Informationsqualität und Wissensgenerierung

3. DGI-Konferenz
66. Jahrestagung der DGI

Frankfurt am Main
8. bis 9. Mai 2014

Proceedings

herausgegeben von
Marlies Ockenfeld

Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V.

© Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI)

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über www.d-nb.de abrufbar.

Dieser Tagungsband enthält die Schriftfassungen aller zur 3. DGI-Konferenz und 66. Jahrestagung der DGI 2014 in Frankfurt am Main eingereichten und angenommenen Vorträge sowie Kurzfassungen der Keynotes und einiger Kurzvorträge.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen und dergleichen in diesem Tagungsband berechtigt nicht zu der Annahme, dass solche Namen ohne weiteres von jedermann benutzt werden dürfen. Vielmehr handelt es sich häufig um gesetzlich geschützte, eingetragene Warenzeichen, auch wenn sie nicht als solche gekennzeichnet sind.

© Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V.
Frankfurt am Main, 2014
Redaktion und Satz: Marlies Ockenfeld
Druck: Digital Print Group, Nürnberg

ISSN 1438-9290 (Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. DGI)
ISSN 1438-9282 (Jahrestagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis e.V. DGI)

ISBN 978-3-925474-73-6
EAN 9783925474736

Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis (DGI-Tagungen) Band 17

---

herausgegeben von Marlies Ockenfeld

3. DGI-Konferenz 2014

Informationsqualität und Wissensgenerierung

66. Jahrestagung mit Mitgliederversammlung

der Deutschen Gesellschaft für Informationswissenschaft
und Informationspraxis e.V. (DGI)

8. bis 9. März 2014
Frankfurt am Main

Veranstalter

Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis e.V. (DGI)
Windmühlstraße 3
60329 Frankfurt am Main
Telefon 069 430313
Telefax 069 4909096
E-Mail zentrale@dgi-info.de
Website www.dgi-info.de

Vorsitzende des Programmkomitees der 3. DGI-Konferenz 2014

Alexander Botte, Frankfurt am Main
Evelyn Dröge, Berlin

Vorwort

„Ich weiß nicht immer, wovon ich rede. Aber ich weiß, dass ich recht habe.“
(angeblich: Muhammad Ali)

Das Zitat von Ali zeigt: Man kann es sich mit dem Wissen also einfach machen! In der Regel aber wollen wir unser Wissen auf verlässliche Information gründen. Denn Information und Wissen werden im Allgemeinen als zwei aufeinander folgende Stufen der Erkenntnis angesehen.

Von der Information zum Wissen ist es manchmal nur ein Schritt: Wenn ich wissen will, wann der nächste Zug nach Frankfurt am Main fährt, reicht mir ein Blick in den Fahrplan. Aber Fahrpläne sind heutzutage in unterschiedlichen Formaten und auf unterschiedlichen Wegen zugänglich. Daher kann auch eine so einfache Anfrage informationsseitig einfacher oder aufwändiger gelöst werden. Dies gilt umso mehr bei komplexen Wissensfragen, die durch informationsmethodisches und -organisatorisches Know-how optimal bearbeitet werden können. Allein die rasanten Veränderungen durch neue innovative Verfahren und Werkzeuge in den offenen Informationsarchitekturen des WWW stellen hier eine kontinuierliche Herausforderung an den professionellen Umgang mit Informationsproblemen:

- Bieten neue Informations- und Kommunikationstechnologien tatsächlich auch einen qualitativen Mehrwert oder kompensieren z.B. automatisierte Anreicherungen von Daten ihre leichte Verfügbarkeit mit Schwächen in der Qualität?
- Wann sind der Aufbau und die Pflege eines umfangreichen Systems des Informationsmanagements sinnvoll und wann ist der Aufwand beim Aufbau größer als der Nutzen?
- Welche Kriterien und Maßstäbe sind angemessen, qualitative Aspekte zu bewerten?
- Können wir allgemeine Evaluationskriterien zulassen, um unterschiedlichen Informations-, Kommunikations- und Wissensquellen gerecht zu werden?

Die 3. DGI-Konferenz beleuchtet diese Fragestellungen aus der Perspektive repräsentativer Projekte und Ansätze, die zugleich wissenschaftlich fundiert und praxisorientiert sind. Wie es bei einem offenen Konferenzprogramm üblich ist, werden in den Vorträgen verschiedene thematische Felder und methodische Herangehensweisen vorgestellt, die jedoch alle unter dem Aspekt relevant sind, welche Bedeutung qualitative Faktoren oder neu eingeführte technische Komponenten für erfolgreiche Informationsinfrastrukturen haben.

Die Langvorträge sind in **vier thematische Sessions** gebündelt, die jeweils Schwerpunkte der Qualität moderner Informationsarchitekturen hervorheben:

Session 1 thematisiert unterschiedliche Aspekte der Recherche bzw. des Information Retrieval.

Session 2 nimmt sich der immer bedeutsamer werdenden Systeme der Informationsorganisation in größeren Betriebseinheiten an.

Session 3 befasst sich mit der Unterstützung des Wissensmanagements durch das Social Web.

Session 4 lenkt den Blick auf die Modellierung moderner Informationsinfrastrukturen.

Dazwischen bietet eine **Session mit Kurzvorträgen** Einblick in neue Verfahren, verfügbare Tools und aktuelle Projektvorhaben, die darauf abzielen, den Weg von der Information zum Wissen weiter zu verkürzen.

Es liegt auf der Hand, dass Informationsqualität der Schlüssel für die Zukunft der Informationsbranche ist. Auf einer **Podiumsdiskussion** wollen wir mit berufenen Professionellen

ergründen, wie Qualität ein perspektivreiches Programm für die Informationswissenschaft und die Informationspraxis werden kann.

Eingerahmt wird die Tagung durch zwei **Keynotes** von Andreas Blumauer und Reinhard Karger, die mit den Themen „Linked Data“ und dem „Internet der Dinge“ zwei immer noch visionäre Trends der Informationswelt aufgreifen, die nicht zuletzt unter Qualitätskriterien einer kritischen Betrachtung würdig sind.

Die Jahrestagung der DGI präsentiert sich 2014 zum dritten Mal als *DGI-Konferenz*.

Der Tagungsband enthält alle Beiträge der Konferenz, die aufgrund des Call for Papers eingereicht wurden und nach einer Begutachtung durch das Programmkomitee angenommen worden sind. Zu den beiden eingeladenen Hauptvorträgen und zum kurzfristig eingeladenen Vortrag über das Ontologiedesign für den Exzellenzcluster „Bild Wissen Gestaltung“ sind Kurzfassungen enthalten. Dies gilt auch für fünf der sieben Kurzvorträge.

Die Reihenfolge der Beiträge in diesem Band entspricht dem zeitlichen Verlauf der Vorträge während der Konferenz. Hinzukommt ein Literaturbericht zu Informationsanalyse, Informationsqualität und Business Intelligence, der vom Programmkomitee angenommen worden war, jedoch nicht als Vortrag gehalten werden konnte. Die Zusammenfassungen der Kurzbeiträge finden sich am Schluss. Ein **Personenverzeichnis** beschließt den Band. Es soll die Kontaktaufnahme zu Autoren, Moderatoren und Referenten erleichtern.

Programmkomitee und Redaktion freuen sich auf einen anregenden Informations- und Meinungsaustausch und wünschen allen, die in diesem Jahr nicht nach Frankfurt am Main kommen können, eine anregende Lektüre, die Lust auf den Besuch künftiger DGI-Tagungen weckt.

Darmstadt, den 10. April 2014

Alexander Botte, Evelyn Dröge, Marlies Ockenfeld

Inhaltsverzeichnis

Einführung 9

Alexander Botte, Evelyn Dröge & Marlies Ockenfeld

KEYNOTE

Linked Data – das Ende des Dokuments? 11

Andreas Blumauer

NUTZUNGSKONTEXTE UND RECHERCHE

Recherche-Unterstützung durch Ontologie-Visualisierung im EnArgus2.0-Projekt 15

Hermann Bense, Ulrich Schade, Frederike Ohrem & Lukas Sikorski

Patent Retrieval und Patent Mining: sind die Anforderungen eingelöst? 25

Julia Maria Struß, Daniela Becks, Thomas Mandl & Christa Womser-Hacker

Google Scholar versus Fachdatenbanken. Wie erfolgreich nutzen Studierende verschiedene Suchwerkzeuge? 37

Nikolas Leichner, Johannes Peter, Anne-Kathrin Mayer & Günter Krampen

INFORMATIONSORGANISATION UND QUALITÄT

Die Modellierung von Interdisziplinarität – Ontologiedesign für den Exzellenzcluster „Bild Wissen Gestaltung“ 47

Christian Stein

Konzeption einer Architektur für flexible Suche und Informationsvernetzung in der Fahrzeugentwicklung 49

Markus Zoier, Bernd Fachbach, Manfred Rosenberger, Nadja Marko, Selver Softic & Alexander Stocker

Von der Dokumentation zum organisationalen WissensPerformanceSystem 65

Klaus Mak, Christian Meurers, Johannes Göllner & Robert Woitsch

SOCIAL KNOWLEDGE MANAGEMENT

Social Media als Quelle berufsbezogener Information in der Unterrichtsvorbereitung 85

Angela Vorndran, Silvana Keutz, Daniela Olenik, Aida Raimbekova & Marc Rittbergera

The Tweet and the City: Comparing Twitter Activities in Informational World Cities 101

Thorsten Förster, Lennart Lamerz, Agnes Mainka & Isabella Peters

Enterprise 2.0 in Bibliotheken und Informationseinrichtungen – Neue Anforderungen an Wissensmanagement in Zeiten des Social Web 119

Sarah Baldin & Beate Renate Somorowsky

MODELLIERUNG UND INFORMATIONSINFRASTRUKTUREN

From ESE to EDM and Beyond: How Europeana Provides Access to its Cultural Heritage Objects 129

Steffen Hennicke, Evelyn Dröge, Violeta Trkulja & Julia Iwanowa

Fachinformationssysteme als Basis der Wissenschaftsforschung. Verwendungspotentiale infrastruktureller Dienste für ein indikatorenbasiertes Forschungsmonitoring am Beispiel der Bildungsforschung 141

Ute Sondergeld & Andreas Kempf

Entwicklung eines Informationssystems für Lehrpläne - die Curricula Workstation 153

Jessica Drechsler, Robert Strötgen & Esther Chen

KEYNOTE

Wenn Informationswelten und Weltwissen zusammenwachsen: Das Internet der Dinge erreicht die Welt der Gegenstände 161

Reinhard Karger

INFORMATIONSQUALITÄT UND BUSINESS INTELLIGENCE

Business Information Quality und organisationale Wissensgenerierung: Informationsanalyse, Informationsqualität und Business Intelligence 165

Ronald Billen

KURZVORTRÄGE INFORMATIONSQUALITÄT UND WISSENSGENERIERUNG

Datenqualität in der Deutschen Digitalen Bibliothek 164

Francesca Schulze

Semantische Online-Suchsysteme. Prozesse, Methoden und Komponenten 165

Andrea Horch

Hocheffiziente Tiefenerschließung multilingualer Textkollektionen mittels intelligentCAPTURE 186

Manfred Hauer

Deutsches Textarchiv Qualitätssicherung (DTAQ) 187

Frank Wiegand

Aufbau von Terminologien aus Dokumentenkorpora - wie ein wenig Mathematik und Sprachverarbeitung Tage und Wochen manueller Arbeit vermeiden hilft..... 188

Stefan Geißler

Mehrwert schöpfen - Metasuche und Analyse von Fachinformationen..... 189

Manfred Jaksch

Referenten, Moderatoren, Autoren..... 190

Notizen

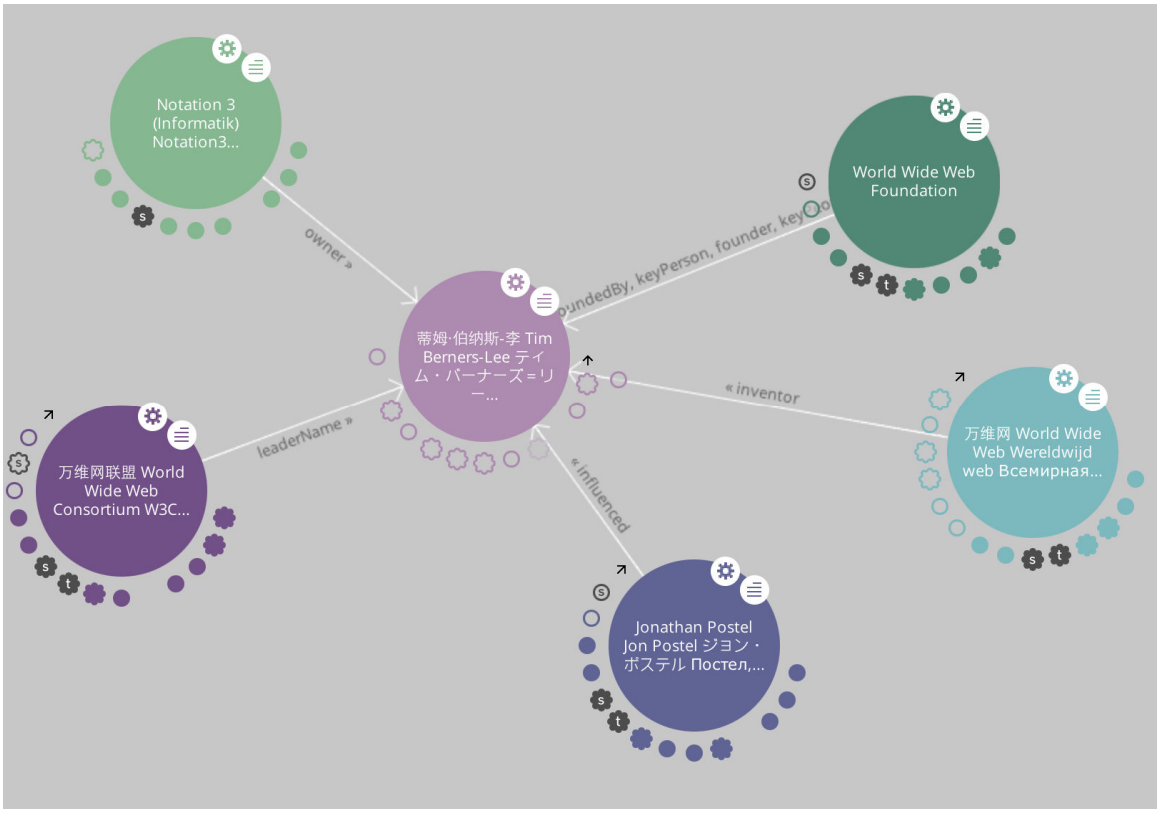
Linked Data – das Ende des Dokuments?

