die Mikro-Analyse ansetzen.
Schritt 3: Das User Tracking des Mikro-Mining Ansatzes lässt sich in diesem Szenario über einen konkreten Suchmaschinen-Suchbegriff aktivieren. Die angeschlossene Mikro-Analyse versucht, alle Websitebesucher, die mit einem konkreten Suchbegriff auf eine Seite gefunden haben, inkl. der weiteren Session-Informationen anzuzeigen. Die User, die über ihre IP-Adresse oder andere eindeutige Merkmale authentifiziert werden konnten, werden mit allen ihren weiteren Transaktionen (vgl. User-Tracking-Protokoll in Abbildung 3) angezeigt. Damit endet das kombinierte Analyse-Szenario.
Die Kombination der beiden Analysemethoden erlaubt, die mit der Makro-Analyse abgeleiteten Tendenzen mit einer Testgruppe zu prüfen und damit noch genauer auf das Verhalten des Users zu fokussieren.
Auf der anderen Seite kann die Mikro-Analyse verwendet werden, um „Micro-Trends“ zu entdecken. In einem zweiten Schritt würde eine erweiterte Makro-Analyse das

[7] Drill down (engl.: tiefer bohren oder graben) meint im Zusammenhang mit Logfile Analysen, dass dem Nutzer über ein Funktionsmenü ein weiterer tieferer Analyseschritt zur Verfügung steht.

Ausmaß des Trends abschätzen. Vorteil dieses Szenario ist die Fähigkeit der Mikro-Analyse kleine aber aussagekräftige Benutzersegmente zu identifizieren, die mit einer Makro-Methode unsichtbar bleiben.

5. Ausblick

„Transaction log files allow us to look at the behaviour of millions of people, but the aggregation misses the detail and the detail can add to the impressions and thoughts about user behaviour." (Nicholas & Huntington 2003)

Ziel des vorliegenden Papers war es zwei methodisch sehr unterschiedliche Webserver Analyseansätze zu beschreiben und die Kombination abgestufter Analysemethoden für künftige Untersuchungen vorzuschlagen. Durch die Kombination beider Analysen (Makro- und Mikro-Mining) wird beispielsweise ein völlig unterschätzter Aspekt des „User Information Retrieval" aufgedeckt: zu den „klassischen" und statischen Informationen, wie der geographischen Herkunft der Benutzer und den Einstiegsseiten, kann das Benutzer-Verhalten als neue dynamische Komponente hinzugefügt werden.

Neue Ansätze zur Analyse von Webserver Logfiles werden immer notwendiger um genauere und stabilere Maße des Websitegebrauchs zu entwickeln. Dies gilt für wissenschaftliche Logfile-Untersuchungen und kommerziell orientierte Verfahren gleichermaßen. Gerade die kombinierten Analysen versprechen hier neue Ergebnisse und tiefere Einsichten in das Benutzungsverhalten zu liefern.

Neben der Reduzierung der Fehleranfälligkeit von Logfile Analysen, sind der gesteigerte Komfort der Websitebesucher, die Reduzierung der Suchzeit und damit letztlich die Besucher- bzw. Kundenzufriedenheit die wichtigsten Ziele künftiger Entwicklungen.

Literatur

Bjöneborn, Lennart; Ingwersen, Peter (2001): Perspectices of webometrics. In: Scientometrics, Vol. 50, pp. 65-82.

Brody, Tim; Harnad, Stevan (2005): Earlier Web Usage Statistics as Predictors of Later Citation Impact. Technical report. URL: http://eprints.ecs.soton.ac.uk/10647/ (access date 14 August 2005)

Gutzman, A. (1999): Analysing Traffic on Your E-commerce Site. URL: http://ecommerce.internet.com/solutions/tech_advisor/article/0,,9561_186011,00.html (access date 14 August 2005)

Koch, Traugott; Golub, Koraljka; Ardö, Anders (2004): Log Analysis of User Behaviour in the Renardus Web Service. URL: www.it.lth.se/knowlib/publ/LIDA2004_final.doc (access date 14 August 2005)

Kosala, Raymond; Bockeel, Hendrik (2000): Web mining research: A survey. In: SIGKDD Explorations, Vol. 2, pp. 1-15.

Lawrence, Steve; Giles, C. Lee; Bollacker, Kurt (1999): Digital Libraries and Autonomous Citation Indexing. In: IEEE Computer, Vol. 32 (6), pp. 67-71. URL: http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html (access date 14 August 2005)

Mayr, Philipp (2004a): Entwicklung und Test einer logfilebasierten Metrik zur Analyse von Website Entries am Beispiel einer akademischen Universitäts-Website. (Berliner Handreichungen zur Bibliothekswissenschaft und Bibliothekarsausbildung ; 129). URL: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h129/ (access date 14 August 2005)

Mayr, Philipp (2004b): Website entries from a web log file perspective - a new log file measure. Proceedings of the AoIR-ASIST 2004 Workshop on Web Science Research Methods. URL: http://cybermetrics.wlv.ac.uk/AoIRASIST/mayr.html (access date 14 August 2005)

Nançoz, Christian (2004): mEdit – membership function editor for fCQL-based architecture. Master Thesis, URL : http://diuf.unifr.ch/is/studentprojects/pdf/M-2004_Christian_Nancoz.pdf (access date 14 August 2005)

Nicholas, David, et al. (1999): Cracking the code: web log analysis. In: Online & CD-ROM Review, Vol. 23, pp. 263-269.

Nicholas, David; Huntington Paul. (2003): Micro-Mining and Segmented Log File Analysis: A Method for Enriching the Data Yield from Internet Log Files. In: Journal of Information Science, Vol. 29 (5), pp. 391-404.

Thelwall, Mike (2001): Web log file analysis: Backlinks and Queries. In: Aslib Proceedings, Vol. 53, pp. 217-223.

Thelwall, Mike; Vaughan, Liwen; Björneborn, Lennart (2003): Webometrics. In: ARIST, Vol. 39, preprint. URL: http://www.db.dk/lb/2003preprint_ARIST.doc

Bibliometrie als DataMining-Werkzeug in der Naturwissenschaft

Dirk Tunger, Jülich

Abstract

Dieser Aufsatz veranschaulicht, wie mit Hilfe bibliometrischer Methoden grosse Datenmengen ausgewertet werden können. Es soll gezeigt werden, dass ein Informationsmehrwert nicht nur auf der inhaltlichen Ebene zu finden ist, sondern auch auf einer Metaebene interessante Informationen verborgen sind, die mit metrischen statistischen Methoden ermittelt werden können.
Es wird eine Einführung in das Thema Bibliometrie gegeben und an einem Praxisbeispiel gezeigt, wie diese theoretischen Annahmen in die Praxis umgesetzt werden können.

1. Einführung

Problemstellung
Es ist keine Neuigkeit mehr, dass die produzierten Datenmengen immer größer werden. Das Problem: Die Inhalte selbst sind zu einem sehr großen Teil nicht mehr zu bewältigen. Dies bedeutet in der Praxis: von der inhaltlichen Seite ist das Problem nicht zu lösen.

Probleme bei der Lösung
DataMing ist oft als Rettung angepriesen worden, im Arbeitsalltag von Informationsspezialisten in Bibliotheken ist davon aber bisher wenig angekommen. Sind die entwickelten Tools auf der einen Seite zu speziell, verlangen sie zu viel theoretisches Wissen? Sind die DataMining-Theorien auf der anderen Seite zu allgemein? Fehlen nur best practise-Beispiele? Einige Fragen scheinen noch ungeklärt.

Lösungsweg
Der Einsatz von Bibliometrie (http://www.bibliometrie.de) als Werkzeug für die Datenauswertung ist nicht unmittelbar neu, die damit verbundenen Möglichkeiten scheinen aber bisher nur sehr selten genutzt zu werden. Bibliometrie ist ein Wissenschaftszweig, in dessen Mittelpunkt die statistische Auswertung von wissenschaftlichen Veröffentlichungen steht (Forschungszentrum Jülich, Zentralbibliothek, 2003). Der Begriff „wissenschaftliche Veröffentlichungen“ ist dabei sehr weit gefasst. Er bezieht sich nicht nur auf Veröffentlichungen in Zeitschriften, sondern kann auch Bücher, Webseiten oder Patente einschliessen.

Ziel
Ziel ist es, Möglichkeiten aufzuzeigen, wie mit Hilfe der Bibliometrie Informationen gewonnen werden können, die mit konventionellen Methoden des Information-Retrieval nicht generiert werden können.

2. State of the Art in der Bibliometrie

Einsatzgebiete von Bibliometrie
Bibliometrie läßt sich hervorragend einsetzen, um einzelne Wissenschaftsgebiete oder wissenschaftliche Einrichtungen gezielt zu untersuchen (Ball, R; Tunger, D., 2005). Ziel einer solchen Untersuchung kann sein, die zeitliche Entwicklung eines Themas zu verfolgen:

- Wie viele Artikel wurden zu einem bestimmten Thema veröffentlicht?
- Wie hat sich dieses Publikationsverhalten im Laufe der vergangenen Jahre geändert?
- Wie ist die Resonanz auf ein Thema? Welche Änderungen hat es hier gegeben?

Ebenso lässt sich Bibliometrie für die Wissenschaftsevaluation nutzen (van Raan, A., 2004). In diesem Fall erhält man Antworten auf Fragen wie:

- Welches sind die führenden Einrichtung zu einem Thema?
- Wie wurden die Veröffentlichungen einer bestimmten Einrichtung im Vergleich mit thematisch ähnlich ausgerichteten Einrichtungen wahrgenommen?
- Welches sind die wahrnehmungsstärksten Zeitschriften einer Disziplin.

Weitere Informationen zur Bibliometrie im Fachportal www.bibliometrie.de.

Ein Bezugsrahmen muss gewählt werden
Veröffentlichungszahlen oder Zitationszahlen allein sagen recht wenig aus, wenn Sie nicht in einen vernünftigen Bezugsrahmen gesetzt werden. Aussagen können zum Beispiel getroffen werden, wenn der Bezugsrahmen zu thematisch ähnlichen Einrichtungen, zu Ländern oder zu einer ausgewählten Fachöffentlichkeit gewählt wird.

Wissenschaftskommunikation
Ein Wissenschaftler veröffentlicht im Wesentlichen aus zwei Gründen:

- Zur Problemlösung in seiner Disziplin
- Zur Erhöhung der eigenen Reputation. Dies bedeutet, der Wissenschaftler möchte durch seine Ergebnisse auf sich aufmerksam machen und seine neuen Methoden vorstellen.

In seinen Veröffentlichungen zitiert ein Wissenschaftler demzufolge, um seine eigenen Ergebnisse mit den Ergebnissen anderer Wissenschaftler zu untermauern.

Er zitiert aber auch, um auf vorausgegangene Veröffentlichungen (Ergebnisse) hinzuweisen.

Datenbasis für bibliometrische Analysen

Wissenschaftliche Veröffentlichungen existieren in verschiedenen Formen: Unter anderem in Büchern, Konferenzbänden und Aufsätzen in wissenschaftlichen Zeitschriften. Bei der Messung der Zitationshäufigkeit werden allerdings oft nur wissenschaftliche Zeitschriften beachtet. Dies liegt an der Zusammensetzung der Datengrundlage: Eine Datenbank, die unter Wissenschaftlern als Science Citation Index (SCI) bekannt ist, wertet regelmässig etwa 5900 naturwissenschaftliche Zeitschriften aus. Dies ist der einzige multidisziplinäre Zitationsindex, der zusätzlich zu bibliographischen Angaben auch die Zitationen der Veröffentlichungen auswertet.
Aus den sozialwissenschaftlichen Disziplinen kommen noch einmal etwa 1200 Zeitschriftentitel dazu. Das klingt insgesamt viel, ist es aber nicht: weltweit existieren ca. 120.000 wissenschaftliche Zeitschriften aller Disziplinen.
Ausgewertet und für bibliometrische Analysen zu Grunde gelegt werden also gerade einmal 5% der wissenschaftlichen Veröffentlichungen in Zeitschriften. Von den unzähligen Büchern und Konferenzbeiträgen werden nur die allerwenigsten erfasst.
Im Umkehrschluss bedeutet dies: Auch Zitate werden damit nur aus diesen etwa 5% der wissenschaftlichen Veröffentlichungen komplett erfasst.

Standardisierung der Naturwissenschaft

Für die Naturwissenschaften bestehen bei der Datenauswahl keine Probleme, da diese sehr international ausgerichtet sind: Naturwissenschaftliche Themen sind weltweit von Interesse, die Fragestellungen ähneln sich. Kommunikationssprache ist Englisch und der größte Anteil naturwissenschaftlicher Arbeiten erscheint in Form von Aufsätzen in Zeitschriften. Bücher spielen in den Naturwissenschaften nur eine untergeordnete Rolle.
Man kann sagen, in den Naturwissenschaften herrschen weltweit sehr ähnliche Standards. Dies ist ein grosser Vorteil und ermöglicht auch erst internationale Vergleiche.
In den Geisteswissenschaften sieht es hingegen anders aus: Themen sind teilweise nur von eingeschränkter regionaler Bedeutung und oftmals in Nationalsprachen abgefasst. Daraus entsteht ein Problem: Für internationale Journals ist das Interesse an derartigen Aufsätzen gering, vor allem, wenn der Bezug zu den USA fehlt. Damit ist es schwierig, in den internationalen sozialwissenschaftlichen Journals Beiträge unter deutscher Beteiligung zu platzieren.
Für die Sozialwissenschaften sind damit die Möglichkeiten internationaler Vergleiche nur schwer anwendbar. Hinzu kommt, dass Bücher eine wesentlich größere Bedeutung einnehmen als in den Naturwissenschaften.

Bildung von Indikatoren

Mit Hilfe von Indikatoren kann eine große Anzahl an Veröffentlichungen beurteilt werden. Hierbei findet keine qualitative Beurteilung statt, sondern eine quantitative. Ein Indikator kann beispielsweise die Anzahl der Zitate pro Artikel (Zitationsrate) benennen. Bei der Bildung dieses Indikators ist dies weniger für einen Artikel interessant, als vielmehr für ein Set an Artikeln. Dieses Set kann dann zu Vergleichs-Sets in Bezug gesetzt werden. Auf diese Weise entsteht ein Ranking, das aus verdichteten Daten besteht und einen Überblick in der Bewertung der untersuchten Artikel liefert.

Ebenso kann man mittels Bibliometrie Vernetzungen und Interdisziplinarität in der Wissenschaft aufzeigen.

Ziel bei der Bildung von Indikatoren ist es, eine vergleichbare Umgebung zu erzeugen, die Grössenunterschiede wissenschaftlicher Einrichtungen relativiert.

3. Kombination von DataMining und Bibliometrie

„DataMining ist die Gewinnung impliziter, bislang unbekannter und potenziell nützlicher Informationen aus Daten“ (Witten, I.; Eibe, F., 2001).

Für die professionelle Anwendung von DataMining existieren etliche Programme, die Unterstützung bieten sollen. Ein sehr bekanntes Tool ist „WEKA“ von der neuseeländischen Universität Waikato. Die Erzielung von brauchbaren Ergebnissen mit diesen Programmen hängt aber von der Struktur der Daten und den Kenntnissen der Auswertungsalgorythmen dieser Programme ab.

Aber auch Abseits von speziellen DataMing-Tools lässt sich DataMining betreiben: Bibliometrie ist zwar gleich ein kompletter Wissenschaftszweig, er hat aber genau das oben beschriebene Ziel vor Augen: die Gewinnung von Informationen aus Daten. Die Schwierigkeiten sind die gleichen, die auch beim tool-unterstützten DataMining auftreten: Die Aufbereitung von Daten vor der weiteren Analyse ist zeitaufwendig.

Der Einsatz bibliometrischer Methoden lohnt sich vor dem Hintergrund, eine grosse Anzahl an wissenschaftlichen Aufsätzen auf einmal auszuwerten und daraus die gewünschten zusätzlichen Informationen zu ziehen.

Daten zu Information veredeln – so kann man kurz das Hauptziel von DataMining fassen (Grötker, R., 2002).

4. Bibliometrie als Teil eines Trenderkennungssystems

Bibliometrie ist mehr als nur ein Werkzeug, mit dem Wissenschaftsevaluation betrieben werden kann. Bibliometrie kann als Controlling-Instrument in der Wissenschaft auch zur Trenderkennung genutzt werden.

Es muss an dieser Stelle bemerkt werden, dass Bibliometrie natürlich nur einen Teilbereich in einem Trenderkennungssystem bildet. Neben wissenschaftlichen Veröffentlichungen sind Patente und Konferenzveröffentlichungen für die technologische Entwicklung ebenfalls relevant.
Neben technologischer Entwicklung sind in einem Trenderkennungssystem auch noch weitere Ebenen von Bedeutung (Gomez, P.,1983).
Die

- soziale Ebene
- politische Ebene
- ökonomische Ebene und
- technologische Ebene

müssen in einem Trenderkennungssystem eine Einheit bilden (Rieser, I., 1980).

Blick zurück nach vorn
Wie auch an anderer Stelle, steigt in der Welt der Wissenschaft die Zahl an verfügbaren Inhalten (wissenschaftliche Veröffentlichungen). Die Zeit, einzelne Ergebnisse wahrzunehmen, wird immer geringer. Dadurch wird nur ein Bruchteil der erzielten wissenschaftlichen Ergebnisse intensiv gelesen und weiterverarbeitet.

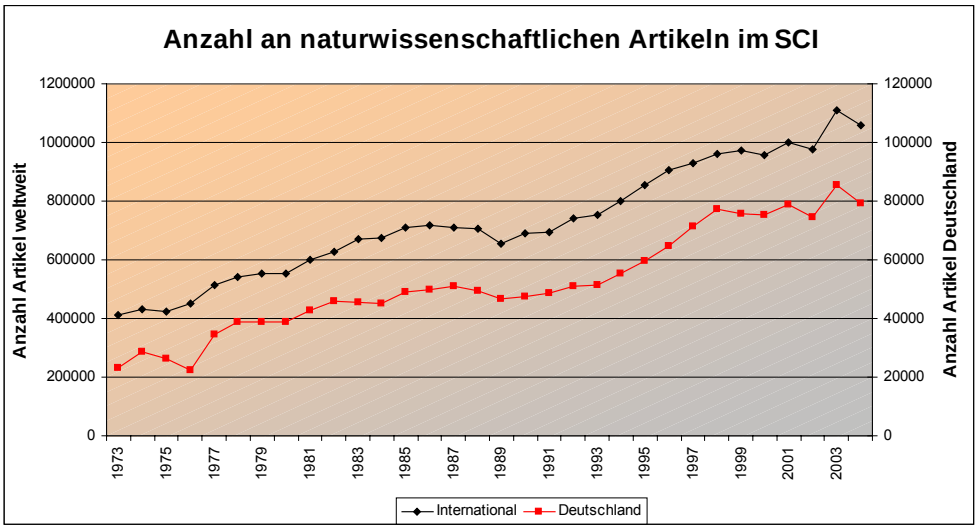


Abbildung 1: Darstellung des zeitlichen Verlaufs der Produktion wissenschaftlicher Artikel (nur im Science Citation Index verzeichnete)

Das Problem liegt darin, einen Überblick über den bisherigen Stand der Forschung zu erhalten. Nur mit einem Überblick ist es aber möglich, ergebnisorientiert und effizient zu forschen. Eine praktisch anwendbare Möglichkeit in der Wissenschaft,

einen Trendscout aufzustellen, der als Trenderkennungssystem fungiert, ist die Durchführung von bibliometrischen Analysen
Mit bibliometrischen Analysen lässt sich beispielsweise die technologische Entwicklung von wissenschaftlichen Disziplinen messen. Die Vorgehensweise ist vom Arbeitsaufwand her vertretbar und vom Ergebnis her sehr aussagekräftig: In der Datenbank „Science Citation Index“ wird zu einem Grossteil wissenschaftlicher Veröffentlichungen die Anzahl der Zitationen verzeichnet.
Aus diesem Grund ist es möglich, zu einer thematisch ausgerichteten Recherche Antworten auf unter anderem drei Fragen zu erhalten:

Vergangenheits-Aspekt: Wie hat sich die Anzahl wissenschaftlicher Veröffentlichungen über einen bestimmten Zeitraum entwickelt?
Gegenwarts-Aspekt: Wie wurden diese Artikel zitiert?
Zukunfts-Aspekt: Gab es bei den Publikationszahlen/ Zitierungen grosse Zuwächse oder Einbrüche?

Der **Vergangenheits-Aspekt** wird gebildet von der *Anzahl der Artikel* in der Datenbank zu einem Thema. Die Betrachtung der Anzahl von Veröffentlichungen schaut in die Vergangenheit, da diese Zahl für den betrachteten Zeitraum nicht mehr zu verändern ist.
Der **Gegenwarts-Aspekt** wird gebildet aus der Anzahl an Zitationen auf die existierenden Artikel. Diese Zahl kann sich täglich ändern, wenn einer der Artikel in anderen Veröffentlichungen zitiert wird.
Der **Zukunftsaspekt** wird gebildet aus den Zuwächsen oder Einbrüchen der Publikationen und Zitationen über einen längeren Zeitraum. Die Veränderung gegenüber einer Vorperiode lässt erkennen, ob das wissenschaftliche Interesse zu- oder abnimmt.
Der Zukunfts-Aspekt ist demnach der wichtigste von den drei Aspekten. Es sind nicht die absoluten Zahlen an Artikeln oder Zitationen, die die Zukunftsaussage tragen, sondern die Veränderung der Zahl an Zitationen im Verhältnis zum Jahr davor.
Man kann einwerfen, dass auch der Zukunftsaspekt auf Zahlen aus der Vergangenheit basiert. Dies ist auch richtig, doch ergibt sich dieses Problem an jeder Stelle, wo von Zukunft die Rede ist. Es ist schlicht unmöglich, Zahlen aus der Zukunft zu erhalten. Aus diesem Grund müssen die ermittelbaren Zahlen in ein Verhältnis gebracht werden, dass sich Zahlen mit Aussagekraft in die Zukunft ergeben. Eine blosse Interpolation wäre wenig sinnvoll: Das Problem ist die Vorhersage von Wendepunkten. Würde man also Interpolation für den Aufbau eines strategischen Radars einsetzen, würde man ein falsches Gefühl der Sicherheit erzeugen.
Die Methode ist für die drei weiteren Ebenen (soziale, politische und ökonomische Ebene) zu übertragen.

Der Vorteil in der vorgestellten Methode liegt darin, immer ganz konkret Entwicklungsträger zu identifizieren. Der zweite Schritt ist die Messung von Resonanz auf diese Entwicklungsträger und die entsprechende Veränderung.
Mit einem einzelnen Indikator kann mit Sicherheit keine Vorhersage von zukünftiger Anwendbarkeit gemacht werden. Der Verbund und die geschickte Kombination mehrerer Entwicklungsträger lässt aber durchaus vernetzte Aussagen zu.

Praxisbeispiel
Die Theorie soll an einem konkreten Beispiel aus der Praxis durchgespielt werden.

Nanotechnologien / Nanomaterialien

Der Begriff geht auf Norio Taniguchi (1974) zurück und beschreibt die Entwicklung von Materialien, die in mindestens zwei Dimensionen kleiner als 100 Nanometer sind. Die Theorie besagt, dass diese Materialien ihre Eigenschaften und ihre Struktur ändern.

Technologische Ebene Die Betrachtung der wissenschaftlichen Artikel zu diesem Thema ist eindeutig: Wurden im Zeitraum 1995 – 1999 weltweit 3190 Artikel zu diesem Thema veröffentlicht, waren es im Zeitraum 2000 – 2004 bereits 9823. Die Zahl der Länder, die sich für dieses Thema interessieren, stieg von 59 auf 80.
China konnte dabei den Anteil seiner Forschung von 9 % auf 17 % fast verdoppeln, während in den USA der Anteil von 34 % auf 28 % zurückging. Deutschland hat ebenfalls einen Rückgang zu verzeichnen, von 13 % auf 10 %. Die Prozentzahlen beziehen sich auf den Anteil der Artikel in der jeweiligen Zeitperiode am weltweiten Output zu diesem Thema. Es hat somit keinen Rückgang der Artikelproduktion oder der Forschung in diesen beiden Ländern gegeben, sondern eine überproportionale Steigerung des Outputs anderer Länder (Weidenfeld, W; Turek, J., 2002).Für einzelne Themenaspekte könnten weitere Untersuchungen angestellt werden, beispielsweise, wie stark das wissenschaftliche Interesse gestiegen ist (gemessen in Form von Zitaten).

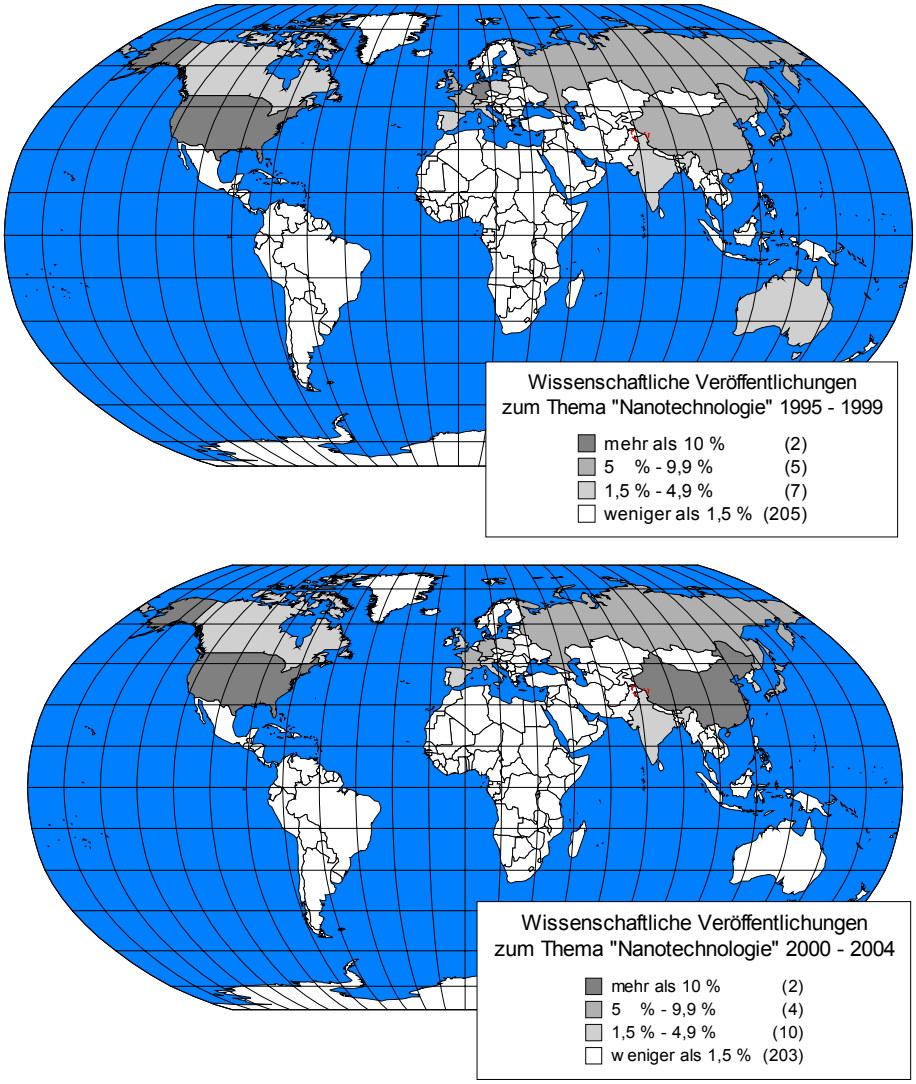


Abbildung 2: Wissenschaftliche Veröffentlichungen zum Thema „Nanotechnologie“. Darstellung des prozentualen Anteils einzelner Länder in zwei Zeitperioden

Ausblick:
Es wurde bereits angedeutet, dass nicht nur die technologische Entwicklung von Bedeutung ist für die Trenderkennung, sondern dass auch weitere Faktoren mit einfliessen müssen. Für die soziale Ebene soll das Beispiel weiter ausgebaut werden.

Soziale Ebene Wie stark wird die Technologie in der Öffentlichkeit wahrgenommen? Wie oft und in welchen Massenmedien wird über Nanotechnologie berichtet?
Der Grafik liegt eine Auswertung der Datenbank GBI (Themenrubrik „Tages- und Wochenpresse“) zu Grunde.
Das Diagramm zeigt, dass eine sehr lange Zeit in der Öffentlichkeit dieses Thema nicht diskutiert wurde. Seit 2000 hat sich dies geändert: Nahezu schlagartig wurde immer öfter und weiter gestreut berichtet. Ein eindeutiges Zeichen für Interesse an diesem Themengebiet.

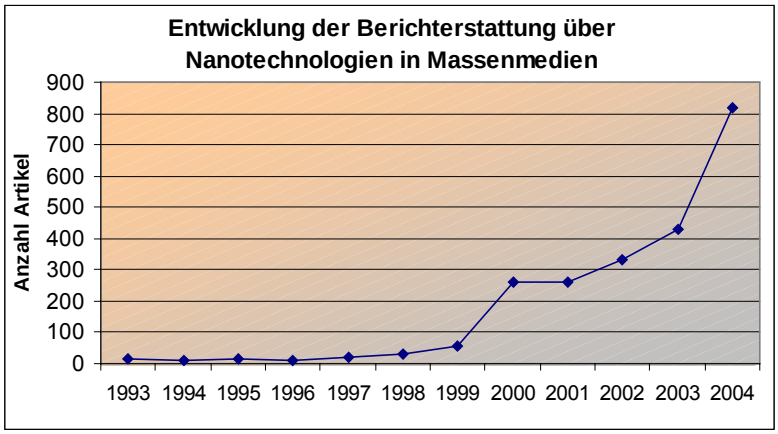


Abbildung 3: Entwicklung der Berichterstattung über das Thema „Nanotechnologie“ in täglich und wöchentlich erscheinenden Massenmedien

Mit metrischen statistischen Methoden können weitere Auswertungen gemacht werden, die Auskunft über folgende Fragen liefern:
Sind Widerstände erkennbar (Bürgerinitiativen oder Vereinsgründungen)? Ist Unterstützung zu erwarten? Derartige Einschätzungen sind durch eine standardisierte Klassifikation und Bewertung von Medienberichten zu erhalten.
Die Betrachtung eines Themas aus verschiedenen Perspektiven hilft, frühzeitig Chancen und Risiken zu erkennen. Dabei reicht eine einmalige Anstrengung nicht aus, in regelmässigen Zeitabständen müssen Veränderungen immer wieder überprüft werden. Die hier vorgestellten Ergebnisse und Methoden sind auch nur als beispielhaft zu verstehen und müssten für ein spezielles Themengebiet weiter konkretisiert werden.
Die Einbeziehung von Experten vereinfacht die Interpretation der Ergebnisse.

Integration von Trenderkennung in die Wissenschaft
In Forschungseinrichtungen sind ein besseres Controlling und eine bessere Steuerung nur über eine stärkere Nutzung bisher ungenutzter Datenbestände zu

erzielen. Die Datenbestände müssen durch Analyse- und Bewertungsverfahren zu Informationen und Wissen umgeformt werden.
Für den Bereich Wissenschaft kann hierzu die bibliometrische Analyse genutzt werden, für den Bereich Gesellschaft/ Soziales kann dies eine Medienresonanzanalyse sein.
Diese Medienresonanzanalyse kann auf Basis ähnlicher statistischer Verfahren die Tages- und Wochenpresse auswerten. Es können Aussagen über den Umfang und die zeitliche Abfolge zu einem Thema gemacht werden. Ebenso kann die Art des Mediums (Reichweite, Meinungsführerschaft etc.) in derartige Analysen miteinfliessen. Eine inhaltliche Bewertung (bezüglich negativer oder positiver Berichterstattung) hilft, die Einschätzung zu verbessern.
Der Bereich Politik spielt mit in den Bereich Gesellschaft hinein und wird teilweise auch über die Massenmedien ausgedrückt. Gesetzgebungsverfahren werden vorher oft in der Presse diskutiert und deren Richtung mit beeinflusst.
Daten zum Bereich Ökonomie werden täglich in riesigen Mengen produziert. Es existieren Daten auf Ebene der Finanzmärkte, aber auch auf ausserbörslicher Ebene. Diese Daten dürfen nicht ohne Beachtung bleiben: was sagen unterschiedliche Experten über die Entwicklung einer Branche aus? Aus welcher Perspektive werden diese Voraussagen getroffen? Welche Daten liegen diesen Bewertungen zu Grunde?

5. Fazit

Es sind immer drei Dinge, die gesucht werden:
1. Träger von Entwicklungen (z.B. wissenschaftliche Veröffentlichungen)
2. Resonanz auf diese Entwicklungsträger (z.B. Zitationen)
3. Veränderung der Resonanz

Kann man entsprechende Datenquellen benennen, die diese drei Faktoren ermittelbar machen, so können entsprechende Indikatoren gebildet werden. Nicht nur das einmalige Erheben der Indikatoren ist der Informationsmehrwert, sondern eine kontinuierliche Beobachtung und eine tiefgehende Interpretation. In letzterem Punkt liegt eine nicht zu unterschätzende Schwierigkeit der Analysen.
In den beschriebenen Methoden wird DataMining nicht in der Form angewandt, dass mit Hilfe von DataMining-Tools Zusammenhänge zwischen den Datensätzen gesucht werden. Vielmehr wird in diesem Fall versucht, eine höhere Aggregationsebene zu erreichen und Daten zu Informationen umzuwandeln, durch das Clustern einer bestimmten Anzahl an Artikeln und das zusammenhängende Auswerten bestehender Datenfelder.

Literatur

Ball, Rafael; Tunger, Dirk: Bibliometrische Analysen – Daten, Fakten und Methoden. Grundwissen Bibliometrie für Wissenschaftler, Wissenschaftsmanager, Forschungseinrichtungen und Hochschulen; Forschungszentrum Jülich, Zentralbibliothek, Reihe Bibliothek Bd. 12, 2005, ISBN: 3-89336-383-1

Forschungszentrum Jülich, Zentralbibliothek: Bibliometric Analysis in Science and Research. Applications, Benefits and Limitations; 2nd Conference of the Central Library; Forschungszentrum Jülich, Zentralbibliothek, Reihe Bibliothek Bd. 11, 2003, ISBN: 3-89336-334-3

Gomez, Peter: Frühwarnung in der Unternehmung. Haupt, Bern, 1983

Grötker, Ralf: Goldgräber in der Datenmine in: Die Zeit 16/2002; http://zeus.zeit.de/text/archiv/2002/16/200216_t-data-mining.xml

Rieser, Ignaz: Frühwarnsysteme für die Unternehmungspraxis. Florentz, München, Wirtschaftswissenschaftliche Forschung und Entwicklung, 1980

van Raan, Anthony: Measuring Science *in*: Moed, H.F.; Glänzel, W; Schmoch, U: Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems; Kluwer Academic Publishers, Dordrecht, 2004, ISBN: 1-4020-2702-8

Weidenfeld, W; Turek, J: Wie Zukunft entsteht. Größere Risiken – weniger Sicherheit – neue Chancen; Gerling Akademie Verlag, München, 2002, ISBN: 3-932425-46-4

Witten, Ian; Frank, Eibe: Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen; Carl Hanser Verlag, München, 2001; ISBN: 3-446-21533-6

Semantische Netze – Wissen professionell organisieren

Technologische Trends beim Einsatz semantischer Netzwerke

Ulrich Bügel, Karlsruhe

Abstract

Erfahrungen beim Einsatz technischer Informationssysteme zur intelligenten Ablage eines Dokumentenbestandes belegen die Notwendigkeit, die eingestellten Inhalte in einem semantischen Netzwerk miteinander zu verknüpfen, um sie so in einen Kontext einbetten zu können. Fortgeschrittene Informationssysteme bieten zur Unterstützung der Vernetzung ein automatisches Beziehungsmanagement auf Basis von Ontologien an. Aus Herstellersicht ist dabei für zukünftige Projekte zu erwarten, dass aufkommende Standards, vor allem im Kontext der Semantic Web Initiative, nicht nur Kompatibilitätsanforderungen stellen werden, sondern auch neuen Technologien zur funktionalen Verbesserung und Automation Vorschub leisten. Um konkurrenzfähig zu bleiben, ist ein schnelles „time to market" erforderlich, was zu einem gehobenen Effizienzanspruch bei der Entwicklung führt und verbesserte Methoden im Engineering erfordert.

1. Einleitung

Der Aufbau technischer Informationssysteme zur intelligenten Ablage eines Dokumentenbestandes (im Sinne des Wissensmanagements) stützt sich meistens auf am Markt verfügbare Werkzeuge wie (Web) Content- oder Dokumentenmanagement-Systeme (CMS bzw. DMS). Am Anfang eines solchen Projektes steht zunächst eine Bestandsaufnahme der Ausgangswissensbasis, welche in Form von Dokumenten vorliegen kann, manchmal aber auch nur in den Köpfen der Wissensträger gespeichert ist. In einem weiteren Schritt wird die Wissensbasis strukturiert, d.h. Wissensträger und Systemanalytiker definieren in einem gemeinsamen Projekt eine Menge von Bausteintypen (sog. „Wissensbausteine"), welche für die Anwendungsdomäne typisch sind (z.B. Dokumente, Personen, Projekte etc.). Für jeden Bausteintyp wird eine Reihe von Attributen zu dessen Beschreibung festgelegt, welche im Layout-Template des Bausteins geeignet dargestellt und platziert werden können und außerdem Gegenstand der Recherche sind. Um konkrete Einträge im Informationssystem erstellen zu können, wird für jeden Bausteintyp eine Eingabemaske zur Erfassung der Attribute erstellt.
Neben einer gezielten Suche über die strukturierten Inhalte werden dem Benutzer auch Navigationshilfen zur Explorierung der Inhalte angeboten. In den meisten Fällen navigiert der Benutzer über eine Hierarchie von Einträgen. In vielen Systemen gibt es

beispielsweise eine „Sitemap“-Funktion, mit der diese Hierarchie angezeigt werden kann, ähnlich der vom Windows-Explorer her bekannten Darstellung des Dateibaumes. Diese Form der Darstellung stößt jedoch gerade bei vielen Anwendungsgebieten technischer Informationssysteme an ihre Grenzen, beispielsweise bei Anwendungen zur wissensintensiven Dokumentation, Forschungsdokumentation oder der Dokumentation von Firmen- und Produktwissen. In solchen Systemen ist eine hierarchisch strukturierte Organisation nicht mehr ausreichend, vielmehr ist der Einsatz semantischer Netzwerke zur Darstellung der Sachverhalte erforderlich, wie im folgenden Kapitel näher ausgeführt wird.

2. Einsatz semantischer Netzwerke

2.1 Grenzen hierarchischer Einordnung

In heutigen (Web-) Content- und Dokumentenmanagementsystemen werden die eingestellten Inhalte meist in eine hierarchische Struktur eingeordnet, die vom Benutzer selbst angelegt werden kann. Diese Form der Einordnung von Einträgen stößt schnell an ihre Grenzen, wie an folgendem einfachen Beispiel bereits deutlich wird:

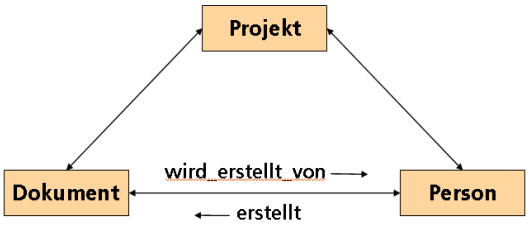


Abbildung 1: Beispiel einer einfachen Vernetzung

Abbildung 1 zeigt drei Wissensbausteine, welche sich nicht in eine Hierarchie einordnen lassen, die dem Beziehungsgeflecht gerecht wird. Personen erstellen Dokumente, Dokumente werden in Projekten erarbeitet und Projekte werden von Personen bearbeitet. Die Darstellung dieses Sachverhaltes durch eine Hierarchie ist nicht möglich – es sei denn durch erhebliche Redundanz, z.B. durch Doppelung von Bausteinen.
Es besteht daher die Anforderung zur Vernetzung der Einträge. In vielen Fällen kann die Semantik der Beziehungen bidirektional beschrieben werden, wie in Abbildung 1 beispielhaft gezeigt. Mit einer steigenden Zahl von Bausteintypen wird das entstehende Netzwerk komplexer.

2.2 Einsatzbeispiele

Erfahrungen in verschiedenen Projekten belegen die Notwendigkeit, Wissensbausteine in einem semantischen Netzwerk miteinander zu verknüpfen, um sie so in einen Kontext einbetten zu können. Die Wissensbausteine wurden jeweils unter Nutzung der hauseigenen CMS-Plattform WebGenesis® (1) als Entwicklungssystem durch programmierte Erweiterungen implementiert. Das Einstellen von Einträgen und deren Vernetzung wurde dezentral von den jeweiligen Wissensträgern selbst durchgeführt.

Kompetenznetzwerk Wissensmanagement:

Ziel war die Aufbereitung und Strukturierung des Themas „Wissensmanagement für Mittelstandsunternehmen in Baden-Württemberg“ (2). Mit den Wissensträgern - Institute aus Baden-Württemberg mit Kompetenz zum Thema Wissensmanagement - wurde ein internetbasiertes Informationssystem konzipiert. Die Ausgangswissensbasis lag in Form von Forschungsberichten sowie Expertenwissen der beteiligten Partner vor. Beispiele für Wissensbausteine sind Produkte, Prozesse, Szenarien, Erfahrungen, Technologien, Methoden, Projekte, Personen, Vorschriften oder Grundlagen. Die Vernetzung dieser Bausteine wurde von den Autoren durch Erstellen von Relationen wie „Technologie ist_realisiert_in Produkt“ oder „Erfahrung ist_über Methode“ realisiert.

Forschungsinformationssystem des BMVBW:

Ziel ist hier die sach- und problemorientierte Aufbereitung der Forschungsergebnisse im Bereich „Verkehr-, Bau- und Wohnungswesen“ sowie „Aufbau Ost“ (3). Wissensträger sind 14 Lehrstühle an deutschen Universitäten mit inhaltlichem und informationstechnischem Koordinator sowie 14 verantwortlichen Redakteuren. Die Ausgangswissensbasis sind Forschungsberichte und Expertenwissen an den beteiligten Lehrstühlen.

Virtuelles Software Engineering Kompetenzzentrum des BMBF:

Hier wurde für Unternehmen ein einfacher und selektiver Zugriff auf Know-how und Erkenntnisse in der Software-Entwicklung bereitgestellt (4). Wissensträger sind Institute mit Erfahrungen auf dem Gebiet des Software Engineering. Die Wissensbasis bestand aus Expertenwissen und verfügbaren Dokumenten.

2.3 Entwicklung semantischer Netzwerke

Die komplexe Vernetzung von Einträgen lässt sich nicht sinnvoll durch Hyperlinks bewerkstelligen, welche fest verdrahtet in die HTML-Darstellung eingefügt sind. Eine solche Verlinkung ist nicht wartbar, erst recht nicht in einem dynamischen CMS mit automatischer Erzeugung der HTML-Darstellung. Es sind daher Möglichkeiten zur Verwaltung der Vernetzung zu schaffen. Technologien, wie sie im Semantic Web (5)

angewendet werden, beispielsweise Ontologien (6), sind dazu hervorragend geeignet.
WebGenesis® bietet zur Unterstützung der Vernetzung ein automatisches Beziehungs-Management auf Basis von – nach Wissensbausteinen konzeptualisierten – Ontologien an. Bei der Erstellung eines Eintrages für einen bestimmten Bausteintyp werden dem Autor zusätzliche Eingabefenster zur Vernetzung mit allen bekannten Zielobjekten angeboten; auch eine Suche unter bekannten Zielobjekten steht zur Verfügung.

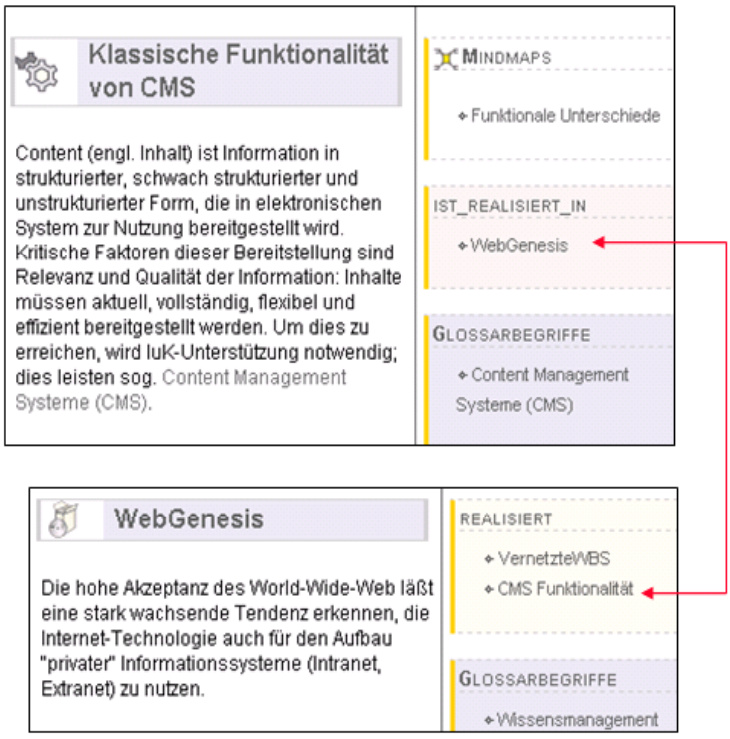


Abbildung 2: Darstellung der Vernetzung

Die erfassten Beziehungen zwischen Einträgen können nun bei der Präsentation der Einträge automatisch mitgeneriert werden. Abbildung 2 zeigt ein Beispiel aus dem Kompetenznetzwerk Wissensmanagement, welches die Vernetzung eines Eintrages vom Typ „Technologie“ („Klassische Funktionalität ...“) mit einem „Produkt“ („WebGenesis“) darstellt.
Die Darstellung der Relation „Ist_realisiert_in“ im oberen Bildschirm indiziert, dass das Produkt WebGenesis® ein CMS ist; per Mausklick auf das Produkt navigiert man zum unteren Bildschirm. Die erfassten Bausteine und Beziehungen können so dem

Benutzer in einer Weise präsentiert werden, welche eine einfache Navigation im vernetzten Informationsbestand zulässt.

3. Zukünftige Anforderungen

Bei allen durchgeführten Projekten wurde zusammen mit den Wissensträgern die Strukturierung der Wissensbausteine erarbeitet. Diese wurden dann als Systemerweiterungen von WebGenesis® implementiert, und schließlich wurden dann die Inhalte über das verteilte Autorensystem dezentral eingestellt und vernetzt.
Obwohl diese Vorgehensweise recht fortschrittlich ist, ist aus Herstellersicht zu erwarten, dass sie bei zukünftigen Projekten kaum mehr wachsenden Ansprüchen genügen wird:

- aufkommende Standards im Bereich des Semantic Web erfordern Anpassungen aus Kompatibilitätsgründen,
- neue Technologien eröffnen Möglichkeiten zur funktionalen Verbesserung und Automation der Vernetzung,
- der gehobene Effizienzanspruch erfordert verbesserte Methoden im Engineering.
- In den folgenden Kapiteln soll dies verdeutlicht werden.

3.1 Standards im Bereich der Semantic Web Initiative

In seinem legendären Artikel (7) definierte Tim Berners Lee das Semantic Web wie folgt:
"The Semantic Web is an extension of the current web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation".
Der Kerngedanke beim Aufbau des Semantic Web besteht darin, automatisch oder semi-automatisch arbeitende Werkzeuge zu entwickeln, welche Web-Seiten, Web-Dienste, Datenbanken etc. mit semantischer Metainformation versehen (annotieren). Die Metainformation kann dann für eine intelligente Suche genutzt und automatisch (d.h. von Anwendungen, Agenten usw.) verarbeitet werden.
Die weltweite Entwicklung des Semantic Web stützt sich auf die Nutzung einer Reihe von Basis-Technologien, welche grundsätzlich auch in geschlossenen oder halboffenen Umgebungen (z.B. Intranets, Extranets) einsetzbar sind. Diese Technologien basieren auf der Nutzung von Ontologien. Dabei kristallisiert sich mit der Web Ontology Language (OWL) ein Standard heraus, welcher vielfältigen Möglichkeiten der Wissensrepräsentation mit unterschiedlichstem Formalisierungsgrad Rechnung trägt (8). Als Resultat dieser Initiative wird es in Zukunft eine Reihe von Fachexperten gemeinsam erstellter und frei verfügbarer Ontologien geben, welche bestimmte Anwendungsdomänen beschreiben. Diese Ontologien können bei

der Erstellung der Konzeptstruktur für die in kommenden Projekten benötigten Wissensbausteine direkt herangezogen oder durch Anpassung genutzt werden.
Dazu muss das Informationssystem in der Lage sein, Ontologien, welche im OWL-Standard formuliert sind, in die interne Darstellung von Wissensbausteinen zu überführen. In WebGenesis® ist dies durch eine OWL-Import/Export-Schnittstelle realisiert.

3.2 Funktionale Verbesserungen

Von der Semantic Web Initiative erwartet man einen Mehrwert vor allem bei der Suche, der Interoperabilität und dem Engineering.

Semantische Abfragen und Suche:

Zukünftige Wissensportale müssen Schnittstellen für formale Abfragen besitzen, welche nicht nur für menschliche Benutzer, sondern auch für Anwendungen oder Agenten zur automatischen Nutzung bereitstehen. Da die Wissensbasis gemäß den Klassen einer Ontologie strukturiert ist, haben auch die Abfragen einen Bezug zu dieser Ontologie; sie beschreibt alle Typen von Wissensbausteinen und deren mögliche Beziehungen (Vernetzung). Im Semantic Web stützen sich ontologiebasierte Abfragen auf eine Wissensbasis mit Beschreibungen von Resourcen (z.B. externe Dokumente, Web-Seiten) in der Sprache RDF (Resource Description Framework) (9), einer Untermenge der Ontologiesprache OWL. Damit das Informationssystem verfügbares externes RDF-Wissen in die Suche einbeziehen kann, muss dieses sich auf die in der Ontologie definierten Klassen beziehen und kann mit Hilfe geeigneter Reader-Programme in die interne Wissensbasis eingebracht werden (Abbildung 3, linker Teil).
In vielen Fällen kann der Benutzer allerdings anfangs gar keine exakte Abfrage formulieren, vielmehr kann er das wonach er sucht, nur unscharf durch bestimmte Begriffe beschreiben. Er möchte – wie in heutigen Informationssystemen und im Web üblich – Suchbegriffe eingeben, und seine Suche nach und nach verfeinern. Darüber hinaus möchte er nicht nur Wissen über Dokumente abfragen, sondern das vollständige Dokument erhalten. Bei dieser konventionellen Schlagwortsuche werden heute Verfahren des „Information Retrieval“ (10) eingesetzt: die Suchbegriffe werden in einer invertierten Liste (dem Suchindex) abgelegt, welcher für jeden Begriff eine Referenz zur Fundstelle mitspeichert (Abbildung 3, rechter Teil). Ziel einer qualitativ verbesserten Suche ist eine kleinere, präzisere Trefferliste und die Einbettung der Suchergebnisse in ihren semantischen Kontext (Beziehungsgeflecht). Dazu kann beispielsweise vorhandenes RDF-Wissen über Resourcen ausgenutzt und in den Suchindex eingearbeitet werden. In Abbildung 3 übernimmt diese Aufgabe ein sog. semantischer Crawler. Alternativ dazu kann der Crawler das benötigte RDF-Wissen auch selbst erzeugen (siehe Kapitel 3.3).

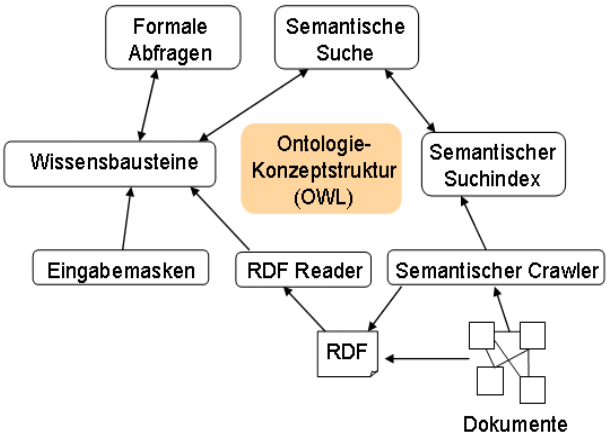


Abbildung 3: Ontologiebasierte Wissensbasis

Interoperabilität:

Der wachsende Bedarf an Informationen löst mehr und mehr die traditionellen Grenzen zwischen Anwendungsdomänen auf. Im Bibliothekswesen stellt beispielsweise die übergreifende Zusammenarbeit zwischen Bibliotheksformen unterschiedlicher Art (z.B. Bibliotheken, Archive, Dokumentationseinrichtungen) und Funktion (Universalbibliothek, Fachbibliothek, Multimedia-Bibliothek, Museum etc.) auch über Landesgrenzen hinweg eine neue Herausforderung dar. Eine semantische Interoperabilität ist dabei grundsätzlich auf der Daten- und der Diensteebene möglich:

- Ontologien können als Zwischensprache zum Austausch von Daten zwischen Informationssystemen genutzt werden. Die Systeme arbeiten dann nicht nur auf syntaktischer Ebene - z.B. über gemeinsame Austauschformate - zusammen, sondern haben auch ein gemeinsames Verständnis von Inhalten. Ontologien für unterschiedliche Domänen werden von den jeweiligen Fachexperten unabhängig voneinander entwickelt; sie legen dazu die identifizierbaren Begriffe, Merkmale, Relationen und Axiome in einer formalen, maschineninterpretierbaren Form fest. Diese können dann mit Hilfe von „Mapping"-Werkzeugen aufeinander abgebildet werden. Semantische Interoperabilität bedeutet beispielsweise, dass unterschiedliche Begriffe mit gleicher Bedeutung ebenso richtig verstanden werden wie gleiche Begriffe mit unterschiedlicher Bedeutung.
- Mit Hilfe von „Semantic Web Services" (11) können Diensteplattformen erstellt werden, welche nicht nur die semantische Suche nach Diensten, sondern auch deren Komposition und Mediation ermöglichen.

Die neue Generation von Informationssystemen muss eine leistungsfähige, funktional hochwertige und generische Daten- und Dienste-Infrastruktur bereitstellen,

um fachbezogene Anwendungsdienste zu ermöglichen, auch über Domänengrenzen hinweg.

Engineering:
Ein semantisches Netz ist die Grundlage zur Spezifikation einer Ontologie: die Knoten beschreiben die Wissensbausteine mit ihren Attributen, die Kanten beschreiben die Vernetzung der Bausteine untereinander. Abbildung 4 zeigt ein Beispiel mit den vernetzten Wissensbausteinen Person, Literatur und Technologie.

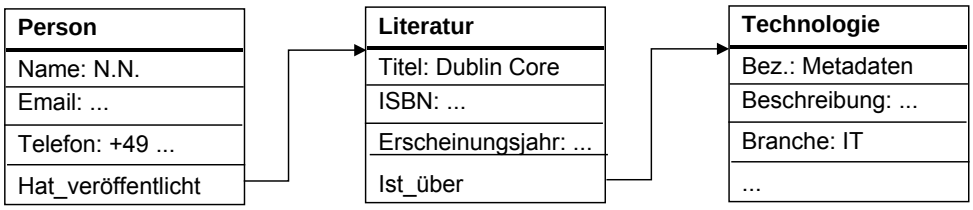


Abbildung 4: Beispiel vernetzter Wissensbausteine

Besonders bei großen Ontologien kann man es sich häufig nicht leisten, alle Klassen von Wissensbausteinen explizit durch Attribute zu beschreiben und in eine Taxonomie (das Vererbungsschema einer Ontologie) einzuordnen. Mit Hilfe der Beschreibungslogik (12) von Ontologien kann ein Großteil der Klassen als sog. abgeleitete Klassen beschrieben werden. Will man beispielsweise festlegen, was ein „Metadaten-Experte“ ist, so definiert man keine neue Klasse, sondern beschreibt mit Hilfe logischer Ausdrücke eine Rolle: Metadaten-Experten sind beispielsweise Personen, welche Literatur über eine Technologie mit der Bezeichnung „Metadaten“ veröffentlicht haben. Zur Klassifizierung eines Metadaten-Experten können auch weitere solcher Beschreibungen herangezogen und ggf. gewichtet werden. Die Semantic Web Sprache OWL unterstützt den Ontologieentwurf mit Beschreibungslogiken.
Diese Technik der logischen Beschreibung kommt einer natürlichen Beschreibung des Sachverhalts nahe („Expertenwissen“), allerdings muss berücksichtigt werden, dass Kenntnisse über die Umsetzung des Expertenwissens in eine formale Beschreibungslogik von der Mehrheit der Autoren nicht „mitgebracht“ werden. Ähnlich wie bei der Eingabe von Regeln in einem KI-System (Künstliche Intelligenz) muss die Umsetzung von einem Wissensingenieur vorgenommen werden, der nach Angaben des Autors das Expertenwissen erfasst und mit formaler Logik beschreibt.
Je nach Formalitätsgrad eignen sich Ontologien somit zum Systementwurf (Engineering). Ein Teil der Applikationslogik kann unabhängig von der jeweiligen Anwendung durch logische Ausdrücke beschrieben werden und durch Standardkomponenten (sog. „Reasoner“ oder „Inferenzmaschinen“), welche in die Applikation eingebunden werden, verifiziert und ausgeführt werden. Damit reduziert sich der „fest

verdrahtete“ Teil von Anwendungen. Gegenüber bekannten Modellen des Software-Entwurfs wie z.B. Entity-Relationship (ER)-Modellen (13) bieten Ontologien wesentlich mächtigere Möglichkeiten zur logischen Beschreibung bestimmter Sachverhalte.
Der Einsatz deskriptiver Logik für den Entwurf gestattet zwar eine präzise und eindeutige Beschreibung von Wissen unabhängig vom Design und der Implementierung eines Informationssystems. Er bringt andererseits ein neues Entwurfsparadigma mit hoher Einstiegsschwelle mit sich und ist daher im konkreten Fall auf Machbarkeit und Sinnfälligkeit zu prüfen.

Einbezug von Diensten:
Inhalte werden in Zukunft nicht nur durch Daten, Texte usw. zur Verfügung gestellt, sondern durch Dienste angeboten. Das Spektrum reicht dabei von Diensten zur Behandlung spezieller Formate und Medien (z.B. Informationsgewinnung aus Bildern, Filmen usw.) über Vermittlungsdienste (z.B. semantische Kataloge) bis zu anwendungsbezogenen Diensten. Die systemübergreifende Nutzung von Diensten zur Informationsbeschaffung erfordert eine Beschreibung der Dienste mit semantischen Metadaten, wie sie z.B. in aufkommenden Standards wie OWL-S (OWL for Services) (14) oder WSMO (Web Service Modelling Ontology) (15) definiert werden. Die Nutzung von Diensten bei der Inhaltsbeschaffung wird daher in Zukunft die Integration solcher Service-Ontologien erfordern.

3.3 Automation

In den bisher auf Basis der Plattform WebGenesis® durchgeführten Projekten (siehe Kapitel 2.2) wurden semantisch vernetzte Wissensbausteine durch programmierte Erweiterungen des Basissystems implementiert. Die Beziehungen zwischen Wissensbausteinen wurden aufgrund von Kenntnissen der Inhalte durch die Autoren definiert und mit Hilfe der Eingabemasken in die Ontologie eingefügt.
Diese Vorgehensweise ist in vielen Fällen zwar ausreichend, es zeichnet sich aber ab, dass für zukünftige Projekte ein schnelleres „time to market“ erforderlich sein wird. Um in Zukunft konkurrenzfähig zu bleiben, müssen für die inhaltliche Kategorisierung und die typisierte Vernetzung alle Möglichkeiten zur Automatisierung ausgelotet werden:

Erweiterbarkeit durch Konfiguration:
Wissensbausteine müssen in Zukunft direkt aus einer in OWL spezifizierten Ontologie heraus generiert werden können. Die Generierung umfasst das Erzeugen des Programmcodes und der zugehörigen Erfassungsmasken.

Klassifikation von Dokumenten:
Bisher werden einzustellende Inhalte vom jeweiligen Autor einem Wissensbausteintyp zugeordnet. Da diese Inhalte oftmals bereits in Form von Dokumenten vorliegen,

kann diese Klassifikation auch durch das Informationssystem selbst automatisch durchgeführt werden, indem es durch Analyse des Dokumenttextes den Wissensbausteintyp bestimmt.

Implizit vorliegende Beziehungen:
Auch der semantische Kontext einzustellender Inhalte kann durch automatische Analyse bestimmt werden, so dass die Autoren bei der Erstellung der Beziehungsgeflechtes unterstützt werden können. Die Analyse kann sich dabei auf Beziehungstypen stützen, welche in der Ontologie spezifiziert sind oder unabhängig von der Ontologie die „semantische Nähe" von Inhalten herausfinden und ggf. Vorschläge zur Erweiterung der Ontologie schaffen.

Hauptaugenmerk aktueller Forschungsarbeiten im Kontext des Semantic Web ist die automatische Erzeugung von Wissen aus allen im Unternehmen vorzufindenden Dokumenten unterschiedlichster Art. Um Wissen automatisch aus unstrukturierten Quellen ableiten zu können, werden Technologien zur Verarbeitung natürlicher Sprache eingesetzt, z.B. statistische Merkmalsextraktion, Abhängigkeitsanalyse, Sequenzanalyse usw. Ein Werkzeugtyp, der gegenwärtig in verschiedenen Open Source Projekten entwickelt wird, sind ontologiebasierte Annotationstools (16), welche Ontologien dazu nutzen, Resourcen automatisch durch RDF mark-up zu beschreiben, und dieses Wissen über das annotierte Dokument der Wissensbasis hinzufügen. Annotierbare Resourcen sind dabei unstrukturierte Dokumente (z.B. Text in Webseiten), Metadaten (z.B. Such-Indexe) oder auch strukturierte Wissensquellen (z.B. Datenbanken, Anwendungen), bei denen das Wissen erst durch Abfragen der gesamten Datenbasis ermittelt werden muss.
Die Vision, die man mit der automatischen Annotation verbindet, besteht darin, dass der Bearbeiter eines Dokumentes sich letztendlich gar nicht damit abgeben muss; vielmehr arbeitet er ausschließlich an seinem Dokument, während ein im Hintergrund laufendes Annotationswerkzeug beim Speichern das Wissen extrahiert, annotiert und der Wissensbasis hinzufügt.

4. Ausblick

Zusammenfassend kann festgestellt werden, dass semantische Netzwerke die Grundlage für künftige Systeme zur Verwaltung von Inhalten bilden werden. Es gibt zwar bereits Produkte, die auf dieser Architektur aufsetzen, die technischen Möglichkeiten sind jedoch bei weitem noch nicht ausgereizt. Strategische Initiativen wie das Semantic Web der W3C leisten dem Einsatz von Methoden Vorschub, welche bisher meist nur Gegenstand von Forschungsprojekten waren und nur begrenzt Zugang zu industriellen Produkten gefunden haben.

Die Integration neuer Technologien in Produkte der Informationsverarbeitung lässt Trends erkennen, welche die Vision einer ganzheitlichen Sicht auf alle Informationsbestände in Unternehmen und Organisationen der Wirklichkeit näher bringen. Dazu gehört beispielsweise die Zusammenführung von Inhalten unterschiedlichster Herkunft und Formate zur Ausgabe auf beliebigen Medien in automatisierter Form: Inhalte werden nach ihrer Erfassung am Arbeitsplatz automatisch dort „abgeholt" und ihrer geplanten Verwendung zugeführt. Diese besteht nicht nur in der Präsentation der Inhalte, sondern kann auch deren Auswertung und Aufbereitung in einer für den Adressaten personalisierten Form (z.B. als Web-Grafiken) umfassen. Die Automation aller organisatorischen Prozesse, mit denen Wissen nutzbar gemacht wird, unterstützt den Übergang von der Informationsverarbeitung zur Wissensverarbeitung.
Die Herausforderung bei der Realisierung dieser Ziele besteht in der Auflösung heterogener Strukturen. Dabei leisten Ontologien aus technologischer Sicht die entscheidende Hilfestellung.

Literatur

Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, 1999.

Berners Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web. Scientific American, 5 (284) – 2001, S. 34-43

Bonn, G.; Kaiser, F.: Das Forschungsinformationssystem des Bundesministeriums für Verkehr, Bau- und Wohnungswesen (BMVBW). Fraunhofer IITB Jahresbericht 2003

DARPA Agent Markup Language (DAML) Homepage: DAML for Services. http://www.daml.org/services/owl-s

DARPA Agent Markup Language (DAML) Homepage: Semantic Web Services. http://www.daml.org/services/

Fraunhofer Institut IITB: Informations-, Wissens- und Community-Management mit WebGenesis®. 2005. http://www.iitb.fhg.de/?2223

Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 1993

Nardi, D. & R.; Brachman, J.: An Introduction to Description Logics. In: The Description Logic Handbook. Baader et al. Cambridge University Press, 2003

OntoWeb Konsortium: A survey on ontology tools. Deliverable 1.3, 2002. http://www.ontoweb.org/download/deliverables/D13_v1-0.zip

Pi-Shan Chen, P.: The Entity-Relationship Model--Toward a Unified View of Data. In: ACM Transactions on Database Systems 1/1/1976 ACM-Press ISBN 0362-5915, S. 9-36

VSEK Konsortium: Software-Engineering-Wissensdatenbank. 2001-2005. http://www.software-kompetenz.de

WiMan Konsortium: Kompetenznetzwerk Wissensmanagement Baden-Württemberg. 2004. http://www.wiman-bw.de

World Wide Web Consortium (W3C®): OWL Web Ontology Language. http://www.w3.org/TR/owl-features

World Wide Web Consortium (W3C®): Resource Description Framework (RDF). http://www.w3.org/RDF/

World Wide Web Consortium (W3C®): Semantic Web. http://www.w3.org/2001/sw/

WSMO Working Group: The Web Service Modelling Ontology. http://www.wsmo.org/

RESIST – Regensburger Signalpfad Informationssystem

Rainer Hammwöhner, Rainer Straub; Regensburg

Abstract

In diesem Beitrag wird ein Projekt vorgestellt, in dem ein Signalpfadinformationssystem entwickelt werden soll, das neue Formen der Wissenschaftskommunikation erlauben soll. Weiterhin soll eine differenziertere Beschreibung von Signalpfaden durch Mehrebenenannotation und Berücksichtigung von Körper oder Zellkompartimenten ermöglicht werden.

1. Einleitung

In diesem Beitrag soll ein im Aufbau befindliches Forschungsvorhaben beschrieben werden, das in Kooperation von Informationswissenschaftlern und Klinikern der Universität Regensburg durchgeführt wird. Ziel dieses Projekts ist die Entwicklung eines Signalpfadinformationssystems neuen Typs, das einerseits einen erweiterten Kreis an Phänomenen erfassen kann und anderseits Instrumente zur Wissenschaftskommunikation zur Verfügung stellt.

Das Auffinden von Signalpfaden[1] ist ein aktives Forschungsgebiet in Biologie, Biochemie und Medizin. Signalpfaddatenbanken dokumentieren die Ergebnisse dieser Forschung. Der Schwerpunkt wird derzeit auf die Erforschung und Dokumentation intrazellulärer Prozesse gelegt. Prozesse auf extrazellulärer und organverbindender Ebene bleiben weitgehend unberücksichtigt, so dass ein Brückenschlag zwischen biochemisch-molekularbiologischer und klinischer Forschung unterbleibt. Wichtige Forschungsergebnisse etwa aus Immunologie oder Endokrinologie können nicht einfach mit den Datenbankinhalten in einen Zusammenhang gebracht werden.

An dieser Stelle setzt das hier skizzierte Forschungsvorhaben an. Zunächst werden Instrumente für die integrierte Beschreibung der aus Zellforschung, Immunologie, Endokrinologie usw. erwachsenden Forschungsergebnisse geschaffen. Dazu sind vorhandene Ontologien zu sichten, zu erweitern und zu integrieren. Beschreibungsformate für intrazelluläre, interzelluläre und gesamtorganismische Phänomene werden entwickelt. Da die Bedeutung von Signalmolekülen je nach Zuordnung zu einem Kompartiment (Körper-, Organ-, oder Zellsegment) wechseln kann, wird eine Erfassung dieser Strukturen erforderlich. Um

den im interdisziplinären Kontext jeweils wechselnden Anforderungen an den Abstraktionsgrad der Beschreibung und die Detaillierung der Modellierung gerecht zu

[1] Signal Transduction Pathways, Metabolic Pathways, Biochemical Pathways

werden, wird die Möglichkeit einer mehrschichtigen Annotation der Daten vorgesehen. So können auch Beschreibungslücken oder Inkongruenzen in der Forschung zwischen den Teildisziplinen abgebildet werden. Mechanismen zur strukturorientierten Recherche in den Datenbeständen und zur Visualisierung der Ergebnisse sind vorgesehen.
Insbesondere in dynamischen Wissenschaftsgebieten stellt der lange Publikationsweg ein hohes Hemmnis für die effiziente Wissenschaftskommunikation dar. Um hier einen Ausweg zu schaffen, wird RESIST mit einer offenen Kommunikationsplattform versehen. Hier können Wissenschaftler Forschungshypothesen und Experimentaldesigns zur Debatte stellen. Es ist wichtig, dass in der Debatte eine exakte Bezugnahme auf die repräsentierten Strukturen ermöglicht wird. Als Folge dieses offenen Charakters wird das Informationssystem Fakten und Hypothesen unterschiedlicher Validität und Reichweite enthalten. Dieser Umstand wird durch die Zuweisung differenzierter Qualitätsmerkmale berücksichtigt.

2. Stand der Wissenschaft und Technik

Im folgenden soll ein kompakter Überblick über die aktuelle Entwicklung in den relevanten Forschungsgebieten vermittelt werden.

2.1 Forschung im Bereich der Signalpfade

Hinsichtlich der zielgruppenspezifischen Erstellung, Aufbereitung und Verbreitung von Informationsinhalten konzentrieren sich die Bemühungen zur Zeit besonders auf intrazelluläre Signalpfade, wie sie für Molekularbiologen und Biochemiker nützlich sind (Signalkaskaden, Signaltransduktionspfade, Reaktionszyklen). Derartige Signalpfad-Datenbanken sind auf der Homepage
„http://home.comcast.net/~natgoodman/Pathway_Web_Sites.htm“
einsehbar. Als Beispiel für ein Signalpfad-Datenbanken hoher Qualität sei hier die Datenbank des *Kyoto Encyclopedia of Genes and Genomes* (KEGG) genannt. Trotz des unbestreitbaren großen Informationsangebotes derartiger Datenbanken bestehen dennoch Nachteile:

1. In den meisten Fällen sind medizinische Themen, die den gesamten Organismus im Blickpunkt haben, von den Betrachtungen ausgeschlossen. Die Kluft zwischen Biochemie/Molekularbiologie (*Genomics* und *Proteomics*) einerseits und klinischen Fragestellungen andererseits wird aufgrund der unterschiedlichen Abstraktionsebenen größer, obwohl genau das Gegenteil wünschenswert wäre. Zwischen Krankheitssymptomen und zellulären Vorgängen wird die stringente Verknüpfung vermisst.

2. In allen uns bekannten Fällen beschreiben die vorhandenen Signalpfad-Datenbanken intrazelluläre Vorgänge ohne Berücksichtigung von Kompartimenten[2]. Man betrachtet die Signalpfade dahingehend, als ob sie in einem einzigen Kompartiment vorhanden wären. Damit wird nicht erfasst, dass ein Signalmolekül in verschiedenen Kompartimenten unterschiedlichen Wirkung haben kann.
3. Die Datenbanken sind nicht für die Nutzer offen, d.h. Nutzer können keine Eingaben vornehmen, um eigene Forschungsergebnisse zur Debatte zu stellen.

Diese Nachteile machen die bisherigen Signalpfad-Datenbanken zu einer Informationsquelle für Biochemiker, Molekularbiologen, Genomiker und Proteomiker, nicht aber für gesamtorganismisch orientierte Pharmakologen, Physiologen, Pathophysiologen, klinisch tätige Ärzte und Studenten der Biologie/Humanmedizin. Die Themen sind zu speziell und zellorientiert. Die Verknüpfungen zwischen gesamtorganismischen Symptomen (z.B. Kopfschmerz) und zellbasierter Forschung sind nicht abgebildet. Hier soll das anvisierte Projekt mit Hilfe einer neuen Form der Signalpfad-Datenbank in einem Internet-basierten, offenen, digitalen Informationssystem Abhilfe schaffen. Hierbei ist es ein Hauptinteresse der Antragsteller, die vorhandenen intrazellulär orientierten Signalpfad-Datenbanken mit dem neuen System zu verknüpfen. RESIST kann insofern als integrierende Metaebene über existierenden Datenbanken betrachtet werden, aus welchen via RESIST Fakten abgerufen werden können. Insofern ersetzt RESIST keinesfalls die bekannten Datenbanken, sondern integriert das neue offene Konzept mit den bekannten Faktendatenbanken. Bei der Entwicklung von RESIST sind verfügbare Open-Source-Projekte im Bereich der Signalpfad-Datenbanken – etwa die Reactome-Datenbank (Joshi-Tope et al. 05) – zu berücksichtigen. So wird einerseits der Entwicklungsaufwand reduziert, die Kompatibilität zwischen bestehenden Ansätzen andererseits erhöht.

2.2 Dokumentation von Signalpfaden

Als Grundlage für die Dokumentation der Signalpfade werden Ontologien – im Sinne der Spezifikationen des Semantic Web – angesehen, wie sie von Forschergruppen und Fachverbänden vorgeschlagen werden (vgl. auch Lewis 2005). Vorschläge für Ontologien in Biologie, Bioinformatik und Medizin liegen vor[3]. Es sind bereits mehrere Ontologien verfügbar[4], die den relevanten Gegenstandsbereich zumindest partiell abdecken. Diese sind gegebenenfalls zu erweitern und an den Bedarf des konzi

[2] Je nach Betrachtungsebene umgrenzte Räume einer Zelle, eines Organs oder des Organismus.

[3] Zusammengetragen etwa unter dem Titel *Open Biological Ontologies* (http://obo.sourceforge.net/, zitiert 28.1.05),

[4] Etwa aus dem Gene Ontology Project (http://www.geneontology.org/, zitiert am 28.1.05), zu berücksichtigen ist aber auch die MeSH-Terminologie, das kontrollierte Vokabular von PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh, zitiert am 28.1.05).

pierten Informationssystems anzupassen[5]. Insbesondere sind funktionelle Aspekte zu berücksichtigen - vgl. (Takai-Igarashi, Mizoguchi 2003) oder (Smith et al. 2005 a) – aber auch einfache anatomische Zusammenhänge müssen zur Repräsentation der Kompartimente berücksichtigt werden (Smith et al. 2005b). Eine Basis für diese Erweiterung kann durch eine Einbettung in die Unified Medical Language geschaffen werden. Die Möglichkeit der Abbildung der Gene Ontology auf UMLS wurde bereits demonstriert (Lomax, McCray 04). Einen Überblick über verschiedene Datenmodelle für die Repräsentation von Biochemical Pathways vermitteln (Deville et al. 03). Ein objektorientiertes Datenmodell für eine objektorientierte Datenbank stellt (Schacherer 01) vor, und gibt damit wertvolle Hinweise für die Realisierung derartiger Systeme.

Auch für die Erfassung von Qualitätsinformationen über die im System verwalteten Fakten – wissenschaftliche Evidenz usw. – muss eine Ontologie bereitgestellt werden. Hier ist auf den von (Karp 04) publizierten Ansatz zu verweisen.

Auf der Basis der zur Verfügung stehenden Ontologien können dann Signalpfade beschrieben, bzw. Publikationen über Signalpfade inhaltlich deskribiert werden. Dazu werden heute bereits Annotationsschemata vorgeschlagen. Hinsichtlich der Beschreibung von Signalpfaden auf unterschiedlichen Ebenen biologischer oder medizinischer Phänomene (intrazellulär, interzellulär, organbezogen, gesamtorganismisch) sind Verfahren der Mehrebenenannotation zu berücksichtigen. Manche der bereits für die Biologie vorgeschlagenen Annotationsschemata schließen Mehrebenenannotationen ein, wobei hier zumeist Abstraktionsebenen unterschieden werden (Battistella et al 2004, Paek et al 2004). Darüber hinaus sind in der Sprachwissenschaft, insbesondere in der Teildisziplin Texttechnologie entwickelte Annotationsverfahren (s. 2.5) zu berücksichtigen.

Hohe Bedeutung kommt bei diesen Aktivitäten der Datenerfassung zu. Damit der aktuelle Forschungsstand angemessen wiedergegeben werden kann, muss eine große Anzahl von Altpublikationen erschlossen, sowie der umfängliche Strom der Neupublikationen mit hoher Vollständigkeit abgedeckt werden. Eine Lösung für das Problem des Erschließungsaufwands wird vielfach in einer semi-automatischen Analyse der Texte gesehen, der eine automatische Faktenextraktion zu Grunde liegt (etwa Koike et al. 04; Karopka et al 04). Die eingesetzten Verfahren sind zumeist in der statistischen Sprachverarbeitung oder in der Erkennung von Satzmustern begründet. Diese Verfahren könnten auch zur Neuindexierung von Altpublikationen herangezogen werden. Es ist jedoch zweifelhaft, ob diese Ansätze zu ausreichend verlässlichen Ergebnissen führen.

Gegenstand der biochemischen Forschung sind, wie bereits oben skizziert, zumeist intrazelluläre oder zumindest vergleichsweise lokale Prozesse. Diese Ausrichtung der Forschung spiegelt sich in der Struktur und Ausarbeitungstiefe der verwendeten Ontologien wider. Dem Kliniker wird so der Zugang zu Forschungsergebnissen, die

[5] Einen Überblick über die Entwicklung und Fusion von Ontologien geben (Ding, Foo 02a und 02b).

auch für ihn relevant sein können, erschwert. Hier ist eine Brücke zwischen den beobachteten Phänomenen und verwendeten Sprachen zu finden.
Die derzeit verfügbaren Signalpfaddatenbanken werden primär durch Erschließung von begutachteten Fachpublikationen aufgebaut. Damit ist das Problem der Validität der aufgenommenen Daten – sieht man von möglichen Erschließungsfehlern ab – auf die Fachgutachter der Publikationsorgane delegiert. Als Folge übertragen sich die Probleme der traditionellen Wissenschaftskommunikation – lange Vorlaufzeiten, z.T. schwer nachvollziehbare Begutachtungsprozesse – auf dieses vom Potential her schnellere und demokratischere Medium.

2.3 Abfragesprachen für Signalpfade

Für die komplexen Graphstrukturen, die als Repräsentationen von Signalpfaden entstehen werden, muss eine adäquate Möglichkeit der Informationssuche geschaffen werden. Während zahlreiche Datenbanken bisher allein konventionelle Methoden der Abfrage – via SQL o.Ä. – anbieten, sind derzeit mächtigere Abfragesprachen in Entwicklung und Erprobung, die auf Verfahren des Graphenabgleichs beruhen (Sohler, Zimmer 2005; Pinter et al. 2005). Diese können auf allgemeine Methoden der Graphensuche zurückgreifen, für die bereits etablierte Werkzeuge zur Verfügung stehen (Giogno, Shasha 2002).

2.4 Visualisierung von Signalpfaden

Der graphischen Darstellung von Signalpfaden kommt eine hohe Bedeutung für das Verständnis dieser komplexen Strukturen zu. Ohne ein quasi-räumliches Modell fällt es schwer, hier einen Überblick zu wahren. Ein Pilotprojekt stellte die Visualisierung der Boehringer Poster dar, eine umfassende Darstellung biochemischer Reaktionszyklen, für die im Rahmen eines vom BMBF geförderten Projekts eine CD-basierte Fassung erstellt wurde (Kanne et al. 99). Die entwickelten Verfahren sind heute im System BioPath verfügbar (Schreiber 02). Mittlerweile ist eine Vielzahl von Signalpfaddatenbanken mit Visualisierungswerkzeugen ausgerüstet, weitere Visualisierungstools sind unabhängig von einer Datenbank verfügbar. Sie beruhen – wenn nicht ohnehin von einer intellektuellen Aufbereitung ausgegangen wird – auf graphbasierten Layoutverfahren – vgl. (Becker, Rojas 2001), (Goesmann et al. 2002), (Krieger et al. 2004).

2.5 Web Services

Es ist offensichtlich, dass das geplante Projekt zahlreiche Anknüpfungspunkte zu anderen Aktivitäten innerhalb der Erforschung biologischer Signalpfade aufweist. Um hier eine gute Anbindung zu erreichen sollen auch weitere Informationsdienstleistungen aus dem Web mit den innerhalb von RESIST verwalteten Daten genutzt werden können. Zu berücksichtigen sind netzbasierte Dienstleistungen, die über gut definierte Schnittstellen als Webservices angeboten werden. Zu denken ist an Dokumentenzugang (Ghandeharizadeh et al. 02), automatische Texterschließung

(Vargas-Vera et al. 02) , Auswertung biologischer Modelle (Rahman et al. 04), Visualisierungen (Rojdestvenski 03) etc.
Hier ist die Systemarchitektur[6] so anzulegen, dass bestehende Dienstleistungen gut integriert werden können. Die Anforderungen an die Schnittstellen müssen berücksichtigt werden, ebenso wie die eventuell erforderliche Anpassung der Inhaltsrepräsentationen.

2.6 Texttechnologische Ansätze

Aus der sich abzeichnenden Notwendigkeit heraus, die Daten auf mehreren Abstraktionsstufen zu beschreiben – etwa biochemische und gesamtorganismisch physiologische Ebene oder quantitative (vgl. Sivakumaran 2003) vs. qualitative Beschreibung – erhebt sich die Frage nach der angemessenen Gestaltung und gegenseitigen Bezugnahme der entsprechenden Deskriptoren. Eine ähnliche Problematik stellt sich in der Linguistik, wenn es um die Beschreibung von Sprach- und Textdaten geht. Auch hier sind mehrere Ebenen der Beschreibung erforderlich (vgl. etwa Sasaki 2004) – etwa eine phonologisch/graphematische Ebene, Syntax, Koreferenz, Semantik und Pragmatik (vgl. Teich, Hansen 01) –, die auch nicht unverbunden nebeneinander stehen dürfen. Hier werden Verfahren der Mehrebenenannotation entwickelt, wobei jede Ebene über ihre Ontologien verfügt, die über Korrespondenzregeln verknüpft werden können. Unterstützung durch Annotationswerkzeuge wird bereitgestellt (z.B. Müller, Strube 2003). Verfahren der Mehrebenenannotation werden aber auch im Image Retrieval genutzt (Fan et al. 2004). Hier ist vor allem die Kombination der Oberflächeneigenschaften des Bildes mit Information über die Bildinhalte (Szenen, Konzepte) von Bedeutung.
Eine weiteres wichtiges Anwendungsgebiet texttechnologischer Methoden liegt im Textmining[7]. Für das hier beschriebene Vorhaben sind Verfahren des Textmining auf mehreren Ebenen von Relevanz: für die Terminologieextraktion im Ontologieaufbau (Buitelaar et al. 2005; Maedche, Volz 2001), für das Auffinden von Belegstellen für bereits in der Datenbank enthaltene Signalpfade, sowie für die Extraktion von Fakten aus Texten (Dickerson et al. 2003).

2.7 Netzgestützte Wissenschaftskommunikation

Die derzeit hauptsächlich genutzten Formen der Wissenschaftskommunikation sind in mancherlei Hinsicht in eine Krise geraten. Wissenschaftliche Zeitschriften können – besonders in aktiven Forschungsgebieten - mit ihrem Erscheinungsrhythmus

[6] Hier sind über die Projektlaufzeit die Ergebnisse der Arbeitsgruppen des W3C zu Beschreibung und Interaktion von Web-Services zu berücksichtigen (http://www.w3c.org/2002/ws/, zitiert am 28.1.05). Besondere Relevanz haben die Ergebnisse der im Kompetenznetzwerk „Neue Dienste, Standardisierung, Metadaten“ vom BMBF geförderten Projekte, insbesondere des Teilprojekts „Generische und komponentenbasierte wissenschaftliche Portale“.

[7] Einen Überblick über die Methoden bietet (Hotho et al. 2005).

immer weniger der schnellen Entwicklung folgen. Ihr ständig steigender Preis stellt wiederum die wissenschaftlichen Bibliotheken vor große Probleme. Als Reaktion wurden Verfahren vorgeschlagen, wie Wissen mit Hilfe der Kommunikationsmöglichkeiten des Internet schneller und flexibler zur Verfügung gestellt werden kann, als dies mit den traditionellen Publikationsorganen möglich sein kann[8]. Diese neuen Ansätze des Knowledge Managements (Kuhlen 2003) beruhen auf dem Aufbau von Wissensportalen durch egalitäre Beteiligung einer interessierten Gruppe[9]. Gegenstand der Arbeit können Enzyklopädien (Stallman 2005) sein oder spezialisiertere Forschungsgebiete. Entscheidend ist, dass einerseits Anreizsysteme zur Teilnahme an derartigen Projekten auffordern sollen, gleichzeitig aber Qualitätsindikatoren den Stellenwert von Beiträgen verdeutlichen müssen (Semar 2004). Als Folge verschwimmen – hinsichtlich der Zugänglichkeit, nicht der Qualitätseinschätzung – die Grenzen zwischen einer begutachteten Fachpublikation und grauer Literatur.
Die Integration der Fachinformation in einem Informationsportal erleichtert zudem, einen Überblick über den aktuellen Forschungsstand zu gewinnen.

3. Vorgehensweise

Eingangs wurden grundsätzlich zwei Problemzonen in der Erschließung und Kommunikation wissenschaftlicher Ergebnisse über Signalpfade identifiziert:

1. Die Erschließungstiefe bisheriger Signalpfaddatenbanken ist nicht ausreichend. Sie orientieren sich primär an intrazellulären Prozessen, geben wenig oder keine Auskunft über funktionale Zusammenhänge und berücksichtigen nicht die Kompartimente, in denen diese Prozesse ablaufen. Des weitern sind keine Instrumente für die Integration klinischer Forschungsergebnisse vorhanden.
2. Die Wissenschaftskommunikation in der Erforschung von Signalpfaden verläuft in vielen Bereichen ineffizient. Als primäres Kommunikationsmittel stehen wissenschaftliche Zeitschriften mit ihrem hohen Qualitätsniveau aber auch langen Vorlaufzeiten zur Verfügung. Ein offene Kommunikationsplattform, die es erlaubt, wissenschaftliche Hypothesen zur Debatte zu stellen könnte hier Abhilfe schaffen.

Aus diesen zwei recht allgemein beschriebenen Problemfeldern lassen sich folgende konkretere Arbeitsbereiche mit jeweils eigenen Methoden ableiten:

1. Formale Grundlagen: Teil des Projekts ist der Entwurf neuer Ontologien und Annotationsschemata. Zu klären ist, wie begriffliche und mereologische Relationen ausgedrückt werden, welche Annotationsschemata für biochemische, physiologische bzw. klinische Phänomene benötigt werden und wie diese unterein

[8] Hier sind auch Internetportale wie etwa vascoda (www.vascoda.de/, zitiert am 28.1.05. vgl. Neuroth; Pianos 02) zu berücksichtigen.

[9] Diese Forschung wird unter dem Titel *K3 – Wissensmanagement über kooperative verteilte Formen* über den Projektträger im DRL NMB+F vom BMBF gefördert.

ander in Beziehung zu setzen sind. Dabei werden auch Inferenzschemata zu definieren sein.

2. Informationsbedarfsanalyse: Vor konkreten Systementwürfen muss eine genaue Analyse des durch diese System zu befriedigenden Informationsbedarfs stehen (Kluck 1997). Durch Umfragen oder Nutzungsstatistiken kann so auf die Konzeptualisierung des Informationsbedarfs durch die Nutzer geschlossen werden, so dass die Strategien der Inhalterschließung und die angebotenen Recherchemöglichkeiten diesem Bedarf angepasst werden können.
3. Akzeptanzanalyse: In Disziplinen, deren Publikationsstandards sehr stark von Impact-Faktoren geprägt sind, werden alternative, informellere Publikationsformen nur schwer Akzeptanz finden. Um Fehlentwicklungen zu vermeiden sind frühzeitig Akzeptanzuntersuchungen durchzuführen. Durch das Angebot vergleichbarer Qualitätsstandards – etwa durch Einsatz von angepassten bibliometrischen Verfahren für das Web (Ball, Tunger 2005) – könnten Widerstände überwunden werden.
4. Korpusanalyse: Der Aufbau der Ontologie erfolgt durch Extraktion von Terminologie und terminologischen Relationen aus einem angemessen ausgewählten Korpus, das zunächst zusammenzustellen ist. Dabei soll soweit als möglich, sowohl was die Analysetools angeht, als auch hinsichtlich der Korpora, auf schon bestehende Ressourcen zurückgegriffen werden (Morgan 2004).
5. Visualisierungsverfahren: Die bestehenden Ansätze zur Visualisierung von Signalpfaden sind zu sichten und so erweitern, dass rein funktionale Zusammenhänge durch Lokalisierungsinformation (Kompartiment) ergänzt wird. Dabei wird, soweit möglich, nicht nur auf die Bildsprache sondern auch auf Abbildungskorpora aus anatomischen Lehrmaterialen zurückgegriffen.
6. Konstruktion eines Informationssystems: Gegenstand des Projekts sind in erster Linie die Entwicklung neuer Erschließungs- und Recherchemöglichkeiten für Signalpfade sowie die Eröffnung neuer Kommunikationswege. Für die Software wird daher eine offene Architektur gewählt, welche die Integration bestehender Komponenten erlaubt. Ziel ist es, die vorgeschlagenen Ansätze an einem experimentellen aber operationalen Prototypen evaluieren zu können.

Die Aufteilung in eher empirisch und eher informationsmethodisch dominierte Fragestellungen entspricht weitgehend der Arbeitsteilung zwischen den Kooperationspartnern aus Medizin und Informationswissenschaft.

Obschon das Projekt hinsichtlich zentraler Fragen Grundlagenforschung leistet – etwa in der Ausgestaltung von Ontologie und Annotationsschemata – ist es dennoch strukturell so angelegt, dass auch der organisatorische Rahmen für eine nachhaltige Einführung eines derartigen Informationssystems von vornherein geschaffen wird. Ein umfänglicher, international besetzter wissenschaftlicher Beirat schafft die Voraussetzung für die Etablierung von Qualitätsstandards. Frühzeitige Einbindung von Verlagen erhöht die Wahrscheinlichkeit, dass das zu konzipierende Informationssystem mit seinen Formen der Wissenschaftskommunikation harmonisch in das

bestehende System integriert wird. Weiterhin ist der langfristige Betrieb zu gewährleisten, für den eine akademische Einrichtung nicht eintreten kann.

4. Projektziele und Stand der Arbeiten

Hauptziel des Projektes ist es, ein neuartiges Informationssystem über Signalpfade nachhaltig zu etablieren. Unabhängig von diesem Globalziel können Teilziele definiert werden, die Ergebnisse von eigenem Wert mit jeweils individuellen Nutzungsmöglichkeiten erzielen werden.

1. Vorschläge für erweiterte Ontologien und Annotationsschemata sind als Standardisierungsvorschläge den zuständigen Gremien vorzulegen und somit dauerhaft zu etablieren.
2. Erschlossene Korpora werden der Fachöffentlichkeit zur Verfügung gestellt.
3. Empirische Untersuchungen zum Informationsbedarf und zur Akzeptanz wissenschaftlicher Kommunikationsformen stellen einen eigenständigen wissenschaftlichen Wert dar.
4. Softwarekomponenten können unter einer offenen Lizenz zur Verfügung gestellt werden.

Zur Zeit befindet sich das Projekt in einer frühen Phase. Vorstudien zu Informationsbedarfs- und Akzeptanzanalyse sind mit einer kleinen Expertengruppe durchgeführt worden. Ihr Ergebnis war ermutigend, bedarf aber noch der Absicherung durch umfangreichere Befragungen. Als nächste Schritte sind die Auswahl von Korpora und die formale Grundlegung der zu erstellenden Ontologie vorgesehen.

Literatur

Ball, Rafael; Tunger, Dirk (2005) Bibliometrische Analysen – Daten, Fakten und Methoden. Schriften des Forschungszentrums Jülich, Reihe Bibliothek, Bd. 12.

Bard, Jonathan B.L.; Rhee, Seung Y. (2004) Ontologies in Biology. Design, Applications and Future Challanges. Nature Reviews, Genetics, Bd. 5, March 2004, S. 213-222.

Battistella, E.; de Souza, J.C.G.; Ferreira, R.A.; Vieira, R.; Mombach, J.C.M.; Lemke, N. (2004) Bioinformatics: A Growing Field for Ontologies. Workshop on Ontologies and their Applications. 28.9.2004, Sao Luis, Brasilien. http://www.ws.onto.ufal.br/Papers/Battistella.pdf (20.1.2005).

Buitelaar, P.; Cimiano, P.; Magnini, B. (2005) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press.

Becker, M.Y.; Rojas, I. (2001) A graph layout algorithm for drawing metabolic pathways. In. Bioinformatics, Bd. 17, Nr. 5, S. 461-467.

Deville, Y.; Gilbert, D.; Helden, J.; Wodak, S. (2003) An Overview of Data Models for the Analysis of Biochemical Pathways. In: Briefings in Bioinformatics, Bd. 4, Nr. 3, S. 246-259.

Dickerson, J.A.; Berleant, D.; Cox, Z; Qi, W.; Ashlock, D.; Wurtele, E. (2003) Creating Metabolic Network Models using Text Mining and Expert Knowledge. http://www.public.iastate.edu/~mash/publications/dickerson03b.pdf, zitiert am 17.7. 2005

Ding, Y.; Foo, S. (2002a) Ontology research and development. Part 1: A Review of Ontology Generation. Journal of Information Science, Bd. 28, Nr. 2, S. 123-136.

Ding, Y ; Foo, S. (2002b) Ontology research and development. Part 2: A Review of Ontology Mapping and Evolving. Journal of Information Science, Bd. 28, Nr. 5, S. 375-388.

Fan, J.;Gao, Y.; Luo, H. (2004) Multi-level annotation of natural scenes using dominant image components and semantic concepts. In Proc. of the 12th annual ACM international conference on Multimedia. S. 540-547.

Goesman, Alexander; Haubrock, Martin; Meyer, Folker; Kalinowski, Jörn; Giegerich, Robert (2002) Pathfinder: Reconstruction and Dynamic Visualization of Metabolic Pathways. In Bioinformatics, Bd. 18, Nr. 1, S. 124-129.

Ghandeharizadeh, S.; Sommers, F.; Joisher, K.; Alwagait, E. (2002) A document as a web service: Two complementary frameworks. In: Chaudhri, A.B. (Hrsg.) et al. XML-based data management and multimedia engineering. Springer, Berlin, S. 450-461.

Giogno, Rosalba; Shasha, Dennis (2002) GraphGrep: A Fast and Universal Method for Querying Graphs. In Proceeding of the International Conference in Pattern recognition (ICPR), Quebec, Canada, August 2002. http://www.cs.nyu.edu/shasha/papers/graphgrep/icpr2002.pdf, zitiert am 18.7.2005

Hotho, Andreas; Nürnberger, Andreas; Paaß, Andreas (2005) A Brief Survey on Text Mining. In LDV Forum, Bd. 20, Nr. 1, S. 19-62.

Joshi-Tope, G.; Gillespie, M.; Vastrik I, D'Eustachio, P.; Schmidt, E.; de Bono, B., Jassal, B.; Gopinath, G.R.; Wu, G.R.; Matthews, L.; Lewis, S.; Birney, E. Stein, L. (2005) Reactome: a knowledgebase of biological pathways. In Nucleic Acids Res. Bd. 33, Nr. 1, Database Issue. S. 428-432.

Kanne, C.-C.; Schreiber, F.; Trümbach, D. (1999) Electronic Biochemical Pathways. In Kratochvíl, J. (Hrsg.), Graph Drawing. Proc. 7th International Symposium. GD'99, Stirín Castle, Czech Republic, September 1999, S. 418.

Karopka, T.; Scheel , T.; Bansemer, S.; Glass, A. (2004) Automatic construction of gene relation networks using text mining and gene expression data. In: Med Inform Internet Med., Bd. 29, Nr 2, S. 169-183.

Karp, P.D.; Paley, S.M.; Krieger, C.J.; Zhang, P. (2004) An Evidence Ontology for Use in Pathway/Genome Databases. Pacific Symposium on Biocomputing. S. 190-201, (http://helix-web.stanford.edu/psb04/karp.pdf, zitiert am 28.1.05).

Michael Kluck (1997): Methoden der Informationsanalyse. In: Buder, M.; Rehfeld, W.; Seeger, TH.; Strauch, D. (Hrsg.): Grundlagen der praktischen Information und Dokumentation. München et al.: K.G. Saur, S. 795-821

Koike, A.; Niwa, Y.; Takagi, T (2004) Automatic extraction of gene/protein biological functions from biomedical text. In: Bioinformatics, epub.

Krieger, Cynthia J.; Zhang, Pelfen; Müller, Lukas A.; Wang, Alfred; Paley, Suzanne; Arnaud, Martha; Pick, John; Rhee, Seung Y; Karp, Peter D. (2004) Nucleic Acids Research, Bd. 32, Nr. 10.

Kuhlen, R. (2003) Change of Paradigm in Knowledge Management – Framework for the Collaborative Production and Exchange of Knowledge. In Hobohm, H.-C. (Hrsg.) Knowledge Management – and Asset for Libraries and Librarians. Collected Papers from LIS Professionals.

Lewis, S.E. (2005) Gene Ontology: Looking Backwards and Forwards. In: Genome Biol. Bd. 5, Nr. 1.

Lomax, J.; McCray, A. (2004) Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics, Bd. 5, Nr. 4, S. 354-361.

Maedche, Alexander; Voltz, Raphael (2001) The Ontology Extraction & Maintainance Framework Text-To-Onto. In ProcWorkshop on Integrating Data Mining and Knowledge Management http://cui.unige.ch/~hilario/icdm-01/DM-KM-Final/Volz.pdf, zitiert am 18.7.2005.

Morgan, Alex (2004) BioNLP-Resources. http://www.tufts.edu/~amorga02/bcresources.html, zitiert am 17.07..2005.

Müller, C.; Strube, M. (2003) Multi-Level Annotation in MMAX. In Proc. of the 4th SigDial Workshop on Discourse and Dialogue, Sapporo, 5-6 July 2003, S. 198-207.

Neuroth, H.; Pianos, T. (2003) VASCODA: A German Scientific Portal for Cross-Searching Distributed Digital Resource Collections. Proc. 7th Int. Conf. on Research and Advanced Technology for Digital Libraries, S. 257-262.

Paek, E.; Park, J.; Lee, K.J. (2004) Multi-layered representation for cell signaling pathways. In Mol Cell Proteomics. Bd. 3, Nr. 10, S. 1009-22.

Pinter, Ron Y.; Rokhlenko, Oleg; Yeger-Loten, Esti; Ziv-Ukelson, Michal (2005) Alignment of Metabolic Pathways. In Bioinformatics, zur Publikation angenommen.

Rahman S.A.; Advani, P.; Schunk, R.; Schrader, R.; Schomburg, D. (2004) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). In: Bioinformatics. Nov 30, epub.

Rojdestvenski, I. (2003) Metabolic pathways in three dimensions. Bioinformatics. Bd. 19, Nr. 18, S. 2436-2441.

Sasaki, F. (2004) Sekundary Information Structuring – A Methodology for the Vertical Interrelation of Information Resources. In Proc. of Extreme Markup Languages, Montréal, Kanada.

Schacherer, Frank (2001) An object-oriented database for the compilation of signal transduction pathways. Dissertation, Technische Universität Carolo-Wilhelmina Braunschweig.

Schreiber, F. (2002) High Quality Visualization of Biochemical Pathways in BioPath. In Silico Biology, Bd. 2, Nr. 6.

Semar, Wolfgang (2004) Incentive Systems in Knowledge Management to Support Cooperative Distributed Forms of Creating and Acquiring Knowledge. In: Arabnia, Hamid; et al. (Hg.): Proceedings of the International Conference on Information and Knowledge Engineering - IKE'04. Las Vegas: CSREA Press, S. 406 - 411

Sivakumaran Sivakumaran, S.; Hariharaputaran, S.; Mishra, J.; Bhalla, U.S. (2003) The Database of Quantitative Cellular Signalling: Management and Analysis of Chemical Kinetic Models of Signalling Networks. Bioinformatics, Bd. 19. Nr. 3, S. 408-415.

Smith, Barry; Ceusters, Werner; Klagges, Bert; Köhler, Jacob; Kumarm, Anand; Lomax, Jane; Mungall, Chris; Neuhaus, Fabian; Rector, Alan L.; Rosse, Cornelius (2005a) Relations in Biomedical Ontologies. In Genome Biology, Vol. 6, Nr. 5, http://genomebiology.com/2005/6/5/R46, zitiert am 17.7.2005.

Smith, Barry; Mehino, Jose L.V.; Schulz, Stefan; Kumar, Anand; Rosse, Cornelius (2005b) Anatomical Information Science. http://ontology.buffalo.edu/anatomy_GIS/FMA-AIS.pdf, zitiert am 17.7.2005.

Sohler, Florian; Zimmer, Ralf (2004) Identifying Active Transcription Factors and Kinases from Expression Data using Pathway Queries. In Bioinformatics, Bd. 20, 1517 – 1521. http://www.bio.ifi.lmu.de/mitarbeiter/sohler/eccb05.pdf, zitiert am 18.7.2005.

Stallman, R. (2005) The Free Universal Encyclopedia and Learning Resource. [http://www.gnu.org/encyclopedia/free-encyclopedia.html], zitiert am 21.1.2005

Takai-Igarashi, Takako; Mizoguchi, Riichiro (2003) Cell Signaling Network Ontology. http://www.bioinfo.de/isb/2003/04/0008/main.html, zitiert am 17.7.2005.