Andreas Blumauer

Dokumente waren bisher stets das zentrale Informationsträger- und Präsentationsmedium zur Wissensvermittlung. Doch dieses rein sequentielle, ausschließlich intellektuell interpretierbare Format reicht in einer zunehmend vernetzten und komplexer werdenden Weltoffensichtlich nicht mehr aus. Warum sich die Bedeutung des Dokuments, vor allem im Sinne einesSpeichermediums, demnächst drastisch ändern könnte, behandelt dieser Artikel.

Semantische Technologien und Semantic Web werden oft in einem Atemzug genannt, obwohl wesentliche Unterschiede bestehen: Geht es in beiden Fällen darum, Informationen *und* ihre Bedeutung verarbeiten zu können, so dreht sich im Semantic Web alles um die Frage, wie Entitäten (Produkte, Organisationen, Orte, etc.) sinnvoll miteinander verknüpft werden können. Die zugrundeliegenden ‚Linked-Data-Technologien' setzen dabei am Paradigma der größtmöglichen *Interoperabilität* an und verwenden im Kern so genannte *Wissensgraphen*.

Zusätzlich zum allgemein bekannten ‚Web of Documents', dessen wesentliches Merkmal Hypertext ist, entwickelt sich also ein ‚Web of Data' (Semantic Web), in dem nicht mehr das Dokument im Mittelpunkt steht, sondern vielmehr Entitäten unterschiedlicher Kategorien, ihre verschiedenen Bezeichnungen (in unterschiedlichen Sprachen) und ihre Relationen zueinander.



Ausschnitt aus einem Wissensgraphen zu Tim Berners-Lee.

Der Grundstein für das Web of Data wurde durch die Linked-Data-Initiative [1] von Tim Berners-Lee gelegt, der ein einfaches Framework, bestehend aus vier Regeln, entwickelt hat, um damit eine weltweite, verteilte Datenbank, eben das Web of Data realisieren zu können. Unter Berücksichtigung der Linked-Data-Prinzipien wurde daraufhin als Nukleus für ein Web of Data die DBpedia [2] veröffentlicht, die inzwischen in 119 Sprachen verfügbar ist. Als ‚semantische Schwester' der Wikipedia bildet sie den Kern der stetig wachsenden ‚Linked Open Data Cloud' (LOD Cloud), einem gigantischen Wissensgraphen, der nun allmählich in Organisationen eindringt und dort um unternehmensspezifisches Wissen erweitert wird.

Überblick über Funktionsweise von Linked-Data-Graphen

Die Basis aller Linked-Data-Graphen bilden Uniform Resource Identifier (URIs), die sämtliche Entitäten des Wissensgraphen eindeutig identifizierbar machen: Beispielsweise lässt sich die ‚Semantic Web Company' im Web of Data mit der URI http://dbpedia.org/resource/Semantic_Web_Company eindeutig referenzieren. Diese Entitäten-orientierte Sichtweise auf Informationen in Kombination mit automatischen Text-Extraktionsverfahren ermöglicht es, Informationsbestände und Dokumente mit Metadaten anzureichern, die quellenübergreifend verstanden werden. Semantische Metadaten werden also konsequent vom Inhalt entkoppelt verarbeitet.

Dieses Grundprinzip macht auch Google für sich nutzbar, indem auf Basis des Google Knowledge Graphs [3] Webinhalte indexiert und verknüpft werden. Damit können beliebige unstrukturierte Informationen als semantischer Graph repräsentiert werden.Für jede Entität werden auf Ebene des Wissensgraphen zunächst Relationen wie z.B. ‚Andreas Blumauer *ist CEO von* Semantic Web Company' oder ‚Semantic Web Company *gehört zur Kategorie* österreichische Softwareunternehmen' hinterlegt. In weiterer Folge können Entitäten mittels voll- oder halbautomatischer Verfahren mit Dokumenten aller Art, z.B. Nachrichten angereichert werden. Daraus ergeben sich so genannte Topic- oder Landing Pages in Form dynamischer Mashups.

Semantic Web in Unternehmen: Einsatzszenarien und Anwendungsfälle

Die BBC als europäisches Leitunternehmen hat 2008 schließlich mit BBC Music [4] das erste kommerzielle Projekt vorgestellt, das unter Verwendung von Linked Data aufzeigt, welche neuartigen Verwertungsstrategien für Medienunternehmen mit Hilfe des Semantic Web möglich werden. Es folgten zahlreiche weitere Medienunternehmen wie Wolters Kluwer oder Elsevier, aber auch Betriebe aus Branchen wie der Automobilindustrie, der Pharmaindustrie oder der öffentlichen Verwaltung [5], die Linked-Data-Technologien intern und extern für verschiedene Aufgaben des Informationsmanagements einsetzen. Insbesondere öffentliche Einrichtungen wie Ordnance Survey (UK), die Europäische Union, die Weltbank oder Bibliotheken wie die Deutsche Nationalbibliothek tragen immer mehr zur Verbreitung von Daten auf Basis von Linked-Data-Standards bei. Das Semantic Web hat also begonnen, Einzug in diverse Branchen zu halten.

Nun können drei grundlegende Szenarien für den unternehmerischen Einsatz von Linked Data unterschieden werden:

1. Daten in die Linked Data Cloud publizieren
   Das Unternehmen publiziert eigene Daten und Inhalte in die Linked Data Cloud und erschließt sich damit neue Distributionswege für digitale Assets.

2. Linked Data als Datenintegrationsprinzip anwenden
   Das Unternehmen verwendet die Linked-Data-Prinzipien und Semantic-Web-Technologien intern, um Datenintegration und Mashups(z.B. für ein Wissensportal) zu realisieren bzw. neue Möglichkeiten einer semantischen Suche zu erschließen.
3. Daten aus der Linked Data Cloud einbinden
   Das Unternehmen konsumiert Daten aus der Linked Data Cloud, um damit z.B. interne Datenbanken oder Inhalte anzureichern.

Fazit

Interne und externe Inhalte sinnvoll und kostenschonend zu verknüpfen, kann nur in einem interoperablen Framework wie dem Semantic Web gelingen. Der Einsatz von Semantic-Web-Technologien im Unternehmen kann daher aus zwei Perspektiven betrachtet werden: aus Sicht der Content-Verwertung und aus Sicht der Prozessunterstützung.

Im Rahmen der Content-Verwertung dienen Semantic-Web-Standards und Methoden der einheitlichen Strukturierung und Anreicherung von Datenbeständen. Sie ermöglichen es, ähnliche oder verwandte Objekte miteinander in Beziehung zu setzen und besser auffindbar zu machen. Bei der Prozessunterstützung nutzen unterschiedliche Services die wohl-strukturierte Datenbasis, um einerseits bestehende Dienste zu verbessern oder neue hervor-zubringen, was häufig unter dem Thema „betriebliches Wissensmanagement“ subsumiert wird. Im Zentrum der aktuellen Entwicklung stehen zwar noch immer Such- und Empfehlungsdienste, die noch stark an der klassischen Dokumentphilosophie haften, jedoch findet allmählich eine Transformation hin zum ‚Web of Data‘ auch innerhalb von Unternehmensgrenzen statt.

Literatur

[1] http://www.w3.org/DesignIssues/LinkedData.html

[2] http://dbpedia.org/

[3] http://www.google.com/insidesearch/features/search/knowledge.html

[4] http://www.bbc.co.uk/music

[5] http://www.w3.org/2001/sw/sweo/public/UseCases/

Notizen

Recherche-Unterstützung durch Ontologie-Visualisierung im EnArgus2.0-Projekt

Hermann Bense, Dortmund, Ulrich Schade, Frederike Ohrem & Lukas Sikorski

Die Energiepolitik und die Energieforschung erfordern in zunehmendem Maße eine hohe Transparenz. In den EnArgus-Projekten wird daher ein Informationssystem entwickelt, das in intuitiver Form Abfragen über die öffentlich geförderten Projekte der Energieforschung unterstützt. Im vorliegenden Artikel wird der mit EnArgus verfolgte Ansatz aufgezeigt. In diesem Zusammenhang wird erläutert, wie man bei einem Suchvorgang die durch die Fachontologie zur Verfügung gestellten Informationen nutzen und wie man mit einer guten Visualisierungskomponente das System evaluieren kann.

Einführung

Das Projekt EnArgus[1] wird vom Bundesministerium für Wirtschaft und Energie aufgrund eines Beschlusses des deutschen Bundestags gefördert. Mit dem Projekt wird das Ziel verfolgt, staatliche Förderpolitik im Bereich der Energieforschung transparenter zu gestalten und die Bewertung von Technologieentwicklungen zu erleichtern (vgl. hierzu das 6. Energieforschungsprogramm der Bundesregierung von 2011). Um dieses Ziel zu erreichen, kooperieren in Anlehnung an die Erstellung des Berichts „Energietechnologie 2050“ (vgl. Wietschel et al., 2010) bei EnArgus Partner aus den Bereichen Energieforschung und Informatik. In der ersten Phase des Projektes (Juli 2011 bis Juni 2013) waren dies das Fraunhofer-Institut für Angewandte Informationstechnik (FIT), das Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE), das Fraunhofer-Institut für System- und Innovationsforschung (ISI), das Fraunhofer-Institut für Umwelt-, Sicherheits- und Energietechnik (UMSICHT), der Lehrstuhl Energiesysteme und Energiewirtschaft (LEE) der Ruhr-Universität Bochum, das Institut für Energie- und Klimaforschung, Abteilung Systemforschung und Technologische Entwicklung (IEK-STE), des Forschungszentrums Jülich und die OrbiTeam Software GmbH (Bonn). An der zweiten, aktuellen Projektphase (bis Juni 2016) sind nach Ausscheiden des IEK-STE zusätzlich das IWAR der TU Darmstadt, die Materialprüfungsanstalt (MPA) der Universität Stuttgart, das Zentrum für Beratungssysteme in der Technik, Dortmund e.V. (ZEDO) und die Dortmunder Verlagsgesellschaft für digitales Publizieren mbh (bense.com) beteiligt. Die Gesamtkoordination hat seitens der Projektpartner das Fraunhofer FIT unter Begleitung des Projektträger Jülich.

In EnArgus wird ein zentrales Informationssystem für Energieforschungsvorhaben aus dem Bereich der Förderung durch Bund und Länder entwickelt. Dadurch erhalten Fachleute, aber auch die interessierte Öffentlichkeit, einen einheitlichen und zentralen Zugang zu Informationen über Energieforschung in der Bundesrepublik Deutschland. In der ersten Projektphase (Juli 2011 bis Juni 2013) wurde ein Demonstrator erstellt und die Nutzbarkeit des Gesamtsystems nachgewiesen. Der Demonstrator deckt im Schwerpunkt die Themen elektrische Energiespeicher, erneuerbare Energien, CCS und Energieeffizienz in der Industrie sowie Brennstoffzellen ab. In der zweiten Projektphase (Juli 2013 bis Juni 2016) wird das System thematisch um weitere Themen der Energieforschung erweitert. Das

[1] Das Projekt EnArgus wird durch das Bundesministerium für Wirtschaft und Energie unter dem Förderkennzeichen 03ET4010A gefördert.