Teich E.& S.Hansen (2001) Towards an integrated representation of multiple layers of linguistic annotation in multilingual corpora. In Online Proceedings of Computing Arts 2001: Digital Resources for Research in the Humanities. Sydney, http://www.coli.uni-sb.de/~hansen/teichhansen-final.pdf, zitiert am 28.1.2005

Vargas-Vera, M.; Motta, E.; Domingue, J.; Lanzoni, M.; Stutt, A.; Ciravegna, F. (2002) MnM: Ontology driven semi-automatic and automatic support for semantic markup. In: Gómez-Pérez, A. (Hrsg) et al., Knowledge engineering and knowledge management. Ontologies and the semantic web. 13th international conference, EKAW 2002, Sigüenza, Spain, October 1-4, 2002. Proceedings. Berlin: Springer. Lect. Notes Comput. Sci. 2473, S. 379-391.

DMG-Lib: ein moderner Wissensraum für die Getriebetechnik

Torsten Brix, Ulf Döring, Sabine Trott; Ilmenau

Abstract

Mechanismen sind unverzichtbare Bestandteile technischer Produkte in zahlreichen Branchen. Das darüber vorhandene Wissen liegt weltweit verstreut in unterschiedlichen Formen vor und ist für Entwickler, Wissenschaftler und Studierende oft schwer oder nur unvollständig zugänglich.

Der Artikel beschreibt die im Aufbau befindliche Digitale Mechanismen- und Getriebebibliothek (DMG-Lib), mit deren Hilfe dieses Wissen zusammengetragen und in nutzergerechter Form als digitale Dokumente, angereichert durch moderne Verfahren der Animation und Simulation, bereitgestellt wird. Dabei werden heterogene Quellen wie Literatur, gegenständliche Modelle und Lehrmaterialien für das genannte Wissensgebiet nach einem einheitlichen Standard erschlossen. Neben der TU Ilmenau arbeiten die getriebetechnischen Lehrstühle der RWTH Aachen und der TU Dresden am Projekt mit.

Notwendigkeit und Zielstellung der DMG-Lib

Schon Mitte des 19. Jahrhunderts begann vor allem in Deutschland die systematische Untersuchung von Mechanismen und Getrieben. Die Notwendigkeit hierfür ergab sich aus dem großen Forschungs- und Lehrbedarf, der vor dem wirtschaftlichen Hintergrund des schnell wachsenden deutschen Maschinenbaus entstand. Besonders hervorzuheben sind die theoretischen Überlegungen und praxisnahen Arbeiten des deutschen Ingenieurs F. Reuleaux [z. B. Reuleaux 1875], der mehr als 1000 Getriebeanordnungen ausführlich beschrieb sowie eine international bekannte Getriebesammlung von über 800 Funktionsmodellen aufbaute, die während des Zweiten Weltkrieges zu großen Teilen verloren ging. Ihm folgten mit bahnbrechenden Arbeiten u. a. L. Burmester [Burmester 1888], M. Grübler, A. Schönflies und H. Alt [Grübler 1917]. Nach dem Zweiten Weltkrieg setzten u. a. W. Lichtenheldt [Lichtenheld 1961] in Dresden, R. Beyer in München, A. Bock [Bock 1959] in Ilmenau und W. Meyer zur Capellen [Meyer] in Aachen die Arbeiten fort. An deutschen Hochschulen und Universitäten gibt es momentan nur noch 12 Lehrstühle mit dem Schwerpunkt Getriebe- und Mechanismentechnik, die trotzdem den internationalen Stand der Technik durch ihre hervorragenden Forschungstätigkeiten entscheidend mitprägen.

Neben dem Lösen kinematischer und dynamischer Problemstellungen bilden Getriebe und Mechanismen mit mehreren Antrieben, seriellen und parallelen Strukturen, gesteuerten Verstelleinrichtungen, nachgiebigen Elementen sowie in

Miniatur- und Mikrobauweise die neuen Untersuchungsobjekte in der Mechanismen- und Getriebetechnik. Dabei kommen u. a. auch Methoden aus den Fachgebieten Maschinenelemente, Konstruktions-, Feinwerk-, Medizin-, Antriebs-, Mess- und Regelungstechnik zur Anwendung. Immer mehr werden Getriebe und Mechanismen als integrale Bestandteile mechatronischer und biomechanischer Bewegungssysteme eingesetzt. Die Bedeutung der Getriebetechnik wird auch durch die Einführung neuer Technologien, wie z. B. der Nanotechnologie, in Zukunft stark zunehmen, da sich neue Anwendungsfelder erschließen.

Obwohl das Wissen über die Mechanismen- und Getriebetechnik nicht nur für den Maschinenbau unentbehrlich ist, können in der Lehre im Allgemeinen nur elementare Grundlagen zur Struktur, Analyse und Synthese von Getrieben und Mechanismen vermittelt werden. Dies wird sich auch nicht ändern, da das Aufgabenspektrum zukünftiger Ingenieure im Bereich Maschinenbau durch neue Technologien und Entwicklungen immer breiter und interdisziplinärer wird. Allein die Computer- und Informationstechnik nimmt mittlerweile einen großen Stundenumfang bei der Ausbildung von Maschinenbau-Ingenieuren ein.

Das vorhandene, umfangreiche getriebetechnische Wissen steht der Öffentlichkeit nur stark eingeschränkt und örtlich weit verstreut zur Verfügung. Es entspricht nicht den heutigen Anforderungen an schnelle Informationsgewinnung. Die zugängliche Fachliteratur (Fachbücher, Fachzeitschriften, Getriebeatlanten, Fachaufsätze etc.) genügt in Inhalt, Umfang und Medium nur noch selten heutigen Ansprüchen. Sehr alte, einzigartige, in nur wenigen Ausgaben vorhandene und der Öffentlichkeit nicht zugängliche Wissensbestände müssen erschlossen, digital aufbereitet und zusammengeführt werden. Hinzu kommt ein immer größer werdender Druck seitens der Industrie, aber auch von Forschungseinrichtungen, auf Kenntnisse über Mechanismen und Getriebe in ihrer gesamten Breite internetbasiert zugreifen zu können, da ausgewiesene Getriebeexperten nicht mehr ausgebildet werden und somit die Anfragen an die entsprechenden Fachgebiete der Hochschulen und Universitäten nicht mehr im vollen Umfang bedient werden können.

Die Bewahrung des erreichten Wissenstandes und der didaktischen Erfahrungen bei der Wissensvermittlung auf dem Gebiet der Mechanismen- und Getriebetechnik ist von sehr großer Bedeutung, da, wie sich in den letzten Jahren zeigte, die Gefahr groß ist, dass mit dem Ausscheiden von Professoren dieses Wissen verloren geht. Zudem werden durch Sparmaßnahmen Lehrstühle mit unterschiedlichen Schwerpunkten zusammengelegt. Als Folge gehen häufig didaktisch wertvolle Lehrmaterialien verloren. Ein Ausweg ist die Sammlung und Veröffentlichung von Lehrmaterialien auf einer geeigneten Internet-Plattform. Diese Plattform sollte auch die Möglichkeit eröffnen, aktuelle Forschungsergebnisse weltweit zu publizieren. Die Forschung und Lehre in den unterschiedlichsten Ingenieurdisziplinen würde durch die Zusammenstellung des Wissens auf dem Gebiet der Mechanismen- und Getriebetechnik mit allen nötigen Querverweisen mit Sicherheit profitieren.

Aus den genannten Gründen begann im Jahre 2004 der Aufbau einer weltweit zugänglichen, digitalen Bibliothek für die Mechanismen- und Getriebetechnik, die den schleichenden Wissensverlust aufhalten soll. Das Ziel der DMG-Lib besteht in der Sammlung, Bewahrung, Systematisierung, Vernetzung und geeigneten Präsentation des umfangreichen Wissens über Mechanismen und Getriebe. Dabei geht es nicht nur um die Bereitstellung von Textdokumenten und Bildern, sondern auch um die Berücksichtigung computergestützter Funktionsmodelle, die in körperlicher Form als Unikate zu Tausenden existieren und der Öffentlichkeit nur sehr eingeschränkt oder überhaupt nicht zugänglich sind. Die Arbeiten werden von der DFG im Rahmen der Förderinitiative „Leistungszentren für Forschungsinformation“ max. mit 2,5 Millionen Euro über fünf Jahre aufgeteilt in zwei Phasen (2 Jahre und 3 Jahre) finanziert. An der TU Ilmenau arbeiten im Rahmen des Projektes die Fachgebiete Konstruktionstechnik, Getriebetechnik, Grafische Datenverarbeitung und Medienproduktion mit der Universitätsbibliothek, dem PATON und dem Universitätsrechenzentrum zusammen. Zudem sind die Getriebelehrstühle der RWTH Aachen und der TU Dresden als Projektpartner am Gesamtvorhaben beteiligt.

Konzept der DMG-Lib

Das umfangreiche, mitunter schon Jahrhunderte alte Wissen über Mechanismen und Getriebe ist heute stark verstreut, teilweise schwer zugänglich und entspricht auch durch die Form seiner Hinterlegung (insbesondere als Text und Bild) nicht den heutigen Anforderungen an eine schnelle Informationsgewinnung. Noch heute sind statische Texte und Bilder die vorherrschenden Formen, in denen technisches Wissen für die Öffentlichkeit formuliert wird, obwohl schon sehr frühzeitig die Bedeutung funktionsfähiger Modelle erkannt wurde. Mittlerweile sind jedoch die technischen Voraussetzungen gegeben, um z. B. gegenständliche Anschauungsmodelle einer breiten Öffentlichkeit zugänglich zu machen.
So erlauben computergestützte Methoden, die Abbildung funktionaler und anderer Eigenschaften von Mechanismen und Getrieben effizient zu realisieren, zu verbreiten und als multimediale Dokumente mit zusätzlichen Informationen, Analyseergebnissen, Animationen, Querverweisen etc. zu versehen [Döring 2005]. Solche Dokumente können mit problemorientierten Suchkriterien effizient abgerufen werden. Somit wird der Übergang von einer statischen zu dynamisch-problemorientierten Bereitstellung von Wissen für ein breites Anwendungsfeld erreicht. Das Leistungszentrum soll diesen Wandel gezielt vorantreiben und auch selbst zu einem Wissensspeicher für die Mechanismen- und Getriebetechnik und angrenzende Ingenieurdisziplinen werden.

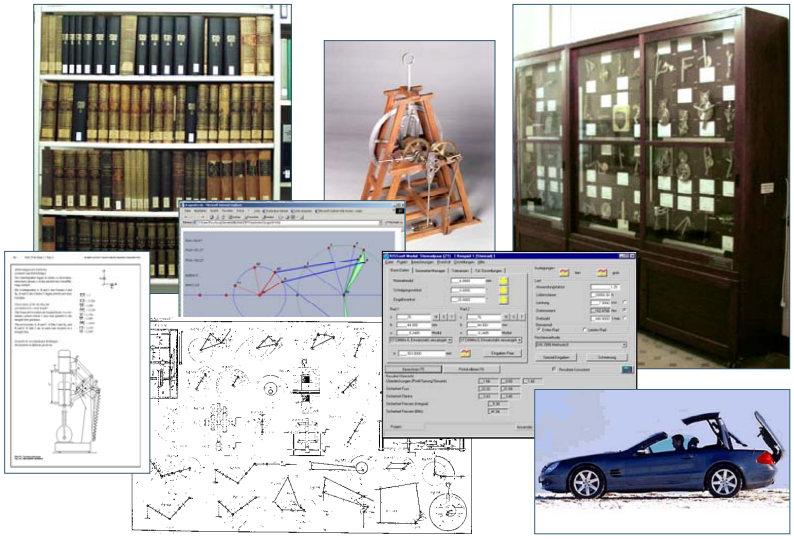


Abbildung 1: Beispiele von Quellen für die DMG-Lib

Durch Digitalisierung und internet-basierte Bereitstellung kann zwar der Zugriff auf die einzelnen Dokumente verbessert werden, eine zielorientierte, nutzerangepasste und damit effiziente Lösungsfindung zu getriebetechnischen Aufgabenstellungen vor allem aus Forschung und Lehre wird so jedoch nicht unterstützt. Deshalb bilden umfangreiche Digitalisierungsarbeiten, wie sie bei vielen Projekten im Vordergrund stehen, nur notwendige Vorarbeiten für die Schaffung eines Leistungszentrums für Forschungsinformation.
Um dem genannten Ziel gerecht zu werden, muss mehr als nur eine Sammlung von digitalisierten Dokumenten oder Links auf relevante Seiten im Internet angeboten werden. Die neue Qualität der Bibliothek besteht insbesondere darin, die Vielzahl von Beschreibungen getriebetechnischer Lösungen, die in den unterschiedlichsten Beschreibungsformen (verbal, analytisch, grafisch, gegenständlich) vorliegen, jeweils zu abstrahieren und in einem einheitlichen Datenformat zu speichern. Die Abstraktion besteht im Ermitteln des technischen Prinzips. Durch die einheitliche Beschreibung einer großen Menge technischer Prinzipien und die darauf aufbauenden Analysen sowie die systematische Speicherung der Ergebnisse wird ein Wissensspeicher aufgebaut, der sich sehr effizient nach geeigneten Lösungen durchsuchen lässt. Dies ist besonders wichtig für diejenigen Nutzer, die auf der Suche nach Lösungen eines bestimmten getriebetechnischen Problems sind.
Hieraus ergeben sich weitere Zielstellungen der DMG-Lib, die sich wie folgt zusammenfassen lassen:

- constraint-basierte Modellierung von Getrieben und Mechanismen als Ausgangsbasis für die Generierung weiterer Beschreibungsformen [Döring 2005],
- Bereitstellung unterschiedlicher Beschreibungsformen (verbal, analytisch, bildlich) für Mechanismen und Getriebe zur Sicherstellung der Anpassbarkeit und somit der allgemeinen wie auch längerfristigen Nutzbarkeit:
  - verbale Beschreibung von Verwendungszweck und Nutzen,
  - Pixelbild für visuellen Endruck,
  - animiertes Pixelbild, Java-Animation, Flash-Animation, VRML vornehmlich zum Erkennen des kinematischen Verhaltens,
  - andere Beschreibungen zur Kopplung mit vorhandenen Analyse-, Synthese- und Optimierungsprogrammen,
  - constraint-basierte Beschreibung,
- plattformübergreifende Repräsentation im Internet für eine breite Öffentlichkeit in unterschiedlichen und auch anpassbaren Formen speziell für Lehre, Selbststudium, Forschung und Produktentwicklung,
- Aufbau von Wissensbasen, die eine Strukturauswahl bzw. Typsynthese unterstützen,
- Bereitstellung automatischer Zugriffsmöglichkeiten auf die Bibliotheksinhalte durch Nutzung unterschiedlichster, anwendungsbezogener Deskriptoren / Metadaten (z. B. über einen OAI-PMH Service) und
- Unterstützung der Forschung und Entwicklung besonders bei Synthese- und Optimierungsproblemen.

Das Vorgehen von der Auffindung relevanter Quellen bis zu deren Bereitstellung im Internetportal der DMG-Lib zeigt Abbildung 2.

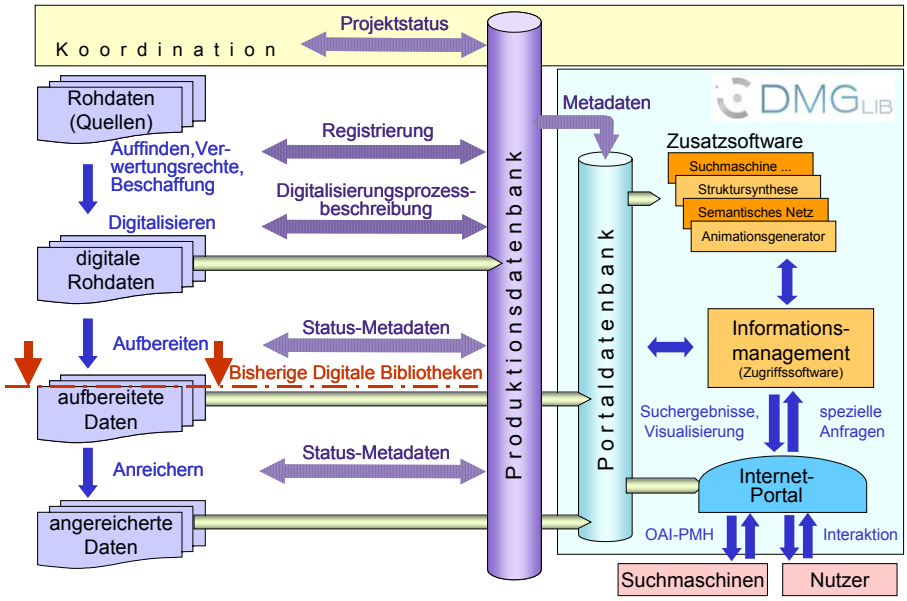


Abbildung 2: Konzept der Aufbereitung der Quellen für das Internetportal der DMG-Lib

Die Quellen der DMG-Lib (Abbildung 1) sind sehr umfangreich und verschiedenartig. Sie umfassen z. B. Funktionsmodelle, Getriebekataloge, technische Reporte, Forschungsberichte, Fachbücher, Fachaufsätze, Videos, Fotos etc.
Diese Originalquellen werden beschafft, digitalisiert und in geeignete Formate konvertiert. Im Gegensatz zu vielen anderen Projekten im Bereich digitaler Bibliotheken, die lediglich die digitalen Rohdaten zugänglich machen, erfolgt im vorliegenden Projekt die sehr wichtige Phase der Aufbereitung und Anreicherung der digitalen Rohdaten mit Zusatzinformationen wie beispielsweise verbalen Beschreibungen, animierten Abbildungen oder constraint-basierten Modellen. Somit lassen sich Bewegungsvorgänge der Mechanismen erkennen und weitergehende Simulationen sowie Analysen vornehmen. Darüber hinaus wird die Nutzung der constraint-basierten Modelle in externen Analyse-, Synthese- und Optimierungsprogrammen möglich sein. Nur so kann ein zielorientierter und effizienter Zugriff auf die Inhalte der DMG-Lib gewährleistet werden.
Über ein Internetportal erfolgt der Zugriff auf die digitale Bibliothek, die es Forschern, Konstrukteuren, Studierenden und sonstigen Interessierten ermöglicht, im gesamten Wissensgebiet unter verschiedensten Aspekten recherchieren zu können.

Realisierung

Für die Realisierung dieses anspruchsvollen Konzeptes ist eine konsequente Zusammenarbeit von Getriebefachleuten, IT-Spezialisten und Bibliothekaren notwendig. Nur dadurch ist es möglich, das fachspezifische, heterogene Quellenmaterial nutzergerecht zu sammeln, aufzubereiten und zur Verfügung zu stellen.

Aufbereitung des Quellenmaterials

Folgendes Quellenmaterial wird verwendet:

- Literatur aus dem Bereich der Mechanismen- und Getriebetechnik (Monographien, Zeitschriftenaufsätze), die aus verschiedenen Bibliotheken und Privatsammlungen beschafft wird,
- körperliche Getriebemodelle der TU Dresden und der RWTH Aachen,
- Photos und Dias von Getrieben, die bei den Projektpartnern vorhanden sind,
- technische Darstellungen (Skizzen, technische Entwürfe, technische Prinzipien) und Berechnungsvorschriften,
- Lehrmaterialien der am Projekt beteiligten Lehrstühle.

Diese Materialien werden digitalisiert und aufbereitet.

So werden z. B. die Literaturquellen mit 300 dpi Auflösung und 256 Graustufen als TIFF-Dateien abgelegt. Für die Erfassung in der Produktionsdatenbank werden Dublin Core Metadaten verwendet [Dublin Core]. Zusätzlich zu diesen Metadaten erfolgt eine Klassifikation der Dokumente nach getriebetechnischen Gesichtspunkten.

Zur Aufbereitung der gescannten Rohdaten ist eine Struktur-/Layout- und Texterkennung erforderlich. Für die Erkennung der physikalischen Struktur (Textblöcke, Bilder usw.) sowie der Textzeichen wird die kommerzielle Software ABBYY-Finereader benutzt, die in ein selbst entwickeltes Applikationsgrundgerüst mit dem Namen AnAnAS (**An**alyse-**An**reicherungs-**A**ufbereitungs-**S**oftware) integriert wurde (Abbildung 3). Die Identifizierung der logischen Struktur (Überschriften, Bildunterschriften) erfolgt zunehmend automatisiert über eine Software, die im Rahmen des Projektes entwickelt wird. Die Speicherung der Metadaten durch AnAnAS basiert auf dem METS-Standard [METS]. Hierbei werden verschiedene Metadaten kodiert: administrative (z. B. wer hat das Dokument gescannt und woher kommen die Daten), deskriptive (z. B. Dublin Core) und strukturelle (Verknüpfung der Inhalts- mit den Metadaten, z. B. Inhalts- und Abbildungsverzeichnis). Im Ergebnis der Struktur-/Layouterkennung wird die logische Seitenstruktur in XML-Dateien gespeichert.

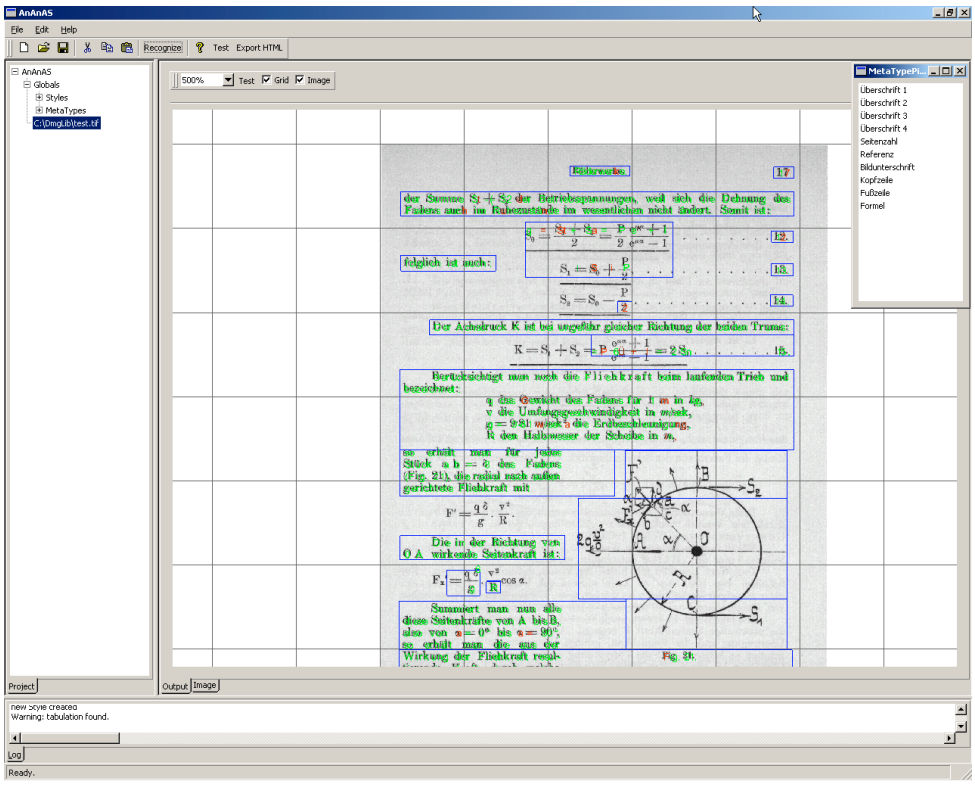


Abbildung 3: Ergebnis der Layoutanalyse in AnAnAS – Text der Seite mit gescanntem Hintergrundbild

Für die Anreicherung der gescannten Dokumente wird ein Animationsgenerator sowie der Export in CAD-/Analysesoftware entwickelt. Grundlage für den Export bzw. die Animationserstellung bildet jeweils ein XML-basiertes Dateiformat zur Beschreibung der abgebildeten Mechanismen [Döring 2005]. Diese abstrakte Modellbeschreibung liefert auch eine Vielzahl von Suchkriterien, zum Teil direkt (z.B. die Anzahl der verwendeten Bauteile), zum Teil aber auch indirekt über eine Simulation und die Analyse der Simulationsergebnisse. So lassen sich Informationen über die Eigenschaften generieren, die sich auf die Funktion der Getriebe beziehen (z.B. das Übertragungsverhalten), was eine problemorientierte Suche wesentlich unterstützt.
Zu einzelnen Modellen, Videos, Bildern und Literaturstellen können durch Getriebefachleute Beschreibungen, Querverweise und Kommentare gegeben werden. Bzgl. der Anreicherung der Literatur ist diese Funktionalität Teil von AnAnAS. Für die anderen Quellen erfolgen Eintragungen direkt über die Benutzungsschnittstelle der Produktionsdatenbank.
Eine erste Modelldatenbank wurde mit MySQL realisiert. Gemeinsam mit allen Projektpartnern wurde die Struktur der Datenbank erarbeitet, um sicherzustellen,

dass alle notwendigen Metadaten (für alle heterogenen Quellen) erfasst werden und so der Datenbestand zur Bearbeitung erweiterter Suchanfragen auf dem DMG-Lib-Server herangezogen werden kann.

Portal

Das Portal bildet die internetbasierte Verbindung zwischen den Nutzern und der DMG-Lib. Für die benutzergerechte Konzeption und Umsetzung des Internetportals erfolgt auch eine Evaluierung der Usability, die sich formal an dem von Deborah Mayhew entwickelten Usability Engineering Lifecycle [Mayhew 1999] orientiert. Dieser ist für die Konzeption der DMG-Lib-Portals sehr gut geeignet, da er speziell auf Prozesse der Softwareentwicklung zugeschnitten ist. Er spezifiziert Aufgaben und Techniken des Usability Engineering in einem iterativen Prozess und beschreibt damit eine Folge logisch zusammenhängender Aktivitäten zum Erreichen der Usability, die als Vorlage für die sich anschließende programmtechnische Umsetzung dienen.

Mit einer Anforderungsanalyse und Experteninterviews wurde ein konzeptionelles Modell unter Berücksichtigung der zu integrierenden Quellen entworfen und ein Demonstrator (High Fidelity Prototyp) des konzeptionellen Modells des Portals erstellt. Ferner wurden Screen Design Standards entwickelt und Screen Design Prototypen umgesetzt (Abbildung 4).

Parallel zur interaktiven Suche bzw. dem interaktiven Browsen duch die Getriebesammlung bietet ein OAI-PMH Service Zugriff auf den Bestand der DMG-Lib.

Abbildung 4: Startseite des Internetportals

Weitere Applikationen

Parallel zur Entwicklung des User Interface des Internetportals der DMG-Lib entstehen derzeit weitere Applikationen, die in das Interportal eingebunden werden und z. B. unterschiedliche Zugänge zur DMG-Lib erlauben, wie ein Zeitstrahl und ein virtueller Rundgang in der Getriebesammlung.

Der Zeitstrahl enthält die Möglichkeit, über historisch bedeutende Persönlichkeiten aus der Getriebe- und Mechanismentechnik und deren Werke auf die Inhalte der Bibliothek zugreifen zu können. Er wird als interaktive, multimediale Applikation in das Internetportal integriert.

Weitere Arbeiten betreffen die zweckmäßige Visualisierung von Getriebeklassen unter verschiedenen Aspekten, wie z. B. strukturelle oder funktionale, die für den Nutzer der DMG-Lib frei wählbar sind. Dabei werden u. a. Einteilungen nach Reuleaux [Reuleaux 1875] oder der IFToMM (International Federation for the Theory of Machines and Mechanisms) [IFToMM] sowie anderer Autoren und Expertengruppen berücksichtigt. Dem Nutzer soll die Visualisierung von Getrieben helfen, nach unterschiedlichen Kriterien die Getriebewelt zu ordnen und einen systematischen Überblick über die Vielzahl der Getriebe zu erhalten. Das Problem liegt dabei im Aufstellen der Klassifikationen und anderer begrifflicher Abhängigkeiten, da unter den Experten der Getriebetechnik unterschiedliche Meinungen und Ansichten existieren. Hierzu erfolgen auch Untersuchungen zum Einsatz semantischer Netze.

Verflechtung mit anderen Digitalisierungsprojekten
Das Projekt DMG-Lib hat seine Aufnahme ins und seine Mitarbeit am Projekt „Portal digitalisierter Drucke“ [PDD] signalisiert. Dazu wurde der Kontakt mit einem der Ansprechpartner (Verbundzentrale des GBV) aufgenommen.

Zusammenfassung und Ausblick

Ziel des DFG-Projektes DMG-Lib ist der Aufbau eines Leistungszentrums für Forschungsinformation zur Sammlung, Systematisierung, Sicherung und geeigneten Repräsentation von Information und Wissen aus der Mechanismen- und Getriebetechnik als digitale, interaktive Bibliothek.
Wesentliche Qualitätsmerkmale des Projektes sind die Erfassung heterogener Quellen in einheitlichem Datenformat, die Aufbereitung und Anreicherung der digitalen Rohdaten mit Zusatzinformationen wie verbalen Beschreibungen, animierten Abbildungen oder constraint-basierten Modellen sowie die benutzergerechte Konzeption und Umsetzung des DMG-Lib-Internetportals.
Die Freischaltung des Internetportals mit einer repräsentativen Quellenauswahl ist für den 1.2.2006 geplant [DMG-Lib].

Literatur und Quellen

[Bock 1959] Bock, A.: *Grundlagen der Getriebelehre* (Sonderdruck aus dem Werkleiter–Handbuch). – Ilmenau, 1959

[Burmester 1888] Burmester, L.: *Lehrbuch der Kinematik*. – Berlin, 1888

[DMG-Lib] *Digitale Mechanismen- und Getriebebibliothek*. – http://www.dmg-lib.de

[Döring 2005] Döring, U.; Brix, T. und Reeßing, M.: *Application of Computational Kinematics in the Digital Mechanism and Gear Library DMG-Lib*, CD-ROM Proceedings of CK 2005, International Workshop on Computational Kinematics. - Cassino (Italien), 2005

[Dublin Core] *The Dublin Core Metadata Initiative.* – http://dublincore.org

[Grübler 1917] Grübler, M.: *Getriebelehre - Eine Theorie des Zwangslaufs und der eigenen Mechanismen.* – Berlin, 1917

[IFToMM] *IFToMM dictionaries online (German)* http://www.ocp.tudelft.nl/tt/cadom/IFToMM/web/online/1031.html (Stand 28.7.2005)

[Lichtenheld 1961] Lichtenheld, W.: *Konstruktionslehre der Getriebe*. – Berlin 1961

[Mayhew 1999] Mayhew, Deborah J.: *The Usability Engineering Lifecycle*. – San Francisco, 1999

[METS] *Metadata Encoding and Transmission Standard* – The Library of Congress, 2005 – http://www.loc.gov/standards/mets (Stand 28.7.2005)

[Meyer] Institut für Getriebetechnik und Maschinendynamik, Aachen – http://www.igm.rwth-aachen.de/deutsch/forsveroeffentlichungen/index.php (Stand 28.7.2005)

[PDD] *Portal Digitalisierte Drucke* – http://selene.hbz-nrw.de/pdd (Stand 28.7.2005)

[Reuleaux 1875] Reuleaux, F.: *Lehrbuch der Kinematik*. – Braunschweig, 1875

Text-Fakten-Integration in Informationssystemen

Maximilian Stempfhuber, Bonn

Abstract

Trotz zunehmender Integration bislang disparater Informationsangebote herrscht bislang die Aufteilung unterschiedlicher Informationstypen (z. B. von Primärinformationen in Form von statistischem Zahlenmaterial und Sekundärinformationen in Form von Texten) auf spezialisierte Informationssysteme vor. Aus Sicht des Nutzers ist dies ein unbefriedigender Zustand, da zur Befriedigung seines Informationsbedürfnisses häufig Informationen unterschiedlichen Typs aus mehreren Quellen notwendig sind. Am Beispiel der Markforschung wird gezeigt, wie eine Integration heterogener Informationen auf konzeptueller, informationswissenschaftlicher und ergonomischer Ebene möglich ist und wie neuen Nutzergruppen der einfache Zugang zu diesen Informationen ermöglicht werden kann.

1 Einleitung

Mit zunehmender Verbreitung standardisierter und zugleich offener Kommunikationsstandards im Internet, die sowohl in Bezug auf Ausdrucksmächtigkeit als auch hinsichtlich ihrer technischen Robustheit und Integrierbarkeit in bestehende IT-Landschaften das „wilde“ Publizieren von Information nachhaltig in Richtung einer serviceorientierten Informationsarchitektur (s. Erl 2005) verschoben haben, stellt sich die Frage, inwieweit die Sicht der Nutzer dadurch bereits berücksichtigt wurde. Zweifellos stellen anwendungsneutrale Kommunikationsprotokolle wie SOAP[1] und die damit realisierten WebServices[2] (s. Weerawarana et al. 2005) genauso wie spezialisierte Protokolle und Formate im Kontext verteilter Informationssysteme (z. B. OAI-PMH[3], das Dublin Core Metadata Element Set[4] oder das Beschreibungsformat der Data Documentation Initiative[5]) eine wichtige Voraussetzung dafür dar, die sichtbaren und „unsichtbaren“, aus Datenbanken gelieferten Inhalte (s. Bergman 2001) des World Wide Web aufzufinden und zu konkreten Informationsdienstleistungen zu kombinieren. Erste Resultate aus dem Bereich der wissenschaftlichen

[1] Simple Object Access Protocol, s. http://www.w3.org/2000/xp/Group/

[2] s. http://www.w3.org/2002/ws/

[3] Open Archive Initiative Protocol for Metadata harvesting, s. http://www.openarchives.org/documents/

[4] s. http://www.dublincore.org/documents/

[5] s. http://www.icpsr.umich.edu/DDI/

Informationsversorgung – wie zum Beispiel das Wissenschaftsportal vascoda[6] – zeigen das Potential dieser integrativen Ansätze und lassen erahnen, welche Möglichkeiten diese Technologien für Digitale Bibliotheken bieten.

Doch stellt die bloße Verfügbarkeit technologischer Lösungen oder Standards bereits eine hinreichende Bedingung dafür dar, dass darauf aufbauende Lösungen aufgabenangemessen sind, den Nutzer also bei der Befriedigung seines Informationsbedürfnisses adäquat unterstützen? Zieht man die Ergebnisse jüngster Benutzerbefragungen heran (s. IMAC 2002, RSLG 2002, Poll 2004), so wird schnell klar, dass das Ziel noch nicht erreicht ist. Neben den teilweise widersprüchlichen – aber dennoch verständlichen – Forderungen nach fachlich ausgerichteten, aber dennoch interdisziplinär verknüpften Angeboten, nach umfassendem Zugriff auf alle potentiell relevante Information bei gleichzeitiger Vermeidung des „information overflow“ und nach direktem Zugriff auf die gesuchte Information ist vor allem der Wunsch hervorzuheben, an einer Stelle integriert auf unterschiedlichste Informationstypen zugreifen zu können. Gerade diese Forderung wird bislang kaum erfüllt, sind doch die verfügbaren Informationssysteme zum einen geprägt vom Nachweis von Sekundärinformationen in Form von Literatur und zum anderen – soweit überhaupt vorhanden – von einer Spezialisierung auf einzelne Informationstypen. So nachvollziehbar dies aus Anbietersicht ist – schließlich handelt es sich ja bei vielen entweder um Produzenten/Vermittler von Primär- (z. B. statistische Ämter oder Datenarchive) bzw. Sekundärinformationen (z. B. Bibliotheken und Fachinformationseinrichtungen) –, für den Benutzer ergibt sich daraus die wenig befriedigende Situation, dass er mehrere, spezialisierte Informationsdienste aufsuchen, seine Anfrage wiederholt formulieren und mittels unterschiedlicher Anfragesprachen und Benutzungsoberflächen umsetzen muss, und dass es ihm schließlich obliegt, die Ergebnisse unterschiedlicher Quellen einzusammeln, zu bewerten und zu integrieren.

Am Beispiel der Marktforschung wird im Folgenden dargestellt, welche komplexen Zusammenhänge zwischen dem Informationsbedürfnis eines Nutzers und den dafür relevanten Informationstypen bestehen und warum die bloße Anwendung standardisierter Modelle und Technologien aus dem Bereich des Information Retrieval hier noch keine qualitativ befriedigenden Ergebnisse garantiert. Darauf aufbauend wird ein informationswissenschaftliches Modell für die nutzerorientierte Integration heterogener Informationstypen vorgestellt und seine konkrete Umsetzung anhand des Verbandsinformationssystems ELVIRA[7] erläutert. Der Schwerpunkt liegt dabei auf den eingesetzten Verfahren zur Behandlung der semantischen Heterogenität in der Datenbasis, sowohl auf der Ebene der Wissensorganisation als auch der softwareergonomischen Umsetzung.

[6] http://www.vascoda.de

[7] Das ELektronische VerbandsInformations-, Recherche- und Analysesystem (ELVIRA) wurde zwischen 1995 und 2000 gefördert vom Bundesministerium für Wirtschaft (BMWi).

2 Fallbeispiel: Marktforschung

An einem Beispiel aus der Marktforschung soll zunächst dargestellt werden, welche Rolle unterschiedliche Informationstypen und deren Verknüpfung schon bei relativ einfachen Informationsbedürfnissen spielen. Im Vorfeld der Entscheidung, ein Produkt auf einem internationalen Markt einzuführen, könnte sich folgender mehrstufige informationelle Prozess abspielen:

- Um eine generelle Entscheidung für die Produkteinführung in einen Markt zu treffen, wird zunächst das Marktvolumen, d. h. die Größe des Marktes ermittelt. Hierzu werden der Umfang der Produktion des Produkts im Zielland sowie die entsprechenden Exporte und Importe ermittelt (stat. Zeitreihen).
- Scheint der Markt groß genug für ein Engagement, ist zu prüfen, ob rechtliche Bestimmungen (z. B. Aus-/Einfuhrbeschränkungen, Zollbestimmungen usw.) den Markteintritt behindern könnten. Hierzu wird in speziellen Literaturdatenbanken recherchiert (Texte, in Texten eingebettete Fakten).
- Zur Kontaktaufnahme vor Ort werden Ansprechpartner in Behörden und Handelskammern oder auch Anwälte benötigt (Fakten).
- Einen Überblick über den derzeitigen Markt und Kundenwünsche geben z.B. Messedatenbanken oder Zeitschriften (Fakten und Texte).
- Zur Ermittlung der größten Mitbewerber auf dem Markt werden nationale und internationale Statistikdaten benötigt (stat. Zeitreihen).

Aus dem Beispiel ist sofort ersichtlich, dass nicht eine einzelne Recherche nach einer bestimmten Datenart in einer isolierten Datenbank betrachtet werden kann, wenn ein Informationssystem für die Marktforschung konzipiert werden soll, sondern dass nur der Gesamtprozess alle wichtigen Determinanten liefert. Um empirische Belege für die Wechselwirkung zwischen Informationstypen zu erhalten, wurden im Projekt ELVIRA II, in dem es speziell um die Zusammenführung von statistischen Daten und Texten innerhalb von Informationssystemen ging, weit über 100 Anfragen von Markforschern an die Statistikabteilungen von zwei Industrieverbänden analysiert. Dadurch sollten zum einen typische Informationsbedürfnisse ermittelt werden, zum anderen aber auch herausgefunden werden, ob ein Zusammenhang zwischen Fragestellungen und gewünschtem Informationstyp besteht und welche Arten von Information bei der Beantwortung der Anfragen verwendet werden. Die Auswertung der Anfragen sowie zusätzlicher Interviews mit Marktforschern ergab, dass enge Wechselwirkungen zwischen Faktendaten und Texten bestehen und nur die Verbindung beider eine adäquate Informationsbasis schafft:

- Nutzer haben bereits bei der Formulierung ihrer Anfrage – und damit auch bei der Auswahl des Informationssystems – die Vorstellung eines „Prototyps“ an Information (z. B. einer Zeitreihe oder eines Zeitschriftenartikels), der ihrem Informationsbedürfnis vermutlich am besten gerecht wird.
- Da unterschiedliche Informationstypen einen spezifischen zeitlichen Bezug zum gewünschten Betrachtungszeitpunkt haben (Zeitschriften greifen i. d. R.

sehr schnell aktuelle Ereignisse auf, während Statistiken teilweise erst Monate nach der Datenerhebung publiziert werden), bestimmt u. U. der Aspekt der Aktualität die Wahl des Informationstyps.

- Häufig führen Datenlücken bzw. die Nichtverfügbarkeit von Faktendaten zum Rückgriff auf Texte, in denen die gewünschten oder ähnliche Informationen enthalten sein könnten.
- Nicht alle Phänomene in statistischen Daten können alleine auf der Basis ihrer selbst erklärt werden. Häufig sind hierzu Hintergrundinformationen aus Texten notwendig (z. B. über Änderungen in der Statistik, Wahlen, Terroranschläge, Naturkatastrophen usw.).

3 Integrationsmodell für Primär- und Sekundärinformationen

Die empirischen Befunde belegen deutlich, dass der nutzerseitigen Integration von unterschiedlichen Informationstypen eine hohe Priorität bei der Entwicklung neuer Informationssysteme zukommen muss. Diese Forderung wird in jüngster Zeit von vielen Seiten verstärkt erhoben, wenn auch mit unterschiedlicher Schwerpunktsetzung. So fordert z. B. die Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (s. KVI 2001) einen generell besseren Zugang zu statistischen Daten, der durch die Schaffung von Forschungsdatenzentren erreicht werden soll, die sich auf die Aufbereitung und Dokumentation von Primärdaten konzentrieren und den qualitativ verbesserten, aber dennoch singulären Zugriff auf ihr eigenes Informationsangebot zum Ziel haben. In die gleiche Richtung zielen sowohl Überlegungen der DFG (siehe DFG 2005) zur besseren Einbindung von Primärinformationen in den Prozess des elektronischen Publizierens sowie mittlerweile in der Praxis am Beispiel der Naturwissenschaften im Test befindliche Verfahren zur Verbesserung der Zitierfähigkeit von Primärinformationen (siehe Lautenschlager & Sens 2003).

Neben datentypsspezifischen Problemen bei der Aufbereitung und Zugänglichmachung von Informationen (z. B. durch Anonymisierung von Umfragedaten in scientific und public use files) muss das Hauptaugenmerk bei der Konzeption von Informationssystemen auf der Modellierung der Informationsbedürfnisse und Suchstrategien der zukünftigen Nutzer liegen. Das Modell muss dabei die Nutzeranforderungen mit der Datengrundlage abgleichen und aufzeigen, wo und wie beide Ebenen in Einklang zu bringen sind.

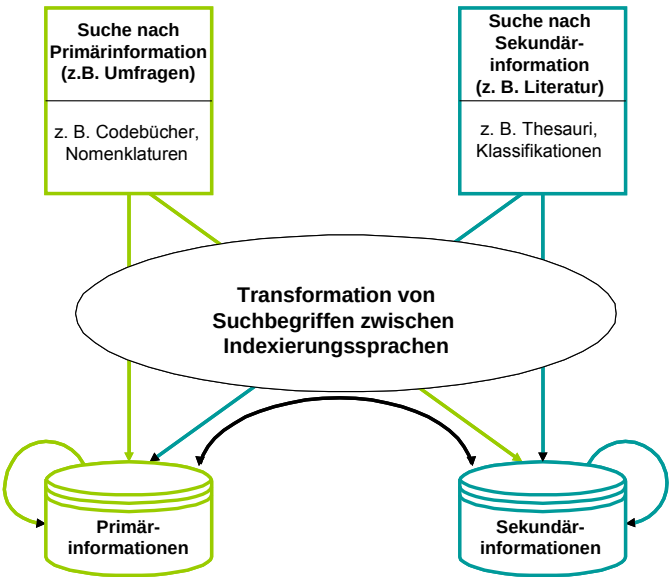


Abbildung 1: Modell für die Text-Fakten-Integration

Die Abbildung 1 zeigt das Grundschema des für ELVIRA entwickelten und mittlerweile für andere Anwendungsfälle angepassten Modells der Text-Fakten-Integration (s. Krause et al. 1997), in dem die Ebene der Nutzer (Anfragen nach Primär- oder Sekundärinformationen) mit den unterschiedlichen Datenarten verbunden ist und Relationen die möglichen Verbindungen beider Ebenen aufzeigen. Das Modell berücksichtigt auf Nutzerseite die Tatsache, dass der Informationssuchende bereits bei der Anfrageformulierung einen Prototyp des gewünschten Ergebnisses (also z. B. einen Zeitschriftenartikel oder eine statistische Zeitreihe, ggf. in Form einer Grafik) „vor Augen hat“ und ihm daher spezialisierte Möglichkeiten zur Anfrageformulierung angeboten werden müssen. Eng verbunden mit der Art der Anfrageformulierung sind die Indexierungsvokabulare, die zur Suche in den Daten verwendet werden können. Üblicherweise herrschen im Bereich von Primärdaten Nomenklaturen (z. B. von statistischen Ämtern) und Codebücher vor, während bei Sekundärdaten sowohl kontrollierte Vokabulare (z. B. Thesauri und Klassifikationen) als auch durch automatische Indexierung erzeugte, nicht kontrollierte Vokabulare zur Verfügung stehen.

Für den Nutzer bedeutet dies, dass er neben dem Recherchezugang auch das Vokabular wählen muss, mit dem er seine Anfrage formuliert. Abhängig von der Datenlage und der Komplexität der Suche tritt der Fall auf, dass der Nutzer sowohl für einen Informationstyp seine Anfrage mehrfach stellen muss (z. B. für die Suche in unterschiedlichen Statistikdatenbanken mit jeweils eigener Nomenklatur), als auch bei der Suche nach unterschiedlichen Informationstypen. In solchen Fällen gilt es,

den kognitiven Aufwand beim Benutzer durch geeignete systemseitige Verfahren zu unterstützen und die wiederholte Anfrageformulierung unnötig zu machen.
Gleichzeitig sollte es dem Nutzer ermöglicht werden, die von ihm präferierte – meist die ihm am besten bekannte – Rechercheoberfläche zu nutzen, unabhängig vom gewünschten Informationstyp. Im Text-Fakten-Integrationsmodell von ELVIRA ermöglichen dies Transformationen, die Nutzeranfragen zwischen Indexierungsvokabularen und auch zwischen Retrievalmodellen (z. B. Boolesches Modell vs. Vektorraummodell) abbilden, in dem sie Anfrageterme zwischen Vokabularen abbilden und die für das Retrievalmodell des jeweiligen Zielsystems notwendige Syntax erzeugen. Dabei werden Domänenrestriktionen ausgenutzt, um die Komplexität zu senken und gleichzeitig die Präzision zu erhöhen.
Anhand des Informationssystems ELVIRA wird nun im Folgenden dargestellt, wie die einzelnen Bestandteile des Modells in der Praxis operationalisiert und dazu genutzt werden können, der semantischen Heterogenität von Text- und Faktendaten zu begegnen, ohne einerseits eine Standardisierung zwingend vorzuschreiben (was in der Praxis nicht umsetzbar wäre) oder durch unspezifische Verfahren (z. B. eine reine automatische Indexierung) ein nur suboptimales Ergebnis zu erreichen.

4 Das Informationssystem ELVIRA

Das Ziel der Entwicklung des Informationssystems ELVIRA (s. Scheinost et al. 1998) war in seiner ersten Phase (1995/1996) die Erleichterung des Zugangs zu statistischen Daten insbesondere für kleine und mittelständische Unternehmen, die im Gegensatz zu Großunternehmen die von statistischen Ämtern und Industrieverbänden zur Verfügung gestellten Daten nur unzureichend in der Marktforschung einsetzten. In der zweiten Projektphase (1997 - 2000) erfolgte die prototypische Weiterentwicklung zu einem Marktinformationssystem, das neben statistischen Daten auch weitere, in der Marktforschung benötigte Informationen unter einer einheitlichen Benutzungsoberfläche integriert. Seit 1997 wird ELVIRA den Mitgliedsfirmen des Hauptverbands der Deutschen Bauindustrie (HVBi) und des Zentralverbands der Elektrotechnik und -elektronikindustrie (ZVEI) zur Verfügung gestellt. Seit dieser Zeit wurde der über ELVIRA verfügbare Datenbestand von ca. 2 Mio. Zeitreihen auf über 18 Mio. Zeitreihen aus aller Welt ausgeweitet. Über 750 Nutzer aus Mitgliedsfirmen der beiden Industrieverbände, aus Ministerien, Banken und Unternehmensberatungen nutzen mittlerweile das System, das nach dem Projektende vom Informationszentrum Sozialwissenschaften in Kooperation mit den Verbänden laufend erweitert und verbessert wird.

4.1 Semantische Integration von Fakten und Texten

Das informationswissenschaftliche und softwareergonomische Konzept von ELVIRA beruht auf dem in Abbildung 1 dargestellten Modell zur Text-Fakten-Integration.

Dieses Modell wird in ELVIRA benötigt, da die z. B. im Bereich Elektroindustrie verfügbaren Daten aus 30 Beständen mit jeweils eigenen Nomenklaturen verschlagwortet sind, so dass insgesamt 46 Nomenklaturen zur Anfrageformulierung verwendet werden müssen. Die Zahl der kontrollierten Vokabulare erhöht sich noch bei Berücksichtigung relevanter Textbestände, wenngleich hier aber in der Praxis weniger unterschiedliche Thesauri berücksichtigt werden müssen.
Die Abbildung zwischen den Indexierungsvokabularen im Bereich der Primärdaten erfolgt über einen zentralen Thesaurus, der – als Vorstufe einer im Aufbau befindlichen Ontologie – die über 33.000 Einträge der 46 Nomenklaturen auf ca. 900 Deskriptoren reduziert. Die Deskriptoren des Thesaurus, die mit ca. 19.000 zusätzlichen Synonym-Relationen angereichert sind, verweisen auf die ursprünglichen Nomenklaturpositionen, so dass eine zentrale Vermittlungsstruktur entstanden ist, die zwischen Einträgen unterschiedlicher Nomenklaturen (Nomenklatur-Thesaurus-Nomenklatur) genauso transformiert wie zwischen dem Sprachgebrauch des Nutzers und der kontrollierten Sprache der Nomenklaturen (Synonym-Deskriptor-Nomenklatureintrag).

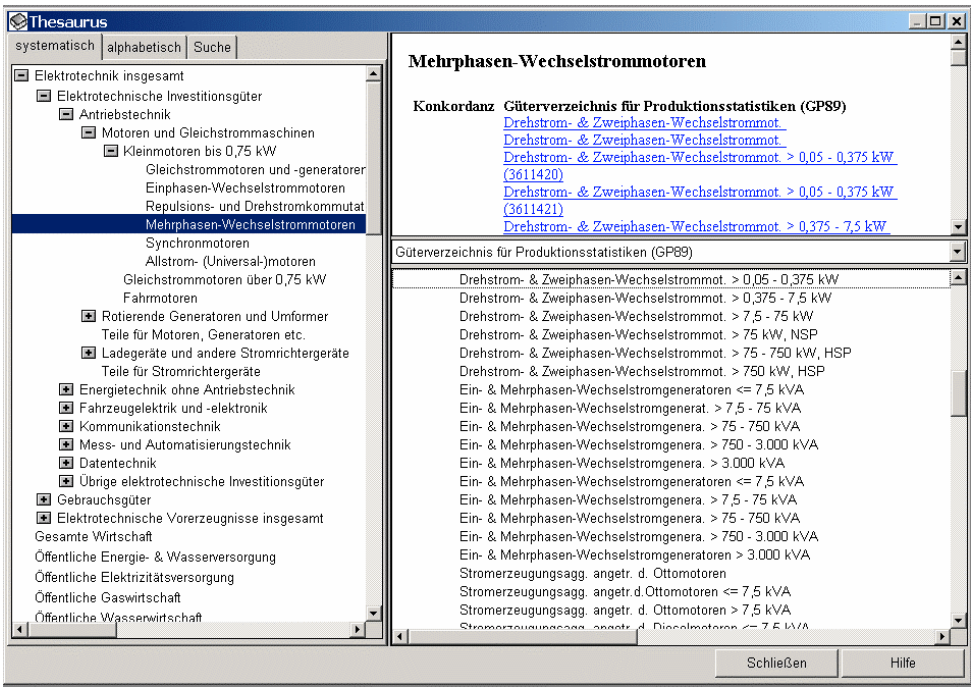


Abbildung 2: Systematische Ansicht des Thesaurus in ELVIRA

Die Abbildung 2 zeigt die systematische Ansicht des Thesaurus (links), die Detailanzeige des Deskriptors zusammen mit den Verweisen auf Nomenklatureinträge und Synonyme (rechts oben) und dem Ausschnitt jeweils einer Nomen

klatur, zu der der Deskriptor relationiert ist (rechts unten). Der Nutzer hat damit die freie Wahl, seine Suchanfrage direkt in der Nomenklatur, über den systematischen Zugang zum Thesaurus oder über eine Suche im Thesaurus zu formulieren.

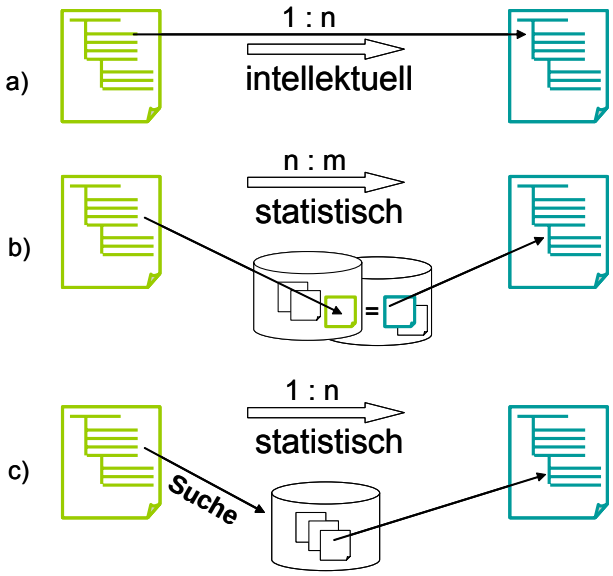


Abbildung 3: Arten der Termtransformation zwischen Indexierungsvokabularen

Zwischen den Indexierungsvokabularen für Primär- und Sekundärdaten werden unterschiedliche Transformationsverfahren verwendet. Die Abbildung 3 zeigt verfügbare Varianten, die je nach Datenlage verwendet werden. Neben der intellektuellen Erstellung von Crosskonkordanzen (Abbildung 3a) zwischen Indexierungsvokabularen, wobei jeweils ein Term des Ausgangsthesaurus mit einem oder mehreren Termen des Zielthesaurus verbunden wird, spielen hier vor allem auch statistische Verfahren eine Rolle, da die Ausgangsterme (aus Produktnomenklaturen) häufig sehr branchenspezifisch sind, wogegen das Indexierungsvokabular für Texte meist auf einem unspezifischeren, aber dennoch fachlichen Niveau angesiedelt ist (z. B. beim Standardthesaurus Wirtschaft). Statistische Verfahren können dann eingesetzt werden, wenn entweder gleiche, aber unterschiedlich erschlossene Dokumente in zwei Datenbeständen identifiziert werden können (Parallelkorpora, Abbildung 3b), oder wenn das Vokabular des Ausgangsthesaurus für eine Freitextsuche in einem unterschiedlich indexierten Textbestand (Metadaten und Volltext) verwendet werden kann (simulierte Parallelkorpora, Abbildung 3c). In beiden Fällen lässt sich zu einem Ausgangsterm eine Gruppe von Termen des Zielthesaurus bestimmen, deren Auftretenswahrscheinlichkeit im Kontext einer größeren Zahl indexierter Dokumente oberhalb eines datenbestandsspezifisch zu definierenden Cut Off-Werts liegt. Diese Auftretenswahrscheinlichkeiten können

dann anstatt von Crosskonkordanzen genutzt werden, Faktenanfragen in Textanfragen und vice versa zu transformieren.
Darüber hinaus werden die in ELVIRA aufgebauten Termtransformationen auch für die Termerweiterung bei Freitextsuchen benutzt. Dies ist notwendig, da die Sprache der amtlichen Statistik – und damit der Nomenklaturen zu den Primärdaten – sehr stark von der in Texten verwendeten Sprache abweicht. Eine Vielzahl der in Nomenklaturen verwendeten „komplexen" Begriffe wird z. B. in Texten durch umgangssprachliche Ausdrücke ersetzt. Eine Suche nach „Waschvollautomat", dem Begriff der amtlichen Statistik, führt z. B. im Informationsverbund Econdoc[8] zu keinem, in den Datenbanken der GBI[9] (Freitextsuche) zu 57 und bei Google zu ca. 170.000 Treffern, die Suche nach dem gebräuchlicheren Begriff „Waschmaschine" führt dagegen zu 37 bzw. 6.622 und ca. 600.000 Treffern.

4.2 Aufgabenadaptive Unterstützung des Nutzers

Um dem Nutzer je nach Informationsbedürfnis eine für den gewünschten Informationstyp optimierte Benutzungsoberfläche (s. Stempfhuber 2003) anbieten zu können, wurde in ELVIRA sowohl ein Zugang für die Faktenrecherche als auch für die Textrecherche implementiert, die über die oben beschriebenen Transformationsmodule so gekoppelt sind, dass jeder Zugang sowohl für die Suche nach Texten als auch nach Fakten verwendet werden kann.
Die Abbildung 4 zeigt die Benutzungsoberfläche für die Faktensuche. Sie visualisiert die Sacherschließung der Faktendaten durch jeweils ein Schlagwort der Bereiche „Thema", „Branche/Produkt" und „Land" durch drei entsprechende Vorlagelisten (die Liste für Länder ist in Abbildung 4 verdeckt), die so synchronisiert sind, dass nach Wahl eines Themas nur noch die Branchen und Länder zur Verfügung stehen, zu denen Daten gefunden werden können. Die Auswahl des Nutzers in den Nomenklaturen wird in einer Zustandsanzeige am oberen Bildschirmrand widergespiegelt, so dass auch beim Navigieren in den umfangreichen Listen der aktuelle Status der Anfrage ortskonstant angezeigt wird. In der Zustandsanzeige können gleichzeitig bereits ausgewählte Suchbegriffe gelöscht oder mittels Freitextsuche und Eingabe von Kennziffern Nomenklatureinträge gesucht oder direkt ausgewählt werden.

[8] http://www.econdoc.de
[9] http://www.gbi.de

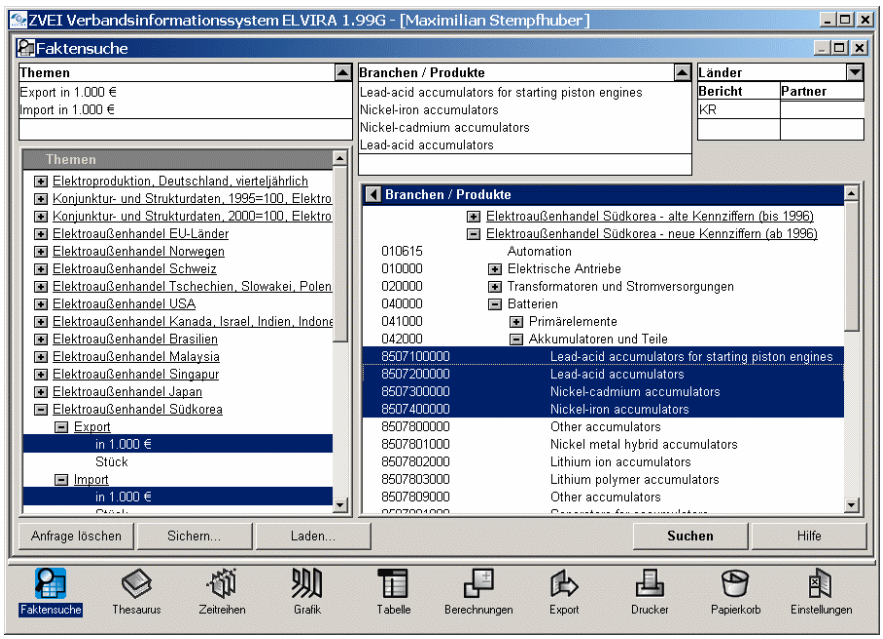


Abbildung 4: Adaptive Benutzungsoberfläche für die Faktensuche

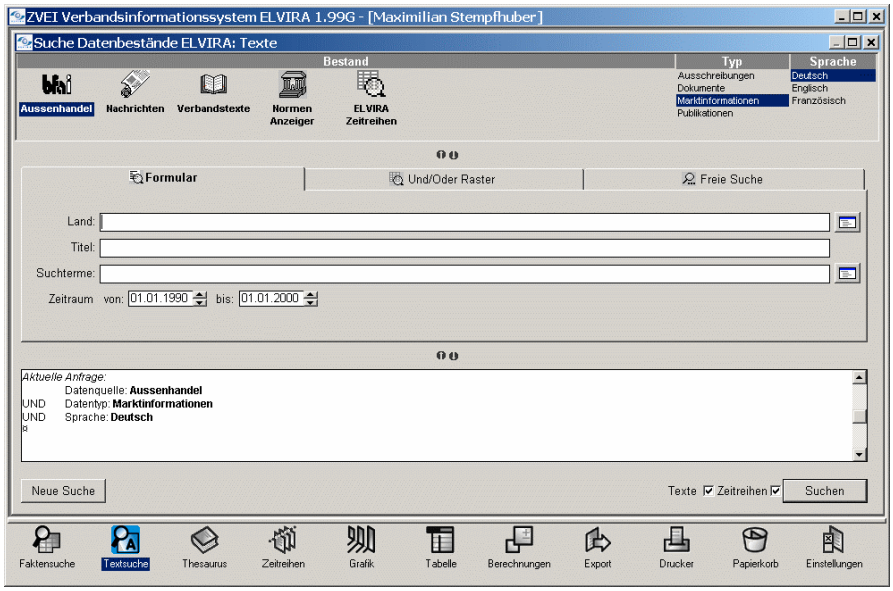


Abbildung 5: Benutzungsoberfläche zur Textsuche

Zur Suche nach Texten stellt ELVIRA mehrere alternative Eingabemöglichkeiten zur Verfügung (Abbildung 5), die analog zur Faktensuche über eine Zustandsanzeige (Abbildung 5, unten) gekoppelt und repräsentiert werden. Neben einer Formularsuche existieren ein Und-/Oder-Gitter, in dem komplexere Boolesche Ausdrücke eingegeben werden können, und eine Freitextsuche.

5 Resümee und Ausblick

Mit dem Informationssystem ELVIRA liegt ein System vor, mit dem heterogene und unterschiedlich erschlossene Informationen in großem Umfang sowohl auf technischer und semantischer, aber auch auf softwareergonomischer Ebene integriert werden können. Waren beim Projektende von ELVIRA im Jahr 2000 noch kaum Textdatenbestände verfügbar, die mit überschaubarem technischen und finanziellen Aufwand über das Internet zugänglich waren, so hat sich diese Situation gerade auch mit der Verfügbarkeit des Wissenschaftsportals vascoda und der daran angeschlossenen Informationsdienstleister grundlegend gewandelt. Der aktuelle Forschungs- und Entwicklungsschwerpunkt im Kontext von ELVIRA liegt daher auf der Integration dieser Sekundärinformationen mit den bereits in großem Umfang in ELVIRA verfügbaren Primärinformationen. Neben den rein technischen Aspekten spielt hier vor allem eine weitere Verfeinerung und Evaluierung des Modells zur Text-Fakten-Integration und der damit einhergehenden Transformationsverfahren eine zentrale Rolle, die vor allem auch im Bereich der empirischen Sozialwissenschaften eine hohe Relevanz besitzt (Krause & Stempfhuber 2005).

Literatur

Bergman, Michael K. (2001): The Deep Web: Surfacing Hidden Value http://www.brightplanet.com/technology/deepweb.asp

DFG (2005): DFG-Positionspapier: Elektronisches Publizieren. März 2005. http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/pos_papier_elektron_publizieren_0504.pdf

Erl, Thomas (2005): Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall PTR.

IMAC 2002: Projekt Volltextdienst. Zur Entwicklung eines Marketingkonzepts für den Aufbau eines Volltextdienstes im IV-BSP. IMAC Information & Management Consulting. Konstanz. September 2002.

Krause, Jürgen; Mandl, Thomas; Stempfhuber, Maximilian (1997): Text-Fakten-Integration in ELVIRA. Bonn: IZ Sozialwissenschaften, IZ-Arbeitsbericht Nr. 12. http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/pdf/ab12.pdf

Krause, Jürgen; Stempfhuber, Maximilian (2005): Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen. In: Datenfusion und Datenintegration. Tagungsberichte Band 10, IZ Sozialwissenschaften (erscheint).

KVI (2001): Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik. Herausgegeben von Walter Müller. Baden-Baden: Nomos, 2001.

Lautenschlager, Michael; Sens, Irina (2003): Konzept zur Zitierfähigkeit Wissenschaftlicher Primärdaten. http://www.std-doi.de/Paper-Konzept-Primdaten.pdf

Poll, Roswitha (2004): Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung, Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft. In: ZfBB 51 (2004), S. 59 - 75.

RSLG (2002): Researchers' Use of Libraries and other Information Sources: Current Patterns and Future Trends. Final Report / Education for Change Ltd.; SIRU, University of Brighton & The Research Partnership, 2002 http://www.rslg.ac.uk/research/libuse/. In Kurzform erläutert in: „Research Support Libraries Group: Final Report". In: The New Review of Academic Librarianship, 8 (2002), S. 3 - 86.

Scheinost, U.; Haas, H.; Krause, J.; Lindlbauer, J. (Hrsg.) (1998): Marktanalyse und Marktprognose: das ZVEI-Verbandsinformationssystem ELVIRA. Bonn: IZ Sozialwissenschaften. 206 S. (Forschungsberichte; 2).

Stempfhuber, Max (2003): Objektorientierte Dynamische Benutzungsoberflächen ODIN: Behandlung semantischer und struktureller Heterogenität in Informationssystemen mit den Mitteln der Softwareergonomie. Bonn: IZ Sozialwissenschaften. 337 S. (Forschungsberichte; 6).

Weerawarana, Sanjiva; Curbera, Francisco; Leymann, Frank; Storey, Tony; Ferguson, Donald F. (2005): Web Services Platform Architecture: SOAP, WSDL, WS-Policy, WS-Addressing, WS-BPEL, WS-Reliable Messaging, and More. Prentice Hall PTR.

Evaluierung von Visualisierungsverfahren bei der webbasierten Suche

Claus Arnold, Christian Wolff; Regensburg

Abstract

Der Beitrag berichtet über die Ergebnisse einer empirischen Untersuchung, bei der eine klassische Suchmaschine mit textorientierter Such- und Ergebnisdarstellung mit Recherchesystemen verglichen wurde, die den Benutzer durch Verfahren der Informationsvisualisierung unterstützen. Dabei widmete sich die Untersuchung nicht der Bewertung der eigentlichen Retrievaleffektivität, da bei unterschiedlichen Datengrundlagen keine Vergleichbarkeit herzustellen war, sondern konzentriert sich auf die Frage, ob Darstellungsformate und Interaktionsmöglichkeiten den Benutzer bei der Durchführung unterschiedlicher Typen von Recherchen unterstützen. Aus den Ergebnissen werden Thesen zu besseren Ansätzen der Evaluierung von Visualisierungskomponenten im Information Retrieval hergeleitet.

1 Einleitung

Informations- und Konzeptvisualisierung ist ein Forschungsfeld, auf dem in den vergangenen 20 Jahren eine Vielzahl experimenteller Systeme zur Unterstützung der Informationserschließung entwickelt wurde. Mittlerweile haben solche Systeme den Status reiner Forschungsprototypen überschritten und stehen für die Verwendung bei der webbasierten Suche im Rahmen von Suchmaschinen zur Verfügung. In der Forschungsliteratur zeigen sich allerdings erhebliche Defizite, was die Begründung für den Einsatz von Visualisierungsverfahren angeht: Meist werden relativ pauschale Argumente („ein Bild sagt mehr als tausend Worte“) herangezogen, um den Einsatz von Visualisierungsverfahren plausibel zu machen; gleichzeitig liegen bisher nur sehr wenige empirische Studien zum Einsatz von Informationsvisualisierungsverfahren vor. Klar zu sein scheint nur, dass individuelle Präferenzen der Benutzer für die Nutzbarkeit textuell bzw. graphisch ausgerichteter Darstellungsformate von Bedeutung sind (vgl. Lohse 1991, Wolff 1996: 177ff) und dass gleichzeitig die gleiche Informationsmenge, dargestellt in unterschiedlichen Formaten (z. B. textuell oder tabellarisch versus visuell) zu unterschiedlichem Aufwand bei der Beantwortung einer Frage führen kann:

"Two representations are informationally equivalent, if all the information in the one is also inferable from the other, and vice versa. Each could be constructed from the other. Two representations are computationally equivalent, if they are informationally equivalent and, in addition, any inference that can be drawn easily and quickly from

the one can also be drawn easily and quickly from the information given explicitly in the other, and vice versa." (Larkin & Simon 1987: 67).
Der Definition von Card et al. folgend ("Information Visualization: The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." (Card et al., 1999: 7) ist Informationsvisualisierung grundsätzlich von der Datenvisualisierung abzugrenzen, bei der der Wertebereich von Messdaten in das visuelle Medium transformiert wird; bei der Informationsvisualisierung liegt dem Visualisierungsprozess eine (intellektuelle) Zuordnung visueller Merkmale zu sprachlichen Konzepten zugrunde, z. B. die Darstellung und Anordnung von Dokumentsymbolen in einer Dokumentlandkarte oder die Anordnung und Verknüpfung von Begriffslabeln in einem Graphen.

2 Visualisierung und Information Retrieval

Auch im Information Retrieval haben Ansätze zur Visualisierung mittlerweile eine breite Vielfalt an Forschungsprototypen hervorgebracht (Übersichten hierzu finden sich etwa bei Chen 1999, Eibl 2003, Wild 2004). Die Visualisierungsansätze lassen sich dabei zum einen nach den verwendeten visuellen Mittel (z. B. Landkarten, Graphen, 3D-Metaphern, Visualisierung von Dokumentinhalten), zum anderen nach der Zuordnung zu einer bestimmten Phase des Retrievalprozesses klassifizieren: Die wesentlichen Schritte sind dabei

- Die Unterstützung der Anfrageformulierung (z. B. das System DeViD, vgl. Eibl 2003 oder das hier getestete System WebBrain, s. u. Kap. 3),
- die visuelle Unterstützung der Darstellung von Ergebnismengen, insbesondere durch Dokumentlandkarten (vgl. Chen et al. 1998),
- Die Visualisierung einzelner Dokumentinhalte (z. B. das TileBars-System, vgl. Hearst 1995).

Auffällig ist dabei, dass die Begründungen für die Einführung visueller Darstellungsmittel sehr heterogen und in der Regel ohne vertieften Rückbezug auf die wahrnehmungspsychologischen Grundlagen erfolgen. Einige Beispiele aus der Literatur sollen dies belegen.
Für das System *ThemeMaps*, bei dem auf der Basis von Kohonen-Maps Dokumentlandkarten erzeugt und dargestellt werden, nehmen die Autoren ohne weiteres in Anspruch, dass die graphische Darstellung für den Benutzer leicht navigierbar sei:
"... a Kohonen self-organizing map (SOM)-based algorithm can successfully categorize a large and eclectic Internet information space...into manageable subspaces that users can successfully navigate to locate a homepage of interest to them." (Chen et al, 1998: 82)
Für *IslandsInterface*, ebenfalls ein System dass eine Landkartenmetapher für die Dokumentdarstellung nutzt, stellen die Autoren auf die entfallende Notwendigkeit des

Erlernens einer Abfragesyntax und die durch die visuelle Darstellung gegebene Übersichtsfunktion ab:
"First, syntax-free creation of queries might eliminate errors, confusion, and lost time in the case of users unfamiliar with the particular query syntax. Second, being able to modify previous queries might help the user improve a query and reduce the proliferation of queries. Third, being able to take in, at a glance, the collection of all queries created so far during a search session might help the user formulate strategies for bringing the session to a successful conclusion."
(Brooks & Campbell, 1999: 15)
Die schnelle Analyse und Interpretation von Begriffsverteilungen in Dokumenten durch eine abstrakte visuelle Darstellung in Form sog. TileBars nimmt Hearst für ihr System an:
System: TileBars; "This paper argues for making use of text structure when retrieving from full text documents, and presents a visualization paradigm, called TileBars, that demonstrates the usefulness of explicit term distribution information in Boolean-type queries. [...] The patterns in a column of TileBars can be quickly scanned and deciphered, aiding users in making judgments about the potential relevance of the retrieved documents". (Hearst, 1995: 9)
Neben der in der Regel knappen theoretischen Begründung der Vorteile visueller Formate fällt auch das fast völlige Fehlen empirischer Studien zur Visualisierung im Information Retrieval auf. Nur wenige Systeme sind bisher evaluiert worden (vgl. Eibl & Mandl 2002: 163, Ogden et al. 1998), gleichzeitig handelt es sich bei IR-Systemen mit Visualisierungskomponenten bisher in der Regel um experimentellen Prototypen, die für spezifische und für einen allgemeinen Benutzerkreis kaum geeignete Domänen bzw. Dokumentkollektionen zum Einsatz kommen.

3 Vergleich von Visualisierungsansätzen im Web Retrieval

In der vorliegenden Studie wurde daher versucht, erste empirische Ergebnisse zum Einsatz von Visualisierungsverfahren im Web Retrieval zu gewinnen. Dabei dienten folgende Überlegungen als Motivation:

- Die Verwendung von Websuchmaschinen ist mittlerweile eine der wichtigsten Anwendungen im World Wide Web.
- Erfolgreiche Suchmaschinen wie Google (Google 2005) setzen seit geraumer Zeit auf eine stark textorientierte Darstellung die auf (nicht-textuelle) Visualisierungselemente weitestgehend verzichten.
- Es existieren eine Reihe von Suchmaschinen, die – anders als etwa Google – mit Visualisierungstechniken arbeiten; damit existiert eine frei zugängliche Anzahl von Systemen mit und ohne Visualisierungstechniken, die einem Benutzertest unterworfen werden können.

- Als Testpersonen sollten typische Benutzer des World Wide Web herangezogen werden, die nicht unbedingt gleichzeitig auch Experten für Informationserschließung sind.

3.1 Systemauswahl

Durch die Vorbedingung, dass nur im World Wide Web frei zugängliche Systeme getestet werden sollten, die zumindest näherungsweise über eine vergleichbare Datengrundlage verfügen, schränkte sich die Auswahl der verfügbaren Systeme stark ein. Nach einer ersten Sichtung von Visualisierungskomponenten bei WWW-Suchmaschinen wurden schließlich Google[1] als Referenzsystem sowie zwei Suchmaschinen mit Visualisierungskomponente ausgewählt: WebBrain und Kartoo (vgl. WebBrain 2005, Kartoo 2005).

3.1.1 WebBrain

Der Suchmaschinenaufsatz Webbrain bietet dem Benutzer als Visualisierung einen dynamischen Konzeptbrowser, bei dem eine hierarchische top-down-Auswahl von Konzepten jeweils das aktuelle gewählte Konzept in den Mittelpunkt der Darstellung rückt und jeweils Ober- und Unterbegriffe sowie benachbarte Konzepte und Querverweise darstellt (recht bzw. links von den in der Mitte dargestellten ausgewählten Begriffen). Abbildung 1 zeigt die Benutzeroberfläche von Webbrain für die Begriffskette „Computers – Internet – Searching“ bei hinreichender Verfeinerung der ausgewählten Begriffskette werden Treffer (Links) im unteren Teil des Bildschirms dargestellt, alternativ steht dem Benutzer auch die Möglichkeit einer direkten Begriffseingabe in ein Textfeld zur Verfügung.

[1] Auf eine nähere Vorstellung der Referenzsuchmaschine Google wird hier verzichtet.

Abbildung 1: Benutzerschnittstelle von WebBrain

3.1.2 Kartoo

Bei der Suchmaschine Kartoo erfolgt der Ersteinstieg in die Suche in „traditioneller“ Weise durch die Eingabe eines Begriffes in ein Textfeld, die Visualisierung konzentriert sich im nächsten Schritt auf eine graphische Darstellung der zum eingegebenen Begriff gefundenen Links, die in einer Karte angeordnet werden. Größe und Position der Darstellung einzelner Ergebnisse sollen die Struktur der Egebnismenge für den Suchenden besser verständlich machen, zusätzlich werden automatisch weitere Suchbegriffe in einer Liste dargestellt (linke Spalte).
Abbildung 2 zeigt das visuelle Suchergebnis für den Suchbegriff „Internetsuche“ bei Kartoo.

Abbildung 2: Benutzerschnittstelle von Kartoo

3.2 Testdesign

Für die Evaluierung im Information Retrieval und insbesondere die Bewertung ihrer Effektivität existieren seit langem etablierte Verfahren (vgl. Sparck Jones 1981, Robertson & Hancock-Beaulieu 1992). Typischerweise wird untersucht, inwieweit es bei einer Suche gelingt alle relevanten Treffer der Dokumentkollektion zu finden (*recall* bzw. Erschöpfungsgrad einer Anfrage) bzw. in welchem Verhältnis relevante zu nicht-relevanten Treffern in einer Ergebnismenge stehen (*precision* oder Genauigkeit einer Anfrage). Da die drei untersuchten Systeme zwar alle auf das World Wide Web als Datengrundlage zugreifen, gleichzeitig aber eine vergleichbare Größe bzw. ein vergleichbarer Inhalt des von den einzelnen Suchmaschinen erfassten Subsets des World Wide Web nicht sichergestellt werden konnte, kam eine solche – sicher sinnvolle – Effektivitätsbewertung zunächst nicht in Betracht. Das Testdesign wurde daher stärker auf die subjektive Bewertung software-ergonomischer Merkmale der Suchmaschinen durch die Versuchspersonen abgestellt.

3.2.1 Aufgaben

Die Formulierung geeigneter Rechercheaufgaben erfolgte auf der Basis der von Kang & King 2003 entwickelten Klassifikation von Suchproblemen im World Wide Web in

- Topic relevance tasks
- Homepage finding tasks und
- Service finding tasks.

Den insgesamt 15 Studenten, die als Versuchspersonen an dem Test mitwirkten, wurde je ein vorformuliertes Rechercheproblem für eine *topic finding task* (Informationen zu neueren wissenschaftlichen Erkenntnissen über den Planeten Saturn) bzw. eine *service finding task* (Finden von Websites, auf denen Eintrittskarten zur Fußball-Weltmeisterschaft 2006 erworben werden können) vorgelegt. Eine dritte Recherche konnte zu einem frei gewählten Suchproblem durchgeführt werden. Aufgrund der relativ geringeren Komplexität einer *homepage finding task* wurde dieser Rechercheтypus nicht in den Test miteinbezogen. Die Testpersonen sollten die Recherche mit den Suchmaschinen iterativ solange durchführen (verfeinern), bis ein aus ihrer Sicht befriedigendes Testergebnis vorlag.

3.2.2 Datenerhebung

Neben der Erfassung einschlägiger demographischer Daten sowie der Messung der für die jeweilige Aufgabenlösung mit einer Suchmaschine benötigten Zeit und der durchgeführten Interaktionsschritte umfasste der Test als zentrales Element einen Fragenkatalog, der sich wesentlich an der einschlägigen Norm zur gebrauchstauglichen Gestaltung von Dialogsystemen (ISO-EN 9241, Teil 10) orientiert. Dazu gehörten Fragen zu folgenden Aspekten der untersuchten Systeme:

- Aufgabenangemessenheit
- Selbstbeschreibungsfähigkeit
- Steuerbarkeit
- Erwartungskonformität
- Individualisierbarkeit
- Lernförderlichkeit
- Eignung für Wahrnehmung und Verständnis
- Eignung für Benutzerbeteiligung
- Sonstiges

Die Fragen waren von den Versuchspersonen jeweils auf einer vierteiligen Skala zu bewerten; insgesamt wurden den Versuchspersonen 25 Bewertungsfragen vorgelegt. Die nachfolgenden Beispiele zeigen exemplarisch einige der gestellten Fragen:

- War ersichtlich, was auf dem Bildschirm dargestellt werden sollte (also durch Symbole, Wörter, Texte)? (Frage 2 zur Selbstbeschreibungsfähigkeit)
- Waren die Verknüpfungen (Links bzw. Symbole) deutlich voneinander zu unterscheiden (z. B. durch genügende räumliche Abgrenzung)? (Frage 4 zur Steuerbarkeit)
- Waren die Informationen leicht zu lesen bzw. gut zu erkennen? (Frage 18 zu Wahrnehmung und Verständnis)

3.3 Wesentliche Ergebnisse

Bei den Messparametern *Zeit* und *Interaktionsschritte* zeigt sich jeweils eine hochsignifikante Überlegenheit von Google gegenüber den Systemen Webbrain und

Kartoo. Abbildung 3 zeigt die mittleren Bearbeitungszeiten der drei getesteten Systeme:

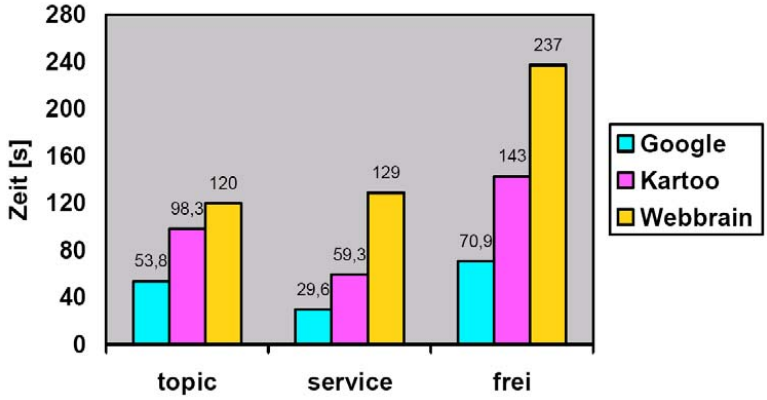


Abbildung 3: Mittlere Bearbeitungszeiten der drei getesteten Systeme nach Aufgabentyp (Arnold 2005: 105)

Auch der bei mittleren Zahl der Interaktionsschritte zeigt sich ein ähnliches Bild: Google liegt mit 5,4 – 7 Schritten deutlich vor Webbrain (10,4 – 16 Schritte) und Kartoo (12-29,2 Schritte).
Auch bei fast allen Bewertungsfragen ergibt sich für Google eine bessere mittlere Bewertung als für die beiden Suchmaschinen mit Visualisierungskomponente. Man hätte immerhin vermuten können, dass interaktive Visualisierungen auf die Benutzer ansprechend wirken, aber auch bei einer hierauf abzielenden Bewertung („War das System für Sie ansprechend? Motiviert es Sie, es weiterhin zu benutzen?“) liegt Google mit einer mittleren Bewertung von 1 (eindeutig „Ja“) deutlich vor Kartoo (2,2) und Webbrain (3,0). Lediglich bei der Frage nach der individuellen Anpassbarkeit der Systeme (Personalisierungsmöglichkeiten) liegt Kartoo etwas vor Google. Bemerkenswert ist dabei, dass die Versuchspersonen bei der subjektiven Bewertung der Qualität der Ergebnismengen Google nur knapp vor Kartoo sehen (Mittelwerte 1,3 bzw. 1,6), während WebBrain auch hier deutlich abfällt.
In der Nachbefragung der Versuchspersonen zu den getesteten Systemen wurden immerhin eine Reihe von Vorteilen der Visualisierung genannt, die sich allerdings hauptsächlich auf die Suchmaschine Kartoo beziehen. Dazu gehören neben der visuellen Hervorhebung relevanter Dokumente und der allgemein ansprechenden Graphik vor allem die in der Visualisierung bei Kartoo vorhandene thematische Gruppierung der Dokumente in der Treffermenge, die immerhin sechs Versuchspersonen als Vorteil nannten.

Zieht man in Betracht, dass die Benutzer angaben, mit Google mit Abstand die meisten Erfahrungen gesammelt zu haben und damit auch regelmäßig zu arbeiten, kann nicht überraschen, dass diese Suchmaschine durchgängig deutlich besser abschneidet als die Vergleichssysteme mit ihren visuellen Unterstützungsverfahren.

3.4 Methodische Probleme

Das auf den ersten Blick eindeutige Ergebnis der Untersuchung kommt vermutlich zu einem guten Teil aufgrund im Rahmen dieser Studie nicht lösbarer methodischer Probleme zustande: Die Versuchspersonen hatten erhebliche Erfahrungen mit der Suchmaschine Google bzw. mit textorientierten IR-Systemen. Der erfolgreiche Umgang mit Informationsvisualisierung kann durch einen – vermutlich längerfristigen – Lernprozess vermutlich nicht unwesentlich gefördert werden.

Gleichzeitig liegt bei Google eine breitere Dokumentbasis zugrunde, was auch die subjektiven Bewertungen der Systemergonomie beeinflussen kann, selbst wenn die Retrievaleffektivität nicht Gegenstand der Untersuchung war.

Zudem ist eine „pauschale“ Gegenüberstellung von Systemen – und damit die Bewertung des gesamten Retrievalprozesses möglicherweise nur bedingt geeignet, Vor- und Nachteile von Visualisierungen herauszuarbeiten, die sich nur auf einzelne Prozessschritte im Information Retrieval (bzw. auf Ergebnismengen in einer Darstellung als Dokumentlandkarte) beziehen.

Schließlich ist vorstellbar, dass für Retrievalexperten, die z. B. als professionelle Information Broker arbeiten und daher auch mehr Zeit in das Erlernen des Umgangs mit visuellen Formaten investieren können, Vorteile der Interaktion mit visuellen Formaten besser erschlossen werden können.

4 Fazit

In einer Studie zur software-ergonomischen Bewertung visueller Suchmaschinen konnte kein Vorteil gegenüber textorientierten Systemen wie Google festgestellt werden, auch quantitative Parameter wie Recherchezeit und Interaktionsschritte sprechen klar für die textorientierte Herangehensweise. Es wäre allerdings verfehlt, dies pauschal als Argument gegen visuelle Verfahren zu werten. Wie die voranstehend geschilderten methodischen Probleme deutlich machen, steht die Forschung sowohl zu den kognitiven Grundlagen von Informationsvisualisierungen im Information Retrieval als auch zu ihrer empirischen Bewertung noch am Anfang. Es ist daher anzuregen, geeignete Testformate zu entwickeln, bei denen sehr viel präziser einzelne Visualisierungsmerkmale für einen bestimmten Prozessschritt des Information Retrieval auf der Basis gleicher Dokumentkollektionen und Retrievalfunktionen getestet werden können. Eine technische Umsetzung könnte dabei – wie bei Metasuchmaschinen üblich – durchaus auf der Grundlage bestehender Systeme wie Google erfolgen.

Literatur

Arnold, Claus (2005). Visualisierung im Information Retrieval. Magisterarbeit, Universität Regensburg, Institut für Informations-, Medien- und Kulturwissenschaft.

Broder, Andrew (2002). „A Taxonomy of Web Search.“ SIGIR Forum 36(2) (2002), 3-10.

Brooks, Martin; Campbell, Jennifer (1999). “Interactive Graphical Queries for Bibliographic Search.” Journal of the American Society for Information Science (JASIS), 50(9), 814-825.

Card, Stuart; Mackinlay, Jock; Shneiderman, Ben (Hrsg.) (1999). Readings in Information Visualization. Using Vision to Think. San Francisco: Morgan Kaufman.

Chen, C. (1999): Information Visualisation and Virtual Environments. London u. a.: Springer.

Chen, Hsinchun and Houston, Andrea L. and Sewell, Robin R. and Schatz, Bruce R. (1998). “Internet Browsing and Searching: User Evaluation of Category Map and Concept Space Techniques.” Journal of the American Society for Information Science, Special Issue on AI Techniques for Emerging Information Systems Applications 49(7), 582-603.

Eibl, M. (2003). Visualisierung im Document Retrieval : theoretische und praktische Zusammenführung von Softwareergonomie und Graphik Design. 2. Aufl., Bonn: Informationszentrum Sozialwissenschaften.

Elzer, P.; Krohn, U. (1997). „Visualisierung zur Unterstützung der Suche in komplexen Datenbeständen.“ Proceedings of the HIM’97, Dortmund 1997, 27-38.

Google (2005). Google Search Engine Homepage. Mountain View/CA: Google Inc. http://www.google.de [letzter Zugriff: September 2005].

Hearst, Marti A. (1995). “TileBars: Visualization of Term Distribution Information in Full Text Information Access.” Proceedings of the CHI’95: Conference on Human Factors in Computing Systems: Mosaic of Creativity, Denver/CO, May 1995, 59-66.

Kang, I.-H.; Kim, G. (2003). “Query Type Classification for Web Document Retrieval.“ In: Proc. SIGIR ’03, July/August 2003, Torionto, 64-71.

Kartoo (2005). Kartoo Metasuchmaschine Homepage. http://www.kartoo.de [letzter Zugriff: September 2005].

Lohse, Gerald Lee (1991A). A Cognitive Model for Understanding Graphical Perception. Ph.D. Thesis, University of Michigan.

Mandl, T.; Eibl, M. (2001). „Evaluating Visualizations: A Method for Comparing 2D Maps." In: Smith, M.; Salvendy, G.; Harris, D.; Koubek, R. (eds.) (2001). Usability Design and Interface Evaluation: Cognitive Engineering, Intelligent Agents and Virtual Reality. Proc. HCI Intl 2001 (9th Intl Conference on Human-Computer Interaction), New Orleans, August 2001. Mahwah/NJ, London: Lawrence Erlbaum, Vol.1, 1145-1149.

Ogden, William; Davis, Mark; Rice, Sean (1998). Document thumbnail visualizations for rapid relevance judgments: When do they pay off? In: Proc. of the 7th Text Retrieval Conference Trec-7, 599-612.

Robertson, Stephen E.; Hancock-Beaulieu, M.M. (1992). „On the Evaluation of IR Systems." In: Information Processing & Management 28(4) (1992), 457-466.

Sparck Jones, Karen (ed.) (1981). Information Retrieval Experiment. London et al.: Butterworths.

WebBrain (2005). WebBrain 2.0 Search Engine Homepage. Santa Monica/CA: TheBrain Technologies Corporation. http://www.webbrain.com [Letzter Zugriff: September 2005].

Wild, F. (2005). „Visuelle Verfahren im Information Retrieval." In: Information in Wissenschaft und Praxis 56(1) (2005), 29-34.

Wolff, Christian (1996). Graphisches Faktenretrieval mit Liniendiagrammen: Gestaltung und Evaluierung eines experimentellen Rechercheverfahrens auf Grundlage kognitiver Theorien der Graphenwahrnehmung. Konstanz: UVK [Schriften zur Informationswissenschaft, Band 24].

Anwendungen des semantischen Wissens über Konzepte im Information Retrieval

Iryna Gurevych, Heidelberg

Abstract

Die im Folgenden beschriebenen experimentellen Arbeiten befassen sich mit der Integration des semantischen Wissens über Konzepte in das *Information Retrieval* (IR) . Verschiedene Arten des Wissens werden aus dem lexikalisch-semantischen Wortnetz GermaNet extrahiert. Einerseits wurde geprüft, ob die Performanz eines IR-Systems durch eine Erweiterung der Anfrage mit Wörtern aus automatisch generierten Definitionen der Konzepte verbessert wird. Andererseits wurde ein neuartiges *Information Retrieval* Modell getestet, dem lexikalisch-semantische Verwandtschaft der Wörter zugrunde liegt. Die Ergebnisse aller Experimente wurden auf einem IR-Datensatz in der Domäne „Elektronische Berufsberatung" mit einem herkömmlichen IR-System verglichen und automatisch ausgewertet. Dabei ergab sich: 1) eine Erweiterung der Anfragen durch Hyponyme ist besonders nützlich; 2) das semantische IR-Modell schneidet auf der gleichen Ebene ab, wie das tf*idf IR-Modell.

1. Einführung

Semantisches Wissen wird oft für die Verbesserung von *Information Retrieval* Systemen benötigt. Die Forschungshypothese ist, dass dieser Typ des Wissens die Qualität gegenwärtiger *Information Retrieval* Systeme erhöhen wird, denen statistische Methoden und Zeichenkettenvergleiche zugrunde liegen. Bisherige Forschungsarbeiten konnten das nicht überzeugend und konsistent beweisen (vgl. z.B. Evens, 2002).

Für die Einbindung semantischen Wissens in das *Information Retrieval* existieren verschiedene Möglichkeiten:

- **Anfrageerweiterung**. Hierbei werden die in der Anfrage enthaltenen Terme mit ihren semantisch verwandten Begriffen erweitert. Als Ergebnis können Dokumente gefunden werden, die den ursprünglichen Anfrageterm nicht enthalten, jedoch mit ihm semantisch verwandte Begriffe. Wichtig ist hierbei die Methode, mit deren Hilfe semantisch verwandte Begriffe für einen Anfrageterm bestimmt werden (Voorhees, 1994; Mandala et al., 1998).
- **Konzepte statt Wörter indexieren**. Hierbei wird der Index nicht auf der Wortebene erzeugt, sondern auf der Konzeptebene. Wörter werden in der Vorverarbeitung auf Begriffe abgebildet, so dass Anfragen dann gegen

einen Konzept-Index abgeglichen werden (Gonzalo et al., 1998; Voorhees, 1999).

- **Semantische Verwandtschaft als Modell des *Information Retrievals*.** Bei diesem Verfahren werden klassische *Information Retrieval* Modelle durch ein alternatives Modell ersetzt. Das alternative Modell berechnet die Relevanz eines Dokumentes in bezug auf eine Anfrage aufgrund der semantischen Verwandtschaft der darin enthaltenen Wörter (Smeaton, 1999).

In diesem Beitrag beschreiben wir Ergebnisse unserer Pilot-Studie, die sich semantischen Wissens über Begriffe in einem *Information Retrieval* System bedient. Das System greift auf das semantische Wissen in GermaNet zu, dem lexikalisch-semantischen Wortnetz für die deutsche Sprache (Kunze, 2004). In diesem Wortnetz sind Nomen, Verben und Adjektive als *is-a* Hierarchien repräsentiert. Zusätzliche Informationen auf lexikalischer (Wort-) Ebene, z.B. Synonymie, Antonymie, Derivation, und semantischer (Wortbedeutungs-) Ebene, z.B. Meronymie, Assoziation können sprachverarbeitenden Anwendungen zur Verfügung gestellt werden.

Zwei Serien von Experimenten wurden durchgeführt. In der Serie haben wir die Anfrage des Benutzers mit verwandten Begriffen erweitert. Für die Anfrageerweiterung wurden verschiedene Arten der aus GermaNet erzeugten künstlichen Definitionen der Wortbedeutungen (Konzepte) herangezogen. Automatisch generierte Definitionen wurden zum ersten Mal für die Berechnung semantischer Verwandtschaft von zwei Wörtern eingeführt (Gurevych, 2005b). Wichtige Parameter für die Erzeugung der Definitionen sind Typen der semantischen Beziehungen in GermaNet, z.B. Hyperonymie, Synonymie, Assoziation und der Verwandtschaftsgrad (Distanz in Kanten vom Konzept zu verwandten Konzepten). Für die Berechnung semantischer Verwandtschaft zwischen zwei Konzepten hat sich die Hyperonymie-Beziehung bis zum Verwandtschaftsgrad 3 als besonders hilfreich erwiesen. In der vorliegenden Arbeit galt es herauszufinden, welche Parameter für die Erzeugung künstlicher Definitionen der Konzepte im *Information Retrieval* Kontext besonders hilfreich sind.

Für die zweite Serie von Experimenten wurden verschiedene Maße der semantischen Verwandtschaft entwickelt und evaluiert. Semantische Verwandtschaft wurde als eine beliebige Art der semantischen oder assoziativen Beziehung zwischen zwei Begriffen definiert (Gurevych & Niederlich, 2005b). Ein Maß semantischer Verwandtschaft (Lin, 1998) wurde im semantischen *Information Retrieval* für die Bestimmung der relevanten Dokumente benutzt. Aus diesem Verfahren resultierende Ergebnisse wurden mit denen eines konventionellen IR Modells verglichen und anhand eines *Gold Standards* evaluiert.

Die in dieser Arbeit beschriebenen Methoden wurden mit dem Wortnetz GermaNet erprobt, können jedoch auf beliebige konzeptuelle Netzwerke und Taxonomien übertragen werden. Insbesondere sind für fachspezifische Gegenstandsbereiche, wie z.B. Bibliothekwissenschaften oder Biomedizin, solche Ressourcen oft bereits vorhanden und auf spezielle Domänen zugeschnitten. In diesem Fall kann auch eine

fachspezifische Ressource, die domänenspezifisches Wissen im großen Umfang und Detail repräsentiert, statt einer allgemeinsprachlichen Ressource eingesetzt werden.
Das in dieser Arbeit beschriebene *Information Retrieval* System wurde in der Domäne „Elektronische Berufsberatung“ erprobt und evaluiert. In dieser Anwendung werden Aufsätze mit Beschreibungen beruflicher Interessen von Personen (Interessenprofile) als Anfragen aufgefasst und gegen eine Dokumentenkollektion mit natürlichsprachlichen Beschreibungen der Ausbildungsberufe automatisch abgeglichen.[1] Ein Interessenprofil gegeben, liefert das System eine nach Relevanz geordnete Liste der Berufe zurück. Die Performanz des Systems wird mit Hilfe der TREC Evaluierungsmetriken gemessen.[2] Der *Gold Standard* für die Evaluierung wurde durch ein wissensbasiertes System simuliert (Gurevych, 2005a), dem die Datenbank „Interesse:Beruf“ der Bundesagentur für Arbeit zugrunde liegt.[3] Das Evaluierungsmaß berechnet gemittelte Präzision für alle relevanten Dokumente (gemittelt über Anfragen).
Das vorliegende Papier wird die folgende Struktur haben: In Kapitel 2 wird die Anwendungsdomäne „Elektronische Berufsberatung“ näher beschrieben. Insbesondere werden die Aufgabe und die in Experimenten eingesetzten Testdaten charakterisiert. Kapitel 3 stellt dann die experimentellen Arbeiten detailliert vor. Künstliche Definitionen der Konzepte und Maße semantischer Verwandtschaft werden erläutert. Anschließend beschreiben wir die Einbindung dieser Wissensarten in das *Information Retrieval*. Experimentelle Ergebnisse werden vorgestellt und diskutiert. Im letzten Kapitel 4 fassen wir den Stand der Arbeiten knapp zusammen und erarbeiten ausgehend von den Ergebnissen Richtlinien für zukünftige Forschungsarbeiten.

2. Elektronische Berufsberatung

In der vorliegenden Arbeit wurde elektronische Berufsberatung als Anwendungsdomäne für das semantische *Information Retrieval* gewählt. „Berufsberatung ... wird meistens von Schülern in Anspruch genommen um zu erfahren, welcher Beruf zu ihnen passt, welche Anforderungen und Kenntnisse gefordert werden. Aber auch Arbeitslose, die aus gesundheitlichen Gründen arbeitslos geworden sind oder im erlernten Beruf keine Perspektive mehr sehen, informieren sich über Umschulungsmöglichkeiten“.[4] (Wikipedia, 2005)

[1] Bundesagentur für Arbeit. BERUFENet. http://berufenet.arbeitsamt.de/, Nürnberg, Germany.
[2] TREC. http://trec.nist.gov/overview.html
[3] Bundesagentur für Arbeit. Interesse:Beruf. http://www.interesse-beruf.de/, Nürnberg, Germany.
[4] http://de.wikipedia.org/wiki/Berufsberatung

Personen, die auf der Suche nach einem zu ihnen passenden Beruf sind, können in der Regel von einem speziell ausgebildeten Experten beraten werden. Ein Berufsberater besitzt umfangreiches Domänenwissen über die Berufswelt. Diese Art der Berufsberatung ist jedoch teuer (Personalkosten), nicht immer zugänglich (Öffnungs- und Wartezeiten), und schlecht reproduzierbar (unterschiedliche Berater geben verschiedene Empfehlungen aus).

Auf der anderen Seite stehen den Ratsuchenden automatisierte Berufsberatungsmöglichkeiten wie z.B. das bereits erwähnte Programm „Interesse:Beruf“ zur Verfügung. Dabei wird der Benutzer gebeten, die auf ihn oder sie zutreffenden Schlagwörter aus drei Kategorien (Was? Wo? Womit?) auszuwählen. Das System greift dann auf eine Datenbank zu, in dem zu jedem Beruf relevante Schlagwörter von Fachexperten vergeben wurden, und berechnet zu jedem Beruf die Anzahl der Treffer. Eine nach Anzahl der Treffer geordnete Liste der Berufe wird dem Benutzer zurückgeliefert. Diese Art der Berufsberatung ist für den Benutzer jederzeit über ein Web-Interface zugänglich und die Ergebnisse sind, zumindest für die jeweilige Version der Datenbank, reproduzierbar. Nichtsdestotrotz ist die Freiheit des Benutzers in der Beschreibung seiner persönlichen Interessen durch eine fest vorgegebene Menge der Schlagwörter stark reduziert. Ein weiterer Nachteil dieser Methode ist der erhebliche manuelle Aufwand für die Pflege der Datenbank mit Zuordnungen einzelner Berufe und Schlagwörter. Oft werden neue Berufe geschaffen, und noch öfter verändert sich die Beschreibung eines bestehenden Berufes, so dass die Schlagwörter geändert werden müssen.

Im vorliegenden Papier schlagen wir einen alternativen, sprachbasierten Zugang zu der Berufsberatung vor. Der Zugang wird als sprachbasiertes Beratungssystem implementiert. Die Aufgabe der Berufsberatung wird als eine *Information Retrieval* Aufgabe definiert. Gegeben eine Anfrage (Interessenprofil), sollen Dokumente (Berufsbeschreibungen) nach ihrer Relevanz für die Anfrage geordnet werden. Ein derartiges System würde dem Benutzer ermöglichen, die Berufsberatung jederzeit in Anspruch zu nehmen und dabei frei in der sprachlichen Formulierung beruflicher Interessen und Vorstellungen zu bleiben. Für die Pflege der Datenbank fällt kein zusätzlicher Aufwand an. Benötigt werden lediglich ausführliche Beschreibungen der einzelnen Berufe, die in der Regel bereits vorliegen.