System nutzt die offiziellen Datenbanken zur Forschungsförderung. Bei inhaltlichen Abfragen (Beispiel: Wie hoch war die Forschungsförderung zur Windkraft in NRW zwischen 2001 und 2010?) ohne Nutzung des Systems werden mittels Stichwortsuchen (im Beispiel wäre das die Suche nach dem Stichwort „Windkraft“) die Datenbanken direkt ausgewertet. Dieses Vorgehen liefert nicht alle einschlägigen Projekte, so dass die Beantwortung der Frage nur unzureichend gelingt. Um diesem Problem entgegenzuwirken, nutzt das EnArgus-System eine Fachontologie, in der Wissen aus den Bereichen Energie, Energieforschung und Energieforschungsförderung hinterlegt ist. In der Ontologie ist das Fachwissen formal repräsentiert, so dass es im Gesamtsystem genutzt werden kann (zum generellen Nutzen von Ontologien als Mittel zur formalen Repräsentation von Wissen vgl. etwa Staab & Studer, 2004). Durch Nutzung der durch die Ontologie bereitgestellten Informationen kann ein Nutzer die Recherche nach Projekten präzisieren, so dass genauere und mehr relevante Ergebnisse gefunden werden. Die Ontologie bietet zu dem vom Nutzer eingegebenen Suchbegriff in einer Auswahlliste diejenigen Begriffe an, die in der Ontologie mit dem Suchbegriff verbunden sind. Dies sind Synonyme, alternative Bezeichner (Label), Hyponyme, das Hyperonym und weitere semantisch verwandte Begriffe, insbesondere verwandte Begriffe, die sich über die Teil-Ganzes-Relation ergeben. Der Nutzer kann dann aus diesen Begriffen diejenigen auswählen, die in der Suche mit berücksichtigt werden sollen. Die umgesetzte ontologie-basierte Suche hat also den Vorteil, dass der Nutzer eine genaue Kontrolle über die semantischen Aspekte hat, die bei der Suche berücksichtigt werden sollen.

Da die Wissenszusammenhänge in der Ontologie relativ komplexer Natur sind, war und ist es nicht einfach sie aufzubauen bzw. sie korrekt und konsistent zu erweitern. Der Aufbau der Ontologie kann nur durch eine enge Zusammenarbeit zwischen den Fachexperten der Energieforschung und den Experten für Informationsverarbeitung erfolgen. Die Fachexperten verfügen über das relevante Fachwissen, und die Experten für Informationsverarbeitung wissen, wie eine Ontologie zu strukturieren und aufzubauen ist. Um diese Zusammenarbeit zu unterstützen und um zu gewährleisten, dass die Fachexperten die in der Ontologie abgelegte Information adäquat auf ihre Korrektheit hin überprüfen können, muss die Ontologie visualisiert werden können. In der zweiten Phase des Projektes wird daher die Visualisierungskomponente neu entwickelt. In der ersten Projektphase wurde eine Ontologie-Visualisierung eingebunden, die die Taxonomie der Ontologie als Baum zeigte (siehe Abbildung 1). In einer vergleichenden Evaluation, die mit den Fachexperten zu Beginn der zweiten Projektphase durchgeführt wurde, ergab sich, dass stattdessen die Visualisierung mittels so genannter Hyperbolischer Bäume (Hypertrees) bevorzugt wird. Die „Hypertrees“ (vgl. Abbildung 3 für ein einfaches Beispiel) erlauben ein grafisches Navigieren in der Ontologie, das als intuitiver und daher „leichter“ erachtet wurde. Auch das Auffinden von Fehlrepräsentationen, der wichtigste Nutzen der Visualisierung, wurde verbessert, was zum Teil daran liegt, dass in den „Hypertrees“ auch die Relationen angezeigt werden, die nicht taxonomischer Art (Ober- und Unterbegriffe) sind. Die Unterscheidung der Relationen zwischen den Begriffen wird in den gewählten „Hypertrees“ zudem farblich repräsentiert. Auch dies unterstützt die Evaluation der Ontologie durch die Fachexperten.

Ausgangspunkt einer jeden Evaluation ist ein Text, den der entsprechende Fachexperte verfasst hat. Die Gesamtheit dieser Texte ist im Gesamtsystem als Wiki hinterlegt, welches auch Nutzer einsehen können, um etwa Begriffe, die aus der Ontologie für die Suche vorgeschlagen werden, nachzuschlagen. Auf der Grundlage dieser Texte wird die Ontologie durch die Experten der Informationsverarbeitung gefüllt, so dass der evaluierende Fachexperte dann durch einen Vergleich zwischen seinem Text und der visualisierten Ontologie Punkt für Punkt überprüfen kann, ob die Ontologie inhaltlich korrekt aufgebaut ist. Dabei

wird die Navigation durch die Ontologie extensiv genutzt, wobei die farblichen Kodierungen (s. unten für Details) die Überprüfung erleichtern.

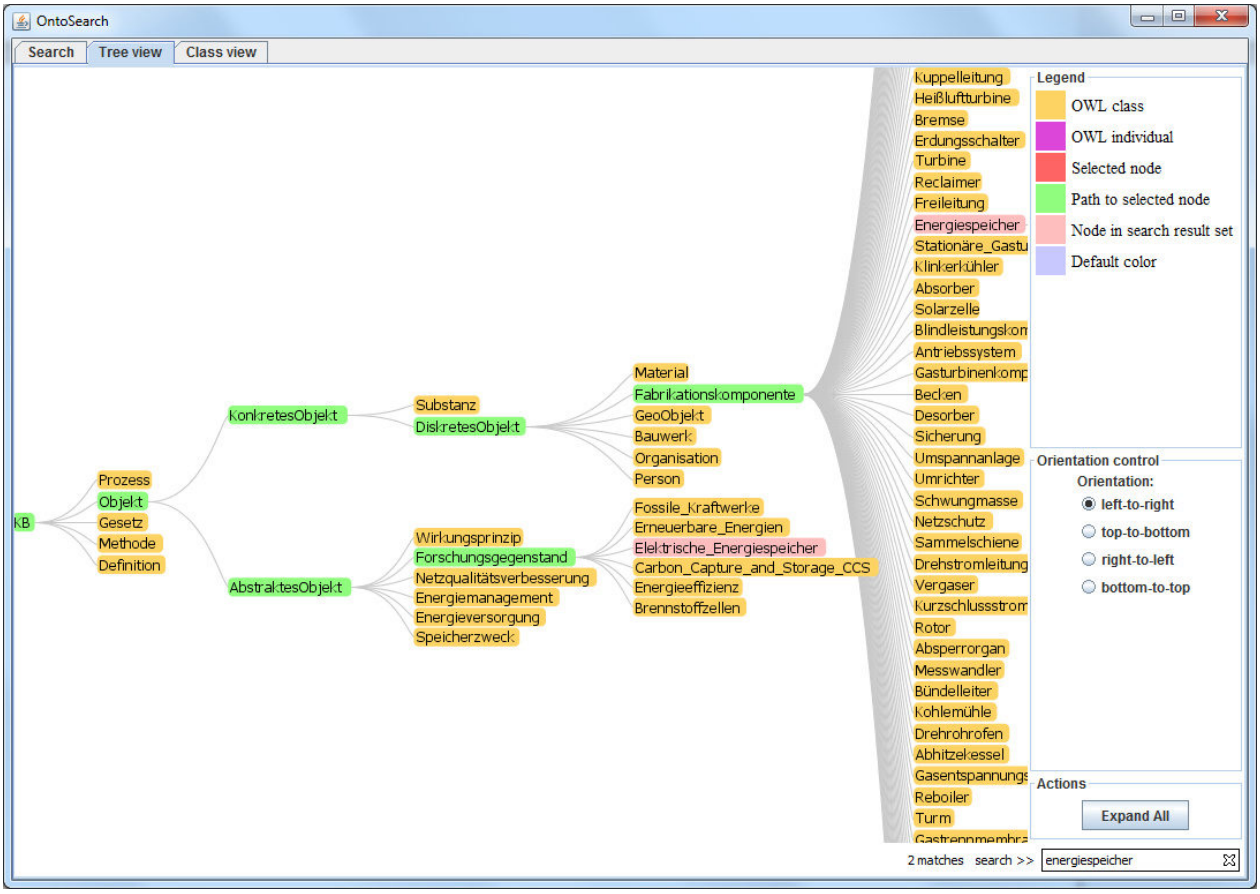


Abbildung 1: Taxonomie im Ontologie-Viewer der ersten Projektphase.

Im Folgenden wird zunächst beschrieben, wie die Ontologie für die angestrebte Visualisierung aufbereitet wird. Im Anschluss daran wird in mehreren Abschnitten, geordnet nach steigender Komplexität möglicher Anfragen, aufgezeigt, wie die Visualisierung über die Evaluation hinaus auch als Recherche-Unterstützung genutzt werden kann. Dies schließt eine nähere Betrachtung der hyperbolischen Bäume ein, da diese für die Visualisierung genutzt werden. Der Beitrag endet dann mit einer Zusammenfassung.

Aufbereitung und Anreicherung der Ontologie für die Visualisierung

In der mittels Protégé entwickelten Fachontologie wird Wissen formal repräsentiert, wobei zwischen Klassen und Instanzen („EnArgus“ ist beispielsweise eine Instanz der Klasse der Projekte) sowie zwischen Attributen und Relationen („Fördersumme“ ist ein Attribut der Klasse der Projekte, welches für die Projektinstanzen mit einem numerischen Wert belegt wird; „wird_gefördert_durch“ ist eine Relation, da ihr Wertebereich eine andere Klasse ist, hier die Klasse der projektfördernden Institutionen) unterschieden. In der Fachontologie selbst werden diese Unterscheidungen in der Namensgebung nicht getroffen (Bense & Haarmann, 2013). Um die Unterschiede aber durch die Visualisierung aufzeigen zu können, werden die Ontologieobjekte zunächst mit Bezeichnern wie „SubclassOf“, „DataProperty“, „ObjectProperty“ etc. versehen. Danach wird die gesamte Fachontologie über einen OWL-Importer in das KB-CMS (Knowledge-Based Content-Management-System) Ontology4

übernommen. Hier werden die Ontologieobjekte nach speziellen Konventionen umkodiert, so dass sie dann grafisch unterschiedlich dargestellt werden können:

- Klassen ⇨ rote Ellipsen
- Attribute ⇨ dunkelgrüne Achtecke
- Attributwerte ⇨ hellgrüne Quadrate
- Relationen ⇨ blaue Verbindungslinien
- ähnliche Begriffe ⇨ orange Kreise

Ontologiegestützte Suche

Die ontologiegestützte Suche ermöglicht Abfragen, die weit über die Möglichkeiten einer reinen Volltextsuche hinausgehen. Dies soll an einigen Beispielen erläutert werden.

Eine Google-Suche nach „Energiespeicher“ liefert im Wesentlichen Ergebnisse zurück, in denen der Begriff „Energiespeicher“ als solches vorkommt. Eine Suche nach dem entsprechenden englischen Begriff „energy store“ (oder „energy storage“) würde Ergebnisse zurückliefern, in denen die Begriffe „energy“ UND „store“ vorkommen. In beiden Sprachvarianten werden orthografische Fehler dabei in gewissen Grenzen toleriert bzw. korrigiert. Bei einer solchen konventionellen Volltextsuche werden in der Regel aber nicht die Dokumente gefunden, die semantisch ähnliche Begriffe beinhalten, sondern nur die Dokumente, die den Suchbegriff selbst beinhalten. Eine Suche nach „Energiespeicher“ verfehlt also Seiten in anderen Sprachen als Deutsch, etwa englische Seiten über „energy storage“, ebenso wie Seiten, die sich ausschließlich auf spezifische Energiespeicher, etwa „Batteriespeichersysteme“ beziehen. Dies ist unter Umständen genau so gewollt und kann im Bedarfsfall durch weitere Abfragen abgefangen werden. Im Kontext des EnArgus-Projektes ist dies aber problematisch, weil hier in aller Regel nicht nach einer spezifischen Seite oder einer spezifischen Information gesucht wird, sondern alle dem Suchbegriff entsprechenden Projekte gefunden werden sollen, über deren Gesamtheit dann eine statistische Auswertung erfolgt. Zusätzlich zu der Möglichkeit, zu vorgegebenen Suchbegriffen semantisch ähnliche Begriffe in die Suche miteinzubeziehen, bietet eine (Fach-)Ontologie auch die Option, bestimmte Fragen direkt aus dem in der Ontologie hinterlegten Wissen heraus zu beantworten, sofern der Suchbegriff in der Ontologie hinterlegt ist. Solche Fragen sind etwa

- die Frage nach der Oberklasse (Was ist ein Energiespeicher?),
- die Frage nach Unterklassen (Welche Arten von Energiespeichern gibt es?),
- die Frage nach alternativen Bezeichnungen (Synonyme, Labels, Übersetzungen),
- die Frage nach Attributen (Welche Merkmale zeichnen einen Energiespeicher aus?),
- die Frage nach dem Einsatzzweck (Wozu dient ein Energiespeicher?) und
- die Frage nach Bestandteilen (Woraus besteht ein Energiespeicher?).

Abbildung 2 zeigt für das Beispiel des Suchbegriffs „Energiespeicher“ eine mit der Bibliothek graphviz (Gansner, Koutsofios & North, 2009) erstellte statische Visualisierung zu einigen der aufgeführten Fragen, und wie sich diese auf einen Blick schon aus der Visualisierung heraus beantworten lassen. Der Suchbegriff „Energiespeicher“ wird in der linken Hälfte angezeigt. Die Verortung des Begriffs in der Begriffstaxonomie der Ontologie

– dies ist der durch die Relationen „ist Oberklasse von“ (<>is) bzw. „ist Unterklasse von“ (<>SubClass) aufgespannte Baum – ergibt sich dadurch, mit welchen Klassen (lies: mit welchen anderen Begriffen) das Konzept „Energiespeicher“ verbunden ist. Klassen sind nach unserer Konvention als rote Ellipsen zu erkennen, so dass sich aus Abbildung 2 ergibt, dass „Energiespeicher“ nach der Fachontologie eine „Fabrikationskomponente“ ist und dass als Arten von Energiespeichern in der Fachontologie „Druckluftspeicher“, „Kurzzeitspeicher“, „Batterien“ und „Wasserstoffspeicher“ repräsentiert sind. Des Weiteren ist erkennbar, dass nach der Ontologie die Klasse „Energiespeicher“ über die Attribute (nach unserer Konvention als dunkelgrüne Achtecke visualisiert) „Leistung“, „Energiedichte“, „Investition“, „Speicherdauer“ und „Energiemenge“ verfügt. Falls nun die Fachontologie überprüft wird, ergeben sich bei dem Blick auf den in Abbildung 2 visualisierten Teil unmittelbar die Fragen, ob der Begriff „Kurzzeitspeicher“ und ob das Attribut „Investition“ korrekt eingeordnet wurde. Eventuell sollte „Energiespeicher“ besser in „Kurzzeitenergiespeicher“ und „Langzeitenergiespeicher“ unterteilt werden, wobei dann „Druckluftspeicher“, „Batterien“ und „Wasserstoffspeicher“ unter „Langzeitenergiespeicher“ zu subsumieren wären. Des Weiteren könnte das Attribut „Investition“ zur Oberklasse „Fabrikationskomponente“ verschoben werden, so dass „Energiespeicher“ dieses Attribut erben würde. An diesem Beispiel ist gut zu erkennen, dass die Visualisierung dabei hilft, mögliche Schwächen der Ontologie auszumachen.

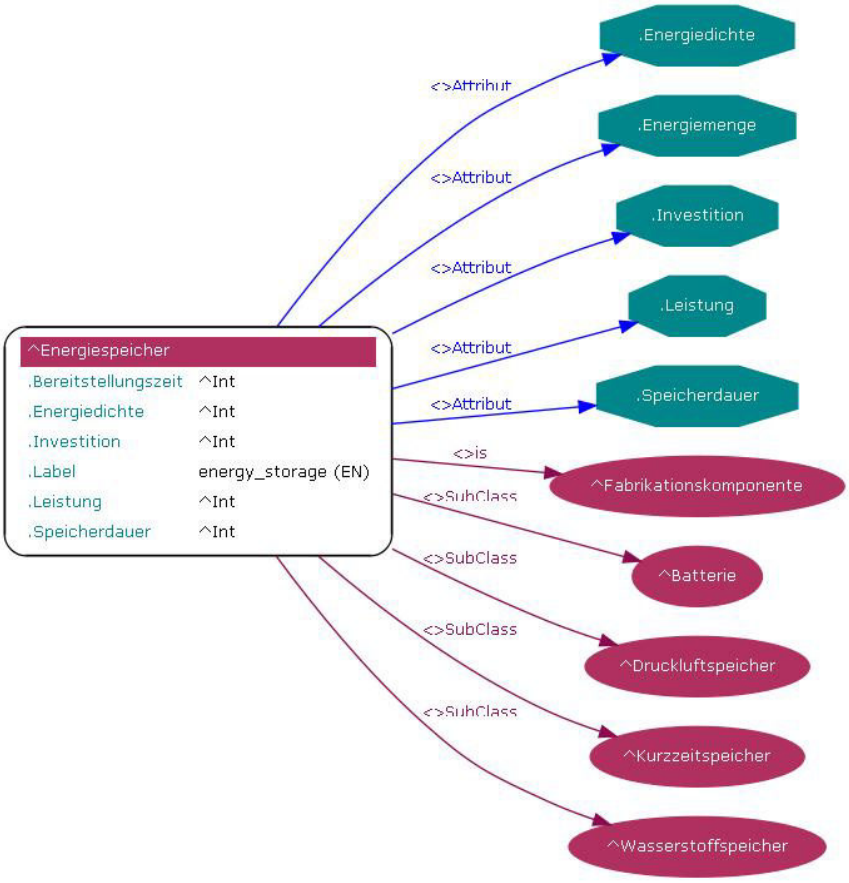


Abbildung 2: Beispiel des Suchbegriffs „Energiespeicher“.

Erweiterte Suche

Eine Sucherweiterung ist in zweierlei Hinsicht möglich. Lässt man bei der Suche zu, dass auch Ontologieobjekte gefunden werden, in deren Bezeichnung die Begriffe „Energie“ und „Speicher“ vorkommen, die aber von dem exakten Begriff „Energiespeicher“ abweichen können, so erhält man zusätzlich als Ergebnisse, die ontologische Repräsentation des Fakts, dass elektrische Energiespeicher einen Gegenstand der Energieforschung darstellen:

- (^Elektrische_Energiespeicher, <>is, ^Forschungsgegenstand)
- (^Forschungsgegenstand, <>SubClass, ^Elektrische_Energiespeicher)

Darüber hinaus gibt es die Möglichkeit, im “Hypertree“ einfache, datenbankähnliche Abfragen durchzuführen. Z.B. liefert die Suche nach „Zweck+Verbesserung“ folgende Ergebnisse für Komponenten, die die Netzwerkqualität verbessern:

- (^Lithium-Ionen-Batterie, <>Einsatz**zweck**, ^Netzwerkqualität**verbesserung**)
- (^Supraleitende_Spule, <>Einsatz**zweck**, ^Netzwerkqualität**verbesserung**)
- (^Doppelschichtkondensator, <>Einsatz**zweck**, ^Netzwerkqualität**verbesserung**)

Darstellung und Funktionsweise der Hyperbolischen Bäume (Hyperbolic Trees, Hypertrees)

Hyperbolische Bäume sind grafische Repräsentationen von Baumstrukturen (für ein zum Teil expandiertes Beispiel mit Suchradius 1 s. Abbildung 3). Der Ausgangsbegriff (die Wurzel) wird mittig dargestellt und die nächstliegenden Äste und Verzweigungen werden gleichmäßig, kreisförmig um den Mittelpunkt angeordnet. Durch Anklicken eines Begriffes erfolgt die Neufokussierung. Der Suchradius kann jeweils neu vorgegeben werden und bestimmt die Schrittweite, in der Konzepte ausgehend vom aktuell fokussierten Begriff gefunden werden. Bei einem Suchradius von 2 werden also nicht nur die unmittelbar verbundenen Konzepte gefunden, sondern auch die Konzepte ihrer ausgehenden Verbindungen. Die dabei verwendete Javascript-Implementierung greift direkt auf die ontologischen Relationen in der mySQL-Datenbank zu, in der die Ontologie gespeichert wurde. Selbst bei der Anzeige mehrerer hundert Kanten aus der Ontologie erfolgt die Anzeige so praktisch in Realzeit. Im Vergleich zu klassischen Triplestore-Ansätzen wurde die Basisrelation, mit der die Wissenselemente repräsentiert werden, um eine vierte Dimension erweitert. Über diese Dimension werden zeitliche Aspekte repräsentiert. Die zeitliche Gültigkeit einer Relation, beispielsweise (>FKIE, <>Projekt, >Enargus2, (2013, 2016)), lässt sich dadurch mit nur einem Datenbankzugriff abfragen. Die Speicherung der Ontologien in SQL-Datenbanken hat zudem den Vorteil, dass sie kein spezielles Know-how in anderen Ontologie-Abfragesprachen wie SPARQL voraussetzt und damit einer wesentlich größeren Entwicklergemeinde zugänglich ist.

Im Vergleich mit statischen Visualisierungsmethoden wie graphviz (s. Abbildung 2) bieten Hyperbolic Trees den Vorteil, dass sie interaktiv nutzbar sind. Die Darstellung ähnelt der von Mindmaps (Buzan & Buzan, 2002). Zusätzlich können die Bedeutungen der Verbindungen zwischen zwei Elementen als Legende angezeigt werden (s. Abbildung 4). Aus der Legende heraus können ebenfalls andere Elemente der Ontologie direkt in den Fokus der Visualisierung gesetzt werden. Die Platzausnutzung bei Verwendung von Hyperbolic Trees ist wesentlich effizienter als bei der linearen Darstellung von Baumansichten (s. Abbildung 1). In der linearen Darstellung können die Elemente nicht gleichmäßig im Radius um ein Element verteilt werden. Vielmehr werden sie dort in einer Ebene übereinander oder

nebeneinander dargestellt, so dass bei einer größeren Anzahl von darzustellenden Elementen deutlich mehr „gescrollt“ werden muss.

Hyperbolic Trees sind ebenfalls ein wichtiges Instrument bei der Entwicklung von Ontologien. So lässt mit ihnen sich sehr schnell überprüfen, ob die Wissenselemente der Ontologie richtig angeordnet sind (Unterklasse- / Oberklassebeziehung) und ob die Wissenselemente über die korrekten Relationen miteinander in Beziehung gesetzt wurden. Durch die Verwendung unterschiedlicher Symbole für Klassen, Attribute, Relationen usw. ist ebenfalls eine schnelle Überprüfung auf die richtige Verwendung der Typen möglich.

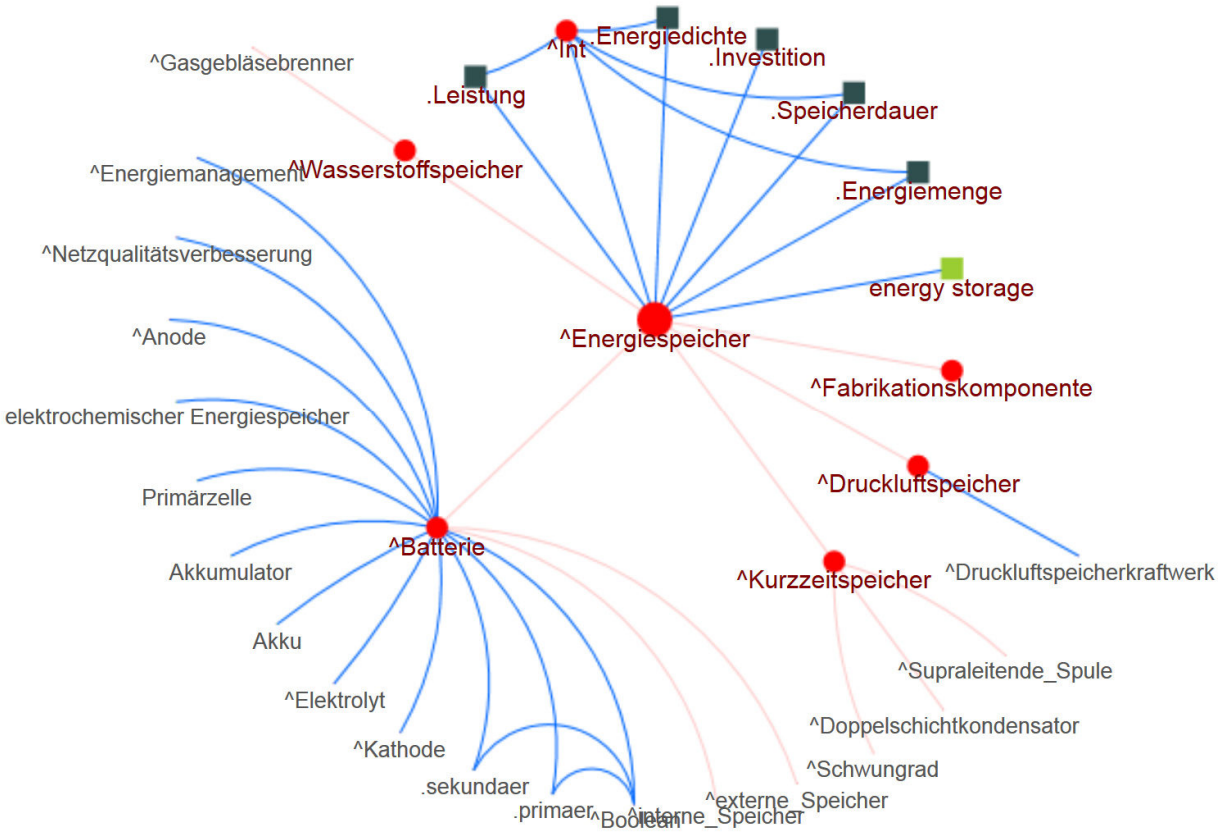


Abbildung 3: Hyperbolischer Baum mit Klasse „^Energiespeicher“ im Fokus, Suchradius 1, teilexpandiert.

Erweiterte Suche mit vergrößertem Suchradius

Durch die Nutzung vergrößerter Suchradien werden Suchergebnisse einbezogen, die einen entsprechend größeren semantischen Abstand zum Ausgangspunkt der Suche haben. Das folgende Beispiel (Abbildung 4) zeigt alle Ontologieobjekte, die von „Energiespeicher“ über einen Suchradius von 2 erreicht werden können. Über solche Bilder können Fachexperten die Ontologie aus einer umfassenderen Perspektive betrachten und evaluieren.