Für die Evaluierung eines *Information Retrieval* Systems wird ein Testdatensatz benötigt, der folgende Komponenten enthält: eine Kollektion von Dokumenten, eine Kollektion von Themen,[5] und eine Menge von Relevanzurteilen für jedes Thema in bezug auf alle Dokumente, die *Gold Standard* genannt wird.

[5] Der Begriff „Thema“ soll im Kontext des *Information Retrievals* vom Begriff „Anfrage“ unterschieden werden. Mit dem Thema wird eine natürlichsprachliche Beschreibung des Informationsbedürfnisses eines Benutzers gemeint. Hingegen wird unter „Anfrage“ eine Menge von Suchtermen verstanden, auf die das ursprüngliche Thema als Folge einer automatischen Vorverarbeitung abgebildet wurde.

Die Domäne der elektronischen Berufsberatung eignet sich wegen ihrer speziellen Eigenschaften besonders gut für unsere Experimente. Auf der einen Seite liegt eine Dokumentenkollektion mit den Beschreibungen von etwa 1.800 Ausbildungsberufen, z.B. Altenpfleger/in, Elektroniker/in, und weiteren 4.000 Berufen vor, z.B. Informatiker/in, Maschinenbauingenieur/in, vor. Diese Berufsbeschreibungen werden in der Datenbank BERUFENet verwaltet, die in der Einführung bereits erwähnt wurde. Informationen über einzelne Berufe werden im XML-Format repräsentiert. Das Datenmodell sieht ca. 60 verschiedene Datenfelder vor, wie z.B. Inhalte der Berufsausbildung, Aufgaben und Tätigkeiten im jeweiligen Beruf, Angaben zu erwarteten Kompetenzen eines Bewerbers. Auf der anderen Seite liegt die Datenbank „Interesse:Beruf“ vor. In (Gurevych, 2005a) wurde ein Verfahren vorgestellt, das einen *Gold Standard* für die Evaluierung der Ergebnisse des *Information Retrievals* aus dieser Datenbank automatisch generiert. Dafür sollen lediglich die Themen, d.h. Interessenprofile, mit zutreffenden Schlagwörtern aus der Datenbank annotiert werden. Es ist nicht mehr nötig, einzelne Dokumente im Hinblick auf ihre Relevanz zur Anfrage manuell zu beurteilen. Das wäre nicht nur aufwendig, sondern würde spezielle Fachkompetenz auf dem Gebiet der Berufsberatung erfordern.
Da „Interesse:Beruf“ für 578 Ausbildungsberufe in Deutschland entwickelt wurde, setzt sich unsere Testkollektion aus den Beschreibungen dieser Berufe zusammen. Um einige Beispiele für die Eingaben in das System zu sammeln, wurden Interessenprofile in einem Experiment mit 30 Versuchspersonen erhoben. Die Teilnehmer am Experiment haben kurze Texte über ihre beruflichen Interessen und Erwartungen geschrieben (Beispiel 1), die eine Eingabe in das IR-System darstellen:
„Ich würde gerne mit Tieren arbeiten, sie behandeln, für sie sorgen, aber ich kann kein Blut sehen und ich habe zu viel Mitleid mit kranken Tieren. Andererseits arbeite ich besonders gerne am Computer, kann programmieren in C, Python und VB und könnte mir daher auch in der Software-Entwicklung einen passenden Beruf vorstellen. Ich kann mir nur schlecht vorstellen in einem Kindergarten, als Sozialberater oder als Lehrer zu arbeiten, da ich mich nicht besonders gut durchsetzen kann.“

3. Semantisches Wissen im *Information Retrieval*

3.1 Baseline Information Retrieval System

Die Ergebnisse des semantischen *Information Retrievals* werden im Folgenden mit einem herkömmlichen Baseline-System verglichen. Das implementierte Baseline-System basiert auf dem erweiterten booleschen Modell (Salton et al., 1983). Im ersten Schritt werden die Dokumente in der Testkollektion unter der Verwendung einer allgemeinen deutschen Stoppwortliste und des *Stemming* indexiert. Im zweiten Schritt werden Anfragen gegen den Index abgeglichen und Dokumente in relevante und irrelevante eingeteilt. Die relevanten Dokumente werden anschließend nach der Gleichung 1 geordnet.

$$\sum_{t\_in\_d} tf(t\_in\_d) \times idf(t)$$

Gleichung 1: *tf(t_in_d)* ist der Termfrequenzfaktor des Termes (*t)* im Dokument *(d)*, und *idf(t)* ist die inverse Dokumentenfrequenz des Termes.

3.2 Experimente mit der Anfrageerweiterung

Eine Erweiterung der Anfrage im *Information Retrieval* wird dadurch motiviert, dass Eingaben der Benutzer auf einer Seite und Dokumente in der Kollektion auf der anderen Seite oft ein unterschiedliches Vokabular, d.h. Wortschatz, benutzen. In der Berufsberatung werden Interessenbeschreibungen eher informell formuliert. Berufsbeschreibungen weisen dagegen eine formale Ausdrucksweise auf. Für den gleichen Begriff werden verschiedene Wörter benutzt, z.B. kann der Benutzer das Wort „Brötchen“ verwenden, wobei „Backwaren“ in der Berufsbeschreibung auftritt.
Dokumente werden als Erstes unter Verwendung der Java-basierten IR-Bibliothek Lucene indexiert.[6] Interessenprofile werden vorverarbeitet (Tokenisierung, Wortartenerkennung), und entweder als eine Menge von Nomen (N) oder als eine Menge von Inhaltswörtern (Nomen, Verben, Adjektive, Adverbien - NVAA) repräsentiert. Daraus resultierende Anfragen an das IR-System werden unter Zugriff auf die Java-basierte GermaNet API mit den Begriffen aus automatisch erzeugten Definitionen expandiert. Dazu wurden verschiedene Typen der semantischen Relationen mit unterschiedlichem Verwandtschaftsgrad verwendet. Auf diesen Repräsentationen wurde das in Lucene implementierte IR-Verfahren (s. Abschnitt 3.1) angewandt und die Ergebnisse mit Hilfe der TREC Evaluierungssoftware ausgewertet (Tabelle 1).

[6] http://lucene.apache.org/java/docs/

Typ sem. Beziehung	Verwandt-schaftsgrad	N	Differenz	NVAA	Differenz
Antonymie	1	0,3153	-0,0034	0,3204	-0,0161
	3	0,3153	-0,0034	0,3204	-0,0161
	alle	0,3153	-0,0034	0,3204	-0,0161
Holonymie	1	0,3238	0,0051	0,3411	0,0046
	3	0,3194	0,0007	0,3378	0,0013
	alle	0,3194	0,0007	0,3378	0,0013
Hyperonymie	1	0,3250	0,0063	0,3280	-0,0085
	3	0,3070	-0,0117	0,3378	0,0013
	alle	0,3163	-0,0024	0,3224	-0,0141
Hyponymie	1	0,3480	0,0293	0,3530	0,0165
	3	0,3459	0,0272	0,3509	0,0144
	alle	0,3492	0,0305	0,3544	0,0179
Meronymie	1	0,3188	0,0001	0,3386	0,0021
	3	0,3173	-0,0014	0,3368	0,0003
	alle	0,3173	-0,0014	0,3368	0,0003
Synonymie	1	0,3100	-0,0087	0,3235	-0,0130
	3	0,3037	-0,0150	0,3175	-0,0190
	alle	0,3100	-0,0087	0,3208	-0,0157
Alle	alle	0,3471	0,0284	0,3586	0,0221
Baseline		0,3187		0,3365	

Tabelle 1

Bei den beiden Arten der Anfragerepräsentation (N oder NVAA) kommt es zu ähnlichen Ergebnissen. Die Ergebnisse für die letztere Systemkonfiguration sind allgemein etwas besser, als für die erste. Bei der Bestimmung der optimalen Parameter zur Erzeugung der künstlichen Definitionen wird deutlich, dass lediglich Hyponyme mit steigendem Verwandtschaftsgrad eine Verbesserung des *Information Retrieval* hervorrufen. Z.B. wird das Konzept „Computer“ durch folgende Begriffe erweitert: „Laptop Notebook Mainframe Minicomputer PC Workstation Macintosh Apple Client Gateway Server“. Der in Lucene implementierte *Stemming*-Mechanismus verursacht einige Fehler. So wird „Blut“ auf den Stamm „blu“ abgebildet, und dieser wird dann nicht nur mit dem ursprünglichen „Blut“, sondern auch mit der „Bluse“ assoziiert. Als Folge wird die Anfrage teilweise mit irrelevanten Begriffen erweitert, z.B. „Top Hemdbluse Seidenbluse Baumwollbluse Leinenbluse Röschenbluse Blutkonserve Konserve“.

Ein weiterer Nachteil der bestehenden Systemarchitektur ist, dass keine Wortlesartendisambiguierung vorgenommen wird. Das bedeutet, dass die Anfrage nicht mit der Definition der aktuellen Bedeutung eines Wortes erweitert wird, sondern mit den

Definitionen aller seinen Bedeutungen. Relevante Forschungsarbeiten haben bisher gezeigt, dass eine Disambiguierung der Wortlesarten keine bemerkbaren Verbesserungen des *Information Retrievals* geleistet hatte (vgl. Sanderson (1994). Der Grund ist insbesondere, dass die Wortlesartendisambiguierung perfekt sein muss, um positive Auswirkungen auf IR zu haben. Die aktuelle Situation auf diesem Gebiet ist jedoch, dass die Genauigkeit der Algorithmen etwa 60-70% beträgt. Diese können somit nicht mit Gewinn in das sprachbasierte Beratungssystem integriert werden.

3.2 Semantische Verwandtschaft der Wörter

In gegenwärtigen *Information Retrieval* Systemen wird die Relevanz eines Dokuments in Bezug auf eine Anfrage auf der Grundlage des booleschen, probabilistischen oder des Vektorraummodells bestimmt. Anfragen und Dokumente werden als Mengen von Index-Termen repräsentiert. Zwischen den einzelnen Wörtern existierende lexikalische und semantische Relationen werden nicht berücksichtigt. Dadurch werden relevante Dokumente, die ein anderes Vokabular benutzen als die Anfrage, nicht gefunden. Eine Alternative zu zeichenkettenbasierten Verfahren des *Information Retrieval*s ist eine Approximierung der Relevanz eines Dokuments in bezug auf eine Anfrage mit Hilfe von Maßen lexikalisch-semantischer Ähnlichkeit. Das bedeutet, dass die relevantesten Dokumente der Anfrage semantisch am ähnlichsten sind. Die Zwischenrepräsentation für das *Information Retrieval* wird verbessert, indem Wörter der natürlichen Sprache auf lexikalische Konzepte in GermaNet abgebildet werden. Die Bestimmung semantischer Ähnlichkeit bezieht das in GermaNet modellierte lexikalische Wissen, Domänen- und Weltwissen sowie die Ergebnisse einer umfassenden Korpusanalyse mit ein.

Existierende Ansätze zur Berechnung semantischer Verwandtschaft können in drei verschiedene Klassen eingeteilt werden: *wörterbuchbasierte* Maße (Lesk, 1986; Gurevych, 2005b), *distanzbasierte* Maße (Hirst & St-Onge 1998; Leacock & Chodorow, 1998), und *informationsgehaltsbasierte* Maße (Resnik, 1995; Lin, 1998).

Wörterbuchbasierte Maße beruhen auf der Annahme, dass in Wörterbuchdefinitionen semantisch verwandter Wörter viele Wortüberlappungen zu finden sind. Die in unseren Experimenten benutzte Ressource, GermaNet, enthält nur wenige Definitionen. Dagegen findet man in einem semantischen Netzwerk semantisch verwandte Konzepte. In Gurevych (2005b) wurde ein neues Verfahren vorgeschlagen, das diese Eigenschaft des Netzwerks ausnutzt und künstliche Definitionen eines Konzeptes nach vorgegebenen Parametern generieren kann. Die Wortüberlappung für die Bestimmung semantischer Verwandtschaft wird dann nicht auf echten, sondern auf automatisch generierten Definitionen berechnet. Dies ermöglicht die Bestimmung semantischer Verwandtschaft zwischen zwei Konzepten.

Distanzbasierte Maße suchen in der Regel nach dem kürzesten Pfad zwischen zwei Konzepten. Bis auf die Arbeit von Sussna (1993), wird bei der Berechnung des kürzesten Pfades lediglich die *is-a* Relation berücksichtigt. Das Maß semantischer Verwandtschaft ist im einfachsten Fall die Pfadlänge selbst oder eine Funktion

davon, vgl. z.B. Leacock & Chodorow (1998). Hirst & St-Onge (1998) gewichten die Pfade abhängig von ihrer Beschaffenheit, d.h. abhängig von verschiedenen Typen der im Pfad enthaltenen semantischen Relationen.
Die dritte Klasse der Methoden zur Bestimmung semantischer Verwandtschaft, informationsgehaltsbasierte Maße, benutzt die Struktur der GermaNet-Hierarchie und statistische Korpusauswertungen. Statistische Auswertungen dienen der Bestimmung des Informationsgehalts eines Konzeptes. Resnik (1995) definiert semantische Ähnlichkeit zwischen zwei Wörtern w_1 und w_2 als den maximalen Informationsgehalt ihres nächsten gemeinsamen Oberknoten *c* (Gleichung 2). c_1 und c_2 sind Konzepte (Wortbedeutungen), die w_1 und w_2 entsprechen. $S(c_1,c_2)$ ist eine Menge der Konzepte, die sowohl c_1 als auch c_2 subsumieren. $-\log p\,(c)$ ist der Informationsgehalt eines Konzeptes. Die Wahrscheinlichkeit *p* wird als die relative Frequenz der Wörter in einem Korpus berechnet. Lin (1998) definiert semantische Ähnlichkeit mit Hilfe des Informationsgehalts und eines Modells aus der Informationstheorie. Sein Maß (Gleichung 3) wird manchmal als universales Maß semantischer Ähnlichkeit genannt, da es anwendungs-, domänen-, und ressourcenunabhägig ist.

$$sim(c1,c2) = \max_{c \in S(c1,c2)} \left[-\log p(c) \right]$$

Gleichung 2

$$sim(c1,c2) = \frac{2 \times \log p(lcs(c1,c2))}{\log p(c1) + \log p(c2)}$$

Gleichung 3

Gurevych & Niederlich (2005b) führten eine komparativ ausgerichtete Evaluierung der Maße semantischer Verwandtschaft mit einem deutschen Datensatz (65 aus Substantiven bestehende Wortpaare) durch. Die Evaluierung ergab, dass wörterbuchbasierte und informationsgehaltsbasierte Maße semantischer Verwandtschaft etwa gleich gut abschneiden. Wörterbuchbasierte Maße ermöglichen Vergleiche zwischen Wörtern unterschiedlicher Wortarten. Informationsgehaltsbasierte Maße erreichen zwar eine etwas bessere Performanz, der Vergleich ist bei diesen Maßen dafür immer innerhalb nur einer bestimmten Wortart möglich, da einzelne Wortarten in GermaNet in verschiedenen Taxonomien modelliert sind. Ebenso stellte sich als problematisch heraus, von unterschiedlichen Methoden berechnete Werte auf einen einheitlichen numerischen Bereich abzubilden, da die Verfahren unterschiedlicher Natur sind.
Da die Methode von Lin (1998) in der Evaluierung mit den deutschen Wortpaaren gut abgeschnitten hatte, wurde dieses Verfahren in das IR-Modell eingebunden. Die vom Verfahren berechneten Werte liegen auf der Skala von 0 bis 1. Die Repräsentation der Texte wurde auf eine Menge von Nomen begrenzt. Zusätzlich wird in der Vorverarbeitung eine domänenspezifische Stoppwortliste angewandt. Die Stoppwortliste bezieht sich auf die Berufsberatungsdomäne und enthält 137 Einträge, z.B. „Abschlussbezeichnung“, „Ausbildungszeit“. Dokumente und Anfragen werden als eine Menge der dort vorhandenen GermaNet-Konzepte dargestellt: *K_d* oder

K_q={k₁, ..., kₙ}. Für jedes Paar *K_d* und *K_q* wird eine zweidimensionale Matrix *(#K_d x #K_q)* erstellt, in der # für die Anzahl der Elemente in der Menge steht. Dann wird für jedes Konzeptpaar ein Wert semantischer Verwandtschaft berechnet. Die Relevanz eines Dokuments *Rl(d,q)* zur Anfrage wurde nach zwei Gleichungen (4 und 5) ermittelt.

$$Rl(d,q) = \frac{\sum_{i=1}^{\#K\_d} \sum_{j=1}^{\#K\_q} rel(i,j)}{\#K\_d \times \#K\_q}$$

Gleichung 4

$$Rl(d,q) = \frac{\sum_{i=1}^{\#K\_d} \sum_{j=1}^{\#K\_q} rel(i,j)}{\#K\_q}$$

Gleichung 5

rel (i,j) steht für einen einzelnen Wert der semantischen Verwandtschaft zwischen zwei Konzepten. Da eine große Anzahl der Wortpaare einen geringen Wert der semantischen Verwandtschaft hat und im *Information Retrieval* Kontext nur eng verwandte Wortpaare von Bedeutung sind, wurde ein Schwellwert implementiert. Wortpaare mit einem Verwandtschaftswert unter diesem Schwellwert wurden bei der Berechnung des Gesamtwertes *Rl(d,q)* nicht berücksichtigt. Ergebnisse dieser Experimente sind in bezug auf verschiedene Schwellwerte und zwei Gleichungen in der Tabelle 2 dargestellt.

Schwellwert	Gleichung 3	Differenz	Gleichung 4	Differenz
kein	0,224	-0,0947	0,2813	-0,0374
0,3	0,2555	-0,0632	0,3035	-0,0152
0,5	0,2543	-0,0644	0,3114	-0,0073
0,7	0,2614	-0,0573	0,3129	-0,0058
0,8	0,281	-0,0377	0,3292	0,0105
0,9	0,2787	-0,04	0,3162	-0,0025
Baseline	0,3187			

Tabelle 2

Die Analyse der Ergebnisse macht deutlich, dass der auf semantischer Verwandtschaft basierende IR-Ansatz ungefähr gleich gut abschneidet, wie das konventionelle IR-System. Nur in einer Systemkonfiguration (Gleichung 5, Schwellwert 0,8) sind die Ergebnisse leicht besser als die Baseline. Es ist zu erkennen, dass der Schwellwert eine wichtige Rolle für die Endergebnisse spielt. So verbessern sich die Ergebnisse des semantischen IR-Systems mit steigendem Schwellwert. Das deutet darauf hin, dass die Berücksichtigung nur sehr eng verwandter Wörter, z.B. Synonyme oder direkte Hyponyme eine positive Auswirkung auf das Information Retrieval hat.
Eine detaillierte Auswertung der log-Dateien offenbarte eine Reihe der Fehlerquellen im System. Das semantische IR berücksichtigt nur Wörter, für die ein Verwandtschaftswert berechnet werden konnte. Einige Wörter, insbesondere viele zusam

mengesetzte Begriffe (Komposita) konnten nicht auf das GermaNet abgebildet werden, da Komposita eine offene Wortklasse im Deutschen sind und in einer statischen Wissensquelle wie GermaNet nicht im großen Umfang repräsentiert werden können. Folglich wurden Komposita bei der Berechnung semantischer Verwandtschaft nicht berücksichtigt.
Ähnlich wie in Experimenten mit Anfrageerweiterung führte das *Stemming* zu Fehlern in der Abbildung auf GermaNet. Daher entstehen manchmal sinnlose Verbindungen, z.B. bei „Blut – Schulzeit“ ergibt sich der Wert 0,85, da „Blut“ fehlerhaft auf „Blüte“ abgebildet wird und beide durch den Oberbegriff „Lebensphase“ miteinander verknüpft werden. Um derartigen Fehlern vorzubeugen, sollte das *Stemming* während der Abbildung auf GermaNet in der Anwendung vorzugsweise durch morphologische Analyse ersetzt werden. Des Weiteren scheint eine Repräsentation des Dokuments durch eine Menge der Nomen unzureichend (zu oberflächlich) zu sein. Wenn im Interessenprofil das Wort „Lehrer“ vorkommt und eine Berufsbeschreibung die Schulfächer des jeweiligen Ausbildungsberufes auflistet, werden die Schulfächer und „Lehrer“ fehlerhaft in einen engen Zusammenhang gebracht („Lehrer – Mathematik“ 0,756, „Lehrer – Biologie“ 0,739, „Lehrer – Seminare“ 0,70). Um diesen Effekt zu vermeiden, wäre es notwendig, semantische Struktur der Berufsbeschreibungen zu erkennen und Interessenprofile nur mit bestimmten Abschnitten, z.B. Aufgabenbeschreibungen des Berufes, zu vergleichen.
Eine weitere grundsätzliche Schwierigkeit (für das semantische IR und das Baseline-Verfahren) ist, dass die Anfragen nicht als präzise definierte Schlagwörter vorliegen, sondern als natürlichsprachlich formulierte Interessenprofile. Diese werden als *bag-of-words* aufgefasst und auf eine Menge der Nomen abgebildet. Verben, z.B. „Ich backe gerne“, oder Adverbien, z.B. „Ich arbeite gerne handwerklich“ werden nicht berücksichtigt. Negationen und verschiedene andere linguistische Mittel um negative Einstellungen einer Person wiederzugeben können zur Zeit nicht behandelt werden, was oft dazu führt, das unerwünschte Berufe zurückgeliefert werden. Ebenso wäre es wünschenswert, einzelne in der Anfrage vorkommende Begriffe nach ihrer Wichtigkeit für den Text zu gewichten. Dazu könnten fortgeschrittene Vorverarbeitungsverfahren, z.B. lexikalische Ketten, eingesetzt werden.

4. Diskussion

In diesem Beitrag wurden Möglichkeiten beschrieben, das semantische Wissen über Konzepte im *Information Retrieval* zu nutzen. Das zugrunde liegende lexikalisch-semantische Wissen wurde aus dem deutschen Wortnetz GermaNet bezogen. Experimentelle Arbeiten wurden an einem Testdatensatz in der Domäne „Elektronische Berufsberatung“ durchgeführt. Dabei wurde die Berufsberatung als eine *Information Retrieval* Aufgabe definiert.

Die Ergebnisse der semantisch angereicherten IR-Systeme wurden gegen einen automatisch erzeugten *Gold Standard* abgeglichen. Zum Vergleich wurde ein konventionelles IR-System herangezogen, das sich Zeichenkettenvergleiche und statistischer Wahrscheinlichkeiten bedient. Experimente mit automatisch erzeugten Definitionen der Konzepte haben ergeben, dass die Hyponymie-Relation sich als nützlich im Kontext des *Information Retrieval* erweist. Alle anderen semantischen Relationen haben sich dagegen als eher schädlich erwiesen. Es gilt zu überprüfen, ob das Hinzufügen einer Komponente zur Wortlesartendisambiguierung die Ergebnisse positiv beeinflussen kann.

Die semantische Verwandtschaft der Wörter wurde als Modell für die Bestimmung der Relevanz eines Dokuments eingesetzt. Dieses Modell hat eine ähnliche Performanz ermöglicht wie herkömmliche Ansätze. Um das System zu verbessern, müssen auf der einen Seite die Maße semantischer Verwandtschaft verbessert werden (z.B. im Hinblick auf die Behandlung mehrerer Wortarten). Auf der anderen Seite müssen zusätzliche Komponenten in der Vorverarbeitung eingesetzt werden (z.B. für die Kompositaanalyse, Wortartenerkennung). Die Aufbereitung der natürlichsprachlichen Interessenprofile als eine IR-Anfrage sollte verbessert werden, indem relevante Inhaltswörter sowie positive/negative Präferenzen eines Benutzers bestimmt werden.

Wichtig ist für zukünftige Arbeiten, die am Fallbeispiel „Elektronische Berufsberatung" erzielten experimentellen Ergebnisse auf einem größeren und neutralen Datensatz zu verifizieren. Für die deutsche Sprache ist mit der sozialwissenschaftlichen Datenbank GIRT eine solche Möglichkeit gegeben (Kluck, 2004).[7] Die Stärken und Schwächen des semantischen *Information Retrievals* können dann in einem Vergleich mit anderen aktuellen Systemen besser verstanden werden.

Danksagung

Wir danken der Klaus Tschira Stiftung für die finanzielle Unterstützung und der Bundesagentur für Arbeit für die zur Verfügung gestellten Daten. Hendrik Niederlich möchte ich besonders für seine wertvollen Beiträge zu dieser Studie danken, die er als Praktikant und Diplomand am EML Research geleistet hat.

Literatur

Evens, Martha: 2002. Thesaural relationships in information retrieval. In Rebecca Green, Carol Bean, and Sung Hyon Myaeng, editors, *The Semantics of Relationships. An Interdisciplinary Perspective*, Chapter 9. Kluwer Academic Publishers, Dordrecht.

[7] http://www.gesis.org/Forschung/Informationstechnologie/GIRT4.htm

Gonzalo, Julio; Verdejo, Felisa; Chugur, Irina and Juan Cigarran: 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING-ACL '98 Workshop Usage of WordNet in Natural Language Processing Systems,* Montreal, Canada, August.

Gurevych, Iryna and Hendrik Niederlich: 2005a. Computing Semantic Relatedness in German with Revised Information Content Metrics. In *Proceedings of "OntoLex 2005 – Ontologies and Lexical Resources" IJCNLP'05 Workshop*, Jeju Island, Republic of Korea, October 15, 2005. *To appear.*

Gurevych, Iryna and Hendrik Niederlich: 2005b. Computing semantic relatedness of GermaNet concepts. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Applications of GermaNet II*, pages 462–474. Peter Lang.

Gurevych, Iryna: 2005a. Automatically generating a task-based information retrieval test collection. Technical Report, EML Research, Heidelberg, 2005.

Gurevych, Iryna: 2005b. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005),* Jeju Island, Republic of Korea, October 11–13, 2005. *to appear.*

Hirst, Graeme and David St-Onge: 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database and some of its applications*, pages 305–332. Cambridge, MA: The MIT Press.

Kluck, Michael: 2004. Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation. Bernard Bekavac, Josef Herget, Marc Rittberger, editors, Informationen zwischen Kultur und Marktwirtschaft. *Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004)*, Chur, 6.-8.Oktober 2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 247 – 268.

Kunze, Claudia: 2004. Lexikalisch-semantische Wortnetze. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, and H. Langer, editors, *Computerlinguistik und Sprachtechnologie. Eine Einführung*, pages 423–431. Heidelberg, Germany: Spektrum Akademischer Verlag, 2nd edition.

Leacock, Claudia and Martin Chodorow: 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. Cambridge: MIT Press.

Lesk, Michael: 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation,* Toronto, Ontario, Canada, June, pages 24–26.

Lin, Dekang: 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* San Francisco, Cal., pages 296–304.

Mandala, Rila; Tokunaga, Takenobu; Tanaka, Hozumi; Okumura, Akitoshi und Kenji Satoh: 1998. Ad hoc retrieval experiments using WordNet and automatically constructed thesauri. In *Text REtrieval Conference*, pages 414–419.

Resnik, Phil: 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, Canada, 20–25 August 1995, volume 1, pages 448–453.

Salton, G., Fox, E. and H. Wu: 1983. Extended boolean information retrieval. Communications of the ACM, 26(11): 1022-1036.

Sanderson, Mark: 1994. Word sense disambiguation and information retrieval. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 142 - 151, Springer-Verlag New York, Inc., New York, NY, USA.

Smeaton, Alan F.: 1999. Using NLP or NLP resources for information retrieval tasks. In Tomek Strzalkowski, editor, *Natural language information retrieval*, pages 99–111. Kluwer Academic Publishers, Dordrecht, NL.

Sussna, Michael: 1993. Word sense disambiguation for free text indexing using a massive semantic network. In Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM'93), Arlington, Virginia.

Voorhees, Ellen M.: 1994. Query expansion using lexical-semantic relations. In SIGIR'94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Voorhees, Ellen M.: 1999. Natural language processing and information retrieval. In Information Extraction: Towards Scalable, Adaptable Systems, pages 32–48, London, UK. Springer-Verlag.

Sprechen Sie "europäisch"?

Wissensbasierte Search-Engine "dandelon.com" revolutioniert wissenschaftliche Bibliotheken

Manfred Hauer, Neustadt

Abstract

Suchen Sie doch mal nach „Bluthochdruck Ernährung" in „dandelon.com“ - und anschließend in einem Bibliothekskatalog in ihrer Nähe! Oder nach „Gärten Häuser"! Oder einem anderen Thema. Schlagen Sie diese Seiten kurz zu, Browser ein und ausprobieren. „dandelon.com" versteht ihre Frage, ergänzt sie automatisch um Synonyme und übersetzt in bis zu 16 Sprachen, sortiert die Ergebnisse nach Ranking - mit einem Klick auch nach Jahr oder Autor. Wenn kein Ergebnis im ersten Ansatz findbar ist, optimiert sich die Frage von alleine, dehnt den Suchraum sprachlich noch weiter aus.

Was ist da anders als bei Google und ziemlich anders als bei einem Bibliothekskatalog?

Zum besseren Verständnis: Fast alle großen Bibliotheksmanagementsysteme basieren heute auf relationalen Datenbank-Management-Systemen und haben eine Nichttrefferraterate von ca. 40 % bei thematischen Recherchen - und 90 % aller Recherchen suchen nach Themen, nicht nach harten Fakten (ISBN, Jahreszahl, Verlag etc.). Die heutigen Studenten waren 10 Jahre alt, als das Internet 1995 in die Breite ging. Sie können sich an die Zeiten davor nicht mehr erinnern. Eine Zeit, in der Bibliotheken noch das „Weltwissen" durchaus gut verzeichneten. Als diese Generation 15 war, kam Google in Mode und wurde zum Synonym für Suchen. Das neue einfache, unkomplizierte Suchen passt nicht mehr zu den strengen bibliothekarischen Erfassungs- und Erschließungsmethoden und meist etwas spröden Katalogen.

Schauen wir uns Unternehmen an, so ist das mit den relationalen Systemen dort nicht viel anders und die Dateien in den Dateisystemen sind ein Platten fressender Dschungel. Information Retrieval und Normalisierung sind verschiedene Sichten auf die Welten – keine kann die andere bislang ersetzen.

Mit „dandelon.com“ gehen einige Bibliotheken seit über zwei Jahren einen neuen, gemeinsamen Weg. Collaboration, Knowledge Sharing, Content Management und ähnliche Schlagworte spielen hinein. Es geht um die Integration der beiden Technologien. In einem wissenschaftlichen Buch ist das Inhaltsverzeichnis in der

Regel eine sehr hochwertige Information über den Inhalt. Kurz, 3-4 Seiten lang, alle wesentlichen Fachbegriffe kommen hier vor. Im Dialog mit Bibliotheken entstand das Programm intelligentCAPTURE, das in einem hoch performanten Workflow das Scanning von Inhaltsverzeichnissen, dessen OCR und dessen linguistisch-statistische Auswertung abarbeitet. Ein suchbares PDF und eine sprachlich optimierte und gewichtete Menge von Deskriptoren steht danach zur Verfügung, um sowohl die relationalen Bibliothekskataloge anzureichern und dort die Suchbarkeit stark zu erhöhen als auch als Tauschware in „dandelon.com“ anderen Bibliotheken oder direkt Wissenssuchenden bereitzustehen. Diese ganze Kommunikation zwischen unterschiedlichen Systemen und Welten nutzt Agenten, die mit Z 39.50, SOAP, anderen Webservices und Formaten wie MAB, MARC und XML sich untereinander unterhalten. Was Amazon & Co zur Verfügung steht, Daten direkt von Verlagen, wird auch hier im internationalen ONIX-Format ergänzt. Content von bisher über 70 Verlagen, denn ohne deren OK dürften die farbigen Titelseiten gar nicht genutzt werden. Das Urheberrecht müssen diese öffentlich-rechtlichen Bibliotheken wahren, die Freiräume größer Player haben sie nicht.

Gerade auch bei den theoretisch über 8.000 elektronisch spiderbaren Zeitschriften mit ca. 1 Mio. Aufsätzen pro Halbjahr spielt die Juristerei intensiv mit, denn hier geht es nicht mehr um die Anzeige von Inhaltsverzeichnissen, sondern direkt um die Artikel selbst. Faktisch sind in „dandelon.com" erst wenige, sprich gut 100.000 Titel verarbeitet worden. Der Workflow dazu hat nur noch einen „Schlitz", in den elektronische Inhaltsverzeichnisse mittels der internationalen Zeitschriftenagentur Swets eingeworfen werden. Die Artikel in fast jedem Format zu öffnen, zu lesen und zu beschreiben ist ein gänzlich automatisierter Prozess. Sie dann bei der Suche anzuzeigen ist aber genauso wie bei Bibliothekskatalogen oder auch Google Scholar und Print eine Frage von Authentifizierung, Verträgen und somit von Geld. Micro Billing hat hier noch wenig Einzug gehalten - oder ist keineswegs „micro".

Die technische Lösung ist modularisiert, eine mächtige „Capturing"-Anwendung, intelligentCAPTURE, die Daten heranholt, abgleicht, konvertiert, extrahiert, maschinell „versteht". Ein Modul zum Management semantischer Netzstrukturen zur Repräsentation sprachlichen Wissens, IC INDEX. Und ein Information Retrieval-Modul, intelligentSEARCH. Alle Module basieren auf Lotus Notes Client oder auf dem Lotus Domino Server von IBM und darin integrierten Programmen. Eine Weiterentwicklung in Richtung des gerade erst von IBM vorgestellten UIMA (Unstructured Information Management Architecture) und IBM Omnifind steht für die nächsten Jahre an.

Die Württembergische Landesbibliothek, obwohl an dem Projekt gar nicht beteiligt, bezeichnete es als einen Quantensprung im Bibliothekswesen. Und empirische Messungen mit 75 Studenten an 295 Medien bestätigen diese Einschätzung. Natürlich springen hier nicht Quanten, dennoch Information Retrieval hat Parallelen, auch hier ist der Zustand eines Systems nicht für jeden Betrachter gleich. Bedeutung und Verstehen spielen eine wichtige Rolle – und dies ist auch für eine einzige Person

nicht konstant. Das System steht mit dem Betrachter über Information im Austausch. Es wechselt laufend seine Zustände. Zu philosophisch?
Jedenfalls weisen die Messungen in eine klare Richtung, Bibliothekskataloge werden mit Daten aus intelligentCAPTURE deutlich besser und durchschlagende Verbesserungen bringt intelligentSEARCH, die Technik hinter dem Wissenschaftsportal „dandelon.com“ und bisher drei fachspezifischen, kleineren Portalen. Zwei davon noch im Test.
Gleiches könnte ebenso auf Forschungsdokumentationen in Unternehmen, Patentsammlungen oder andere Dokumentationen angewendet werden. eMails zu erschließen und allgemein Festplatten mit einem Volltextindex zu erschließen, ist derzeit aber nicht das Ziel dieser Entwicklungen, reicht doch hierfür schon die Search-Engine von Domino oder ein Google Desktop.
Insbesondere „dandelon.com“ bezieht seine Stärke aus dem kollaborativen Ansatz, in dem alle angeschlossenen Partner zusammenarbeiten, immer mehr Bibliotheken, bisher ein Bibliotheksservice-Zentrum – zuständig für Bibliotheken in acht deutschen Bundesländern, ein internationaler Buchhändler, Verlage, Entwickler von semantischen Netzstrukturen, Computerlinguisten und das eigentliche Integrations-Software-Team. Ohne Replikation und Realtime Collaboration, sprich IBM Instant Messaging wäre das gemeinsame Projekt wohl nicht möglich.

Best Practices in Digital Libraries: possible avenues of Indo-German Collaboration

Areti Ramachandra Durga Prasad, Bangalore (Indien)

Abstract

This paper is a result of a study conducted to examine areas of collaboration for Indo-German Digital Library (DL) Initiatives. Paper outlines the important issues in digital libraries and describes the projects that have definite implications for Indian initiatives. Digital preservation is the aim of digital repository projects that helps preserve valuable documents and information. Indian and German projects are involved in digitization, archiving and preservation. Standards for metadata and interoperability adapted for DLs enable resource discovery and they should adhere to world standards. Several issues pertaining to sustainability and success of preservation projects including open standards and formats are discussed. Areas of research and development such as personalization and semantic web that enrich digital libraries with customized patron services are mentioned.

Introduction

The area of digital libraries grew in parallel with the web technologies. An examination of the DL trends and projects reveal that the nature of the projects by content and services has followed the practices of the electronic services of the decade before. The patterns follow the cultural trends of the electronic services era in each country. In India there is a keen interest and awareness in the area of digital libraries with International and national conferences and workshops being conducted. Some projects deal with digitization, few handle online theses and dissertations and yet others with archiving facilities to journal subscriptions. Prestigious projects such as the National Manuscripts Mission under the Indira Gandhi National Centre for Arts (IGNCA) and different projects under the Million Books Project mainly focus on digitization aspects from hard copies to digital. However, the major concern is regarding standards and adapting the available technologies for digital repositories.

Best Practices for Indian Digital Libraries

The field of Digital Libraries has many facets, right from digitization to planned information services online. Broadly, the issues involved can be dealt under the following areas:

Digitization:
Digitization entails digitizing the resources that are in print or other physical media and making them available as digital library collections. Digitization of print is especially undertaken to preserve rare documents that are out of print and of collections such as manuscripts and paintings that are rare and valued resources. However, there are several issues to be solved before building Digital libraries of digitized material.

Issues in Digitization:
Some issues that need to be answered before taking up digitization are: [Wisser, 2005]

1. **Selection of material:** which documents should be digitized?
2. **Choice of technology:** which medium and what kind of equipment should be used?
3. **Presentation of Information:** how should the information be presented and how will it be stored?
4. **Access Mechanism:** what will be the mechanism of access be like to make a document available?
5. **Copyright and DRM**: Not the least of the issues is the issue of copyright to put the content into digital repositories and managing the digital rights to the content once it is made available. Copyright seems to be the major obstacle to building digital collections. However, knowledge of open access material and availability of resources in open access systems may be the way to address copyright issues. As more and more members of the academic community are becoming aware of and encouraging open access material it will be accepted and perhaps become a norm of academic and research publishing.

Digitization involves digitizing text as well as other formats such as images and tables. Further for the scanned data to be computer processible, the textual data has to be recognized using Optical Character Recognition (OCR) tools. While for English languages the OCR tools are fairly mature and efficient for other languages especially the Asian languages they are far from satisfactory. The area of digitization and making digital repositories encompasses the following:

- tools and techniques for text, image capture
- text conversion,
- bibliographic description (metadata)
  document management and
- provision of online access.

The Center for Retrospective Digitization of library materials in Göttingen (GDZ) (http://gdz.sub.uni-goettingen.de/en/index.html), Göttingen State and University Library (SUB have evaluated techniques that relate to the above points. The retro-

conversion of print to digital documents is based on the guidelines and recommendations of the technical working group as outlined in its final report.
In India digital library projects deal with retro-conversion of data from print to digital and some projects deal with digitally born documents. The National Manuscripts Mission (NMM) of India has undertaken a massive projects of identification of old and rare manuscripts in various languages and scripts across and digitizing the same. So far about 5 million such resources have been identified. In this regard it is ideal to identify ancient and medieval literature of Indian origin available across European and in particular German. The other projects are working as the partners of the Million Books projects of the Carnegie Mellon University and are involved in digitization of books. These projects are also digitizing some books in Indian languages.

E-publishing:
E-publishing is one of the popular applications of digital libraries though in India it is yet to take off in formal projects. However, open access publishing is gaining importance with scientific publications in open repositories. The German projects, Exile Press (http://deposit.ddb.de/online/exil/exil.htm) [2] [3] and Digital Peer Publishing (DiPP) (http://www.dipp.nrw.de/) are interesting German projects in the area of e-publishing.

1. DRTC's LDL *(https://drtc.isibang.ac.in)* is a production system of open access articles, theses and dissertations and presentations in the domain of library and information science. The model allows for online publishing with a review process incorporated. Presently it enlists members from 14 countries and publications from Indian, European and American professionals. In most of the cases DRTC is associated with digital repositories of many institutions and organization of India such as
2. Indian National Science Academy (INSA) (http://drtc.isibang.ac.in/insa)
3. National Chemical Laboratory (NCL) (http://dspace.ncl.res.in/dspace/index.jsp)
4. ETD@IISc (http://etd.ncsi.iisc.ernet.in/)
5. IIT, Delhi (http://eprint.iitd.ac.in/dspace/)
6. INFLIBNET (http://dspace.inflibnet.ac.in)
7. ISI, Bangalore (http://www.library.isibang.ac.in:8080/dspace)
8. University of Hyderabad (http://202.41.85.207:8080/dspace)
9. LDL: Librarians' Digital Library (https://drtc.isibang.ac.in)
10. NIT, Rourkela (http://dspace.nitrkl.ac.in/dspace/)

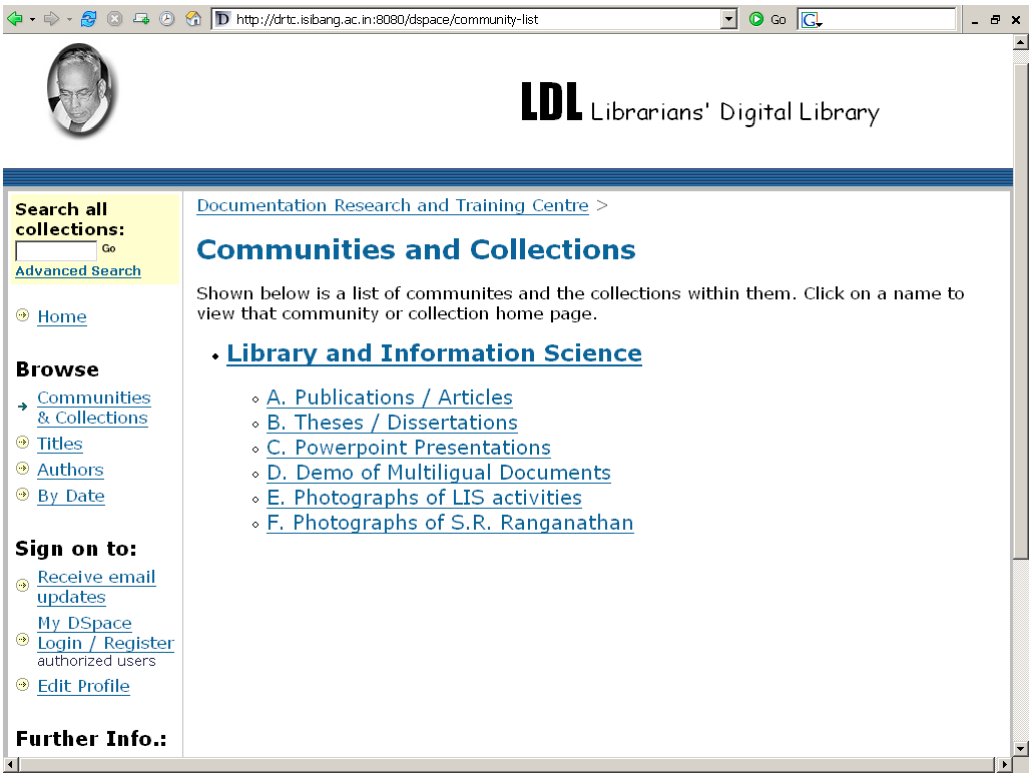


Figure 1: DRTC's Librarians' Digital Library (LDL)

Metadata Harvesting and Interoperability

Metadata standards ensure exposing information within digital repositories to the web patrons through search engines. Presently, though there are few digital repositories in India and the initiative at DRTC/ISI offers its SDL *(http://drtc.isibang.ac.in/sdl)* as OAI-compliant search facilities across DLs. One of the major problems with such harvesters is that they can harvest repositories, which are OAI standard compliant only. However, it is generally accepted that many open access journals and databases are not OAI compliant as OAI itself is a recent concept and a standard. For example, a preliminary study conducted by DRTC in identifying open access journals in the field of library and information science revealed that about 90 online journal offer free access and non of them are OAI-compliant. Some journal articles have been harvested through some manipulation of the metadata supplied with a few of DLIB magazine's articles.

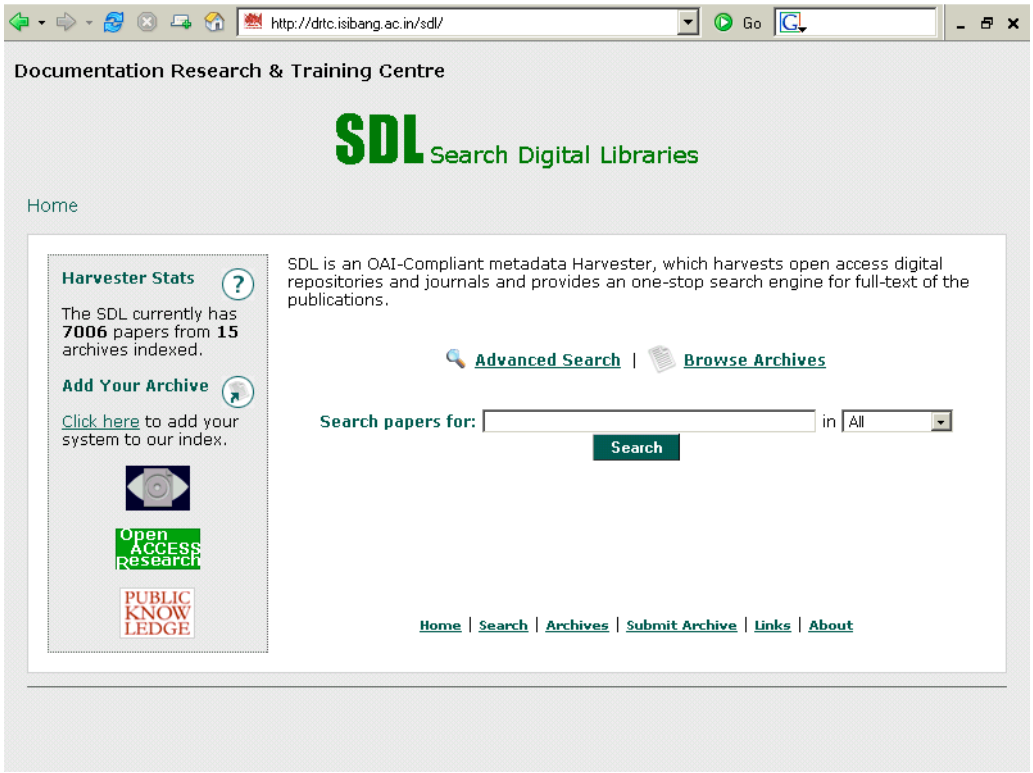


Figure 2: DRTC's Search *Digital Library* (SDL) Service

Long Term Preservation:

The purpose of preservation is to ensure protection of information of enduring value for access by present and future generations. For decades libraries and archives has played a central role in the preservation of documents [Conway, 1994]. The information storage media have changed over a period of time with the advancement of technology. Preservation work encompasses the preservation of traditional as well as digital documents or objects. Both the types of documents demand distinct preservation measures. By preservation, original artifacts and other traditional documents are protected from further deterioration and perhaps improvement of their physical condition to continue their use. Digital preservation involves *'the managed activities necessary for ensuring both the long-term maintenance of a byte stream and continued accessibility of its contents'* [5].

Preservation Issues:

Digital preservation poses some challenge before digital libraries. The rapidly changing technology impacts preservation and results in the need for constant up-gradations in equipment and software migrations. A rather intimidating problem is

need for the personnel to get acquainted with handling new technology and keeping track of newer trends as well as being able to adapt to those. The unprecedented expectation of the professionals has given rise to apprehensions and question sustainability of preservation projects. Various preservation issues identified by the Digital Preservation Coalition (www.dpconline.org) are as below [6]:

Technological Issues

- Digital Media
- Changes in Technology
- Authencity and Context
- Scale
- Strategies

Organizational Issues

- Costs
- Expertise
- Organizational structure
- Roles
- Selection

Legal Issues

- Intellectual property rights and preservation
- Legal deposits of electronic publications
- Other statutory requirements
- Access and security
- Business models and licensing
- Stakeholders, contract and grant conditions, and moral rights
- Privacy and confidentiality
- Investment in deposited materials by the repository

Open standards and formats:

Presently long term preservation of digital documents is a highly speculative area. Many digital library initiatives have been launched with the objective of preserving information for posterity. Paradoxically, the question remains how long the digital documents remain. The issues involved in preservation are:

- How long the various digital media last
- How long the standards used in digital libraries last
- How long the various software and hardware currently used remain in the future

Preservation has many challenging issues pertaining to formats, tools and hardware factors. The only solution to the perpetual of preservation is to follow open source software; more importantly open source software that are based on and use open standards and protocols so that there may be some way of migration when the existing tools and formats become out dated. KOPAL (Co-operative development of

long-term digital information archive) (http://www.kopal.langzeitarchivierung.de/) has identified the nuances of this issue. As the project proposes to use open standards like TIFF, Tex, it can be hoped the information can be extracted and can have migration path to any future standards, even if the software which can read TIFF or Tex files become outdated in the future. In addition, as KOPAL integrates from the start several different partners at different locations, long-term preservation will be ensured to a great extent. Nestor is another initiative for digital preservation. The goal of Nestor is a permanent distributed infrastructure for long-term preservation and long-term accessibility of digital resources in Germany. As the perspective of current and future archive users is central to the project, the emphasis is put on long-term accessibility of digital resources. [7]

Digital Libraries Research and Development

Research and development trends follow projects focusing on personalization for information access, visualization, metadata crosswalks, multilingual retrieval and semantic web. More emphasis is in the area of semantic web aiming at semantic retrieval on the web.

Personalization:
In the Web information environment personalization is an effort to relieve information overload and provide users tailor-made information best suiting to their information needs. In more generic terms, personalization involves a process of gathering user-information during interaction with the user, which is then used to deliver appropriate content and services, tailor-made to the user's needs. The aim is to improve the user's experience of a service [Bonett, 2001]. Many web service providers like Yahoo!, Google, Netscape and many others have started customization services. In traditional libraries also information personalization or tailor-made information services is not a new concept. Selective Dissemination of Information (SDI) service is one such example. Modern day digital libraries are also aiming at providing personalized user interface along with personalized information services. Cornell University Library's MyLibrary (http://mylibrary.cornell.edu/MyLibrary/Main) service one of such services which brings under one umbrella a whole variety of personalizing tools whereby users can make use of only a selected set of information in a plethora of information.

Semantic Web:
The basic problem of semantic (meaningful) information retrieval still persists in Web information retrieval scenario though technologies have been developed to integrate search from several digital libraries at a time. The idea of Semantic Web is proposed by Tim Berners-Lee of World Wide Web Consortium (W3C) who describes the

Semantic Web as “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation” [Berners-Lee et.al, 2001]. Basically it’s an effort to make the information available on Web machine processible so that web information retrieval systems can retrieve meaningful information, which will result in many web services.
Typical usage scenarios for Semantic Technologies in Digital Libraries include among others user Interfaces and human-computer interaction (displaying information, allowing for visualization and navigation of large information collections), user profiling (taking into account the overall information space), personalization (balancing between individual and community-based personalization), user interaction [Sure and Studer, 2005].