Würde man die gleiche Anfrage nach „Energiespeicher“ an eine klassische Suchmaschine stellen, so würden als Suchergebnis in der Regel nur Fundstellen gelistet, die den Begriff als Text enthalten. Das hier vorgestellte Verfahren schlägt über die Visualisierung mittels Hyperbolic Trees darüber hinaus Begriffe zur semantische Sucherweiterung vor, die u.a. folgende semantische Relationen zwischen Ausgangsbegriff und Begriffsvorschlägen

abdecken: „ist-ein(e) (Unterklasse/Oberklasse) „besteht-aus“, „ist-Bestandteil-von“, „ist-alternativer-Bezeichner-von (Label)“, „steht-in-Beziehung-zu (z.B. findet-statt-in) usw.

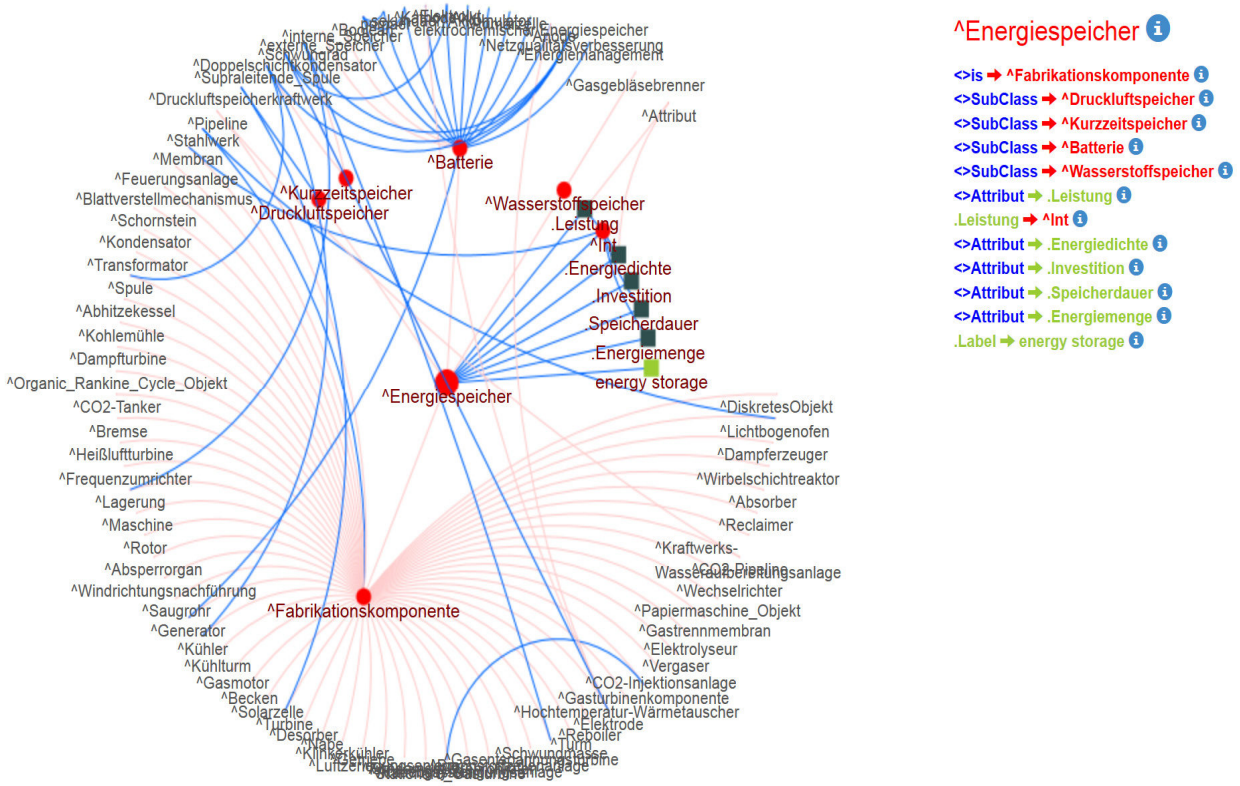


Abbildung 4: Hyperbolischer Baum mit „Energie Speicher“ im Fokus und einem Suchradius von 2.

Zusammenfassung

In dem EnArgus-Projekt wird ein Informationssystem entwickelt, auf welches über ein Internetportal zugegriffen werden kann. Dabei ist eine der Kernfunktionen des Informationssystems die ontologiebasierte Suche in Datenbanken, in denen öffentlich geförderte Vorhaben zur Energieforschung abgelegt sind. Der Benutzer bekommt, im Vergleich zur bekannten Google-Suche, ergänzende Informationen angeboten, die im Kontext seiner Anfrage relevant sind und die daraufhin in die Suche mit eingebunden werden können. Bei einer Suche nach dem Begriff „Energiespeicher“ lassen sich so etwa auch Fragen wie „Was ist ein Energiespeicher?“ oder „Welche Arten von Energiespeichern gibt es?“ beantworten. Um einen besseren Überblick über die ontologischen Klassen und Relationen zu erhalten, kann zudem eine Suchunterstützung mittels „Hypertrees“ genutzt werden, die die konventionelle textbasierte Recherche ergänzt. Das Navigieren innerhalb dieser Hypertrees ist intuitiv, schnell und erfordert in der Regel keine weitere textuelle Eingabe. Der vorliegende Artikel erläutert diese Suchunterstützung und zeigt an einer Beispielsuche auf, wie mögliche Schwachstellen in der Ontologie mittels der Visualisierung anhand der Hypertrees erkennbar werden. Durch diese Suchunterstützung kann also zum einen eine wesentlich bessere Qualität der Suchergebnisse erreicht werden. Zum anderen können Ergebnisse erschlossen werden, die bei rein textueller Suche nicht so schnell

erkennbar sind, was die stets notwendige Evaluation von Ontologien durch Fachexperten erheblich vereinfacht und damit unterstützt.

Literatur

Bense, H. & Haarmann, B. (2013). *A Richer Notation for the Representation of Ontological Knowledge.* International Conference on Computer, Electrical, and Systems Sciences, and Engineering (ICCESSE 2013), London.

Buzan, T. & Buzan, B. (2002). *Das Mind-Map-Buch. Die beste Methode zur Steigerung Ihres geistigen Potentials*. Landsberg: MVG, 5. Auflage.

Gansner, E.R., Koutsofios, E. & North, S. (2009). *Drawing graphs with dot.* (Download: http://www.graphviz.org/pdf/dotguide.pdf)

Ohrem, F., Sikorski, L. & Haarmann, B. (2013). *EnArgus: A Knowledge-based Search Application for Energy Research Projects.* Proceedings of the World Academy of Science, Engineering & Technology, Rome.

Staab, S. & Studer, R. (Eds.) (2004). *Handbook on Ontologies in Information Systems.* Berlin: Springer.

Wietschel, M., Arens, M., Dötsch, C., Herkel, S., Krewitt, W., Markewitz, P., Möst, D., Scheufen, M. (2010). *Energietechnologie 2050 – Schwerpunkte für Forschung und Entwicklung.* ISI-Schriftenreihe Innovationspotenziale. Karlsruhe: Fraunhofer ISI.

Web-Links

[EnArgus] *Central information system for energy research funding* https://enargus.fit.fraunhofer.de/ [17.02.2014]

[Hypertree] *A Hypertree (Hyperbolic Tree) allows to show and navigate concepts with a special tree browser.* http://o4fkie.cms2web.com/deutsch/Hypertree/ [17.02.2014]

[Ontology4] *Knowledge Based Content Management System (KB-CMS)* http://www.ontology4.us [17.02.2014]

[Protégé] *The protégé portal of the open source ontology editor and knowledge-base framework* http://protege.stanford.edu/ [17.02.2014]

[Energieforschungsprogramm 2011] *6. Energieforschungsprogramm der Bundesregierung* http://www.bmwi.de/DE/Themen/Energie/Energieforschung-und-Innovationen/6-energieforschungsprogramm.html [17.02.2014]

Notizen

Patent Retrieval und Patent Mining: Sind die Anforderungen eingelöst?

Julia Maria Struß, Daniela Becks, Thomas Mandl, Michael Schwantner & Christa Womser-Hacker

Dieser Artikel gibt einen Überblick über die sprachlichen Besonderheiten von Patenten und die daraus resultierenden Herausforderungen an inhaltliche Erschließungsverfahren. Es werden Techniken und Verfahren sowie Aufgabenszenarien aus dem Bereich des Information Retrieval, des Text Mining und abschließend auch des Trend Mining in diesem Kontext dargestellt. Ein besonderer Schwerpunkt liegt dabei auf Arbeiten im Rahmen der Evaluierungsinitiativen NTCIR und CLEF.

1 Einleitung

In den letzten Jahren hat die Anzahl der Patentanmeldungen sowohl im europäischen Raum als auch weltweit deutlich zugenommen. So verzeichnet das Europäische Patentamt im Jahr 2012 zum dritten Mal in Folge einen neuen Rekord bezüglich der Anzahl der Patentanmeldungen, aber auch in China, Japan und Korea werden stark steigende Zahlen verzeichnet (vgl. Europäisches Patentamt, 2013a). Dabei spielen in Europa vor allem die Bereiche der digitalen Kommunikation und der elektrischen Maschinen mit zweistelligen Wachstumsraten eine große Rolle, die nach der Medizintechnik die Plätze zwei und drei hinsichtlich des Gesamtvolumens der Anmeldungen einnehmen (vgl. Europäisches Patentamt, 2013b). Gleichzeitig ist zu beobachten, dass viele in Patenten publizierte Erkenntnisse nicht mehr anderweitig veröffentlicht werden. So gibt The Thomson Corporation (2007, S. 5) an, dass 70 bis 90 Prozent der in Patenten veröffentlichten Information ausschließlich dort zu finden ist. Diese Entwicklungen machen es wichtiger denn je, effektive inhaltliche Erschließungsmechanismen für diese spezielle Textsorte zu finden.

Im Folgenden wird ein Überblick über die sprachlichen Besonderheiten von Patenten und die daraus resultierenden Herausforderungen an inhaltliche Erschließungsverfahren im Rahmen von Patent Retrieval und Patent-Mining-Kontexten gegeben. Ziel des Artikels ist es, einen Überblick über existierende Techniken und Verfahren sowie verschiedene Aufgabenszenarien in diesen Bereichen aufzuzeigen, wobei ein besonderer Schwerpunkt auf den Arbeiten im Rahmen der Evaluierungsinitiativen CLEF und NTCIR liegt.

2 Sprachliche Besonderheiten von Patenten

Patentschriften zählen im Hinblick auf die Anforderung im Rahmen des Patent Retrieval und des Patent Mining zu einer schwierigen und komplexen Textsorte. Die Schwierigkeit resultiert in erster Linie aus der Patentterminologie und den von der Allgemeinsprache abweichenden linguistischen Strukturen. Auch weisen Patentschriften unterschiedliche Bestandteile mit verschiedenen sprachliche Stilen auf: Die Ansprüche sind durch juristische Sprache geprägt, wohingegen die Beschreibung Ähnlichkeiten mit wissenschaftlichen Artikeln aufweist.

Im Folgenden sollen nun einige der genannten Besonderheiten näher beschrieben und erläutert werden. Im Unterschied zu anderen Textsorten sind in Patentschriften beispielsweise für die zentralen Begriffe (die Erfindung und ihre Bestandteile) verschiedene Bezeichnungen oder Formulierungen zu finden. So hat Dederding (1982, S. 87) in einer

deutschen Auslegeschrift bspw. 14 Entsprechungen für ein entsprechendes Kompositum finden können (vgl. Becks, 2013, S. 109). Sowohl für Retrievalsysteme als auch für Text-Mining-Systeme ergibt sich daraus, dass der Zusammenhang zwischen diesen verschiedenen Varianten erkannt und entsprechend gehandhabt werden muss. Zusätzlich sind die gewählten Terme häufig abstrakter bzw. allgemeiner oder „kreativer" als solche in wissenschaftlichen Artikeln, um einen möglichst weiten Geltungsbereich für die patentierten Ansprüche zu erreichen (vgl. Nanba et al., 2008b, S. 325). Weiter ist in Patentschriften die durchschnittliche Satzlänge ausnehmend groß: Verberne, D'hondt, & Oostdijk (2010, S. 16) berichten auf Basis von Untersuchungen des englischen MAREC-400000 Subkorpus, das aus ca. 67.000 Ansprüchen (claims) besteht, dass der Median der Satzlänge bei 22 Token und die durchschnittliche Länge bei 53 Token liegt. Diese langen Sätze, die mit komplexen Satzstrukturen einhergehen, können sich insbesondere bei der Verwendung syntaktischer Parser als problematisch erweisen (vgl. Verberne et al., 2010, S. 20), die vielfach auf Zeitungskorpora mit Texten, die weniger komplexe Satzstrukturen aufweisen, trainiert sind. Erschwerend kommt hinzu, dass Ansprüche nicht aus Teilsätzen, sondern vorwiegend aus einer Aneinanderreihung von Nominalphrasen bestehen, wohingegen traditionelle Parser ersteres erwarten (vgl. Verberne et al., 2010, S. 19). Auch finden sich grammatikalische Konstruktionen, die in der Allgemeinsprache nicht verwendet werden, ebenso wie neue Wortschöpfungen (vgl. European Patent Office, 2012).

3 Patent Retrieval

Die Anforderungen und Aufgaben im Patent Retrieval weisen neben den sprachlichen Besonderheiten in dieser Domäne einige weitere Charakteristika im Vergleich mit klassischem Information Retrieval auf. So reichen einem Patentrechercheur je nach Rechercheanlass[1] nicht das eine oder einige wenige Dokumente aus, die ein Informationsbedürfnis befriedigen, wie das bei einer klassischen Recherche im Internet der Fall ist, sondern er sieht sich insbesondere bei einer Genehmigungsrecherche vor der Aufgabe, (nahezu) alle relevanten Dokumente finden zu müssen. Ein einziges nicht gefundenes Dokument könnte bspw. in diesem Kontext schon zu einer Patentverletzung führen (vgl. Becks, 2013, S. 16).

Einige Evaluierungsinitiativen wie die NTCIR[2] und CLEF[3] haben sich der interessanten Herausforderungen der Patentdomäne angenommen und Tasks zum Patent Retrieval angeboten. Bei der dritten NTCIR-Runde[4] wurde zunächst ein Pilottask durchgeführt, bei dem die Aufgabe der Teilnehmer entweder monolingual oder cross-lingual ausgehend von Zeitungsartikeln, die über technische oder kommerzielle Produkte berichteten, darin bestand, Patente zu diesen entsprechenden Technologien nachzuweisen (vgl. Iwayama et al., 2002, S. 2). Die Teilnehmer mussten also in der Lage sein, begriffliche Entsprechungen für die technologischen Terme zwischen den Zeitungsartikeln und den Patentschriften zu finden, die wie eingangs beschrieben häufig abstrakter bzw. allgemeiner sind als die Terminologie der Allgemeinsprache. Der erfolgreichste Ansatz mit einer Average Precision von ca. 0,27 (vgl. Itho et al., 2002, S. 4) verwendet neben Pseudo-Relevance-Feedback einen Ansatz, den die Autoren *Term Distillation* nennen. Dabei wird für die einzelnen Bestandteile der Query die

[1] Eine Übersicht über die verschiedenen Rechercheanlässe ist bei Trippe and Ruthven (2011), Hunt, Nguyen, and Rodgers (2007) und Becks (2013) zu finden.
[2] NTCIR (NII Test Collection for IR Systems) Project: http://research.nii.ac.jp/ntcir/index-en.html
[3] CLEF (Conference and Labs of the Evaluation Forum und vor 2010 Cross-Language Evaluation Forum): http://www.clef-initiative.eu/
[4] http://research.nii.ac.jp/ntcir/ntcir-ws3/patent/cfp-en.html

unterschiedliche Termfrequenz in der Quell- und Zieldomäne berücksichtigt (vgl. Itho et al., 2002, S. 2f).

Im darauf folgenden Jahr wurde der Fokus der Task dahingehend verändert, dass nicht mehr das Szenario des unerfahrenen Endnutzers, sondern ein Rechercheanlass aus dem professionellen Recherchekontext gewählt wurde, namentlich die Nichtigkeitsrecherche (*Invalidity Search*) (vgl. Fujii et al., 2004, S. 1). Es wird ein Korpus aus ca. 1,7 Millionen Patentschriften aus fünf Jahren bereitgestellt sowie die entsprechenden englischen Abstracts; als Topics werden abgelehnte Patente verwendet, für die bereits die Entgegenhaltungen vorlagen, die zur Ablehnung führten, und entsprechend als Relevanzurteile verwendet werden konnten (vgl. Fujii et al., 2004, S. 2). Diese Relevanzurteile wurden für 34 der 101 Topics durch Patentrechercheexperten weiter ergänzt. Die beste Mean Average Precision (MAP) mit einem Wert von ca. 0,17 bzw. ca. 0,27 für die 34 Haupttopics wurden von zwei Systemen erzielt, die mit verschiedenen Termgewichten in Bezug auf die Ansprüche (Claims) experimentiert, aber ebenfalls verschiedene Dokument- und Absatz-basierte Scores für das Filtern und Ranken der Dokumente verwendet haben (vgl. Fujii et al., 2004, S. 5). Bedingt durch die Art der Task können für einige Topics sehr wenige relevante Dokumente verfügbar sein, daher regen die Organisatoren an, andere Maße für die Evaluierung zu entwickeln und die MAP für diese Art von Szenario zu überprüfen.

Im folgenden Jahr wird die Task auf gleiche Weise fortgeführt, ergänzend jedoch auch eine *Passage Retrieval Task* angeboten und die Korpora um weitere fünf Jahrgänge von Patentschriften bzw. Abstracts ergänzt (vgl. Fujii et al., 2005, S. 2). Um eine größere Anzahl an Topics (1189) zu generieren, wird auf zusätzliche manuelle Relevanzurteile verzichtet und lediglich die angegebenen für die Ablehnung verantwortlichen Patente herangezogen, die sich für die gewählten Patente auf maximal fünf beläuft (vgl. Fujii et al., 2005, S. 2). Das beste System erzielt eine MAP von ca. 0,195 für die Pflichtruns, für die optionalen Runs erzielte das beste System, das die Suche jeweils auf die IPC-Klassen des Ausgangspatents beschränkt, eine MAP von ca. 0,21 (vgl. Fujii et al., 2005, S. 5f.). Eine zusätzliche Evaluierung, die die Topics danach differenziert, ob die relevanten Patente jeweils in mindestens einer derselben IPC-Klassen enthalten sind, bzw. nach dem Namen des Anmelders, ergab, dass für beides die Ergebnisse für die Topics, die diesen Kriterien entsprechen, jeweils signifikant höher sind (vgl. Fujii et al., 2005, S. 5, 7).

Für die Passage Retrieval Task wurden manuelle Relevanzurteile erstellt, die sowohl einzelne Absätze als auch Absatzgruppen umfassen konnten. Für die Evaluierung wurde der „*Combinational Relevance Score*“ (CRS) entwickelt, der die Rangposition angibt, bei der ein Nutzer ausreichend Gründe für die Wertung eines Dokuments als relevant bzw. teilweise relevant hat (vgl. Fujii et al., 2005, S. 4). Die CRS und die MAP-Werte, die im Vergleich ebenfalls angegeben werden, weisen unterschiedliche Systeme als die jeweils besten aus, zu bedenken ist jedoch, dass in der MAP keine kombinierte Relevanz mehrerer Paragraphen abgebildet werden kann (vgl. Fujii et al., 2005, S. 7). Fujii & Ishikawa (2005) segmentieren zunächst die Ansprüche und erweitern die für die Query selektierten abstrakteren Terme (Substantive) aus den Ansprüchen jeweils mit konkreteren bzw. spezifischeren Termen aus den zugehörigen Abschnitten der Beschreibung und setzen zusätzlich Pseudo-Relevance-Feedback auf Dokumentenebene ein (vgl. Fujii & Ishikawa, 2005, S. 3). Für das Ranking der Paragraphen innerhalb eines Patents werden diese als separate Dokumente aufgefasst. Der Ansatz erreicht einen CRS-Wert von 10,91 sowie eine MAP von 0,48 bzw. 0,46[5] (vgl. Fujii et al., 2005, S. 9). Mase et al. (2005) nutzen einen n-Gramm Ansatz und erweitern die Query durch Terme aus den Ansprüchen eines Patents und erreichen nach eigenen Berechnungen

[5] mit bzw. ohne Paragraphen, die nur in Kombination relevant sind

einen CRS-Wert von 13,2 sowie eine MAP von 0,56 bzw. 0,52[4] bei Berücksichtigung auch teilweise relevanter Dokumente (vgl. Mase et al., 2005, S. 4).

Aber auch bei CLEF wurden Tracks zum Patent Retrieval (Intellectual Property Tracks[6]) durchgeführt. Im ersten Jahr der CLEF-IP Tracks wird eine Task zur Prior Art Suche angeboten, für die ein Korpus mit ca. 2 Millionen Patentdokumenten automatisch auf Basis der Zitationen zwischen Patenten mit Relevanzurteilen für die Evaluierung versehen werden (vgl. Roda et al., 2010, S. 387). Kritisiert wurde dabei zum einen die Methode, die für die Erstellung der Relevanzurteile verwendet wurde und zum anderen die Verwendung von bereits erteilten Patenten für das Szenario der Prior Art Suche, die von Experten mit Patentanmeldungen assoziiert werden (vgl. Roda et al., 2010, S. 387). Dennoch ist die Pilottask mit 15 teilnehmenden Teams erfolgreich angelaufen (vgl. Roda et al., 2010, S. 386).

In den CLEF-IP Tracks ist bereits eine große Varianz an verschiedenen Retrievaltechniken zum Einsatz gekommen, jedoch nach wie vor mit nur mäßigem Erfolg. So lagen die beste MAP 2009 bei 0,11 (vgl. Roda et al., 2010, S. 400)[7]. Eine gute Übersicht über die in den Jahren 2009 bis 2011 verwendeten Methoden und Ansätze ist in Jürgens et al. (2012) zu finden. Die Autoren schlagen aufgrund der nach wie vor mäßigen Erfolge der eingesetzten Techniken vor, das Wissen um den Suchprozess in die Ansätze zu integrieren und dies auch für die Gestaltung realistischer Tasks in Evaluierungsinitiativen heranzuziehen (vgl. Jürgens et al., 2012).

4 Text Mining

Text Mining beschäftigt sich mit der Extraktion von Wissen aus Dokumentenkollektionen, setzt also i.d.R. nach dem Retrievalprozess an oder geht von einer vorhandenen Dokumentenmenge aus. Grundlegend werden dabei Muster und Häufigkeiten bestimmter Gruppen von Wörtern in verschiedenen Textkollektionen betrachtet. Zu beachten ist, dass die gewonnenen Erkenntnisse immer auf der Gesamtschau vieler Texte und der Häufigkeitsverteilung von Begriffen in diesen Texten basieren und der Nutzer das Wissen durch Verwendung entsprechender Werkzeuge in einem interaktiven Prozess gewinnt. (vgl. Mandl, 2013, S. 183).

Die Herausforderungen durch die Verarbeitung natürlicher Sprache, die in allgemeinsprachlichen Texten wie Zeitungsartikeln z.B. im Umgang mit Vagheit und heterogenen Darstellungsformen bei Eigennamen liegen (vgl. Mandl, 2013, S. 183), sind bei der Verarbeitung von Dokumentenkollektionen aus Patentschriften durch die zuvor beschriebene große Varianz im Vokabular einer Patentschrift und deren größere Abstraktheit ungleich größer. Im nachfolgenden Kapitel werden die Tasks der Evaluierungsinitiativen, die sich mit dieser Thematik befassen, sowie die jeweils besten Ansätze vorgestellt.

4.1 Patent Mining

Zum ersten Mal wird eine Patent Mining Task bei NTCIR im Jahre 2004 angeboten, bei der die Aufgabe in der Generierung sogenannter *Patent Maps* bestand, die für eine Technologie die beiden Dimensionen „zu lösende Probleme“ sowie „Lösungen“ aufspannen und mit den

[6] http://www.clef-initiative.eu/track/clefip

[7] Das Ergebnis von Lopez, Romary, and Laurent (2009) mit einer MAP von 0,27 kann in diesem Zusammenhang nicht berücksichtigt werden, da sie die Zitatinformation genutzt haben, die auch für die Erstellung der Relevanzurteile genutzt wurden (vgl. Roda, Tait, Piroi, and Zenz (2010, pp. 398, 400)).

entsprechenden Patenten befüllt werden (vgl. Fujii et al., 2004, S. 4). Es wurde jedoch zunächst nur eine subjektive, manuelle Bewertung der beiden eingereichten Systeme vorgenommen.

Diese Task wird im folgenden Workshop in Form zweier Subtasks fortgeführt. Zum einen besteht die Aufgabe darin, Patente in sogenannte *Themes* (Technologiebereiche) zu klassifizieren, die einem Set an japanischen *File Index* (FI)-Codes entsprechen, die wiederum eine Erweiterung der Internationalen Patentklassifikation (IPC) darstellen (vgl. Iwayama et al., 2005, S. 2). Die zweite Task besteht darin, Patente mit vorgegebenen Technologiebereichen in sogenannte *File Forming Terms* (F-Terms) zu kategorisieren, die im japanischen Patentamt verwendet werden. Ein F-Term besteht dabei immer aus einem sogenannten *Viewpoint* wie „*purpose*“ oder „*means*“, die aber je nach Technologiebereiche durchaus variieren können, und einem zugehörigen Element, das einen Aspekt des *Viewpoints* näher beschreibt (vgl. Iwayama et al., 2005, S. 2, IPDL 2008)[8]. Diese Kategorisierung kann dann genutzt werden, um *Patent Maps* zu erstellen. Für die Task wurden fünf Technologiebereiche und ca. 500 zugehörige Patente pro Technologiebereich ausgewählt, denen durchschnittlich ca. 4,8 F-Terms zugeordnet sind. In beiden Subtasks ist jeweils ein k-Nearest-Neighbor (k-NN) Ansatz der erfolgreichste. Kim et al. (2005) verwenden einen k-NN Ansatz unter Berücksichtigung der Struktur der Patentdokumente, wobei sie die Ähnlichkeitsberechnung separat für verschiedene Bestandteile des Patents durchführen. Neben den vorgegebenen Strukturen wie Abstract und Anspruch nutzen sie noch eine weitere Unterteilung der Beschreibung in sechs weitere Sektionen, dazu zählen u.a. Methode und Zweck der Erfindung (vgl. Kim et al., 2005, S. 4). Sie erzielen so eine MAP von 0,69 und ein F-Maß von 0,27 in der ersten Subtask (vgl. Iwayama et al., 2005, S. 8). Murata et al. (2005) verwenden einen modifizierten k-NN Ansatz zusammen mit BM25 für die Ähnlichkeitsberechnung unter Berücksichtigung der Abstracts, der Ansprüche und den jeweils ersten Paragraphen zur technischen Domäne der Erfindung sowie der Methode (vgl. Murata et al., 2005, S. 3–4) und erzielen eine MAP von 0,5 und ein F-Maß von 0,44 (vgl. Iwayama et al., 2005, S. 9).