Semantic Web Enabler Technologies:
Hybrid models are built for semantic analysis and representation. The technologies include the following:

- XML
- RDF/RDFS
- Ontologies
- Annotations
- Inference Engines

Initiatives on Semantic Digital Library

Semantic Interoperability of Metadata and Information in unLike Environments (SIMILE) :
SIMILE (http://simile.mit.edu/) is a joint project conducted by the W3C, MIT Libraries, and MIT CSAIL. It seeks to enhance inter-operability among digital assets, schemata/vocabularies/ontologies, metadata, and services. A key challenge is that the digital collections which must inter-operate are often distributed across individual, community, and institutional stores. It aims to be able to provide end-user services by drawing upon the assets, schemata/vocabularies/ontologies, and metadata held in such stores. Software developed under this project to achieve its objectives:

- Longwell
- Haystack
- Gadget
- RDFizers
- Welkin
- Piggy Bank
- Semantic Bank

SIMILE project also aims at enhancing the features of DSpace open source digital library software to support arbitrary schemata and metadata through the application of RDF and semantic web technologies.

Protégé

Protégé ***(http://protege.stanford.edu/index.html)*** is a Java based open source ontology editor and tool for knowledge base framework, which allows users to construct domain ontologies, customize data entry forms, and enter data. It is also a platform, which can easily be extended to include graphical components such as graphs and tables, media such as sound, images, and video, and various storage formats such as OWL, RDF, XML, and HTML.

The Protégé Application Programming Interface (API) makes it possible to for other applications to use, access, and display knowledge bases created with Protégé.

Virtual Solider (http://www.virtualsoldier.net/), BioMediator (http://www.biomediator.org/), eMystics (http://emystics.org/) , SAGE Project (http://www.sageproject.net/) are some of the examples of Protégé implementation.

Conclusion

India is beginning to realize the importance of open access to information, digital libraries, ETD repositories and digital preservation. As some of German institutes and organization have gone through this initial phase, Indian professionals will greatly benefit from the German experience. At the same time, Indian professionals have rich experience in traditional librarianship, which if well adapted to web environment, greatly enriches the web. This has been obvious with the emphasis on metadata and ontologies in the recent past and both the topics are traditionally dealt by library and information science professionals.

In summary, Indian library professionals would be willing to collaborate with German counterparts in the following areas:

1. Institutional Repositories
2. Digital preservation
3. Digitization: best practices
4. Open standards
5. Personalization and visualization
6. Semantic web
7. Web ontologies
8. Metadata and crosswalks
9. Cross Language Information Retrieval
10. ETD repositories

Acknowledgement

The author acknowledges with thanks the support extended by Max Muller Bhavan-New Delhi (Goethe Institute) for conducting this study to facilitate Indo-German collaborations in Digital Libraries

References

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 2001(5). Available at http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Bonett, Monica. (2001) Personalization of web services: opportunities and challenges. Ariadne Issue no. 28, 22 June, 2001. Available at http://www.ariadne.ac.uk/issue28/personalization/intro.html

Conway, Paul. (1994). Digitizing preservation. Library Journal, (February 1, 1994): 42-45.

Current activities of Die Deutsche Bibliothek in the field of electronic publications http://www.nla.gov.au/padi/topics/DDB.html

Digital Preservation Coalition. Preservation issues. Available at http://www.dpconline.org/graphics/digpres/presissues.html

Exile Collections of Die Deutsche Bibliothek http://www.goethe.de/kug/prj/bib/bil/spezb/en49544.htm

Nestor Network. Network of expertise in long-term storage of digital resources http://www.langzeitarchivierung.de/index.php

Sure, York and Studer, Rudi. (2005). Semantic Web technologies for digital libraries. Library Management, Volume 26, Numbers 4-5, April 2005, pp. 190-195(6).

Trusted digital repositories: attibutes and responsibilities. An RLG-OCLC Report. May 2002. Available at http://www.rlg.org/legacy/longterm/repositories.pdf

Wisser, Katherine M. (2005). Guidelines for Digitization. 2005 Edition. At http://www.ncecho.org/Guide/toc.htm

Wissenschaftliche Kooperation und Kommunikation durch GridComputing

Grids als Plattform für eScience: Chancen und offene Fragen

Heinz-Gerd Hegering, München

1. Begrifflichkeiten, Einführung

Der Titel des Beitrages verlangt zunächst einmal die Klärung der Begriffe Grid, eScience und Plattform. Alle drei Begriffe sind im Gebrauch; sie sind aber unscharf definiert und werden üblicherweise mehrdeutig verwendet.

Grids

Der Begriff Grid entstand zunächst in den 90er Jahren im Umfeld des Höchstleistungsrechnens, als man zunächst Rechnerkapazitäten zusammenzuschalten versuchte, um mehr Leistungskapazität auf der Grundlage von Metacomputing zu erzielen. Ziel war also koordiniertes Resource Sharing, um große Rechenprobleme lösen zu können. Angegangen wurde dieser Ansatz von Forschergruppen in der Teilchenphysik und anderen naturwissenschaftlich orientierten Disziplinen, deren Mitglieder räumlich verteilten unterschiedlichen Institutionen angehörten, die aber eine gemeinsam interessierende wissenschaftliche Fragestellung beantworten wollten. Natürlich wurde bald klar, dass dieses Vorgehen grundsätzlich erweitert werden kann und nicht auf Rechner als Ressourcen beschränkt bleiben muss. Somit kann der Begriff Grid nach Foster weiter gefasst werden (Foster 1999, 2004):

„The Grid is an emerging infrastructure that will fundamentally change the way we think about and use computing … The word GRID is chosen by analogy with the electric power grid which provides pervasive access to power … Grid means coordinated resource sharing in dynamic, multi-institutional virtual organizations … It means the ability to discover, allocate, and negotiate the use of network-accessible capabilities, be they computational services offered by a piece of software, or storage space provided by a storage system …"

Grid Computing bedeutet also gemeinsames, zielorientiertes Nutzen von Ressourcen in koordinierter Weise auf der Basis einer Infrastruktur, bestehend aus vernetzten Ressourcen und einer adaptiven Softwarearchitektur, die die erforderliche Koordinierungsfunktionalität bereitstellt. Diese adaptive Software wird oft als Grid Middleware bezeichnet. Zusammen mit den Ressourcen und deren steuernder Betriebssoftware bilden sie die Grid-*Plattform.*

eScience
Der Begriff wurde von John Taylor im Zusammenhang mit der britischen eScience-Initiative 2003 aufgebracht (Hegering et al. 2003): *„In the future, eScience will refer to the large scale science that will increasingly be carried out by stronger distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists …".*
Der Begriff eScience geht also deutlich über den Begriff Grid hinaus. Er bezeichnet vernetztes Forschen auf der Basis von Grids und Grid-Methoden, wobei der Ressourcen-Begriff gegenüber den oben definierten Grids noch erweitert werden kann: zur gemeinsamen Nutzung im vernetzten Kontext bieten sich auch Modelle, Verfahren, Simulationen, veredelte oder vorverarbeitete Daten, Programme usw. an. Bei eScience handelt es sich um die Gestaltung und Nutzung neuartiger Internet-basierter wissenschaftlicher Arbeitsumgebungen auf der Basis neuartiger Kooperationsmodelle. Grids im obigen Sinne dienen dabei als „enabling systems", also als Plattformen für eScience.

2. Chancen von eScience und Grids

Grids betreffen die Aufgabenstellungen, einen adaptiven Software- und Organisationsrahmen für eScience bereitzustellen. In Essenz sind damit Aspekte betroffen wie Virtualisierung von Ressourcen und Virtualisierung von Organisationen sowie Unterstützung von flexiblen Kollaborationsmöglichkeiten und gemeinsamer Ressourcennutzung. Die systemtechnisch geeignete Unterstützung dieser Aspekte würde im Sinne von eScience die Bildung von spontanen oder geplanten Teams und Communities erleichtern, somit auch einer Förderung von Kooperation und Interdisziplinarität in der Forschung dienlich sein, national wie international, innerhalb der Wissenschaft und zwischen Wissenschaft und Wirtschaft. eScience bedeutet, dass man auf der Basis von Grid-Plattformen die Gemeinsamkeit im Sinne einer Schichtung von Diensten und Ressourcen soweit wie möglich und sinnvoll nutzt und anbietet, damit möglichst viele Fachdisziplinen davon profitieren können, insbesondere in Kooperationsprojekten. Eine eScience-Plattform soll begriffen werden können als gemeinsamer, standardisierter „Werkzeugkasten", mit dem eScience-Anwendungen sich zusammenbauen lassen. Diese Vision lag auch dem folgenden Leitbild für ein eScience-Framework zugrunde, das in den Jahren 2003 und 2004 innerhalb der deutschen D-Grid-Initiative ((Hegering et al. 2003) & (eScience 2004)) entwickelt wurde.
eScience (digitally enhanced science) ist gemäß (eScience 2004) die Bezeichnung für eine Arbeitsweise in der Wissenschaft, die durch gemeinsame, kooperative Entwicklung, Öffnung und Nutzung ihrer Ressourcen und Projekte eine wesentliche

Steigerung der Qualität und Leistungsfähigkeit erreicht. Ressourcen sind wissenschaftliche Verfahren einschließlich Expertise, Software, Datenbestände, Rechner, Kommunikationsnetze und andere wissenschaftliche Geräte. In zahlreichen Disziplinen ermöglicht eScience erst bestimmte Formen der wissenschaftlichen Arbeit, die die Bearbeitung neuer Zielstellungen ermöglichen und so zu völlig neuen Erkenntnissen führen können.

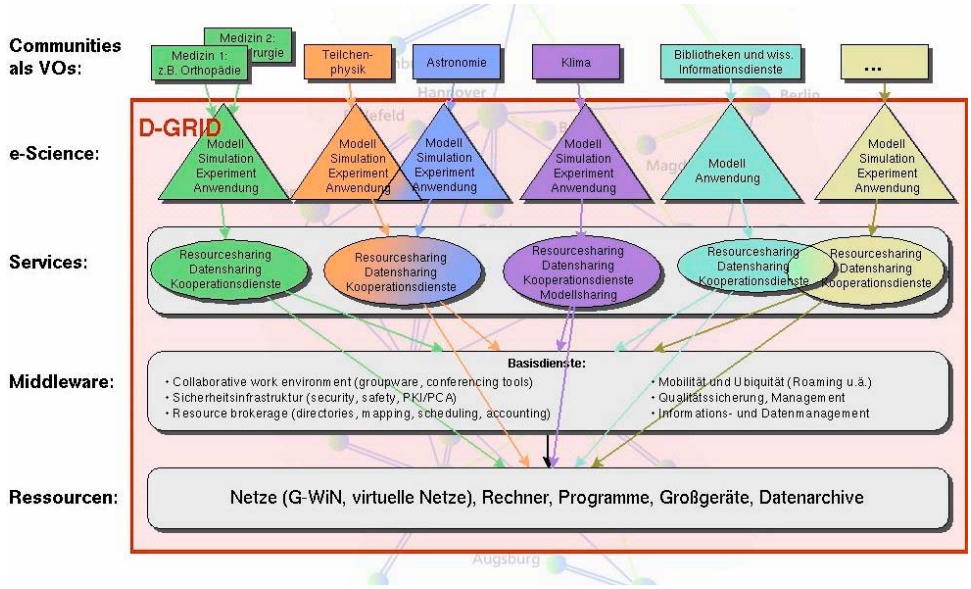


Abbildung 1: D-GRID

Mit dem Übergang zur gemeinsamen Entwicklung und Bereitstellung von Ressourcen, insbesondere Verfahren und Daten, verschiebt sich zugleich die Wertschöpfung in der Wissenschaft. Es kann erreicht werden, dass ein abnehmender Teil der wissenschaftlichen Arbeit der Erzeugung von Verfahren und Daten gewidmet wird, die bereits an anderem Ort vorhanden sind; damit werden Kräfte für die Gewinnung neuer Erkenntnisse frei.

Das obige Bild verdeutlicht inhärentes Synergiepotential: die Bündelung und Nutzung von Middleware, Services und eScience-Methoden über bisher oft noch isoliert agierende Wissenschafts-Communities hinweg. Das Framework erleichtert auch das Einbringen von Ressourcen, Informationen und Datenständen zur Nutzung durch neue Communities.

eScience ersetzt nicht bisheriges Forschungsvorgehen, sondern will neue Gestaltungsräume für wissenschaftliches Arbeiten erschließen. Die bisherigen positiven Erfahrungen vieler Grid-Communities ermutigen dazu.

3. Herausforderungen bei Grids

Charakteristika im Grid-Umfeld

Um Herausforderungen bei Grids beleuchten zu können, ist es zunächst erforderlich, nochmals die Aspekte aufzuführen, die Grids zu einem Paradigmenwechsel werden lassen.

Virtualisierung von Organisationen: Grid-Communities sind in der Regel dynamisch bildbare, oft temporäre Gruppen, deren Mitglieder üblicherweise von verschiedenen realen, rechtlich unabhängigen Organisationen stammen. Communities sind durch eine gemeinsame „Interessenslage" charakterisiert (gemeinsame Ziele, Aufgaben, Ressourcen).

Virtualisierung von Ressourcen: Grid-Anwendungen arbeiten auf i.d.R. verteilten Ressourcen, auf die die Communities system-, orts- und organisationsübergreifend zugreifen möchten. Dies verlangt u.a. ein community-angepasstes Maßschneidern von Diensten und Management-Randbedingungen einer Grid-Infrastruktur.

Flexibilisierung von Kooperationsformen durch entsprechende Groupware, Conferencing Tools etc.

Bereitstellen einer Infrastruktur im Grid-Umfeld

Die Ressourcenanbieter eines Grids sind i.d.R. verschiedene formale Organisationen mit unterschiedlichem Rechts- und Finanzierungshintergrund, die nicht notwendig Mitglieder einer Community sind. Dies bedingt besondere Absprachen und Verfahren im AAA-Bereich (Autorisierung, Authentifizierung, Abrechnung). Selbst bei identischen Ressourcen können diese Regelungen Community-spezifisch verschieden sein.

Grid-Ressourcen sind sehr verschiedenartig (z.B. Rechensysteme, Speicher, Software, Datenarchive, Messgeräte und Experimente, Netzdienste mit diversen QoS-Anforderungen, spezielle Dienste) und in sich oft heterogen durch verschiedene technische Systemumgebungen (z.B. Hardware-Architekturen, Betriebs- und Dateisysteme, Datenformate, Software-Versionen). Noch komplexer wird der Aspekt Ressourcen-Bereitstellung durch die Tatsache, dass in einem Grid i.d.R. die Anbieter zu verschiedenen, unabhängigen Organisationen gehören. Das bedeutet unterschiedliche Verfügbarkeit, verschiedene Betriebs- und Nutzungskonzepte. Grid-Middleware hat hier für die Schaffung der gewünschten Transparenz bzw. der nötigen Abbildungen zu sorgen.

Ein weiterer Problemkreis betrifft die Ressourcenvergabe. Dies berührt nicht nur die Klärung und Durchsetzung von Regeln zur „acceptable use policy", sondern auch die Entwicklung geeigneter Abrechnungsmodelle. Letzteres schließt die Frage nach Grid-Verrechnungseinheiten (Grid-Währung) ein; die Heterogenität der Ressourcen setzt ein geeignetes Ressourcen-Benchmarking voraus. Hinzu kommt, dass Grid-Ressourcen (freie Kapazitäten, Nutzungsbedingungen, Zugangsregelungen) oft potentiellen Interessenten nicht bekannt sind. Somit werden geeignete Verzeichnis-

und Brokerdienste erforderlich mit der zusätzlichen Möglichkeit, nutzer- oder communityspezifische Service Level Agreements zu vereinbaren.

Betrieb einer Grid-Infrastruktur

Wegen der Ressourcenheterogenität kommt der Lösung des Interoperabilitätsproblems auf allen Ebenen des Grid-Modells Bedeutung zu. Entsprechendes gilt für die Aktualität und Zuverlässigkeit der Broker und Ressourcenverzeichnisse.

Die Ressourcenanbieter wollen i.d.R. autark und autonom bleiben in Bezug auf ihre lokalen IT-Infrastrukturen samt ihrer Betriebskonzepte, Grid-Anwendungen laufen aber auf verteilten Ressourcen. Diese Situation erschwert ein umfassendes Anwendungsmonitoring, aber auch das Einrichten von Prozessen des Incident oder Problemmanagement. Ein sonst übliches zentrales IT-Management ist im Grid-Umfeld inadäquat. Entsprechend ist eine stärkere Unterstützung von flexiblen und adaptiven Betriebsabläufen gefragt: automatisiertes Konfigurieren, selbständiges Recovery, „self-healing“, Verfahren eines intelligenten Lastausgleichs seien als Beispiele genannt.

Ein verteiltes Management legt wegen der Autonomie und Verschiedenartigkeit der Anbieter, Nutzer und Communities die Errichtung eines Qualitätssicherungssystems sowie eines Grid Operation Center nahe. Diese müssen Aufgaben übernehmen wie Evaluierung, Zertifizierung, Ressourcen-Benchmarking, Interoperabilitätstests, Versionenpflege, Tests von Musterabläufen, Anstoßen von Managementprozessen u.dgl.

Unverzichtbar für eine Grid-Plattform ist ein umfassendes Sicherheitskonzept, das neben der AAA-Problematik heterogener Ressourcen von autonomen Anbietern die Herausforderungen der dynamischen und temporären virtuellen Organisationen berücksichtigen.

Virtuelle Organisationen im Grid-Umfeld

Communities sind virtuelle Organisationen (VOs). VOs müssen IT-gestützt gebildet und verändert werden können. VOs sind nicht nur charakterisiert durch ihre Mitglieder oder die von ihnen genutzten Ressourcen, Methoden und Dienste, sondern insbesondere auch durch die VO-Policies, die den Nutzungszugang, die Gruppenzugehörigkeit, die Sicherheitsaspekte, die Dienstgüte und die Abrechnung steuern. Entsprechend müssen Grid-Verzeichnisse Eigenschaften und Rollen der zertifizierten Personen und Gruppen für AAA-Zwecke enthalten und Grid-Operationen Policy-Spezifikationen und deren Durchsetzung unterstützen sowie Zielgruppen-spezifischen Support technisch und organisatorisch vorsehen. Dass Werkzeuge für Groupware und Teleconferencing unverzichtbar sind, ist offensichtlich.

Grid-Management: Sonstige Aufgaben

Um den zuvor genannten Herausforderungen zu begegnen, müssen nicht nur technische Managementsysteme für Grids entwickelt werden. Begleitend und vorbe

reitend bedarf es einer Standardisierung von Ressourcen- und Dienstbeschreibungen sowie der Entwicklung von Metriken für Grid-Verrechnungseinheiten. Es fehlt an der Klärung der vielfältigen rechtlichen Fragen im Zusammenhang mit der Grid-Ressourcen-Bereitstellung. Um die Nachhaltigkeit einer allgemeinen Grid-Plattform für z.B. eScience zu sichern, ist der Entwurf einer Rahmenvereinbarung für alle Grid-Anbieter samt Qualitätssicherungsmethoden ebenso sinnvoll wie der Entwurf von Musterverträgen für bilaterale Beziehungen zwischen Nutzer und Anwender. Noch gibt es zudem wenig Erfahrung mit Betriebs- und Nutzungsmodellen oder Zugangs- und Vergaberegeln für eine allgemeine Grid-Infrastruktur. Natürlich wird eine solche nicht aus nur einem Grid bestehen.

Forschungs- und Entwicklungsaufgaben

Die D-Grid-Initiative hat unter Berücksichtigung internationaler Grid-Projekte für die Vorbereitung eines deutschen eScience-Förderprogrammes in mehreren Arbeitskreisen ein Forschungs- und Entwicklungsprogramm [4] aufgestellt, in dem folgende Aufgabengruppen genannt sind:

1. Communities: Mechanismen zur Bildung Virtueller Organisationen, Beschreibung von VO-Policies und SLAs, Simulationen des Verhaltens von Community Grids.
2. Middleware/Services: Verbesserung der Gridfähigkeit (Job- und Ressourcen-Monitoring, Job-Accounting), Verbesserung der Ausführungseffizienz (Ressourcen-Scheduling), Verbesserung des Zuverlässigkeits- und Automatisierungsgrades, Verbesserung von Sicherheit und Vertrauenswürdigkeit, Daten- und Informationsmanagement.
3. Ressourcen: Standardisierung von Ressourcenbeschreibungen, Ressourcenerfassung, Ressourcen-Metrik, Ressourcen-Qualitätssicherung, Ressourcen-Sicherheit.
4. Netze: Erweiterung von VPN-Techniken, Verbesserung der Mobilität in Grids, Alternative Transportprotokolle.
5. Querschnitts- und Integrationsaufgaben: Betriebsmodelle, Sicherheit, Clearings, Nutzersupport etc.

4. Schlussbemerkungen

Die D-Grid-Initiative hat konkret zu einer Förderprogramm-Ausschreibung und entsprechenden Projektanträgen geführt. Einige Community-Projekte und ein Integrationsprojekt starteten ihre Arbeit im Spätsommer 2005. Natürlich müssen weitere Ausschreibungen folgen, denn eScience und die zugrunde liegende Grid-Infrastruktur bergen riesige Herausforderungen.

Es ist auch klar, dass alle Arbeiten im internationalen Kontext gesehen werden müssen, denn eScience macht nicht an nationalen Grenzen halt. Bereits die dazu

nötigen internationalen Kooperationen geschehen unter Zuhilfenahme von Grid-Verfahren. International werden die Anstrengungen und Förderprogramme für eScience und Grids intensiviert, wohl wissend, dass noch ein Stück Weg gegangen werden muss bei der Beantwortung vieler offener Fragen, um letztlich die Chancen, die eScience bietet, nutzen zu können.

Literatur

Foster, I. and C. Kesselman: The Grid. Morgan Kaufman, 1999

Foster, I. and C. Kesselman: The Grid 2. Morgan Kaufman, 2004

Hegering, Hiller, Maschuw, Reinefeld, Resch: D-Grid: Auf dem Weg zur eScience in Deutschland. Strategiepapier, www.d-grid.de, 2003

eScience in Deutschland, F&E-Rahmenprogramm 2005-2009, www.d-grid.de, Juli 2004

X-WiN – Netzressource im GRID

Von ersten Ansätzen der Wissenschaftsnetze zu modernen Kollaborationswerkzeugen

Karin Schauerhammer, Berlin

Abstract

In einem kurzen geschichtlichen Abriß der Wissenschaftsnetze werden vor allem die Rolle der Wissenschaftsnetze als Kollaborationswerkzeuge skizziert. GRIDs als relativ neues Werkzeug wissenschaftlicher Arbeit wird weitgehend am Beispiel des deutschen GRID Programmes „D-Grid“ dargelegt. Schließlich wird am Beispiel eines konkreten Grid Projektes aus der wissenschaftlichen Welt, dem Grid der Teilchenphysik (LCG), das Potential des Grids, die Bedeutung der Wissenschaftsnetze für die Implementierung von Grids und die Rolle des LCG als Prototyp des neuartigen Werkzeuges dargestellt.

1. Vom Rechner zum Kommunikationsinstrument Netzwerk

An einigen Wissenschaftseinrichtungen in den USA gab es in den 60ger Jahren erste Überlegungen, Rechner miteinander zu vernetzen. J.C.R. Licklider, ein Wissenschaftler vom MIT, veröffentlichte z.B. gemeinsam mit Taylor 1967 einen der ersten Artikel zum Thema Rechnernetze: "The Computer as a Communications Device", in dem das Potential von vernetzten Rechnern als Kommunikationsinstrument auch zur Kollaboration in der Wissenschaft beschrieben wurde. Ende der 60er Jahre wurde im Auftrag des U.S. Verteidigungsministeriums das Computernetzwerk ARPAnet entwickelt, welches am 1.9.1969 an der University of California, Los Angeles in Betrieb genommen wurde. Die ARPAnet Technologie wurde später zur Grundlage für das heutige Internet.

Das ARPANET sollte es ermöglichen, Computer von verschiedenen wissenschaftlichen Instituten des Landes zu verbinden, um gemeinsam Ressourcen nutzen zu können. Damit wurden erstmals Wege zu einer netzgestützten Kommunikation und Kollaboration in der Wissenschaft eröffnet.

Die Anzahl der Rechner und die Anzahl der Nutzer in diesem Netz stiegen ständig an. Neue Dienste wie Email, Usenet/News, Filetransfer (FTP) brachten erheblichen Fortschritt der Kommunikationsmöglichkeiten.

Mit der Entwicklung des WWW (im CERN, also in der Wissenschaft) wurde Mitte der neunziger Jahre eine neue Stufe der Kommunikations- und auch Publikationsmöglichkeiten erreicht; Wissen ist vernetzt und leicht zugänglich für „jedermann“ im

Web erreichbar. Initiativen wie die „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen“ werden die Formen wissenschaftlichen Publizierens beeinflussen.
In Europa entstanden etwas später als in den USA – in den 70ger Jahren - die ersten Rechnernetze für die Wissenschaft, als erstes das britische JANET, später folgten weitere. 1986 wurde die Organisation RARE (Réseaux Associés pour la Recherche Européenne) gegründet, die Aktivitäten zur Rechnervernetzung europaweit koordinierte. RARE rief dazu u.a. ein Projekt namens COSINE (Cooperation for an Open Systems Interconnection Networking in Europe) ins Leben, das den organisatorischen Nukleus der heute verfügbaren Vernetzung der Wissenschaft in Europa bildete.
Das TCP/IP-Protokoll setzte sich als de facto Standard durch. In Europa entstand schließlich ein Datennetz, das multiprotokollfähig war und unter anderem TCP/IP unterstützte. Dieses Netz lief zunächst unter der Bezeichnung EuropaNET. Verschiedene nationale Wissenschaftsnetze, auch das Deutsche Forschungsnetz (der DFN–Verein gegründet 1984 mit dem Zweck für die Wissenschaft in Deutschland eine Kommunikationsinfrastruktur aufzubauen und zu betreiben) schlossen sich an. Ähnliche Entwicklungen fanden in anderen Ländern und auf anderen Kontinenten statt mit dem Ergebnis, daß ein weltweiter Verbund von Wissenschaftsnetzen basierend auf der Internettechnologie sich zu einer unverzichtbaren Infrastruktur für Forschung, Wissenschaft und Lehre entwickelte. Förderprogramme der öffentlichen Hand unterstützten in der Aufbauphase die Entwicklung von Wissenschaftsnetzen, die EU trägt zur Zeit mit beträchtlichen Fördermitteln (etwa 100Mio€ über vier Jahre) zum Aufbau des europäischen Overlaynetzes Geant2 bei.

2. Wissenschaftsnetze und Wissenschaftskommunikation in Deutschland und Europa

Derzeit versorgt das Deutsche Forschungsnetz (DFN) mit seinem Gigabit-Wissenschaftsnetz (G-WiN) ca. 450 Universitäten, Fachhochschulen und wissenschaftliche Institute mit Internetzugängen von 2 Mbit/s bis zu 10 Gb/s. Über diese Zugänge können klassische Netzdienste wie E-mail, News, FTP und WWW aber auch höherwertige Kommunikationsdienste wie der DFN Videokonferenzdienst genutzt werden, ohne die heute eine Kommunikation zwischen Wissenschaftlern national und international nicht mehr denkbar ist. Über die technische Versorgung mit Netzdiensten hinaus bietet der DFN Verein eine Plattform zum Austausch von Informationen sowie Beratungs- und Kompetenzzentren für neue Technologien an wie zum Beispiel eines für den Videokonferenzdienst und eines für das Thema Sicherheit im Netz.
Das im G-WiN transportierte Datenvolumen beträgt derzeit ca. 1 Pbyte/Monat und nimmt mit den Anforderungen und dem Aufkommen neuer datenintensiver Anwendungen (Teilchenphysik, Klimawissenschaften, Astronomie) aus der Wissen

schaftscommunity stetig zu. Diese Entwicklung ist typisch für alle europäischen und auch die US amerikanischen Wissenschaftsnetze (National Research and Education Networks, NRENs). Wegen dieser wichtigen Infrastrukturfunktion ist es nötig, den Ausbau der Wissenschaftsnetze weltweit voranzutreiben. Erst mit der Deregulierung der Telekommunikationsmärkte in Europa und der Produktionsreife und Verfügbarkeit von optischer Netztechnik ist es möglich geworden, den Bandbreitenbedarfen der Wissenschaft wirtschaftlich sinnvoll nachzukommen.
So werden derzeit in Europa zahlreiche sogenannte hybride glasfaserbasierte Wissenschaftsnetze aufgebaut, die es ermöglichen, auf einer Netz-Plattform sowohl den wissenschaftsspezifischen Internet Verkehr zu transportieren als auch dedizierte optische Virtuelle Private Netzwerke (VPNs) bereitzustellen. Typische Anwendungen für solche VPNs sind Verbünde für Datensicherung zwischen Universitäten, internationale Projekte, bei denen die Verteilung von großen Experimentdatenmengen eine Rolle spielt (z.B. Teilchenphysik), Kopplung von Hochleistungsrechnern (DEISA) oder auch die Auswertung von Daten aus Sternwarten, die in sehr vielen europäischen Ländern ihre Daten aufnehmen und zentral auswerten (Astrophysik) wollen.
Die Kostengesetze, die diesen hybriden Netzen innewohnen, ermöglichen es, Bandbreiten bedarfsgerecht zu akzeptablen Entgelten bereitzustellen. Mit dem Start des Geant2 im Juli diesen Jahres und dem Aufbau des X-WiN in Deutschland bis zum Ende diesen Jahres steht der Wissenschaft in Deutschland eine moderne leistungsfähige Netzinfrastruktur zur Verfügung. Im X-WiN werden Anwenderbandbreiten von 10 Gb/s keine Ausnahme mehr sein.

3. Grid – eine neue Methode des netzgestützten wissenschaftlichen Arbeitens

In den vergangenen Jahrzehnten haben internationale Experimente und Simulationsrechnungen in verschiedenen naturwissenschaftlichen Disziplinen, z.B. der Teilchenphysik und der Astronomie, zu einem rasant steigenden Bedarf an Rechen- und Speicherleistung geführt. Die in Experimenten und Messungen anfallenden großen Datenmengen können nur an einigen Standorten, an denen ausreichend Rechen- und Speicherressourcen zur Verfügung stehen, analysiert und gesammelt werden. Es entsteht die Notwendigkeit, eine standortunabhängige und kostengünstige Auswertung sowie Speicherung von Massendaten zu ermöglichen.
Aufgrund dieser Entwicklung entstand Ende der 90er-Jahre die Vision, nicht nur Rechner- und Speicherressourcen, sondern auch Datenbanken, Datenauswertungstools und große Geräte wie z. B. Beschleuniger in der Teilchenphysik über eine neuartige IT (Informationstechnologie)-Architektur bereitzustellen. Diese neue Architektur soll „flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources“ ermöglichen. Foster und Kesselmann nannten dieses Konzept Grid-Computing (Foster I., Kesselmann

C.;1999), (Foster I., Kesselmann C, Tuecke S.; 2001). Für den Betrieb solcher Grids ist die Entwicklung und Bereitstellung von Hochleistungsnetzen zur Verbindung der Reccourcen und Anwendungen eine grundlegende Voraussetzung. Die Heterogenität der Daten, unterschiedliche Hard- und Softwarestandards und die Notwendigkeit einer Sicherheitsinfrastruktur stellen dabei eine besondere Herausforderung für Entwicklungen und Einsatz der Grid-Technologie dar.
In den USA (Cyberinfrastructure Initiative), Großbritannien (e-Science programme), und den Niederlanden (Virtual Lab) wurden um das Jahr 2000 große Förderprogramme der öffentlichen Hand aufgelegt, um die Entwicklung der neuen Grid-Technologie voranzutreiben. In Deutschland gab es erste Grid-Initiativen um das BMBF Projekt Unicore sowie im DFN Verein, der mit seiner G3- Gruppe erste Grid-Projekte mit Unterstützung des BMBF auf den Weg brachte. Im Jahr 2003 gründete sich die D-Grid-Initiative, ein Zusammenschluß der wichtigsten deutschen Wissenschafteinrichtungen (HGF, MPG, Hochschulen, FHG, DFN-Verein), um die Bedürfnisse und Anforderungen der Wissenschaft an die Entwicklung und Nutzung der Grid-Technologie in Deutschland zu formulieren.
Die Initiative legte im Juni 2004 ein „Forschungs- und Entwicklungsrahmenprogramm 2005 – 2009, e-Science in Deutschland“ vor.
In diesem Programm heißt es u.a.:„*Virtualisierung* ist daher der Schlüsselbegriff zukünftiger Grid-Systeme. Im Gegensatz zu den existierenden verteilten Informationssystemen virtualisieren Grid-Systeme der nächsten Generation Daten-, Informations- und Rechendienste, indem sie die technischen Details der konkreten i.a. verteilten Realisierung hinter Oberflächen (interfaces) verbergen. ... Zukünftige Grid-Systeme werden Tausende geographisch verteilter Ressourcen umfassen, die über Weitverkehrsnetze, wie z.B. das Internet, miteinander verbunden sind. Neben der Virtualisierung ist *die kooperative Nutzung* von IT-Ressourcen ein Schlüsselelement zukünftiger Grid-Generationen. Diese ermöglichen vollständig neue Arbeitsformen in Wissenschaft und Industrie.“ (D-Grid Initiative; 2005-2009)
Um dies zu erreichen, will die D-Grid-Initiative in Deutschland die bestehenden -Aktivitäten bündeln, um Synergien für globale, verteilte Wissenschaftskollaborationen auf der Basis netzgestützter Dienste freizusetzen (e-Science). Hierzu ist es notwendig, eine Netz- und Middleware-Infrastruktur zu entwickeln und aufzubauen, die sich an internationalen und europäischen Standards und Projekten (wie z.B. dem EGEE-Projekt) orientiert und dazu kompatible Tools und Dienstleistungen liefert.

Die Architektur dieser technischen Infrastruktur wird in Abbildung 1 dargestellt.

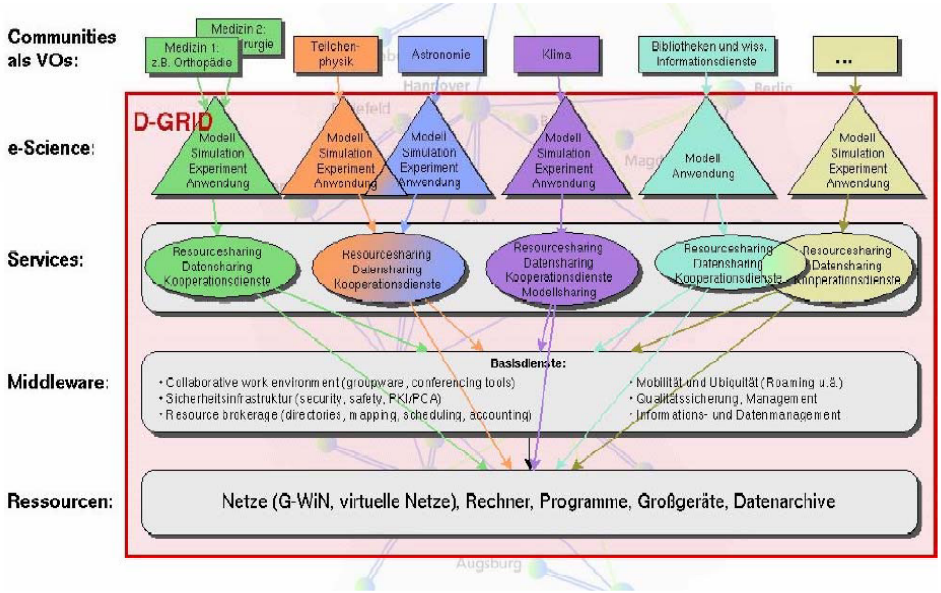


Abbildung 1: D-Grid Framework für e-Science (VO=Viruelle Organisation)

Sie gliedert sich in vier Schichten und basiert auf einer Vielzahl von miteinander zusammenhängenden und/oder aufeinander aufbauenden Grid-Komponenten. In der untersten Schicht befinden sich Ressourcen (Rechner, Programme, Großgeräte, aber auch Datenarchive) und ein die Ressourcen verbindendes leistungsfähiges Netz.
Die Aufgabe der Middleware ist es, Heterogenität und räumliche Verteilung der einzelnen Ressourcen vor dem Nutzer zu verbergen und soweit dies möglich ist, generische Basisdienste wie z.B. eine Authorisierungs- und Authentifizierungs-Infrastruktur, Kollaborationstools (z.B. Videokonferenz-Dienste) und Dienste zum Informations- und Datenmanagement für alle Communities zu erbringen.
Für die Wissenschaftscommunities wird mit diesen Diensten die gemeinschaftliche Nutzung von Daten, Modellen und anderen Ressourcen erleichtert. Im D-Grid müssen verschiedene Rechner- und Systemarchitekturen, Programmsysteme und Datenbestände integrieren werden.
In Deutschland steht ab 2006 für anspruchsvolle Grid-Anwendungen das X-WiN zur Verfügung, was mit seiner Konnektivität zu Geant2, den US amerikanischen wie auch über den sogenannten Global Upstream zu den kommerziellen Netzen, eine Plattform sowohl für den Internet-Verkehr als auch zur Schaltung dedizierter optischer Verbindungen in internationalen Kooperationen bildet. Im Kontext der EU

Infrastruktur-Förderung (6. Rahmenprogramm) ist der DFN Verein an Projekten wie Geant2 und EGEE aktiv beteiligt.
Er betreibt mit VIOLA ein optisches Testbed, das als Verbundprojekt mit Partnern aus Hochschulen, Forschungseinrichtungen, Industrieunternehmen und dem DFN-Verein organisiert ist. Ziel des Projektes ist es, neue Netztechniken und neue Formen der Netzintelligenz einzusetzen und sie integriert mit entsprechend anspruchsvollen Anwendungen zu erproben, um so den beteiligten Partnern Know-How für zukünftige Netzgenerationen zu vermitteln. So soll ein Testnetz aufgebaut werden, in dem Bandbreiten durch Anwendungen dynamisch angefordert werden können. Hierzu werden Untersuchungen und Entwicklungen auf Steuerungs- und Signalisierungsebene hochbitratiger Übertragungsnetze durchgeführt.

4. LCG – LHC Compute Grid

Eine der anspruchvollsten Grid-Anwendungen ist zur Zeit das LCG (LHC Compute Grid) -Projekt, das durch CERN koordiniert vorbereitet wird. Der Large Hadron Collider (LHC), der bis 2007/2008 am CERN aufgebaut wird, ist eines der größten wissenschaftlichen Instrumente der Erde. Wenn er seinen Betrieb aufnehmen wird, werden Datenmengen von jährlich 15 Petabyte anfallen, die von Wissenschaftlern weltweit zu analysieren sind. Die insgesamt mehr als 5000 Wissenschaftler aus ca. 50 Nationen, die an den LHC-Experimenten beteiligt sind, machen das LCG-Projekt zur aktuell größten gemeinschaftlichen Anstrengung der Physik im IT Bereich in der Grundlagenforschung.
Es wird vier Experimente am LHC geben: ALICE, ATLAS, CMS und LHCb.
Aufgabe des LCG-Projektes ist es, eine Datenspeicher- und Analyseinfrastruktur für die Teilchenphysik-Community, die den LHC nutzen wird, aufzubauen und vorzuhalten.
In Abbildung 2 ist ein logisches Modell des LCG dargestellt.

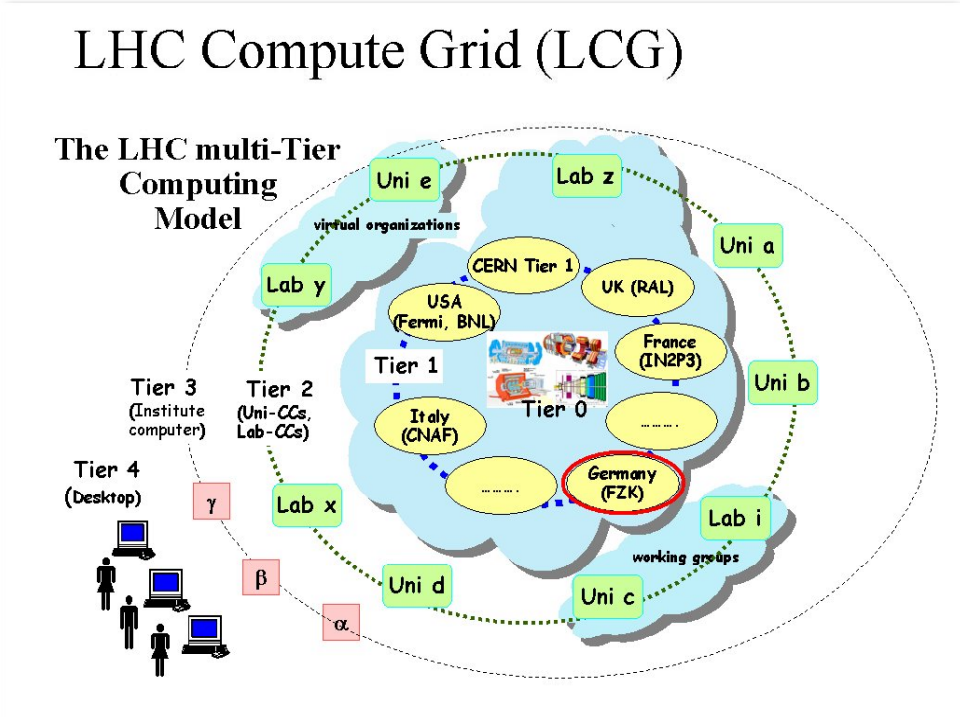


Abbildung 2: Modell des LHC Compute Grid (LCG)

Die Daten der LHC-Experimente werden weltweit nach einem Vier-Ebenen Modell (4 „Tiers“) verteilt. Ein erster Backup der Daten wird im CERN, dem „Tier0“-Zentrum des LCG gehalten. Nach einem Initialisierungsprozess werden diese Daten auf weitere 11 Tier1-Zentren verteilt. Dies sind Rechenzentren mit großen Speicher- und Rechenkapazitäten, die auch einen 24/7-Support für Grid-Dienste anbieten werden. Die Tier1-Zentren stellen Daten und Rechenleistung sowie Kollaborationswerkzeuge für die Analyse für Tier2-Zentren bereit. Einzelne Wissenschaftler greifen über Tier3-Zentren, z.B. lokale Cluster an Universitäten oder sogar individuelle PCs, die am LCG Projekt teilnehmen, zu.
Um diese netzgestützte kollaborative Arbeitsweise zu ermöglichen, ist es zwingend auf einer geeigneten Netzinfrastruktur, die betriebliche Domains in verschiedenen europäischen, den US amerikanischen, kanadischen und taiwanesischen Wissenschaftsnetzen, umfaßt, aufzusetzen. Es muß Vereinbarungen zwischen den Experimenten, dem CERN und den nationalen Zentren geben, die regeln, nach welchen Zugangspolicies und Cost Sharing Modellen die Infrastruktur genutzt werden kann.
Für die Planung des LCG-Netzes wurde eine Arbeitsgruppe aus Vertretern der Experimente, Netzspezialisten der beteiligten NRENs und Geant2 sowie der

Netzgruppen von CERN und Tier1-Zentren gebildet. Der Netzentwurf umfaßt derzeit zunächst das Tier0/Tier1-Netz, da noch keine abgestimmten Anforderungen und Festlegungen zu den Tier2–Tier3-Zentren vorliegen.
Es wurde von folgenden Designkriterien und Annahmen ausgegangen:

- Jedes Tier1 Zentrum wird mit einer dedizierten 10G-Verbindung mit dem Tier0 Zentrum verbunden. Eine weitere 10G-Verbindung wird als Backup zu einem benachbarten T1 Zentrum geführt.
- Es werden kontinuierliche Datenströme von bis zu 10Gbit/s erwartet.
- Aufgrund der einfachen Quelle-Senke Beziehung ist es sinnvoll, den Transport auf Ebene 2 abzuwickeln.
- Sicherheitsprobleme im Netz können durch diesen Ansatz leicht bewältigt werden („trusted sources“)

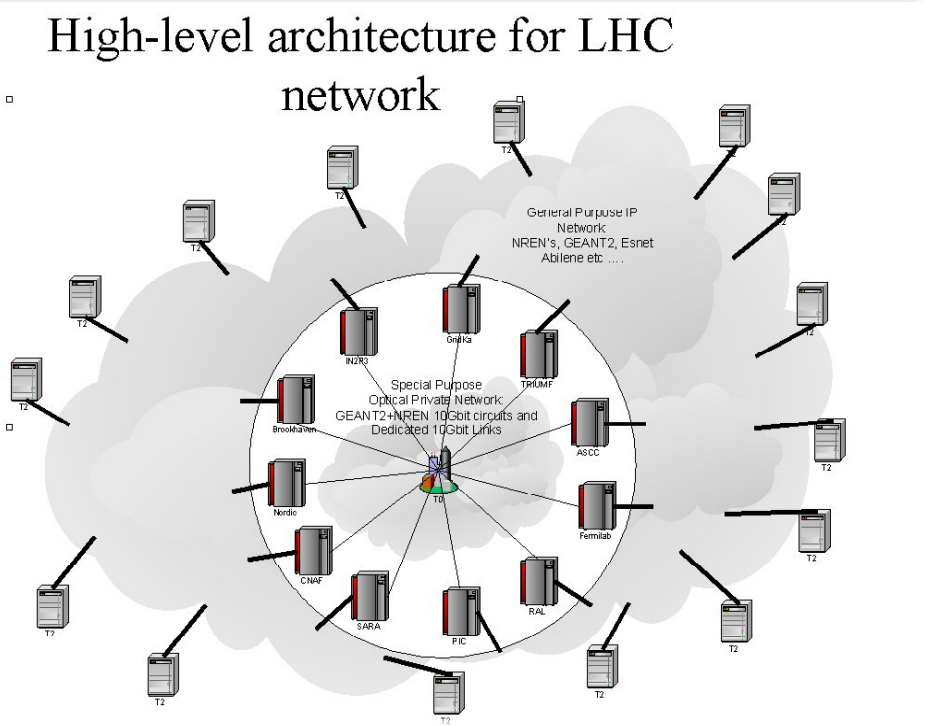


Abbildung 3: Architektur für das LCG-Netz (T0-T1 Netz)

Die Abbildung 3 zeigt das logische Modell für das LCG-Netzwerk; bestehend aus dedizierten 10G-Verbindungen zwischen Tier0 und jedem Tier1. Für ein Backup, der dedizierten Verbindungen Tier0-Tier1, der in Abbildung 3 nicht gezeichnet ist, wird jedes Tier1-Zentrum sich mit mindestens einem anderen Tier1-Zentrum über einen alternativen Weg verbinden. Für den Betrieb des Netzes wird eine spezielle

Routerstruktur (an jedem Tier1 und am Tier0) vorgesehen, die unabhängig vom normalen Internetverkehr der Tier1-Zentren nur für LCG-Zwecke vorgesehen sind. Damit ist eine für diesen Anwendungsfall eventuell nützliche neue Parametrisierung der Router (z.B. Packet Size) möglich, was zur Erhöhung des Datendurchsatzes beitragen wird. Der Betrieb der Router kann zentral erfolgen, was zu einer deutlichen Verbesserung der Zuverlässigkeit des Betriebes führen wird. In einem derartigen Optischen VPN kann bedarfsgerecht die Bandbreite erhöht werden, indem eine weitere 10G-Verbindung zu geringen Mehrkosten geschaltet wird.
Damit ist die Voraussetzung für die Verteilung der Daten auf die Tier1-Zentren weltweit geschaffen. In den an den Experimenten beteiligten Ländern nimmt die Möglichkeit der Verfügbarkeit von optischen Links ständig zu. So wird es in Deutschland mit dem X-WiN ohne Probleme möglich sein, die größeren Tier2-Zentren wie z.B. DESY und GSI mit optischen Verbindungen an das deutsche T1-Zentrum in Karlsruhe oder ggf. an ein weiteres T1 Zentrum heranzuführen.
Dieses Beispiel zeigt exemplarisch, welche entscheidende Rolle moderne Wissenschaftsnetze für neuartige Anwendungen (hier LCG Netz) spielen.

Ausblick

Mit der Entwicklung von Rechnernetzen für die Wissenschaft, die letztlich in die Entstehung des weltweiten Internets mündete, wurde eine grundlegende Änderung des kooperativen Arbeitens in der Wissenschaft und in der Kommunikation zwischen Wissenschaftlern angestoßen.
In verschiedenen Wissenschaftsdisziplinen stellen sich neue „grand challenges"-Aufgaben, die eine Menge von Ressourcen in einer bisher nicht gekannten Komplexität in internationalen Kooperationen benötigen und gemeinsam nutzen werden. So werden Grid-Anwendungen im Bereich der Supercomputer (Vernetzung zu einem „virtuellen" Supercomputer – DEISA Projekt), auf dem Gebiet der Klimaforschung (C3-Grid-Projekt), im Lifesciene-Bereich und im Bereich des wissenschaftlichen Informationsmanagements (Digital Library, Data-Mining Anwendungen) entstehen, die neue, netzgestützte Formen der wissenschaftlichen Zusammenarbeit zum Erreichen ihren Ziele benötigen. Die europäischen Wissenschaftsnetze, sind für die neuen Anforderungen gut gerüstet und können durch neue Netzdienste zur Bewältigung der „grand challenges" in der Wissenschaft beitragen.

Literatur

D-Grid Initiative: e-Science in Deutschland, F&E Rahmenprogramm 2005 – 2009, 6.Juni 2004.

Foster I., Kesselmann C, Tuecke S.: The Anatomy of the Grid: Enabling Scalable Virtual Organisations, Intl. J. Supercomputer Applications, 2001.

Foster I., Kesselmann C.: The Grid: Blueprint for a New Computing Infrastructure, San Francisco, Morgan Kaufmann 1999.

Web Services im CampusGrid

Olaf Schneider, Karlsruhe

Abstract

Virtualisierung heterogener Ressourcen ist eine der zentralen Ideen des Grid-Computing. Im Rahmen des Projekts CampusGrid soll diese Vision im Forschungszentrum Karlsruhe in die Praxis umgesetzt werden, um die Bedürfnisse der wissenschaftlichen Anwender erfüllen zu können. Diese Anwender kommen aus allen Bereichen der am Forschungszentrum etablierten Wissenschaften, zum Beispiel aus der Festkörperphysik, Nanotechnologie, Biophysik, Meteorologie und Bioinformatik.
Web Services werden zurzeit von vielen Fachleuten als neue lingua franca des verteilten Rechnens in organisationsübergreifenden Grids favorisiert. Mit dem Web Service Resource Framework (WSRF) existiert ein allgemein anerkannter Rahmen zur Implementierung von langlebigen Statuseigenschaften, Unterstützung von multiplen Instanzen und anderen Anforderungen komplexer netzwerkzentrierter Applikationen. Auf dieser Basis werden im CampusGrid prototypisch Schnittstellen für verschiedene Anwendungen und Dienste implementiert. Wir beschreiben den Einsatz von Globus Toolkit 4 und Java zu diesem Zweck und berichten über die gewonnenen Erfahrungen.