NTCIR-6 führt nur noch die Subtask zur Kategorisierung der Patente in die sogenannten F-Terms weiter. Für 108 verschiedene Technologiebereiche wurden jeweils ca. 200 Testdokumente für die Task ausgewählt (vgl. Iwayama et al., 2007, S. 368–369). Im Unterschied zum Vorjahr werden bei der Evaluierung die hierarchischen Strukturen der *F-Terms* mit berücksichtigt, das so auch das Werten korrekter über- oder untergeordneter F-Terms in Form von Partial Matches erlaubt. Die besten Systeme erreichten mit einer MAP von ca. 0,49 (Fujino & Isozaki, 2007) bzw. 0,48 (Li et al., 2007) und einem F-Maß von ca. 0,40 bzw. 0,41 vergleichbare Ergebnisse wie das System von (Murata et al., 2005) bei NTCIR-5, das dieses Mal Platz drei einnimmt. Die Ergebnisse für die Partial Matches liegen jeweils ca. zehn Prozentpunkte über den genannten Werten. Fujino & Isozaki (2007) setzen hybride Klassifikatoren ein, die sie auf Basis des Maximum Entropie Prinzips aus einzelnen *Naive Bayes* Klassifikatoren für verschiedene Bestandteile des Patents (Titel, Autoren, Abstract, Ansprüche und Beschreibung) unter Verwendung eines Bag-of-Words Ansatzes mit Beschränkung auf Substantive, Verben und Adjektive kombinieren. Li et al. (2007) beschränken sich bei ihrem *Support Vector Machine (SVM)*-Ansatz ebenso auf Substantive, Verben und Adjektive, verwenden jedoch außerdem die als unbekannt getaggten Terme (vgl. Li et al., 2007, S. 398) und nutzen zusätzlich zum gesamten Inhalt der Patente (mit Ausnahme der bibliographischen Informationen und Klassifikation) auch die Kurzbeschreibungen zu den F-Terms für die Klassifikation. In weiteren Experimenten mit eigenen

[8] Beispiel: Theme: detection and correction of errors (in computers), Viewpoint: means, Elements: code operations, comparison, interleaving (vgl. Iwayama, Fujii, and Kando (2005, p. 2))

Evaluierungsmaßen stellen die Autoren fest, dass die hierarchische Variante der SVM hinsichtlich der exakten Zuordnung zu den F-Terms ungenauer als die einfachere Variante ist, jedoch in Bezug auf die Partial Matches besser ist, da sie F-Terms wählt, die in der Hierarchie näher zum korrekten F-Term stehen (vgl. Li et al., 2007, S. 401).

Die Organisatoren der Task halten noch fest, dass keine signifikanten Unterschiede in den Ergebnissen der Systeme bestehen, die die Beschreibung der F-Terms verwenden bzw. nicht verwenden (vgl. Iwayama et al., 2007, S. 371), wobei Li et al. (2007) dies im Vergleich ihrer verschiedenen Systemversionen schon feststellen konnten. Ebenso ermittelten die Organisatoren auch keine signifikanten Unterschiede zwischen Systemen, die lediglich die Abstracts und die Ansprüche für die Klassifikationsaufgabe verwendeten und denen, die die Volltexte nutzen (vgl. Iwayama et al., 2007, S. 370).

Bei NTCIR-7 wird eine neue Task eingeführt, die sich mit der Einordnung von wissenschaftlichen Artikeln in die Internationale Patentklassifikation (IPC) beschäftigt. Die Teilnehmer müssen zunächst Patente identifizieren, die sich mit derselben Thematik wie der vorgegebene wissenschaftliche Artikel befasst, um auf dieser Grundlage mögliche IPC-Klassen zu ermitteln. Diese Task wird sowohl monolingual für Japanisch und Englisch angeboten als auch cross-lingual zwischen den beiden Sprachen (vgl. Nanba et al., 2008b, S. 326). Die Erstellung der Test- und Evaluierungsdaten erfolgte semi-automatisch, indem die Zitationsinformation aus dem entsprechenden Feld herangezogen wird. Hier ist angegeben, welche wissenschaftlichen Artikel die Autoren zur vorliegenden Erfindung veröffentlicht haben[9]. Aus den automatisch ermittelten Kandidatenlisten werden dann manuell die entsprechenden Publikationen herausgesucht und die IPC-Codes des Patents auf die Publikation übertragen (vgl. Nanba et al., 2008a, S. 328). Die erfolgreichsten Systeme haben auch bei dieser Task wieder einen k-NN-Ansatz verwendet (vgl. Nanba et al., 2008a, S. 329). So erreichen Mase & Iwayama (2008, S. 365) mit ihrem k-NN-basierten Ansatz, der ebenfalls nur Substantive, Verben und Adjektive berücksichtigt und tf-idf-Gewichte verwendet, eine MAP in der monolingualen japanischen Task von 0,44 (vgl. Nanba et al., 2008a, S. 330). Die Autoren experimentieren außerdem noch mit verschiedenen *Query Expansion* (QE)-Methoden, wobei die Verwendung von Patenten für die QE bessere Ergebnisse als die Verwendung weiterer wissenschaftlicher Artikel liefert (vgl. Mase & Iwayama, 2008, S. 366f). Für die englische monolinguale Task schneidet das k-NN-basierte System mit einer Kombination verschiedener Re-Ranking-Algorithmen von Xiao, Cao, & Li (2008) mit einer MAP von ca. 0,49 ab (vgl. Nanba et al., 2008a, S. 330).

Die vorgestellten Ansätze zu den verschiedenen von der Evaluierungsinitiative NTCIR angebotenen Task zeigen, dass k-NN-basierte Ansätze für viele der Probleme am besten geeignet scheinen, jedoch sind auch diese noch weit entfernt von den Ergebnissen vergleichbarer Tasks bei anderen Textsorten wie Zeitungsartikeln oder Webdokumenten.

4.2 Trend Mining im Patentbereich

Ein weiterer Bereich, der vor dem Hintergrund des immer stärkeren globalen Wettbewerbs zunehmendes Interesse in der Forschung findet, ist das Trend Mining auf Patenten. Neben den Herausforderungen, die das Text Mining im Patentbereich mit sich bringt, kommen im Trend Mining noch die Analyse zeitlicher Zusammenhänge und Entwicklungen hinzu. Dabei wird unter einem Trend im Allgemeinen ein Themenbereich verstanden, der über die Zeit an

[9] Im japanischen Patentierungsverfahren darf innerhalb von sechs Monaten vor der Offenlegung eine Veröffentlichung in Form eines wissenschaftlichen Artikels erfolgen, ohne dass dies zur Ablehnung des Patents führt (vgl. Nanba, Fujii, Iwayama, and Hashimoto (2008a, p. 328)).

Bedeutung zunimmt (vgl. z.B. Kontostathis et al., 2004, S. 186). In einem Beispiel spezifizieren die Autoren, dass ein solcher Trend z.B. durch die Anzahl der Treffer zu einem Schlagwort (hier: XML) in der INSPEC Datenbank im Verlauf mehrerer Jahre sichtbar wird. Eine genauere auf den Patentbereich zugeschnittene Definition für das Trend Mining im Patentbereich ist in den Arbeiten i.d.R. nicht zu finden. Dabei ist z.B. durchaus von Interesse, welche Eigenschaften und Verläufe ein Trend aufweisen kann, wie diese sich in verschiedenen Anwendungsdomänen unterscheiden und welche für die potentiellen Nutzer eines solchen Systems von besonderem Interesse sind. Im Folgenden wird ein Überblick über die Arbeiten im Trend Mining gegeben, die sich mit Patentschriften oder Teilen von Patentschriften befassen.

Es sind bereits eine Vielzahl von Techniken in diesem Bereich zur Anwendung gekommen: Die meisten dieser Techniken stammen aus dem Bereich des maschinellen Lernens (vgl. z.B. Shih et al., 2008, Yoon et al., 2011, Pottenger & Yang, 2001, Choi et al., 2011), wobei ein besonderer Fokus auf dem Clustering (vgl. z.B. Chang et al., 2010, Yoon & Park, 2004) liegt, aber auch Kombinationen von Clustering z.B. mit Multidimensionaler Skalierung werden eingesetzt (vgl. Gerken 2012). Vielfach kommen auch Netzwerkanalysen (vgl. z.B. Chang et al., 2010, Choi et al., 2011, Yoon et al., 2011, Tang et al. 2012, Park et al. 2013) oder bibliometrische Methoden (vgl. z.B. Chang et al., 2010) oder Kombinationen mehrerer der genannten Methoden zum Einsatz. In letzter Konsequenz wird die Entscheidung darüber, ob es sich um einen Trend handelt, vielfach dem Nutzer überlassen, unterstützt durch verschiedene Visualisierungstechniken (vgl. z.B. Havre et al., 2002, Lee et al., 2011). Auch wird zur Unterstützung der maschinellen Ansätze häufig noch auf Expertenwissen zurückgegriffen. Dies ist vor allem in der Anfangsphase der Analyse der Fall, hauptsächlich bei der Auswahl relevanter Terme, die in thematischem Zusammenhang zum Untersuchungsfeld stehen. So greifen beispielsweise Chang et al. (2010) auf manuell durch Experten erstellte Termlisten als Ausgangspunkt für die Analyse zurück, aber auch semiautomatisch erstellte Termlisten, bei denen ein Experte aus einer zuvor automatisch zusammengestellten Liste Terme auswählt, kommen zum Einsatz (vgl. z.B. Lee et al., 2011 und Havre et al., 2002). Wie gut die einzelnen Ansätze funktionieren, lässt sich jedoch vielfach nicht in letzter Konsequenz nachvollziehen, da die Funktionsfähigkeit der Ansätze häufig nur an einzelnen Fallstudien demonstriert wird, was dem Fehlen entsprechender Evaluierungsressourcen geschuldet ist, oder nur eine Evaluierung einzelner Teilaufgaben, i.d.R derjenigen ohne Zeitbezug, stattfindet.

Diesem Problem hat sich die Patent Mining Task[10] von NTCIR-8 im Rahmen einer Subtask, die sich mit dem Erstellen von Technologie-Trend-Maps befasst, angenommen. Die Aufgabe für die Teilnehmer besteht dabei zunächst darin, englisch- oder japanisch-sprachige Dokumente (Titel und Abstract wissenschaftlicher Artikel und Patente) zu finden und aus diesen dann elementare Technologien und deren Effekte, die wiederum aus Attributen und Werten bestehen, zu identifizieren und die Dokumente entsprechend klassifizieren (vgl. Nanba et al., 2010, S. 293f). Die besten Ergebnisse liegen mit einem F-Maß von 0,275 für japanische wissenschaftliche Artikel und 0,481 für japanische Patente, sowie Werten von 0,164 für englische wissenschaftliche Artikel bzw. 0,332 für englische Patente. Dabei lösten die Teilnehmer die Aufgabe, die Technologie aus dem Titel eines Dokuments zu extrahieren i.d.R. am besten und das Identifizieren von Effekten in den Abstracts am schlechtesten.

[10] http://www.ls.info.hiroshima-cu.ac.jp/~nanba/ntcir-8/cfp.html

5 Abschließende Betrachtung

Der Artikel hat die verschiedenen sprachlichen Besonderheiten von Patenten und die daraus resultierenden Herausforderungen an inhaltliche Erschließungsverfahren im Rahmen von Patent Retrieval und Patent Mining Kontexten aufgezeigt und einen Überblick über existierende Techniken und Verfahren sowie verschiedene Aufgabenszenarien in diesen Bereichen gegeben.

Die Evaluierungsergebnisse insbesondere der beschriebenen Ergebnisse im Patent Retrieval zeigen, dass die Herausforderungen, die die sprachlichen und strukturellen Besonderheiten der Patente an die Erschließungsverfahren stellen, noch weitgehend ungelöst sind. Neue Herangehensweisen müssen daher entwickelt werden, wie bspw. die Einbeziehung des Wissens um den Suchprozess, wie von Jürgens, Hansen, & Womser-Hacker (2012) vorgeschlagen. Auch für das Patent Mining und Trend Mining in der Patentdomäne ist eine solche Einbeziehung des Nutzers in die Entwicklung und Evaluierung der Systeme anzustreben. Insbesondere der Anwendungskontext der Nutzer im Trend Mining und die daraus möglicherweise resultierenden unterschiedlichen Definitionen eines Trends sind hier zu untersuchen.