1. Einleitung

Seit vernetzte Computer in zunehmendem Maße in Unternehmen und Institutionen die klassischen Großrechnersysteme ersetzen, wird das Konzept des verteilten Rechnens (distributed computing) propagiert und ist mittlerweile auch im praktischen Einsatz vielfach erprobt. Grundlegende Prinzipien sind die Client/Server-Architektur, der Remote Procedure Call (RPC) und eine gemeinsame Datenhaltung. Die Verbreitung von Clustern aus Commodity-Komponenten in der Domäne der klassischen Supercomputer, dem Hochleistungsrechnen[1], erweiterte den Nutzerkreis und die Akzeptanz des verteilten Rechnens und brachte neue Impulse. In Anlehnung an das verteilte Rechnen im Local Area Network (LAN) kann Grid Computing als natürliche Ausdehnung dieses Konzepts auf Wide Area Networks (WAN) aufgefasst werden. Eine frühe Implementierung dieser Idee war die erste Version der Globus-Software (Foster, I., Kesselmann, C., 1997). Der Begriff selbst wurde in Anlehnung an das elektrische Stromnetz (electrical power grid) geprägt (Foster, I., Kesselmann, C., 1998): Rechenleistung wird durch den Nutzer verbraucht wie elektrischer Strom, d.h. die entsprechenden Geräte sind an ein Netz angeschlossen und bedienen sich

[1] High Performance Computing (HPC)

aus diesem im nötigen Maße, während der Vorgang für den menschlichen Nutzer völlig transparent bleibt. Daneben ist der Terminus Metacomputing in der Literatur geläufig. Im Laufe der letzten Jahre wurden die Konzepte verfeinert und durch alternative Visionen ergänzt. So bildet die Virtuelle Organisation (VO) die Gemeinschaft der Nutzer eines Grids und deren soziale Strukturen ab (Foster, I. et al., 2001). Die anhaltende Erfolgsgeschichte des World Wide Web (WWW) hat zur analogen Begriffsbildung „World Wide Grid“ animiert. Diese Bezeichnung bekommt durch die Adoption der Web Services durch die Grid Community und die sich abzeichnende Konvergenz zu einem so genannten Semantic Web zusätzliche Berechtigung. Als grundlegend gilt heute OGSA, die Open Grid Service Architecture (Foster, I. et al., 2002). Damit eng verbunden ist das Konzept der Service Oriented Architecture (SOA), das auch den klassischen Web Services zugrunde liegt. Klassische Web Services lassen einige Eigenschaften vermissen, die für Grid Services wünschenswert oder notwendig sind. Die ursprünglich vorgeschlagene Ablösung des Web Service Standards durch OGSI[2] mündete schließlich in der kompatiblen Erweiterung WSRF[3] (Foster, I. et al., 2004).

Hochleistungsrechnen im Forschungszentrum Karlsruhe startete 1983 mit den gemeinsamen Aktivitäten der Universität Karlsruhe und dem Forschungszentrum. Mit Standort an der Universität Karlsruhe wurde eine Cyber 205 beschafft und gemeinsam genutzt. Erschwerend für die Akzeptanz war die völlig eigenständige Nutzung der Cyber an der Universität und der IBM Rechnerlandschaft im Forschungszentrum, d.h. Dateiverwaltung, Benutzerverwaltung, Jobsprache und vieles mehr waren voneinander unabhängig und in höchstem Maße verschieden. Der Nutzer musste sich jeweils von neuem um seine maschinenspezifische Umgebung kümmern. 1997 änderte sich mit der Beschaffung eines Vektorrechners VP50 für das Forschungszentrum dieser Zustand. Das Rechenzentrum ermöglichte den Benutzern von dem zentralen Rechner des Forschungszentrums aus (IBM 3090 unter MVS) die einfache Nutzung der VP50 und später der VP400EX durch einen so genannten Präprozessor. Jedoch wurde diese Lösung von den Benutzern nur bedingt angenommen, so dass die Nachfolgeinstallationen (Cray J916, VPP300, VPP5000) völlig unabhängig von den IBM (MVS, AIX) und später auch Linux Rechnern betrieben wurden. Die Benutzer mussten ihre Daten mehrfach halten, wenn sie auf mehr als einer Rechnerplattform arbeiten wollten. Andererseits war seit 1996 mit der gemeinsam mit der Universität Karlsruhe betriebenen DCE/DFS-Zelle[4] bereits ein wichtiger Schritt in Richtung verteiltes Rechnen unternommen worden. Der Betrieb von DCE/DFS wurde Anfang 2004 zugunsten eines neu entwickelten Konzepts “Globale Datenhaltung und Benutzerverwaltung” (Schmitz, F., 2005) abgekündigt.

[2] Open Grid Service Infrastructure

[3] Web Service Resource Framework

[4] Distributed Computing Environment / Distributed File System

Im Jahr 2004 wurde das FuE-Vorhaben CampusGrid gestartet mit dem Ziel, ein sich über unterschiedliche Architekturen und Betriebssysteme spannendes Rechen- und Datennetz zu entwerfen und aufzubauen. Dabei sollen die in der derzeitigen Produktionsumgebung gegebenen Randbedingungen bezüglich gemeinsamer Datenhaltung und globaler Benutzerverwaltung sowie Kostenabrechnung berücksichtigt werden, um später einen einfachen Übergang vom Prototyp zum zentrumsweiten Produktionsbetrieb zu ermöglichen. Viele Grid-Projekte stellen den Aspekt der (global) verteilten Ressourcen in den Vordergrund, beschränken sich aber andererseits auf eine eng umgrenzte Palette von Rechner-Architekturen und Betriebssystemen. Dagegen liegt im Projekt CampusGrid der Schwerpunkt auf der heterogenen Rechnerlandschaft und ihrer Anwender, vor allem aus dem Bereich des Hochleistungsrechnens, am Forschungszentrum. Um den Nutzern eine effiziente und problemangepasste Verwendung der heterogenen Ressourcen zu gestatten, ist eine Vermittlungsschicht auf der Basis von Grid-Technologien zu implementieren. Diese „Middleware“ soll nicht von Grund auf neu entwickelt werden, sondern setzt sich aus einer Vielzahl existierender oder in Entstehung befindlicher Grid-Software-Komponenten zusammen, die im Rahmen des Projekts evaluiert sowie für unsere spezifischen Belange adaptiert und verbessert werden. Dabei werden bestehende und in Entwicklung befindliche Grid-Standards sowie "Best Practices" berücksichtigt, um Interoperabilität mit dem künftigen World Wide Grid sicherzustellen. Deshalb sollten alle CampusGrid-Ressourcen als WSRF-konforme Web Services ansprechbar sein.
In Abschnitt 2 erläutern wir das generelle Konzept von CampusGrid und stellen den Stand der Arbeiten kurz vor. In einem eigenen Abschnitt werden die Untersuchungen zu -Middleware und Web Services dargestellt (siehe Abschnitt 3). Schließlich geben wir in Abschnitt 4 eine Einschätzung dieser Erfahrungen und weisen auf offene Fragen hin.

2. CampusGrid-Architektur und Projekt-Status

Wesentliche Voraussetzungen zur Harmonisierung einer heterogenen Rechnerumgebung sind eine einheitliche Benutzerverwaltung, ein leistungsfähiges gemeinsames Filesystem und eine Ressourcen-Verwaltung zum Starten und Überwachen von Jobs, wobei ein intelligenter Broker den geeigneten Compute-Server in Abhängigkeit von Hauptspeicher- und Plattenspeicherbedarf, unterstützter bzw. bevorzugter CPU-Architektur (Vektor, Skalar, MPP, SMP) sowie Antwortzeit bzw. Preisvorstellung automatisch auswählt.
Somit besteht das System aus folgenden Hauptkomponenten:

- Fabric (Hardware)
- Resource Management (Broker, Information System)
- Shared Filesystem, Data Services

- Identity Management, Authentication & Authorization
- User Interface (Web-GUI, CLI, API)

Wichtig ist die Austauschbarkeit der einzelnen Komponenten. Daraus resultiert die Notwendigkeit standardisierter Schnittstellen. Für das Filesystem stellt etwa POSIX einen solchen Standard dar. Für andere Komponenten wie den Resource Broker sind Web Services das bevorzugte Paradigma (Abbildung 1).

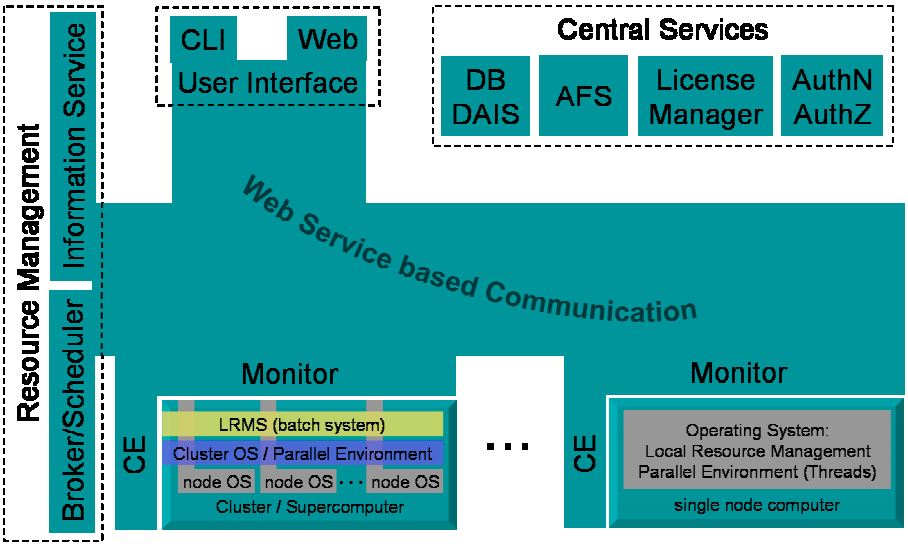


Abbildung 1: Entwurf der CampusGrid-Architektur

Im Rahmen des Projekts wurde die Evaluierung verschiedener Filesysteme vorangetrieben. Ziel war es, derzeit oder in naher Zukunft am Markt verfügbare globale plattformübergreifende Filesysteme auf Eignung als leistungsfähige Lösungen für das CampusGrid zu überprüfen. Erste Ergebnisse liegen in Form von qualitativen Bewertungen sowie Performance-Messungen vor (Brandel, T., Schmitz, F., 2005). Die in Frage kommenden Software-Produkte sind typischerweise ausgelegt für ein SAN[5] und verfügen über ein Metadatenmanagement auf Blockebene mit dedizierten Metadaten-Servern. Neben FibreChannel als Transportschicht wurden auch InfiniBand und iSCSI untersucht. Insbesondere die Möglichkeit, in einem PC-Cluster mit InfiniBand-Interconnect die I/O-Anbindung der Rechen-Knoten ohne zusätzliche physikalische Leitungen sicherzustellen, hat sich als viel versprechend erwiesen (Schwickerath, U., Heiss, A., 2004). Weitere Bewertungskriterien sind:

[5] Storage Area Network

- Verfügbarkeit von Clients für alle Zielplattformen (Windows und verschiedene Unix-Varianten),
- Skalierbarkeit,
- Zuverlässigkeit,
- Geschwindigkeit und
- Management und Handhabung der Komponenten.

Neben einem schnellen Filesystem im SAN kommt OpenAFS als globales, vom geographischen und netzwerktopologischen Standort des Nutzers unabhängiges Filesystem zum Einsatz. Seine besondere Stärke liegt in der sicheren Kommunikation (auch über unsichere Netzwerkverbindungen) durch die nahtlose Integration mit Kerberos.

In Anknüpfung an die derzeit im Forschungszentrum Karlsruhe betriebene „Globale Benutzerverwaltung“ (GBV) wurde ein Konzept für eine verbesserte Anbindung von UNIX-Servern an den verwendeten Verzeichnisdienst Microsoft Active Directory Server (ADS) entwickelt. Dabei erfolgt eine unmittelbare Einbindung aller Ressourcen (ob UNIX oder Windows) in die Kerberos-Domäne des ADS. Clients und Server authentifizieren sich gegenseitig durch Tickets entsprechend dem Kerberos5-Protokoll. Beim Absenden von Batch-Jobs übernimmt eine eigens angepasste Version der Software PSR[6] die Ausstellung der Tickets (Obholz, L., Mayer U., 2005). Für die Nutzer ist damit echtes Single-Sign-On möglich.

In einem Grid kommt normalerweise eine zusätzliche Software-Schicht zum Einsatz, um die Rechnerressourcen zu virtualisieren. Im – meist relativ engen – Verbund eines Rechenzentrums bietet sich daneben die Möglichkeit, die Integration der heterogenen Plattformen auf der Ebene der vorhandenen Batch Queuing und Job Scheduling Systeme vorzunehmen. In der Literatur wird in diesem Zusammenhang teilweise von Distributed Resource Management System (DRMS) im Gegensatz zum auf eine einzige Hardware beschränkten Local Resource Management System (LRMS) gesprochen (Rajic, H. et al., 2003). Dieser Ansatz wurde für das Projekt CampusGrid durch einen LoadLeveler Mixed Cluster verwirklicht (Obholz, L., Mayer, U., 2005). Darin sind Opteron-Commodity-Systeme, Blades und IBM pSeries Server miteinander so verbunden, dass ein einzelner Job im Verlauf seiner Abarbeitung verschiedene Hardware-Plattformen und Betriebssysteme (AIX, verschiedene Linux-Varianten) nutzen kann.

3. Erfahrungen mit Globus Toolkit 4

Mit Globus Toolkit 4 (kurz GT4) ist im April 2005 die neueste Generation der der Globus Alliance[7] erschienen. Wie schon die Vorgängerversionen besteht die

[6] Password Storage and Retrieval System, http://www.lam-mpi.org/software/psr/

[7] http://www.globus.org/

Software aus einer Reihe von Komponenten, die vorgeschlagene Standards etwa des GGF[8] implementieren oder eigene Quasi-Standards für bestimmte Funktionalitäten setzen. GT4 umfasst alle Fähigkeiten früherer Versionen, insbesondere die grundlegenden seit Globus Toolkit 2 (GT2) bekannten Dienste für Sicherheit, Daten-Transfer, Monitoring und Job-Verwaltung (GSI, GridFTP, MDS und GRAM). Das Spektrum wurde beginnend mit Version 3 vor allem um Komponenten erweitert, die eine der Open Grid Service Architecture (OGSA) entsprechende Infrastruktur bieten. Diese „high level" Grid-Services (wie beispielsweise WS-GRAM, MDS4 und RFT) basieren in GT4 vollständig auf der mitgelieferten WSRF-Implementierung, welche in so genannten Web Service Containern gebündelt wird. Genauer gesagt laufen diese Services in einem Java Container, der neben GT4-spezifischen Teilen auch „3rd party"-Software (vor allem aus dem OpenSource-Projekt Apache, zum Beispiel die SOAP Engine Apache Axis[9]) enthält. Hinzu kommen Tools für Entwicklung und Deployment von Web Services. Somit bietet GT4 eine vollständige (wenn auch etwas spartanische) Entwicklungsumgebung für WSRF-konforme Web Services.
Um die GT4-Services im CampusGrid anbieten zu können, wurde ein dedizierter Rechner (FSC-Celsius, Opteron-CPU, 8GB Hauptspeicher, SuSE-Linux) mit Globus Toolkit in der Version 4.0.0 ausgestattet. Folgende Dienste bzw. Komponenten sind auf diesem Host konfiguriert und in Betrieb:

- Java WS Container
- WS-GRAM (Grid Resource Allocation Manager)
- RFT (Reliable File Transfer)
- GridFTP
- WS Authentication & Authorization
- Delegation Service
- GSI (Grid Security Infrastructure)

Der GT4 Java-Container stellt dabei das Hosting-Environment für die Web Services (WS-GRAM, RFT, WS Authentication & Authorization, Delegation Service) zur Verfügung. Zusätzlich können selbst geschriebene Java-basierte Web Services in den Container eingebunden werden. Die dazu nötigen Schritte sind im „Programmer's Tutorial" (Sotomayor, B., 2005) in sehr gut nachvollziehbarer Form beschrieben.
Die „high level" Grid-Services setzen zum Teil auf den Funktionen der GT2-kompatiblen Dienste auf. So erfordert RTF einen funktionsfähigen GridFTP-Daemon, WS Authentication & Authorization und Delegation Service nutzen die Zertifikate und Tools der Grid Security Infrastructure (GSI). Dagegen handelt es sich bei WS-GRAM und den MDS4-Komponenten um unabhängige Neuimplementierungen.

[8] Global Grid Forum, http://www.ggf.org/
[9] http://ws.apache.org/axis/ bzw. http://ws.apache.org/axis2/

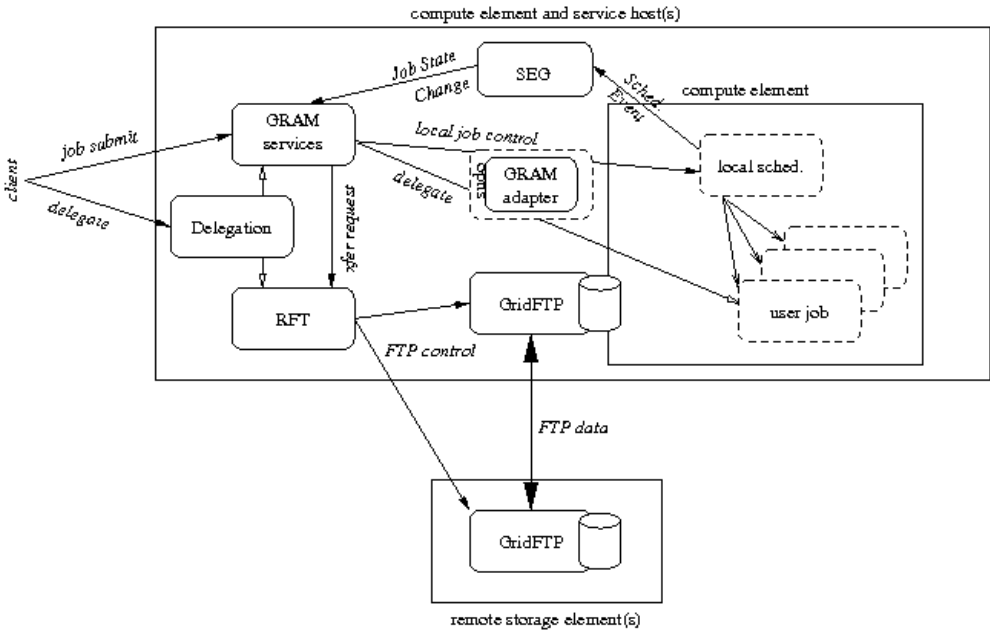


Abbildung 2: Zusammenspiel verschiedener Globus-Komponenten beim Job-Submit (Quelle:http://www-unix.globus.org/toolkit/docs/4.0/execution/key/WS_GRAM_ Approach.html)

Andererseits setzen die GT4 Web Services auch aufeinander auf. So verwendet WS-GRAM für das so genannte File-Staging oder das Streaming der Standard-Ausgabe eines Jobs den Reliable File Transfer. Um diesen zu nutzen ist wiederum eine Weitergabe der Security Credentials mittels Delegation Service erforderlich. Auf diese Weise erfordert bereits das simple Absetzen eines Jobs ein relativ komplexes Zusammenspiel verschiedenen Komponenten (Abbildung 2). Das eigentliche Absetzen des Jobs überlässt WS-GRAM einem so genannten Local Scheduler. In der Default-Konfiguration ist das einfach ein UNIX-Fork. Daneben existieren Scheduler Adapter für gängige Batch Systeme wie PBS, LSF und Condor sowie eine Erweiterungsmöglichkeit um eigene Adapter. In diesem Modell ist der Scheduler Adapter für das Absenden des Jobs zuständig, während die Überwachung des Job Status durch eine weitere Komponente, den Scheduler Event Generator (SEG) erfolgt. Auch diese Komponente stellt eine spezifische Schnittstelle zum lokalen Scheduler dar. Ein einziger GRAM Web Service kann gleichzeitig mehrere verschiedene lokale Scheduler verwalten. Bei der Konstruktion des Jobs durch den „ManagedJobFactoryService“ wird der gewünschte lokale Scheduler im Parameter „FactoryType“ übergeben.

In unserer WS-GRAM-Installation steht als Factory-Type neben „Fork“ auch „PBS“ zur Verfügung. Letzterer greift jedoch nicht auf einen lokal laufenden PBS-Server zu, sondern stellt eine Remote-Anbindung an ein InfiniBand-Cluster mit bis zu 32 Opteron-Knoten zur Verfügung. Der entsprechende PBS-Server läuft auf dem Head-Node des Clusters (Abbildung 3). Für den SEG ist ein Filesharing des PBS-Log-Verzeichnisses erforderlich (derzeit über einen Read/Write-NFS-Mount). Um eine sinnvolle Nutzung des Clusters über WS-GRAM einschließlich File-Staging zu ermöglichen, sind außerdem die Nutzer-HOME-Verzeichnisse via NFS angebunden. Künftig soll dafür das gemeinsame SAN-Filesystem zum Einsatz kommen. Geplant ist ebenfalls die Einbindung des bestehenden LoadLeveler Mixed Clusters als zusätzlicher Factory-Type.

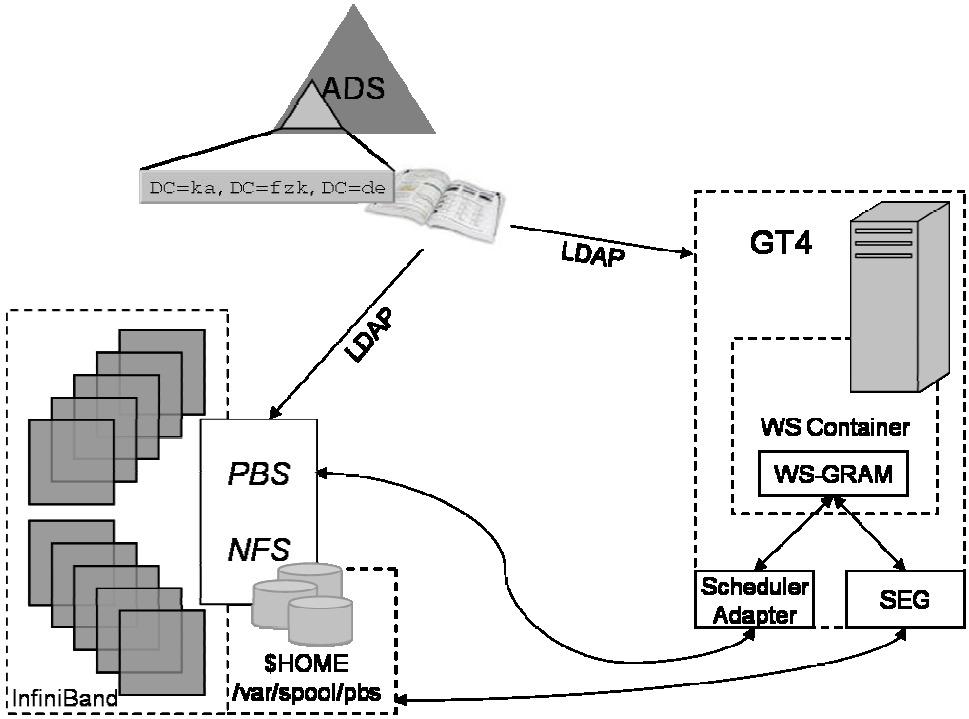


Abbildung 3: Anbindung des InfiniBand-Clusters an den GT4-Host

Die Authentifizierung der Benutzer basiert – entsprechend der Default-Konfiguration von GT4 – auf den X.509-Zertifikaten der Grid Security Infrastructure (GSI). Als Certificate Authority (CA) wird derzeit nur die Zertifizierungsstelle GridKa-CA (Epting, U., 2005) anerkannt, welche am Forschungszentrum Karlsruhe betrieben wird (GermanGrid). Der Server besitzt ein entsprechendes Host-Zertifikat der GridKa-CA. Da derzeit nur lokale Nutzer des Forschungszentrums vorgesehen sind, erfolgt im

grid-mapfile eine Abbildung der Zertifikat-Subjects auf nutzerspezifische UNIX-Accounts. Die nötigen Identity-Informationen können via LDAP aus dem ADS repliziert werden (gleiche Technik wie bei GBV). Das Einpflegen der Nutzer-Zertifikate bzw. ihrer Subjects muss derzeit noch manuell erfolgen. Gleiches gilt für die Eintragung des Nutzers in der Sudo-Konfiguration, die für WS-GRAM erforderlich ist. Lediglich die Aktualisierung der Certification Revocation Lists (CRL) erfolgt automatisiert via cron (genauer gesagt mit einer angepaßten Version des Fetch-CRL-Scripts aus dem EDG-Projekt[10]).
Mit der GT4-Distribution wird ein Kommandozeilen-Client für WS-GRAM ausgeliefert, der in C implementiert ist. Neben der C-API gibt es auch solche für Java und Python. Ein graphischer Client existiert derzeit nicht. Für die Nutzung von Grid-Ressourcen innerhalb von Anwendungsprogrammen sind die APIs von besonderem Interesse, weshalb es eine der Aufgaben im Projekt ist, sich mit diesen Schnittstellen vertraut zu machen.

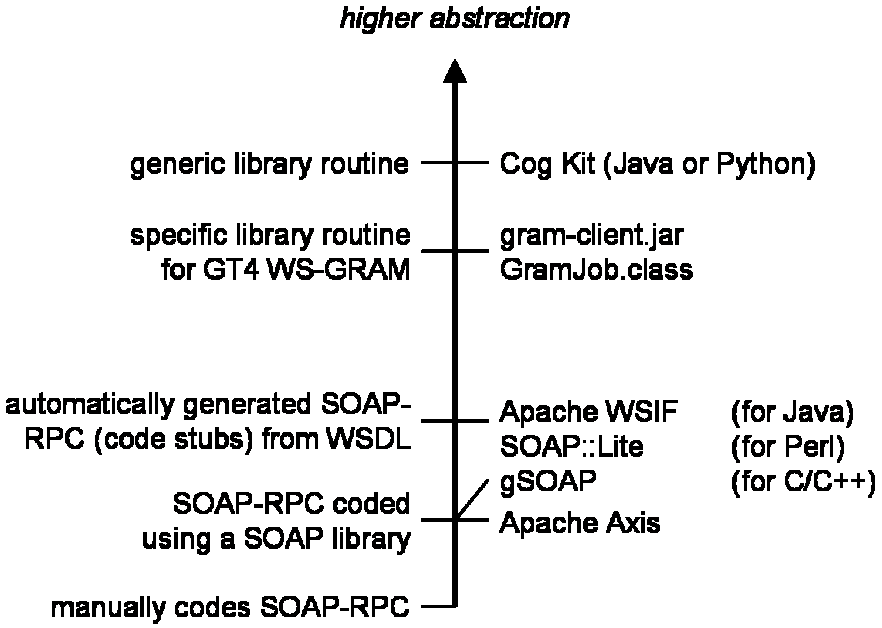


Abbildung 4: verschiedene Abstraktionsebenen (links) und Software-Tools (rechts) zur Implementierung eines WS-GRAM-Clients

Für die Implementierung eines WS-GRAM-Clients gibt es mehrere Möglichkeiten. Sie unterscheiden sich im Wesentlichen durch die zum Zugriff auf den Service verwendeten Software-Tools und damit letztlich durch die gewählte Abstraktions-ebene (Abbildung 4). Auf der untersten Ebene steht der direkt programmierte

[10] European Data Grid (EDG), http://eu-datagrid.web.cern.ch/eu-datagrid/

Remote Procedure Call (RPC) entsprechend der im WSDL-Dokument definierten Schnittstelle. Dabei können (und sollten) Tools verwendet werden, die dem Programmierer das Parsen von XML sowie das Erstellen und Analysieren der SOAP-Messages erleichtern. Auf der nächsthöheren Ebene erfolgt die Generierung des RPC-Codes automatisch aus dem WSDL-Dokument, beispielsweise mittels Apache WSIF[11] (Duftler, M. J. et al., 2001) oder dem Perl-Modul SOAP::Lite[12]. Noch bequemer ist der direkte Zugriff auf geeignete Bibliotheks-Funktionen in der gewünschten Programmiersprache über ein WS-GRAM-API. So enthält die GT4-Distribution eine Java-Klasse GramJob und als Beispiel-Implementierung das Kommandozeilentool GlobusRun. Das Java CoG Kit (Laszewski, G. et al, 2001) stellt eine weitere Abstraktionsebene bereit, die einen einheitlichen Zugriff auf verschiedene Globus-Versionen und andere Grid-Umgebungen (z.B. UNICORE[13]) erlaubt. Der Implementierungs-Aufwand (für einen Einsteiger) sinkt nicht unbedingt bei Verwendung einer höheren Abstraktionsebene, da die entsprechenden Konzepte eine relativ steile Lernkurve mit sich bringen.

Als Anwendungsbeispiel für die Nutzung der WS-GRAM-Schnittstelle im CampusGrid soll die Software COSMOS (Sternberg, U. et al., 2003) dienen. COSMOS ist ursprünglich als Windows-Programm entstanden, mit dem auf einfache Weise Molekülmodelle betrachtet, modelliert und optimiert, sowie Berechnungen von Moleküleigenschaften durchführt werden können. Um für umfangreichere Berechnungen auch Hochleistungsrechner nutzen zu können, ist ein rein kommandozeilenorientiertes Backend-Programm entwickelt und auf diverse UNIX-Plattformen portiert worden. Dessen Nutzung gestaltet sich häufig wie folgt: Eingabedateien (Atom-Koordinaten, Optionen) werden unter Windows erstellt und gespeichert, auf den UNIX-Server transferiert und in einem Batch-Job dem Backend zum Verarbeiten übergeben. Anschließend werden die Ausgabedateien zurück in die Windows-Welt übertragen, um sie im Frontend betrachten und weiterverarbeiten zu können. Die Nachteile dieses Vorgehens sind offensichtlich: Der Benutzer ist gezwungen seine gewohnte Umgebung zu verlassen, wenn er die HPC-Ressourcen nutzen will. Der gesamte Vorgang ist umständlich und wenig benutzerfreundlich. Das gilt in besonderem Maße, wenn eine quasi-interaktive Benutzung wünschenswert ist (so genanntes „computational steering"). Ein typisches Beispiel ist eine längere Zeit (Tage) dauernde Molekül-Dynamik-Simulation, die regelmäßig den aktuellen Zustand des Moleküls als Atom-Koordinaten-File abspeichert. Anhand der Visualisierung eines solchen Zwischenstandes im Frontend kann der Wissenschaftler entscheiden, ob eine Fortsetzung des Simulationslaufes sinnvoll ist oder eine neue Simulation mit geänderten Parametern gestartet werden muss. Eigentlich ist also eine regelmäßige Inspektion der Zwischenstände erforderlich, wofür jeweils ein Datei-Transfer von

[11] Web Service Invocation Framework, http://ws.apache.org/wsif/

[12] http://soaplite.com/

[13] http://unicore.sourceforge.net/

Unix nach Windows nötig ist. Ein erster Ansatz zur Verbesserung der Situation ist das Starten des Batch-Jobs mit dem Backend direkt aus dem Fronend heraus, das auf dem Desktop-PC des Wissenschaftlers läuft. Die Kommunikation erfolgt WS-basiert, im einfachsten Fall durch direkten Aufruf von WS-GRAM. Die passenden Kommandozeileparameter für das Backend müssen in diesem Fall durch Programmlogik im Frontend bereitgestellt werden. Noch eleganter wäre es, die Funktionen des Backend-Programms als Web Service (oder eine Sammlung von Web Services) verfügbar zu machen, die dann vom Frontend als Client aufgerufen werden können.

4. Diskussion und Ausblick

Der Java WS Container hat sich auch über längere Laufzeiten hinweg als stabil erwiesen. Weniger überzeugt haben dagegen die mitgelieferten Scheduler Adapter für das LRMS. Die Implementierung der Befehle zur Jobverwaltung, bei der ein (automatisch generiertes) Perl-Script auf die Kommandozeilentools von PBS zugreift, sorgt für häufig unverständliche und wenig hilfreiche Fehlermeldungen auf der Java-Ebene. Abhilfe könnte ein direkter API-Zugriff auf das LRMS schaffen, soweit entsprechende Schnittstellen vorhanden sind. Es ist zu prüfen, ob hier im Rahmen des Projekts ein eigener Entwicklungs-Beitrag geleistet werden soll. Dabei sollte der Read/Write-Zugriff auf das PBS-Log-Verzeichnis vermieden werden. Zudem wäre so eine Anbindung von LoadLeveler (oder gar NQS auf einem Vektorrechner) relativ kurzfristig möglich.

Zur besseren Beurteilung von Stärken und Schwächen der Globus-Lösung ist ein Vergleich mit anderen WSRF-Implementierungen erforderlich. In Frage kommen etwa WSRF::Lite[14] (McKeown, M., 2004) und WSRF.Net[15] (Humphrey, M., Wasson, G., 2005). Insbesondere bei der Erstellung anwendungsspezifischer Web Services werden positive Erfahrungen mit dem Perl-Modul WSRF::Lite bzw. dem Vorgänger OGSI::Lite berichtet (Chin, J.; Coveney, P. V., 2004).

Monitoring und Resource Discovery sind wichtige Funktionen (crucial challenges) im Grid. Eine Untersuchung der entsprechenden GT4-Komponenten (WebMDS, Trigger, Index) ist deshalb dringend erforderlich.

Die Grid Security Infrastructure der GT4-Default-Installation ist in gewisser Weise orthogonal zur Kerberos-basierten Authentifizierung und Autorisierung im CampusGrid. Ziel ist deshalb eine Integration der GT4-Security-Schnittstellen (GSI, WS Authentication & Authorization) mit Kerberos und ADS, zum Beispiel mittels KX.509 und KCA[16] (Doster, W. et al., 2001, Dussa, T., 2003) oder PKINIT[17].

[14] http://www.sve.man.ac.uk/Research/AtoZ/ILCT

[15] http://www.cs.virginia.edu/~gsw2c/wsrf.net.html

[16] http://www.globus.org/grid_software/security/kx509-and-kca.php

[17] http://www.globus.org/grid_software/security/pkinit.php

Aktuelle Informationen zum Projekt CampusGrid sind über unsere Web-Seiten[18] verfügbar.

Literatur

Brandel, T., Schmitz, F.: *SAN Testbed and StorNext File System*, AIX-Arbeitskreis im Forschungszentrum Karlsruhe, http://hikwww2.fzk.de/hik/orga/hlr/aix-ak/vortrag/snfs.pdf, 2005.

Chin, J.; Coveney, P. V.: *Towards tractable toolkits for the Grid: a plea for leightweight, usable middleware*, http://www.nesc.ac.uk/technical_papers/UKeS-2004-01.pdf, 2004.

Doster, W. et al.: *The KX.509 Protocol*, Technical Report, Center for Information Technology Integration, http://www.citi.umich.edu/techreports/reports/citi-tr-01-2.pdf, 2001.

Duftler, M. J. et al.: *Web Services Invocation Framework (WSIF)*, IBM T.J. Watson Research Center, http://www.research.ibm.com/people/b/bth/OOWS2001/duftler.pdf, 2001.

Dussa, T.: *Kerberos-Based Authentication in Grid Computing*, Diplomarbeit, Universität Karlsruhe (TH), Fakultät für Informatik, 2003.

Epting, U.: *GridKa-CA Certificate Policy and Certification Practice Statement*, Version 1.2, http://grid.fzk.de/ca/gridka-cps.pdf, Forschungszentrum Karlsruhe, 2005.

Foster, I. and Kesselmann, C.: *Globus: A Metacomputing Infrastructure Toolkit*, Intl. J. Supercomputer Applications, 11(2), pp. 115-128, 1997.

Foster, I. and Kesselmann, C. (Eds): *The Grid: Blueprint of a Future Computing Infrastructure*, Morgan Kaufman Publishers Inc., San Francisco, 1998.

Foster, I. et al.: *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*, Intl. J. Supercomputer Applications, 15(3), pp. 200-222, 2001.

Foster, I. et al.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Open Grid Service Infrastructure WG, Global Grid Forum, 2002.

Foster, I. et al.: *From Open Grid Services Infrastructure to Web Services Resource Framework: Refactoring and Evolution*, http://www-128.ibm.com/developerworks/ library/ws-resource/ogsi_to_wsrf_1.0.pdf, 2004.

Foster, I. et al.: *Modeling Stateful Resources using Web Services*, http://www-106.ibm.com/developerworks/library/ws-resource/ws-modelingresources.pdf, 2004.

Humphrey, M., Wasson, G.: *Architectural Foundations of WSRF.NET*, International Journal of Web Services Research, 2(2), pp. 83-97, 2005.

[18] http://www.campusgrid.de/

Laszewski, G. et al.: *A Java Commodity Grid Kit*, Concurrency and Computation: Practice and Experience, vol. 13, no. 8-9, pp. 643-662, http:/www.cogkit.org/, 2001.

McKeown, M.: *OGSI::Lite and WSRF::Lite - Grid Services in Perl*, Second Annual RealityGrid Workshop, http://vermont.mvc.mcc.ac.uk/RealityGridWorkshop2.pdf, 2004.

Obholz, L., Mayer, U.: *LoadLeveler for mixed cluster*, German LoadLeveler User Group (GLLUG) Meeting, AIX-Arbeitskreis im Forschungszentrum Karlsruhe, http://hikwww2.fzk.de/hik/orga/hlr/aix-ak/vortrag/LoadlMixedAFS.pdf, 2005.

Rajic, H. et al.: *Distributed Resource Management Application API Specification 1.0*, Global Grid Forum, http://www.drmaa.org/, 2003.

Schmitz, F.: *Benutzerverwaltung und Datenhaltung einmal anders: LDAP, Windows, VPP5000, Linux-Cluster, Power3+4*, AIX-Arbeitskreis im Forschungszentrum Karlsruhe, http://hikwww2.fzk.de/hik/orga/hlr/aix-ak/vortrag/GBV_und_GDH_im_FZK.pdf, 2005.

Schwickerath, U., Heiss, A.: *First experiences with InfiniBand interconnect*, Nuclear Instruments and Methods in Physics Research, A 534, pp. 130-134, 2004.

Sotomayor, B.: *The Globus Toolkit 4 Programmer's Tutorial*, Globus Documentation Project, http://gdp.globus.org/gt4-tutorial/, 2005.

Sternberg, U. et al.: *COSMOS-Software*, http://www.cosmos-software.de/, 2003.

From UNICORE to UniGrids

Achim Streit, Jülich

Abstract

Die UNICORE Grid Software ermöglicht den stoßkantenfreien, sicheren und intuitiven Zugang zu verteilten Grid Ressourcen. Mit der Anpassung an zukünftige Standards aus der Grid und Web Services Welt wird UNICORE mit anderen Grid Middleware Systemen interoperabel sein. Zu Beginn wurde UNICORE in zwei vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Forschungsprojekten entwickelt. In der Folgezeit entwickelte sich die Software im Rahmen zahlreicher europäischer Forschungsprojekte zu einem ausgewachsenen und wohlgetesteten Grid Middleware System, welches heutzutage im täglichen Produktionsbetrieb an vielen Supercomputer Zentren weltweit eingesetzt wird. Darüber hinaus ist UNICORE eine Basis für zahlreiche europäische und internationale Forschungsprojekte, in denen bereits existierende Komponenten genutzt werden, um weiterführende Funktionalität und höherwertige Services zu implementieren sowie Anwendungen aus den unterschiedlichsten Wissenschaftsbereichen zu unterstützen. Um die fortlaufenden Entwicklungen weiter zu fördern, ist UNICORE als Open Source unter BSD Lizenz auf SourceForge verfügbar. Dort werden auch regelmäßig neue Versionen und Updates der Software hinterlegt.

Einleitung

Ende 1998 wurde das Konzept des "Grid Computing" erstmalig im Buch "The Grid: Blueprint for a New Computing Infrastructure" von I. Foster und C. Kesselman (1) vorgestellt. Bereits anderthalb Jahre zuvor, wurde mit der Entwicklung von UNICORE – Uniformes Interface zu Computer Ressourcen – auf Initiative der deutschen Supercomputer Zentren begonnen, um Nutzer einen stoßkantenfreien, sicheren und intuitive Zugang zu verteilten, heterogenen Ressourcen wie Rechner, Daten und Software zur Verfügung zu stellen. Vergleichbar zum Globus Toolkit ® (2) startete UNICORE noch bevor "Grid Computing" das allseits akzeptierte neue Paradigma im Verteilten Rechnen wurde.

Die UNICORE Idee wurde dem Bundesministerium für Bildung und Forschung (BMBF) präsentiert und anschließend gefördert. Ein erster Prototyp wurde im UNICORE[1] Projekt (3) entwickelt. Die Basis für die aktuelle Produktionsversion wurde im Nachfolgeprojekt UNICORE Plus[2] (4) gelegt, welches im Jahre 2002

[1] BMBF-Förderkennzeichen 01 IR 703, Dauer: August 1997 – Dezember 1999

[2] BMBF-Förderkennzeichen 01 IR 001 A-D, Dauer: Januar 2000 – Dezember 2002

erfolgreich abgeschlossen wurde. Seitdem wird UNICORE in Produktion bei den deutschen Supercomputer Zentren eingesetzt und bildet darüber hinaus eine Basis für zahlreiche europäische Projekte.
Obwohl bereits im ersten UNICORE Projektvorschlag im Jahre 1997 definiert, sind die Ziele und Zielsetzungen der UNICORE Software noch heute gültig:

- Zu vorderst steht das Ziel von UNICORE, die Unterschiede aufgrund verschiedener Hardware-Architekturen, Hersteller-spezifischen Betriebssystemen, inkompatiblen Batch-Systemen, unterschiedlichen Anwendungsumgebungen, historisch gewachsenen Rechenzentrenpraktiken, Namenskonventionen, Dateisystemstrukturen und Sicherheitsregeln – um nur die wichtigsten zu nennen – zu verstecken.
- Gleichzeitig ist Sicherheit ein zentraler Bestandteil des Designs von UNICORE, wobei X.509 Zertifikate für die Authentisierung von Nutzern, Servern und Software sowie für die Verschlüsselung der Kommunikation über das Internet verwendet werden.
- Schlussendlich ist UNICORE für Wissenschaftler und Ingenieure nutzbar, ohne dass Hersteller- oder Rechenzentrums-spezifisches Fachwissen notwendig ist. Eine grafische Benutzerschnittstelle ist verfügbar und hilft Nutzern in der Erstellung und dem Management von Jobs.

Zudem erfüllt UNICORE weitere Grundvoraussetzungen: die Grid Middlware unterstützt alle an den Zentren installierte Betriebssysteme und Batch-Systeme der verschiedenen Hersteller. Im Jahre 1997 waren das große Cray T3E Systeme, NEC und Hitachi Vektorcomputer, IBM SP2s und kleinere Linux Cluster. Heutzutage ist das unterstützte Spektrum deutlich größer und unterstützt neuartige Hardware wie etwa IBM p690 Systeme. Dabei soll die entwickelte Software so wenig wie möglich Veränderungen in der Hardware- und Software-Infrastruktur hervorrufen. Die Wahrung der Autonomie des Zentrums ist ein wichtiger Bestandteil, wenn Fragestellungen der Akzeptanz und Einsetzbarkeit vom Blickwinkel der Systemadministratoren adressiert werden. Zusätzlich unterstützt UNICOREs Sicherheitsmodell Rechenzentrums-spezifische Sicherheitsmaßnahmen, wie etwa Firewalls. Nach dem Ende der initialen Förderung des UNICORE Plus Projektes existierte ein funktionsfähiger Prototyp, der zeigte, dass das anfängliche Konzept funktioniert. Durch die Integration innovativer Ideen und bewährter Komponenten über die Jahre, entwickelte sich der erste Prototyp hin zu einer vertikal integrierten Grid Middleware Lösung.

Architektur von UNICORE

In Abbildung 1 ist die Architektur von UNICORE mit den drei Ebenen Benutzer, Server und Zielsystem dargestellt (5). Die Implementierung aller Komponenten ist in JAVA durchgeführt. Eine detaillierte Analyse von UNICORE (6,7) zeigt, dass die

Architektur dem Open Grid Services Architecture (OGSA) (8) Konzept und dem Paradigma "Everything being a Service" folgt.

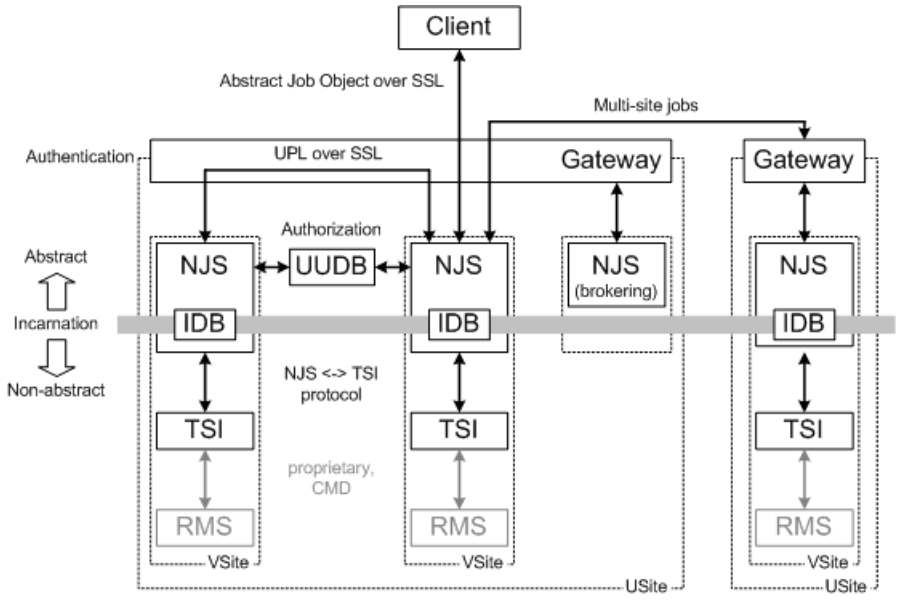


Abbildung 1: Architektur von UNICORE

Benutzer Ebene

Der UNICORE Client stellt eine grafische Benutzerschnittstelle zur Verfügung, die es ermöglicht, den kompletten Umfang an unterliegenden Services anzusprechen. Der Client kommuniziert dabei mit der Server Ebene durch das Senden und Empfangen von Abstract Job Objects (AJO) und Dateidaten über den UNICORE Protocol Layer (UPL), der auf dem SSL Protokoll aufbaut. Das AJO ist die Realisierung von UNICORE's Job-Modell und ist ein zentraler Bestandteil der Philosophie von Abstraktion und Stoßkantenfreiheit. Es enthält System- und Rechenzentrums-unabhängige Beschreibungen der einzelnen Berechnungs- und Daten-Tasks eines Jobs sowie Ressourcen-Informationen und Workflow-Spezifikationen zusammen mit Benutzer und Sicherheitsinformationen. AJOs werden zum UNICORE Gateway in Form von serialisierten und signierten Java Objekten geschickt, wobei ggf. Daten als Bytestream mitgeschickt werden.

Der UNICORE Client unterstützt den Benutzer bei der Erstellung komplexer, untereinander abhängiger Jobs, die ohne weitere Modifikationen an jeder beliebigen UNICORE Site (USite) ausgeführt werden können. Ein UNICORE Job, bzw. genauer eine UNICORE Job-Gruppe, kann rekursiv andere Job-Gruppen oder Tasks sowie

Abhängigkeiten zwischen diesen enthalten, so dass ein Workflow entsteht. Neben der Beschreibung eines Jobs als eine Menge von gerichteten azyklischen Graphen, sind auch konditionale und wiederholende Ausführungen von Job-Gruppen oder Tasks möglich. Zur Überwachung von Jobs ist deren Status in jeder Ebene der Rekursion bis hinunter zum einzelnen Task verfügbar. Ebenso sind detaillierte Informationen zur genauen Analyse potentieller Fehler abrufbar. Am Ende einer Jobausführung ist es möglich, die stdout und stderr Ausgaben der Jobs abzuholen, wobei Daten-Management Funktionen wie Import, Export und Transfer als explizite Tasks im Client zur Verfügung stehen. Dies ermöglicht dem Benutzer auch den Datenfluss, z.B. als Teil eines Workflows von einem zum anderen Zielsystem, von oder zum lokalen Rechner, vor oder nach der Jobausführung oder zur permanenten Datenspeicherung in einem Archiv zu definieren.
Die beschriebenen Funktionen stellen bereits ein effektives Werkzeug zur Nutzung von Ressourcen auf unterschiedlichen Rechner sowohl für Capacity und auch Capability Computing dar, jedoch benutzen viele Wissenschaftler und Ingenieure Programmpakete. Für Anwendungen ohne grafische Nutzerschnittstelle können mit Hilfe eines Toolkits spezifische UNICORE Plugins entwickelt werden. Mit der Zeit sind in den verschiedenen Projekten zahlreiche Plugins entwickelt worden, so dass für nahezu jede wissenschaftliche Anwendung fertig Plugins bereits existieren, z.B. für CPMD (Car-Parinelle Molecular Dynamics (9)), Fluent oder MSC Nastran.

Server Ebene
Die Server Ebene enthält den Gateway und den Network Job Supervisor (NJS). Der Gateway überwacht den Zugang zu einer Usite und akzeptiert sowie authentisiert eingehende UPL Anfragen. Eine Usite repräsentiert eine Site (z.B. ein Rechenzentrum) durch einen symbolischen Namen, der in die URL des Gateways aufzulösen ist. Eine Institution kann Teil mehrer Grids sein und dabei die gleichen oder unterschiedliche Ressourcen anbieten. Der Gateway leitet eingehende Requests zum unterliegenden NJS einer virtuellen Site (Vsite) zur weiteren Verarbeitung. Der NJS repräsentiert dabei Ressourcen mit einem einheitlichen Benutzerschema und ohne zusätzliche Sicherheitsmaßnamen (wie z.B. Firewalls).
Eine Vsite repräsentiert dabei eine bestimmte Menge an Ressourcen an einer Usite, die durch ein NJS kontrolliert werden. Eine Vsite kann daher ein einzelner Supercomputer, z.B. ein IBM p690 System mit LoadLeveler oder ein Linux Cluster mit PBSpro als Ressourcen-Management-System sein. Hierbei können mehrere Vsites in einer Usite existieren. Die Flexibilität des Konzeptes unterstützt verschiedene Architekturen und bewahrt dem Eigner die volle Kontrolle über die eigenen Ressourcen.
Der NJS ist verantwortlich für die Virtualisierung der von ihm verwalteten Ressourcen und führt das Mapping des abstrakten Jobs auf das Zielsystem durch. Dieser Prozess wird als Inkarnation bezeichnet und verwendet die Incarnation Database (IDB). In der IDB sind systemspezifische Daten gespeichert, die die Software und

Hardware Infrastruktur des Systems beschreiben. Zusätzlich führt der NJS für Workflows das Pre- und Post-Staging von Daten durch und autorisiert Benutzer mit Hilfe der UNICORE User Database (UUDB). Typischerweise werden Gateway und NJS auf dedizierten Systemen hinter der Firewall des Zentrums ausgeführt, wobei jedoch das Gateway auch außerhalb bzw. innerhalb einer demilitarisierten Zone platziert werden kann.

Zielsystem Ebene

Das Target System interface (TSI) implementiert die Schnittstelle zum darunterliegenden Supercomputer und dessen Ressourcen-Management-System. Das TSI ist ein zustandsloser Prozess, der auf dem Zielsystem ausgeführt wird und das lokale Ressourcen-Management-System, z.B. LoadLeveler oder PBSpro, eine Batch-System-Emulation auf Linux-Basis (d.h. neue Jobs werden per fork gestartet) oder einen weiteren Grid Resource Manager wie etwas Globus' GRAM (10,11) anspricht.

Single Sign-On

Das Sicherheitsmodell von UNICORE basiert auf der Nutzung permanenter X.509 Zertifikate, die von einer zuverlässigen Certificate Authority (CA) ausgestellt sind, sowie auf SSL-basierter Kommunikation über 'unsichere' Netzwerkverbindungen. Zertifikate werden für den Single Sign-On im Client verwendet, der den Keystore des Benutzers jeweils beim ersten Start öffnen, so dass keine weiteren Passworteingaben mehr durch den Benutzer notwendig sind. An jeder UNICORE Site sind die Zertifikate von Benutzern auf deren lokale Accounts (Standard UNIX gid/uid) gemappt, welche aufgrund verschiedener Namenskonventionen bei den unterschiedlichen Zentren natürlich unterschiedlich sein können. Somit behalten die Zentren die volle Kontrolle über den Benutzerzugang zu den Ressourcen, in dem die Identität jedes individuellen Benutzers, z.B. auf Basis des Distinguished Names (DN), oder anderer im Zertifikat enthaltener Informationen überprüft wird. UNICORE kann auch mit mehreren Benutzer-Zertifikaten umgehen, d.h. Benutzer können in mehreren, nicht miteinander verbundenen Grids teilnehmen. Es ist ebenfalls möglich, dass Benutzer im Client Projektaccounts angeben, um unterschiedliche Projekte und damit Privilegien zu nutzen.

Der private Schlüssel im Zertifikate wird zur Signierung des Jobs und aller darin enthaltenen Teiljobs für den Transport vom Client zu einer Usite und zwischen Usite verwendet. Dies schützt vor Manipulation, während der Jobs auf unsicheren Internetverbindungen übertragen wird. Dadurch ist es möglich, die Identität des Benutzers beim Empfänger zu überprüfen, ohne dabei den auf dem Übertragungsweg liegenden Zwischenstationen vertrauen zu müssen, die den Job weitergeleitet haben.

UNICORE-basierte Projekte

Während der evolutionären Entwicklung der UNICORE Technologie haben sich mehrere europäische und internationale Projekte entschlossen, ihre Grid Software Implementierungen auf UNICORE zu basieren oder UNICORE um zusätzliche Funktionalität zu erweitern. Die Zielsetzungen der Projekte, die UNICORE verwendet haben, ist dabei nicht auf die Informatikgemeinde beschränkt. Zahlreiche andere Wissenschaftsfelder, wie etwa die Biologie oder Chemie, verwenden die UNICORE Technologie als Basis für ihre Arbeiten.

EUROGRID – Application Testbed for European Grid Computing

Im Rahmen des EUROGRID[3] Projektes (12,13) wurde eine Grid-Infrastruktur zwischen führenden europäischen Hochleistungsrechenzentren aufgebaut. Auf der Basis UNICORE wurden anwendungsspezifische Grids integriert, betrieben und demonstriert:

- Bio-Grid für die Bio-Molekular-Wissenschaft
- Meteo-Grid für die lokale Wettervorhersage
- CAE-Grid für gekoppelte Applikationen
- HPC-Grid für die allgemeine HPC Benutzerschaft

Dabei wurde UNICORE um effizienten Datentransfermechanismen, Ressourcen-Brokering-Mechanismen, Werkzeuge und Services für Application Service Provider (ASP), gekoppelte Anwendungen und einen interaktiven Zugang erweitert.

GRIP – Grid Interoperability Project

Unterschiedliche Anwendergruppen and Anwendungsfelder haben mit der Zeit ihre jeweils spezifischen Grid-Lösungen entwickelt. Zielsetzung des GRIP[4] Projektes (15) war zu zeigen, dass sich die unterschiedlichen Ansätze zweier Grid-Lösungen komplementär ergänzen und miteinander interoperieren können. Zwei prominente Grid System wurden daher ausgewählt: UNICORE und Globus™ (16) aus den USA. Im Gegensatz zu UNICORE stellt Globus eine Menge von APIs und Services zur Verfügung, dessen Benutzung mehr Tiefenwissen benötigt. Globus wird in zahlreichen internationalen Projekten verwendet und ist an vielen Zentren installiert. Die Zielsetzungen im GRIP Projekt waren:

- Softwareentwicklung zur Interoperabilität unabhängig voneinander entwickelter Grid Lösungen
- Aufbau und Demonstration prototypischer Inter-Grid Anwendungen
- Beeinflussung internationaler Grid Standards

Die Ergebnisse des GRIP Projektes sowie die gewonnenen Erfahrungen bei der Entwicklung der Interoperabilitätssoftware machen seitdem eine aktive Mitarbeit im

[3] Förderkennzeichen: IST-1999-20247, Dauer: November 2000 – Januar 2004
[4] Förderkennzeichen: IST-2001-32257, Dauer: Januar 2002 – Februar 2004

internationalen Standardisierungsgremium Global Grid Forum (GGF) möglich, in dem Themen wie Sicherheit, Architektur, Protokolle, Workflow, Produktionsmanagement und Anwendungen addressiert werden.

OpenMolGRID – Open Computing Grid for Molecular Science and Engineering
Ziel des OpenMolGRID[5] Projektes (16,17) war die Entwicklung von gridifizierten Anwendungen aus dem Bereich des Biomolekulardesigns. Die Zielsetzungen des Projektes waren:

- Entwicklung von Werkzeugen für den sicheren und stoßkantenfreien Zugang zu verteilen Informationen und Berechnungsverfahren für die Biomolekularwissenschaften auf Basis von UNICORE
- Aufbau und Betrieb einer realen Testumgebung und Implementierung von Referenzapplikationen
- Ausarbeitung von Designgrundsätzen für ein Molekular-Entwicklungs-System der nächsten Generation

UNICORE wurde dabei verwendet, um die Prozesse zu automatisieren, zu integrieren und zu beschleunigen.

DEISA – Distributed European Infrastructure for Scientific Applications
Um die Wissenschaft in Europe auf eine neue Ebene zu stellen, haben sich acht führende europäische Höchstleistungszentren im DEISA[6] Projekt (18) zusammengeschlossen. Die Zentren vereinen ihre Kräfte und Möglichkeiten, um eine verteilte Supercomputing Umgebung mit tera-scaler Leistung aufzubauen. Dies wird durch eine Integration existierender nationaler Hochleistungsplattformen, ein dediziertes Netzwerk und innovative System und Software möglich. Das Projekt hat sich entschlossen, UNICORE als Grid Middleware einzusetzen.
In der ersten Projektphase haben die vier Kernpartner über 4000 IBM Power 4 Prozessoren und 416 SGI Prozessoren zusammengeschaltet und insgesamt eine aggregierte Spitzenleistung von 22 TeraFlops erzielt. Dabei stellt UNICORE den stoßkantenfreien, sicheren und intuitiven Zugang zu diesem Super-Cluster bereit.
Das Forschungszentrum Jülich ist einer der Kernpartner und ist verantwortlich für die Einführung von UNICORE als Grid Middleware bei allen Partnern.

Zukunft von UNICORE

Die bisherige UNICORE Software implementiert eine vertikal integrierte Grid Architektur, die einen stoßkantenfreien Zugang zu den unterschiedlichsten

[5] Förderkennzeichen: IST-2001-37328, Dauer: September 2002 – Februar 2005
[6] Förderkennzeichen: FP6-508803, Dauer: Mai 2004 – April 2009

Ressourcen zur Verfügung stellt. Jede Ressource ist statisch in das UNICORE Grid, durch ein entsprechendes Interface zum Ressource Manager, integriert.
Eine der Vorteile von Web Services ist das Konzept der lose gekoppelten, verteilten Services. Die Verbindung der Idee "Everything being a service" mit den Erfolgen der Gridgemeinde führte zum Begriff der Grid Services, die eine neue Art und Weise des Designs von Grid Architekturen ermöglichen. Die Berücksichtigung von XML und die Bemühungen zur Standardisierung der Open Grid Services Architektur (OGSA) sind ein weiterer Schritt in Richtung interoperabler Grid Infrastrukturen. Ein Demonstrator bestätigte, dass die Architektur von UNICORE mit OGSA/OGSI (Open Grid Service Infrastructure (19)) übereinstimmt, welche die Entwicklung einer OGSA-basierten UNICORE Software weiter vorwärts treibt.

UniGrids – Uniform Interface to Grid Services
Im UniGrids[7] Projekt (20) wird UNICORE an eine neue Architektur lose-gekoppelter Komponenten angepasst, wobei die Ende-zu-Ende Eigenschaften erhalten bleiben. Daher werde die Integrierungsarbeiten von Web Services und UNICORE, die bereits im GRIP Projekt begonnen haben, im UniGrids aufgenommen und fortgeführt. Interoperabilität durch die Anwendung und Beeinflussung von internationalen Standards bilden eine Basis von UniGrids. Das Projekt hat als Ziel, UNICORE mit Interfaces auszustatten, die kompatibel zum Web Services Resource Framework (WS-RF) (21) sind. Dieser Ansatz bietet viele Vorteile für die einfachere Entwicklung neuer Komponenten durch ein Zusammenschalten von nicht-UNICORE Services. Zielsetzungen des Projektes sind daher:

- Entwicklung einer WS-RF-kompatiblen Hosting Umgebung, die für die Bereitstellung von UNICORE Job und Datei Services auf Web Service Basis
- Unterstützung von dynamischen virtuellen Organisationen durch Erweiterung der UNICORE Sicherheits-Infrastruktur, um unterschiedliche Nutzungsmodelle sowie Delegation und kollektive Autorisierung zu ermöglichen
- Entwicklung von Überstzungsmechanismen, z.B. Ressourcen Ontologien für Interoperabilität mit anderen OGSA-basierten Grid Systemen. Unterstützung von Grid Ökonomie durch Entwicklung einer Service Level Agreement (SLA) Umgebung und von Cross-Grid Brokering Services
- Entwicklung und Integration von generischen Softwarekomponenten für die Visualisierung und Steuerung von Simulationen (VISIT (22)), Geräteüberwachung und -kontrolle sowie Werkzeuge für den Zugang zu verteilten Daten und Datenbanken

Die Entwicklungen im UniGrids Projekt führen zu Unicore/GS, der nächsten Generation der UNICORE Software, die auf OGSA aufgebaut ist und durch WS-RF interoperabel ist.

[7] Förderkennzeichen: IST-2002-004279, Dauer: Juli 2004 – Januar 2006

Genauso wie UNICORE wird auch Unicore/GS unter BSD Lizenz als Open Source verfügbar sein und unter http://unicore.sourceforge.net zum Download zur Verfügung stehen.

Literatur

DEISA - Distributed European Infrastructure for Supercomputing Applications. http://www.deisa.org.

Erwin, D. (Ed.): UNICORE - Uniformes Interface für Computing Ressourcen, Final project report (in German). 2000.

Erwin, D. (Ed.): UNICORE Plus Final Report - Uniform Interface to Computing Resources. Forschungszentrum Jülich, 2003.

EUROGRID - Application Testbed for European Grid Computing. http://www.eurogrid.org.

Foster, I. and C. Kesselman (Eds.): The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Fransisco, 1999.

Foster, I. and C. Kesselman: Globus: A Metacomputing Infrastructure Toolkit. International Journal on Supercomputer Applications, 11(2):115{128, 1997.

Foster, I.; Kesselmann, C. ; Nick, J. M. and S. Tuecke: The Physiology of the Grid. In F. Berman, G. C. Fox, and A. J. G. Hey, editors, Grid Computing, pages 217-249. John Wiley & Sons Ltd, 2003.

Globus: Research in Resource Management. http://www.globus.org/research/resource-management.html.

GRIP - Grid Interoperability Project. http://www.grid-interoperability.org.

Huber, V.: Supporting Car-Parrinello Molecular Dynamics Application with UNICORE. In Proc. of the International Conference on Computational Science (ICCS 2001), pages 560-566, 2001.

Mallmann, D.: EUROGRID - Application Testbed for European Grid Computing. In Proc. of Industrial GRID Conference 2001, 2001.

Menday, R. and Ph. Wieder: GRIP: The Evolution of UNICORE towards a Service-Oriented Grid. In Proc. of the 3rd Cracow Grid Workshop (CGW'03), pages 142-150, 2003.

OASIS Web Services Resource Framework (WSRF). http://www.oasis-open.org/committees/wsrf.

OpenMolGRID - Open Computing Grid for Molecular Science and Engineering. http://www.openmolgrid.org.

Romberg, M.: The UNICORE Grid Infrastructure. Scientific Programming, 10(2):149-157, 2002.

Sild, S.; Maran, U.; Romberg, M.; Schuller, B. and E. Benfenati: OpenMolGRID: Using Automated Workflows in GRID Computing Environment. In Proceedings of the European Grid Conference 2005, 2005.

Snelling, D Berghe, .; S. van den; Laszweski, G. von; Wieder, Ph.; Breuer, D. ; MacLaren, J.; Nicole, D. and H.-Ch. Hoppe: A UNICORE Globus Interoperability Layer. Computing and Informatics, 21:399-411, 2002.

Snelling, D.: UNICORE and the Open Grid Services Architecture. In F. Berman, G. Fox, and T. Hey, editor, Grid Computing: Making The Global Infrastructure a Reality, pages 701-712. John Wiley & Sons, 2003.

The Globus Project. http://www.globus.org/.

Tuecke, S.; Czajkowski, K.; Foster, I. ; Frey, J.; Graham, S.; Kesselman, C.; Maquire, T.; Sandholm, T.; Snelling, D. and P. Vanderbilt (eds.): The Open Grid Services Infrastructure (OGSI) Version 1.0, 2003.

UniGrids - Uniform Access to Grid Services. http://www.unigrids.org.

VISIT – A Visualization Toolkit. http://www.fz-juelich.de/zam/visit/.

Identitäts- und Accessmanagement in Service-orientierten Umgebungen

Wilfried Stüttgen, Ratingen

Abstract

Sun Microsystems unterstützt mit den Sun Java Systems Softwareprodukten offene Standards u.a. im Bereich Identitäts- und Accessmanagement. Ein Beispiel für eine offene Service-orientierte Umgebung ist das Projekt „Service-orientierte It-Infrastruktur“ (SOI) der niedersächsischen Hochschulen. Das SOI Konsortium setzt sich zusammen aus: Technische Universität Braunschweig, Fachhochschule Braunschweig/Wolfenbüttel, Technische Universität Clausthal, Universität Hannover, Universität Oldenburg, Sun Microsystems und Niedersächsisches Ministerium für Wissenschaft und Kultur.
Die Ziele des Projekts sind die Vergabe und Verwaltung einer niedersachsenweiten eindeutigen elektronischen Identität für Studierende und Bedienstete der Hochschulen, erleichterte Authentifizierung durch Verwendung eines einzigen Passwortes auf unterschiedlichen Systemen, rollenbasierte Rechte- bzw. Zugriffsverwaltung, die Anbindung der verschiedensten Dienste, sowie die Nutzbarkeit von Identität und Passwort an unterschiedlichen Hochschul-Standorten. Dieses Dokument beschreibt die Voraussetzungen und Maßnahmen zum Aufbau einer solchen Infrastruktur.

Herausforderungen

An den Hochschulen existiert eine heterogene System- und Dienstumgebung. Wesentliche Bestandteile sind Systeme der Hochschulverwaltung für die Administrierung von Studenten- und Mitarbeiterdaten (z.B. Hochschul-Informations-Systeme (HIS), SAP/HR), Systeme der Hochschulrechenzentren zur Verwaltung von Accounts und Zugriffsrechten auf Rechnerplattformen und Infrastruktureinrichtungen, Bibliothekssysteme und eLearning-Applikationen.
Das zukünftige Identitäts- und Accessmanagement und die dafür erforderliche Infrastruktur an einer Hochschule ist auf der Basis hoch-integrierter und hochverfügbarer Dienste zu etablieren. Die wesentlichen Funktionen in den unterschiedlichen Bereichen wie Forschung, Lehre, Bibliothek und Verwaltung werden wesentlicher enger miteinander interagieren. Portale, administrative Systeme, Lernmanagement-Plattformen, Content-Management-Systeme, Bibliotheks- Verwaltungssysteme und digitale Bibliotheken werden miteinander vernetzt, um
einen identischen Service-Level innerhalb und außerhalb des Campus bieten zu können,

qualitativ hochwertige und wettbewerbsfähige Dienste für Studenten und Mitarbeiter im Vergleich zu anderen Hochschulen bereit stellen zu können.
Portale und E-Learning-Angebote bringen eine Hochschule in eine Rolle, die mit einem E-Commerce-Anbieter vergleichbar ist. Eine Verfügbarkeit "rund um die Uhr" wird nicht nur innerhalb, sondern in besonderem Maße auch außerhalb der Hochschule (Studentenwohnheim, "virtueller" Student, Gastwissenschaftler, eigener Wissenschaftler im Ausland) als selbstverständlich angesehen und gefordert werden. Die Themen Erreichbarkeit, Sicherheit, Interoperabilität, Skalierbarkeit und Performance werden zu wichtigen Kriterien für die IT-Infrastruktur.

Systemarchitektur

Im folgenden sind die wesentlichen Komponenten und Methoden, die zum Aufbau der Infrastruktur eingesetzt wurden, beschrieben.

Die Basis von SOI - LDAP und LDAP-Proxy

Der Kern von SOI besteht aus den LDAP-Servern der jeweiligen Standorte und einem zentralen LDAP-Proxy (Abbildung 1). Die LDAP-Server dienen in erster Linie zur Speicherung von lokalen Benutzerdaten und halten zudem Konfigurationsdaten anderer Server-Produkte. Neben den reinen Benutzerdaten sind im LDAP-Server auch Berechtigungsprofile auf Basis von Rollen abgelegt.

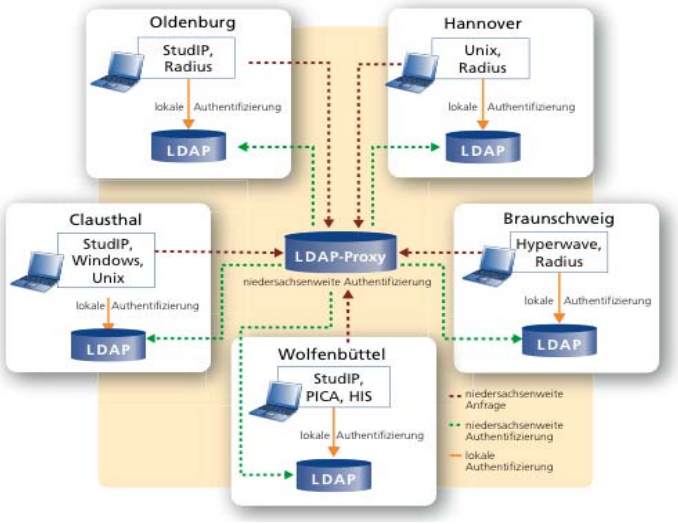


Abbildung 1: LDAP - Systemarchitektur

Die Authentifizierung von Benutzern innerhalb von SOI erfolgt grundsätzlich gegen den LDAP-Proxy, der die Anfrage je nach Zuständigkeit an die LDAP-Server verteilt. Dabei orientiert sich der LDAP-Proxy an der niedersachsenweit festgelegten UID der Form "Hans.Norden@uni-hannover.de". Die Authentifizierung erfolgt grundsätzlich gegen den LDAP-Server der Heimat-Hochschule des Benutzers.
Zur Erhöhung der Dienstverfügbarkeit wird der LDAP-Proxy von einem Watchdog-Prozess überwacht, zudem ist für andere Server der lokale LDAP-Server als Failover-Instanz konfiguriert.
Um unnötiges Routing von Authentifizierungsvorgängen über den zentralen Proxy zu vermeiden, wurde auch der Einsatz von lokalen LDAP-Proxies an den einzelnen Standorten mit Erfolg getestet. In diesem Fall läuft die Authentifizierung erst einmal über den lokalen LDAP-Proxy, der die Anfrage je nach Form der UID direkt an den lokalen LDAP-Server schickt ("Hans.Norden") oder an den LDAP-Proxy eines anderen Standortes sendet ("Hans.Norden@<andere.domain>").

Access Manager

Der Sun Java System Access Manager erweitert klassische LDAP-Rollen um Vererbungsmechanismen und vereinfacht so die Verwaltung von Berechtigungsinformationen. Um die Berechtigung von Benutzern für bestimmte Dienste zu steuern, fragen die betreffenden Applikationen bei der Authentifizierung spezielle, dienstspezifische Indikatorattribute im LDAP-Server ab. Diese Indikatorattribute werden über Rollen an die Benutzer vererbt, müssen also nicht für jeden einzelnen Benutzer verwaltet werden.

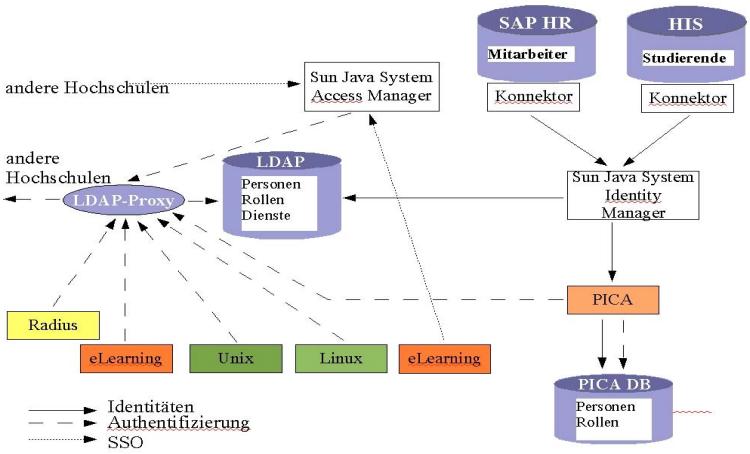


Abbildung 2: Access-Management

Eine weitere für SOI wichtige Funktion ist die des Web Single Sign On (SSO), die einem Benutzer beim Zugang mittels Web-Browser nur eingangs eine einmalige Authentifizierung abverlangt. Diese Funktion kann nicht nur an einem Hochschulstandort, sondern auch übergreifend genutzt werden. Bei übergreifendem (föderiertem) SSO wird die Security Assertion Markup Language (SAML) als offener Standard genutzt, der auch die Interaktion mit den Produkten anderer Hersteller erlaubt. Darüber hinaus ist SAML die Basis der Liberty Alliance (https://www.projectliberty.org/) und Shibboleth (http://shibboleth.internet2.edu/), so dass hier die Freiheit gewahrt bleibt, in nachfolgenden Projektabschnitten eine komplexe Lösung für föderiertes SSO einzusetzen.

Identity Manager

Die automatische und manuelle Provisionierung von Benutzerdaten obliegt dem Sun Java System Identity Manager. Mit seiner Hilfe lassen sich Benutzerdaten automatisch von beliefernden Systemen wie z.B. SAP oder HIS an andere Systeme oder Applikationen verteilen. Dies betrifft sowohl die Neuanlage als auch die Aktualisierung von Benutzerinformationen.

Im Rahmen von SOI verteilt der Identity Manager Benutzerdaten in erster Linie an den lokalen LDAP-Server des Standortes. Je nach Standort ist auch die Replikation dieser Daten an Windows-Systeme bzw. das Bibliotheks-System PICA sinnvoll. Denkbar ist auch die Belieferung weiterer, spezialisierter und noch nicht integrierter LDAP-Server.

Die Web-Oberfläche des Identity Managers ist der Zugangspunkt für die Selbst-Verwaltungsfunktionen, die den Benutzern zur Verfügung stehen. Insbesondere die Verwaltung des oder der Passworte erfolgt auf diesem Weg. Der Identity Manager aktualisiert nach einer Änderung des Passwortes die Accounts des Benutzers auf allen relevanten Systemen. Analog wird mit anderen Benutzerdaten verfahren.

Mit Windows besteht zudem die Option der bidirektionalen Passwort-Synchronisation.

Kategorien angebundener Systeme

Bei der Betrachtung der für SOI-Anbindung ausgewählten Systeme und Applikationen lassen sich drei Kategorien unterscheiden:

1. Applikationen/Systeme mit LDAP-Authentifizierung (E-Learning, Radius, Unix)
2. Applikationen/Systeme mit Web-SSO-Befähigung (Sun Java System Access Manager)

Applikationen/Systeme mit Bedarf für eigene Benutzer-DB (PICA, Windows)

Systeme mit Befähigung für LDAP-Authentifizierung können relativ einfach durch Konfiguration angebunden werden. Ist dies nicht gegeben, liegt meist die Beschränkung auf eine eigene Benutzer-Datenbank vor, gegen die auch authentifiziert wird. Systeme dieser Kategorie müssen mit redundanten Daten versorgt werden.

Die Einbindung in Web-SSO-Szenarien erfordert häufig geringfügige Modifikationen im Bereich des Applikations-Login.
Die Abbildung 3 zeigt eine charakteristische lokale Infrastruktur.

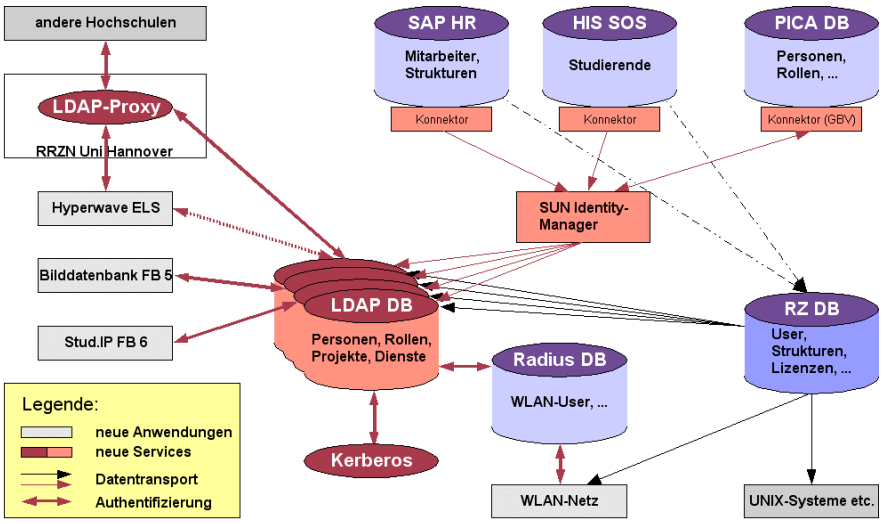


Abbildung 3: Geplante Identity-Management Infrastruktur am Rechenzentrum der TU Braunschweig

Beliefernde Systeme

1. SAP-HR
   Der Identity Manager verfügt über einen Adapter zu SAP-HR. Da derzeit noch keines der SAP-HR-Systeme an den Hochschulen produktiv im Einsatz ist, wird der Adapter für SOI z.Zt. noch nicht genutzt.
2. HIS-SOS und HIS-SVA
   Für SOI wurde ein Adapter für die HIS-Staging-Tabellen der HIS-Version 7 implementiert. Mit seiner Hilfe können von HIS gemeldete Änderungen am Datenbestand automatisch an den Identity Manager übertragen werden. Dies dient sowohl der Automatisierung und damit der Vereinfachung der Datenübernahme zum Semesterbeginn als auch der nachfolgenden Aktualisierung einzelner Benutzerdatensätze.
   Die HIS-Staging-Tabellen werden von HIS-SOS und HIS-SVA beliefert. Derzeit erlaubt die Implementierung der HIS-Staging-Tabellen nur das Lesen von HIS bereitgestellter Daten.
   Mangels eines zugänglichen HIS-Systems an den Hochschulen konnte der Adapter für SOI noch nicht genutzt werden. Eine Simulation der HIS-Staging-Tabellen ist in SOI in Betrieb.

E-Learning-Systeme

1. Stud.IP und Hyperwave
   Test-Instanzen von Stud.IP und Hyperwave sind im Rahmen des Projektes für LDAP-Authentifizierung konfiguriert. Die Authentifizierung läuft in beiden Fällen über den LDAP-Proxy und gibt damit auch Angehörigen anderer Hochschulen die Möglichkeit sich mit ihrer niedersachsenweiten UID und ihrem einheitlichen Passwort bei diesen Applikationen anzumelden.

Rechnerplattformen

1. Unix
   Unix-Systeme können gemäß der SOI-Architektur auf zwei Wegen angebunden werden. Die Verwendung von Pluggable Authentication Modules (PAMs) erlaubt die Authentifizierung von Benutzern gegen einen LDAP-Server. Im Rahmen von SOI arbeitet ein Solaris 10 PAM-LDAP gegen den LDAP-Proxy und ermöglicht so den Rechnerzugang für Angehörige anderer Hochschulen.
   Alternativ können mit Hilfe des Identity Managers auch dedizierte Accounts auf den Rechnern administriert werden. Mischformen beider Ansätze sind konfigurierbar.
2. Windows
   Der Identity Manager kann Accounts in Windows Domains verwalten. Dazu ist die Installation einer DLL auf einem beliebigen Rechner der Domäne erforderlich. Mit Windows besteht die Möglichkeit einer bidirektionalen Passwort-Synchronisation. Ändert ein Benutzer sein Passwort auf Windows, so wird das neue Passwort an den Identity Manager geschickt und auf alle Systeme und Applikationen, die dem Benutzer zugeordnet sind, verteilt.

Infrastruktur

1. Radius
   Der Zugang zu Internet und WLAN wird an den meisten Hochschulen über Radius-Server geregelt. Radius-Server sind im allgemeinen zur Authentifizierung gegen LDAP in der Lage. Weiterhin können Radius-Server UIDs der für SOI verwendeten Form erkennen und entsprechende Verarbeitungsschritte einleiten. Für SOI wird ein Radius-Server zur Authentifizierung gegen den LDAP-Proxy konfiguriert. Andere Radius-Server erkennen die zu SOI gehörigen UIDs und leiten die Anfragen an den entsprechenden Radius-Server weiter. Damit ist auch der Zugang zu Internet und WLAN für Angehörige anderer Hochschulen im Rahmen des SOI-Konzeptes möglich.

Bibliothekssystem

1. PICA
   PICA besitzt eine proprietäre Benutzerverwaltung. Die Fachhochschule Braunschweig/Wolfenbüttel hat zusammen mit der GBV (Gemeinsamer Bibliotheks

verbund) eine Lösung zur Authentifizierung mittels LDAP entwickelt. Das PICA-System wird von der Hochschulbibliothek der TU Braunschweig für die TU Braunschweig, die Fachhochschule Braunschweig/Wolfenbüttel, die Hochschule für Bildende Künste Braunschweig und für die Herzog August Bibliothek Wolfenbüttel betrieben. Der Server befindet sich im Netz der TU Braunschweig, d.h. das Rechenzentrum der Fachhochschule hat keinen direkten Zugriff auf das PICA-System. Der Datenaustausch erfolgt nur über einen Transferbereich in eine Richtung, von der Fachhochschule zum PICA-System. Lesende Zugriffe sind nicht möglich bzw. werden vom Betreiber nicht zugelassen. Für die Authentifizierung zum Benutzerkonto im PICA-System werden Zugangskennung und Passwort verwendet, die im LDAP der Fachhochschule stehen.

Die interne Anmeldeseite von PICA ist mit Hilfe eines Konfigurationsskriptes der GBV auf eine nachgebildete Anmeldeseite der Fachhochschule umgeleitet worden. Die Anmeldung in PICA erfolgt über ein externes PHP-Skript, das eine Authentifizierung gegen den zentralen LDAP der Fachhochschule durchführt. Nach erfolgter Authentifizierung wird aus dem LDAP der Barcode und das PICA-interne Passwort ausgelesen, der entsprechenden PICA-Web-Seite als Parameter übergeben und in der jeweiligen Session gespeichert.

In PICA ist kein durchgehendes und homogenes Session-Handling realisiert, was bei der Funktion „Vormerken von Büchern (Auswahl aus einer Bandliste)“ eine andere Vorgehensweise erfordert, da hier erneut eine Authentifizierung außerhalb der bestehenden Session erforderlich ist.

Aufgrund dessen wurde der Link zum Abruf der Bandliste ebenfalls von der GBV auf das PHP-Skript der Fachhochschule umgeleitet. Hierbei wird die Bandliste als unangemeldeter Benutzer aus dem PICA-System abgerufen, und aus der empfangenen Seite herausgefiltert. Anschließend wird eine neue Seite mit der Bandliste zur Auswahl erzeugt und mit einem Anmeldeformular versehen. Nach erfolgreicher Authentifizierung gegen den LDAP der Fachhochschule werden PICA die Funktionsparameter, das ausgewählte Band und die Anmeldedaten (Barcode und PICA-Passwort) als Parameter übergeben.

Aus Sicherheitsgründen wurde nicht das zentrale LDAP-Passwort in das PICA-System synchronisiert, sondern ein dem Nutzer unbekanntes 16-stelliges Zufallspasswort, das als Parameter des jeweiligen Nutzers im LDAP der Fachhochschule gespeichert ist. Der Barcode des jeweiligen Nutzers wurde ebenfalls im LDAP hinterlegt.

Diskussion

Im folgenden werden die während des Projektes gewonnen Erfahrungen diskutiert.

Offenheit der service-orientierten Umgebung
Die Anforderungen für die Teilnahme einer Hochschule am SOI-Verbund sind relativ gering. Im Normalfall können vorhandene LDAP-Directories einfach genutzt werden. Zudem sind unterschiedliche Stufen der Einbindung möglich. Neben der Teilnahme auf Ebene der reinen LDAP-Authentifizierung über den LDAP-Proxy ist auch die Einbindung in ein standortübergreifendes SSO möglich. Letzteres erfordert mehr Abstimmung zwischen den Hochschulen. Auch auf dieser Ebene ist das Konzept durch die Nutzung von SAML absolut offen und produktunabhängig.

Integration
Viele Applikationen und Systeme an den Hochschulen sind für eine Authentifizierung gegen LDAP-Directories konfigurierbar. SOI deckt also konzeptionell und technologisch weite Teile der Hochschul-Infrastruktur ab.

Transparenz des LDAP-Proxy
Der Einsatz des LDAP-Proxy ist transparent für die angeschlossenen Applikationen und Systeme, d.h. notwendige Anpassungen konzentrieren sich nur auf den Proxy. Damit werden aufwendige Modifikationen an den integrierten Applikationen vermieden.

Flexibilität
SOI ist für die existierenden, heterogenen Hochschul-Umgebungen anwendbar. Die konzeptionelle Offenheit erlaubt unterschiedliche LDAP-Directories, Schemata und Directory Information Trees.

Datenschutz - Verzicht auf Redundanz
Es werden keine permanenten Daten für Nutzer aus anderen Standorten zentral gespeichert. Neben rein technischen Vorteilen wie effizientem Hardware-Einsatz und der Vermeidung von Inkonsistenzen ist dies auch aus datenschutzrechtlichen Gründen äußerst vorteilhaft.

Sicherheit
Trotz der standortübergreifenden Authentifizierung ist ein schreibender Zugriff auf Daten im lokalen LDAP-Directory nur lokal möglich. Spezielle Funktionen im LDAP-Proxy sichern dem lokalen Administrator die Steuerung der Zugriffsrechte von Nutzern aus anderen Standorten auf die lokale Infrastruktur.

Zusätzliche Hardware für LDAP-Proxy
Die Authentifizierung läuft in jedem Fall über ein zusätzliches System, den LDAP-Proxy. Dies bedeutet zusätzlicher Hardwarebedarf und wirkt sich auf die Verarbeitungszeiten aus. Der Einsatz jeweils eines lokalen Proxy pro Standort würde insofern Verbesserungen bringen, als lokale Authentifizierung auch im lokalen Netz
bliebe.

Abhängigkeit
Der Proxy besitzt grundlegende Bedeutung für die Authentifizierung an den Standorten. Sein Ausfall würde die Standorte massiv behindern. Dasselbe gilt allerdings auch für die LDAP-Directories der Hochschulen. In beiden Fällen empfiehlt sich eine hochverfügbare Lösung. Soweit für Applikationen und Systeme eine Failover-Instanz für die Authentifizierung konfiguriert werden kann, sollte diese direkt auf ein lokales LDAP-Directory zeigen.

Anmerkung

Dieses Dokument basiert auf dem Entwurf für den Abschlussbericht „Service-Orientierte IT-Infrastruktur an den Niedersächsischen Hochschulen“ des Konsortiums. Der Abschlussbericht wird voraussichtlich im September 2005 fertig sein und kann hier bezogen werden: http://soi.lanit-hrz.de

Liste der Autoren

Prof. Dr. Claus Arnold
Universität Regensburg
Lehrstuhl für Psychologie
Institut für Psychologie
93040 Regensburg
E-Mail: claus.arnold@psychologie.uni-regensburg.de

Dr. Rafael Ball
Forschungszentrum Jülich
Zentralbibliothek
52425 Jülich
E-Mail: r.ball@fz-juelich.de

Prof. Dr. Norbert Bolz
Technische Universität Berlin
Institut für Sprache und Kommunikation
Fachgebiet Medienwissenschaft
Ernst-Reuter-Platz 7
10587 Berlin

Prof. Dr. Christoph Bläsi
Universität Erlangen-Nürnberg
Institut für Buchwissenschaft
Harfenstr. 16
91054 Erlangen
E-Mail: christoph.blaesi@buchwiss.uni-erlangen.de

Dr.-Ing. Torsten Brix
Technische Universität Ilmenau
Fachgebiet Konstruktionstechnik
Postfach 10 05 65
98684 Ilmenau
E-Mail: torsten.brix@tu-ilmenau.de

Jochen U. Brüning
Universität Konstanz
Fachbereich Informatik und Informationswissenschaft
78457 Konstanz
E-Mail: jochen.bruening@uni-konstanz.de

Ulrich Bügel
Fraunhofer Institut IITB
Fraunhoferstr. 1
76131 Karlsruhe
E-Mail: bgl@iitb.fhg.de

Dr. Hans-Robert Cram
Verlag Walter de Gruyter GmbH & CO KG
Genthiner Str. 13
10785 Berlin
E-Mail: h.r.cram@degruyter.de

Ulf Döring
Technische Universität Ilmenau
Fachgebiet Graphische Datenverarbeitung
Postfach 10 05 65
98684 Ilmenau
E-Mail: ulf.doering@tu-ilmenau.de

Dr. Iryna Gurevych
EML Research gGmbh
Natural Language Processing Group
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
E-Mail: Iryna.Gurevych@eml-r.villa-bosch.de

Prof. Dr. Rainer Hammwöhner
Universität Regensburg
Lehrstuhl für Informationswissenschaft
93040 Regensburg
E-Mail: rainer.hammwoehner@sprachlit.uni-regensburg.de

Manfred Hauer
AGI - Information Management Consultants
Mandelring 238b
67433 Neustadt/Weinstrasse
E-Mail: Manfred.Hauer@agi-imc.de

Prof. Dr. Heinz-Gerd Hegering
Universität München und
Leibniz-Rechenzentrum
Barer Str. 21
80333 München
E-Mail: hegering@lrz.de

Dr. Wolfram Horstmann
Hochschulbibliothekszentrum NRW
Digital Peer Publishing
Jülicher Str. 6
50674 Köln
E-Mail: wolfram.horstmann@gmail.com

Reiner Krause
Max-Planck-Institut für Biogeochemie
Postfach 10 01 64
07701 Jena
E-Mail: rkrause@bgv-jena.mpg.de

Prof. Dr. Jens Krinke
FernUniversität Hagen
Fach Softwaretechnik
58084 Hagen
E-Mail: Jens.Krinke@FernUni-Hagen.de

Prof. Dr. Rainer Kuhlen
Universität Konstanz
Fachbereich Informatik und Informationswissenschaft
78457 Konstanz
E-Mail: rainer.kuhlen@uni-konstanz.de

Rainer Kupsch
Forschungszentrum Karlsruhe GmbH
Institut für Wissenschaftliches Rechnen (IWR)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
E-Mail: rainer.kupsch@iwr.fzk.de

Sören Lorenz
Universität Bielefeld
Lehrstuhl für Neurobiologie
Postfach 10 01 31
33501 Bielefeld
E-Mail: soeren.lorenz@uni-bielefeld.de

Philipp Mayr
Informationszentrum Sozialwissenschaften (IZ)
Forschung & Entwicklung
Lennestr. 30
53113 Bonn
E-Mail: philipp.mayr@bonn.iz-soz.de

Christian Nançoz
Wilerweg 37
3280 Murten
Schweiz
E-Mail: christian.nancoz@elogic.ch

Cordula Nötzelmann
Universitätsbibliothek Bielefeld
Postfach 10 01 91
33502 Bielefeld
E-Mail: cordula.noetzelmann@uni-bielefeld.de

Dr. H. Peter Ohly
Informationszentrum Sozialwissenschaften (IZ)
Internationale Gesellschaft für Wissensorganisation, Deutsche Sektion
Lennestr. 30
53113 Bonn
E-Mail: oh@iz-soz.de

Dr. Areti Ramachandra Durga Prasad
Indian Statistical Institute
Documentation Research and Training Centre
8th Mile, Mysore Road
Bangalore 560 059
India
E-Mail: ard@drtc.isibang.ac.in

Prof. Dr. Ludwig Richter
DIMDI
Projekt German Medical Science
Waisenhausgasse 36-38a
50676 Köln
E-Mail: richter@dimdi.de

Martin Roos
FernUniversität Hagen
58084 Hagen
E-Mail: Martin.Roos@FernUni-Hagen.de

Karin Schauerhammer
DFN-Verein
Geschäftstsstelle
Stresemannstr. 78
10963 Berlin
E-Mail: schau@dfn.de

Olaf Schneider
Forschungszentrum Karlsruhe GmbH
Institut für Wissenschaftliches Rechnen (IWR)
Postfach 36 40
76021 Karlsruhe
E-Mail: Olaf.Schneider@iwr.fzk.de

Han Shucheng
Tohoku University
School of Engineering
Department of Management of Science and Technology,
980-8597 Sendai-shi
Japan
E-Mail: hansc99@163.com

Dr. Maximilian Stempfhuber
Informationszentrum Sozialwissenschaften (IZ)
Abteilung Forschung & Entwicklung
Lennestr. 30
53113 Bonn
E-Mail: stempfhuber@iz-soz.de

Prof. Dr. Rainer H. Straub
Universität Regensburg
Klinik und Poliklinik für Innere Medizin I
93040 Regensburg
E-Mail: rainer.straub@klinik.uni-regensburg.de

Dr. Achim Streit
Forschungszentrum Jülich GmbH
Zentralinstitut für Angewandte Mathematik (ZAM)
52425 Jülich
E-Mail: a.streit@fz-juelich.de

Dr. Wilfried Stüttgen
Sun Microsystems GmbH
Marktentwicklung Forschung und Lehre
Brandenburger Straße 2
40880 Ratingen
E-Mail: wilfried.stüttgen@sun.com

Dr. Sabine Trott
Technische Universität Ilmenau
Universitätsbibliothek
Postfach 10 05 65
98684 Ilmenau
E-Mail: sabine.trott@tu-ilmenau.de

Dirk Tunger
Forschungszentrum Jülich GmbH
Zentralbibliothek
52425 Jülich
E-Mail: d.tunger@fz-juelich.de

Dr. Doris Wochele
Forschungszentrum Karlsruhe GmbH
Institut für Wissenschaftliches Rechnen (IWR)
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
E-Mail: doris.wochele@iwr.fzk.de

Prof. Dr. Christian Wolff
Universität Regensburg
Institute für Medien-, Informations- und Kulturwissenschaft (IMIK)
Universitätsstr. 31
95053 Regensburg
E-Mail: christian.wolff@sprachlit.uni-regensburg.de

Christian Woll
Bergmannstr. 144
50354 Hürth
E-Mail: christian.woll@gmx.de

Prof. Harayama Yuko
Tohoku University
School of Engineering
Department of Management of Science and Technology,
980-8597 Sendai-shi
Japan
E-Mail: harayama-yuko@rieti.go.jp

Register

I

J

K

L

M

S

T

U

V

W

Sponsoren

Wir danken unseren Sponsoren!

HERMIS 3.0

HARRASSOWITZ Electronic Resources Management and Information Solutions

HARRASSOWITZ

Electronic

Resources

Management and

Information

Solutions

In addition to a complete array of global print subscriptions services, HARRASSOWITZ extends its tradition of the highest standards of service to the management of electronic resources for academic and research libraries. Backed by a team of dedicated professionals and state-of-the-art technologies, HERMIS 3.0 meets the challenges of today's libraries with a complete menu of services and products in support of the library's electronic journals workflow:

- Resource identification and evaluation
  - Detailed information on all periodical title variations, print and electronic
  - Notifications of changes in online availability of publications on order with HARRASSOWITZ
  - Notifications of changes in publishers' e-journal policies
- License management
  - License analysis based on library licensing requirements
  - License negotiation with publishers on behalf of libraries
  - Online repository for signed licenses
  - License status in OttoSerials 3.0
- Ordering and payment, renewals and cancellations
  - Price quotes for e-journals and backfiles
  - Traditional ordering, payment, renewal and cancellation services
  - Electronic invoices for orders placed directly with the publisher
- Activation of electronic resources
  - Automatic activation on behalf of the library
  - Activation profile and activation status in OttoSerials 3.0
- Public resource discovery and access
  - A-Z lists, link resolvers, tables of contents service, and MARC records available through HARRASSOWITZ's industry partners
- Technical access management
  - Notification of URL changes
  - Notification to publisher of library IP changes
- Usage tracking
  - Links to COUNTER statistics on publishers' websites

For more information contact:

HARRASSOWITZ
Booksellers & Subscription Agents
65174 Wiesbaden
Germany

Email: service@harrassowitz.de
Web: www.harrassowitz.de
Toll-free (800) 348-6886
(USA/Canada)

HARRASSOWITZ

SERVICES

HARRASSOWITZ
is a book and serial vendor for the academic and research library community, specializing in the distribution of scholarly books, periodicals, e-resources, and music scores. HARRASSOWITZ offers the following products and services:

Subscriptions
Print and electronic journals published anywhere in the world in all subject areas and all languages

Standing Orders
Monographic series and works in parts in all formats and all languages

Monographs
Published throughout Europe in all formats and all languages

Approval Plans
Customized profiles for approval plans and form selection programs for books and music scores

Music Scores
Music scores published in all of Europe, the Middle East, the Far East, and Oceania

Electronic Data Exchange
Exchange of electronic orders, claims and invoices with library management systems

HERMIS
HARRASSOWITZ Electronic Resources Management and Information Solutions for academic and research libraries

OttoSerials
Online management of periodical subscriptions and standing orders

OttoEditions
Online management of monographic orders and approval plans

Publisher Services
Promotion and supply of materials to academic and research libraries

Customer Service
Available from the North American Service Office from 8 a.m. until 5 p.m. Central Time

For more information contact:

HARRASSOWITZ
Booksellers & Subscription Agents
65174 Wiesbaden
Germany

Email: service@harrassowitz.de
Web: www.harrassowitz.de
Toll-free (800) 348-6886
(USA/Canada)

500 12/2004

Schriften des Forschungszentrums Jülich
Reihe Bibliothek / Library

1. **Naturwissenschaft und Technik – nur für Männer?**
   Frauen mischen mit!
   Auswahl-Bibliographie Wissenschaftlerinnen (1999), 28 Seiten
   ISBN: 3-89336-246-0

4. **Schweißen & Schneiden**
   Wissenschaftliche Veröffentlichungen des Forschungszentrums Jülich
   (1997), 16 Seiten
   ISBN: 3-89336-208-8

5. **Verzeichnis der wissenschaftlich-technischen Publikationen**
   des Forschungszentrums Jülich
   Januar 1993 - Juli 1997 (1997), ca. 100 Seiten
   ISBN: 3-89336-209-6

6. **Biotechnologie**
   Wissenschaftliche Veröffentlichungen der Institute für Biotechnologie
   des Forschungszentrums Jülich
   Januar 1992 - Juni 1997 (1997), 48 Seiten
   ISBN: 3-89336-210-X

7. **Verzeichnis der wissenschaftlich-technischen Publikationen**
   des Forschungszentrums Jülich
   1997 bis 1999 (2000), 52 Seiten
   ISBN: 3-89336-260-6

8. **Kompendium Information**
   Teil I: Archive, Bibliotheken, Informations- und Dokumentationseinrichtungen
   Teil II: Ausbildungsstätten, Fort- und Weiterbildungsaktivitäten, Informationsdienste, Presse- und Nachrichtenagenturen, Verlagswesen und Buchhandel, Einrichtungen des Patent- und Normungswesen, Publikationen
   von G. Steuer (2001), 1130 Seiten
   ISBN: 3-89336-286-X

9. **Die Zukunft des wissenschaftlichen Publizierens**
   Der Wissenschaftler im Dialog mit Verlag und Bibliothek
   Jülich, 28. bis 30. November 2001. 40 Jahre Zentralbibliothek. Konferenz und Firmenausstellung
   Tagungsprogramm und Kurzfassungen (2001), 50 Seiten
   ISBN: 3-89336-292-4

10. **Die Zukunft des wissenschaftlichen Publizierens**
    Der Wissenschaftler im Dialog mit Verlag und Bibliothek
    Jülich, 28. - 30.11.2001. Tagungsprogramm und Vorträge (2002), 184 Seiten
    ISBN: 3-89336-294-0 (broschiert)
    ISBN: 3-89336-295-9 (CD)

Schriften des Forschungszentrums Jülich
Reihe Bibliothek / Library

11. **Bibliometric Analysis in Science and Research**
Applications, Benefits and Limitations
2nd Conference of the Central Library, 5 – 7 November 2003, Jülich, Germany
Conference Proceedings (2003), 242 pages
ISBN: 3-89336-334-3

12. **Bibliometrische Analysen – Daten, Fakten und Methoden**
Grundwissen Bibliometrie für Wissenschaftler, Wissenschaftsmanager, Forschungseinrichtungen und Hochschulen
von R. Ball, D. Tunger (2005), 81 Seiten
ISBN: 3-89336-383-1

13. **VIRUS – Sicher im Netz?**
2. Internationale Konferenz zur Virtuellen Bibliothek des Goethe-Instituts Brüssel
herausgegeben von R. Ball, C. Röpke, W. Vanderpijpen (2005), 137 Seiten mit beiliegender CD-ROM
ISBN: 3-89336-377-7

14. **Knowledge eXtended**
Die Kooperation von Wissenschaftlern, Bibliothekaren und IT-Spezialisten
3. Konferenz der Zentralbibliothek, 2. – 4. November 2005 Jülich
Vorträge und Poster (2005), 392 Seiten
ISBN: 3-89336-409-9

Die Konferenz wird unterstützt von:

Hauptsponsor:

Weitere Sponsoren:

- **Elsevier**
- **Harrassowitz Verlag**
- **ImageWare**
- **Institute of Physics Publishing**
- **ProQuest Information and Learning**
- **Swets Information Services Deutschland**

Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft

Band / Volume 14
ISBN 3-89336-409-9

Bibliothek
Library