Literatur

Becks, D. (2013). *Entwicklung eines Framework für die begriffliche Optimierung von Patentanfragen*. Dissertation, Universität Hildesheim. Eggenstein-Leopoldshafen: FIZ Karlsruhe.

Chang, P.-L., Wu, C.-C., & Leu, H.-J. (2010). Using Patent Analyses to Monitor the Technological Trends in an Emerging Field of Technology: a Case of Carbon Nanotube Field Emission Display. *Scientometrics*, *82*(1), 5–19.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C.-H. (2011). SAO Network Analysis of Patents for Technology Trends Identification: a Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells. *Scientometrics*, *88*(3), 863–883.

Dederding, H.-M. (1982). *Wortbildung, Syntax, Text: Nominalkomposita und entsprechende syntaktische Strukturen in deutschen Patent- und Auslegeschriften* (Vol. 34). Erlangen: Palm & Enke.

Europäisches Patentamt. (2013a). *Jahresbericht 2012: Statistiken und Trends. Europäische Anmeldezahlen*. Entwicklungen der letzen fünf Jahre. http://www.epo.org/about-us/annual-reports-statistics/annual-report/2012/statistics-trends/patent-filings_de.html

Europäisches Patentamt. (2013b). *Jahresbericht 2012: Statistiken und Trends. Europäische Patentanmeldungen.* Technische Gebiete mit starken Wachstumsraten. http://www.epo.org/about-us/annual-reports-statistics/annual-report/2012/statistics-trends/patent-applications_de.html

European Patent Office. (2012). *EPO - Patent searching for beginners - EPO Patent Information Beginners Seminar (EPIBS)*. http://www.epo.org/learning-events/events/training/patent-information-training/PI01-2013.html

Fujii, A., Iwayama, M., & Kando, N. (2004). Overview of Patent Retrieval Task at NTCIR-4. In: *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies.*

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-OV-PATENT-FujiiA.pdf

Fujii, A., & Ishikawa, T. (2005). Document Structure Analysis for the NTCIR-5 Patent Retrieval Task. In: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-FujiiA.pdf

Fujii, A., Iwayama, M., & Kando, N. (2005). Overview of Patent Retrieval Task at NTCIR-5. In: *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-OV-PATENT-FujiiA.pdf

Fujino, A., & Isozaki, H. (2007). Multi-label Patent Classification at NTT Communication Science Laboratories. In: *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 381–384). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/28.pdf

Gerken, Jan M. (2012): PatMining – Wege zur Erschließung textueller Patentinformationen für das Technologie-Monitoring. Dissertation. Universität Bremen, Bremen. Projektmanagement und Innovation.

IPDL (2008): Outline of Search Items. F-terms. Industrial Property Digital Library. Online verfügbar unter http://www.ipdl.inpit.go.jp/HELP/pmgs_en/database/format_summary.html#fterm, zuletzt aktualisiert am 27.10.2008, zuletzt geprüft am 11.03.2014.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). ThemeRiver: Visualizing Thematic Changes in Large Document Collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9–20.

Hunt, D., Nguyen, L., & Rodgers, M. (2007). *Patent Searching: Tools & Techniques*. Hoboken, N.J: Wiley.

Itho, H., Mano, H., & Ogawa, Y. (2002). Term Distillation for Cross-DB Retrieval. In: K. Oyama, E. Ishida, & N. Kando (Eds.), *Proceedings of the 3rd NTCIR Workshop.* Tokyo, Japan: Nihon Printing. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-PATENT-ItohH.pdf

Iwayama, M., Fujii, A., & Kando, N. (2005). Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task. In: *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-OV-PATENT-IwayamaM.pdf

Iwayama, M., Fujii, A., & Kando, N. (2007). Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task. In: *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 366–372). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/68.pdf

Iwayama, M., Fujii, A., Kando, N., & Takano, A. (2002). Overview of Patent Retrieval Task at NTCIR-3. In: K. Oyama, E. Ishida, & N. Kando (Eds.), *Proceedings of the 3rd NTCIR Workshop. Evaluation of Information Retrieval, Automatic Text Summarization*

and Question Answering. Tokyo, Japan: Nihon Printing. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-OV-PATENT-IwayamaM.pdf

Jürgens, J. J., Hansen, P., & Womser-Hacker, C. (2012). Going beyond CLEF-IP: The 'Reality' for Patent Searchers? In T. Catarci, S. Forner, A. Peñas, & G. Santucci (Eds.), *Lecture Notes in Computer Science: Vol. 7488. Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative, CLEF 2012, Rome, Italy, September 17-20, 2012.* (pp. 30–35). Berlin, Heidelberg: Springer.

Kim, J.-H., Huang, J.-X., Jung, H.-Y., & Choi, K.-S. (2005). Patent Document Retrieval and Classification at KAIST. In: *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-KimJ.pdf

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). A Survey of Emerging Trend Detection in Textual Data Mining. In: M. W. Berry (Ed.), *Survey of Text Mining. Clustering, Classification, and Retrieval* (pp. 185–224). New York, London: Springer.

Lee, C., Jeon, J., & Park, Y. (2011). Monitoring Trends of Technological Changes Based on the Dynamic Patent Lattice: A Modified Formal Concept Analysis Approach. *Technological Forecasting and Social Change*, *78*(4), 690–702.

Lopez, P., Romary, & Laurent (2009). Multiple Retrieval Models and Regression Models for Prior Art Search. In: Working Notes for the CLEF 2009 Workshop. http://clef.isti.cnr.it/2009/working_notes/lopez-paperCLEF2009.pdf

Li, Y., Bontcheva, K., & Cunningham, H. (2007). SVM Based Learning System for F-term Patent Classification. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 396–402). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/38.pdf

Mandl, T. (2013). Text Mining und Data Mining. In: R. Kuhlen, W. Semar, & D. Strauch (Eds.), *Grundlagen der praktischen Information und Dokumentation. Handbuch zur Einführung in die Informationswissenschaft und -praxis.* (pp. 183–191). Berlin, Boston: De Gruyter Saur.

Mase, H., Matsubayashi, T., & Ogawa, Y. (2005). NTCIR-5 Patent Retrieval Experiments at Hitachi. In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-MaseH.pdf

Mase, H., & Iwayama, M. (2008). NTCIR-7 Patent Mining Experiments at Hitachi. In: *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 365–368). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMN/09-NTCIR7-PATMN-MaseH.pdf

Murata, M., Kanamaru, T., Shirado, T., & Isahara, H. (2005). Using the K Nearest Neighbor Method and BM25 in the Patent Document. In: *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies.*

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/PATENT/NTCIR5-PATENT-MurataM.pdf

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008a). Overview of the Patent Mining Task at the NTCIR-7 Workshop. In: *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 325–332). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMN/01-NTCIR7-OV-PATMN-NanbaH.pdf

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008b). Overview of the Patent Mining Task at the TCIR-7 Workshop. In: *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMN/01-NTCIR7-OV-PATMN-NanbaH.pdf

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2010). Overview of the Patent Mining Task at the NTCIR-8 Workshop. In: *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.*(pp. 293–302). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-PATMN-NanbaH.pdf

Park, Hyunseok; Kim, Kwangsoo; Choi, Sungchul; Yoon, Janghyeok (2013): A Patent Intelligence System for Strategic Technology Planning. In: Expert Systems with Applications 40 (7), S. 2373–2390.

Pottenger, W. M., & Yang, T.-H. (2001). Detecting Emerging Concepts in Textual Data Mining. In: M. W. Berry (Ed.), *Computational Information Retrieval* (pp. 89–105). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. http://dl.acm.org/citation.cfm?id=762544.762552

Roda, G., Tait, J., Piroi, F., & Zenz, V. (2010). CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In: C. Peters, G. M. Nunzio, M. Kurimo, D. Mostefa, A. Penas, & G. Roda (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments. 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers* (Vol. 6241, S. 385–409). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Shih, M.-J., Liu, D.-R., & Hsu, M.-L. (2008). Mining Changes in Patent Trends for Competitive Intelligence. In: T. Washio, E. Suzuki, K. M. Tin, & A. Inokuchi (Eds.), *Lecture Notes in Computer Science: Vol. 5012. Advances in Knowledge Discovery and Data Mining. 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings* (pp. 999–1005). Berlin, Heidelberg: Springer.

Tang, Jie; Wang, Bo; Yang, Yang; Hu, Po; Thao, Yanting; Yan, Xinyu et al. (2012): PatentMiner. Topic-driven Patent Analysis and Mining. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and DataMining. ACM SIGKDD. Beijing, China, August 12–16, 2012. New York, NY, USA: ACM.

The Thomson Corporation. (2007). *Global Patent Sources: An Overview of International Patents* (6th ed.). London: Thomson Scientific. http://ip-science.thomsonreuters.com/m/pdfs/mgr/global_patent_sources.pdf

Trippe, A., & Ruthven, I. (2011). Evaluating Real Patent Retrieval Effectiveness. In: M. Lupu, K. Mayer, J. Tait, & A. J. Trippe (Eds.), *The Information Retrieval Series: Vol. 29. Current Challenges in Patent Information Retrieval* (pp. 125–143). Berlin, Heidelberg, New York: Springer.

Verberne, S., D'hondt, E., & Oostdijk, N. (2010). Quantifying the Challenges in Parsing Patent Claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10)* (pp. 14–21). Milton Keynes.

Xiao, T., Cao, F., & Li, T. (2008). KNN and Re-ranking Models for English Patent Mining at NTCIR-7. In: *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.* (pp. 333–340). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C3/PATMN/02-NTCIR7-PATMN-XiaoT.pdf

Yoon, B., & Park, Y. (2004). A Text-mining-based Patent Network: Analytical Tool for High-technology Trend. *The Journal of High Technology Management Research*, *15*(1), 37–50.

Yoon, J., Choi, S., & Kim, K. (2011). Invention Property-function Network Analysis of Patents: a Case of Silicon-based Thin Film Solar Cells. *Scientometrics*, *86*(3), 687–703.

Google Scholar versus Fachdatenbanken: Wie erfolgreich nutzen Studierende verschiedene Suchwerkzeuge?

Nikolas Leichner, Johannes Peter, Anne-Kathrin Mayer & Günter Krampen

In der Literatur ist belegt, dass Fachdatenbanken Google Scholar überlegen sind, wenn ihr Funktionsumfang ausgeschöpft wird. In dieser Studie wird untersucht, ob Recherchen in Fachdatenbanken auch dann erfolgreicher sind, wenn diese von Psychologiestudierenden durchgeführt werden. Im Rahmen der Evaluation eines Informationskompetenztrainings zeigte sich, dass Studierende mit Fachdatenbanken erfolgreicher recherchierten, wenn sie am Training teilgenommen hatten. Vor der Trainingsteilnahme waren Fachdatenbank-Nutzer nicht erfolgreicher. Die Ergebnisse werden so interpretiert, dass Schulungsbedarf besteht, damit Studierende die Möglichkeiten von Fachdatenbanken ausschöpfen können.

Einleitung

Professionelle Informationskompetenz wird als eine Schlüsselkompetenz wissenschaftlichen Arbeitens angesehen, deren Förderung bereits im schulischen Kontext, spätestens jedoch zu Beginn der Hochschulausbildung erfolgen sollte (vgl. z.B. Hochschulrektorenkonferenz, 2012). Zentraler Bestandteil professioneller Informationskompetenz ist die Fähigkeit zur gezielten und effektiven Recherche nach Fachliteratur unter Verwendung angemessener Suchwerkzeuge (Association of College and Research Libraries, 2010). Voraussetzung hierfür ist, dass nicht nur die Funktionen der jeweiligen Suchwerkzeuge, sondern auch deren Stärken und Schwächen bekannt sind, sodass – je nach Anforderung – das geeignete Suchwerkzeug ausgewählt werden kann.

Die wichtigsten Suchwerkzeuge für Psychologen sind vermutlich die Fachdatenbanken PSYNDEX und PsycINFO® sowie Google Scholar. Üblicherweise wird auf die Fachdatenbanken über die Suchoberflächen von Datenbank-Hosts wie Ovid Technologies oder EBSCO zugegriffen; diese Suchoberflächen erlauben üblicherweise auch, beide Fachdatenbanken parallel zu durchsuchen. Im Folgenden soll untersucht werden, ob Psychologiestudierende, die eine der beiden (oder beide) Fachdatenbanken für die Recherche nutzen, eine Rechercheaufgabe erfolgreicher bearbeiten als Studierende, die Google Scholar nutzen. Dabei wird auch der Frage nachgegangen, inwieweit sich das Suchverhalten von Studierenden durch die Teilnahme an einem Informationskompetenztraining ändert.

Hintergrund

Seit der Einführung im Jahr 2004 ist Google Scholar eines der am häufigsten genutzten Werkzeuge für wissenschaftliche Literaturrecherchen, obwohl Information Professionals dieser Ressource kritisch gegenüber stehen (Walters, 2011; Hightower & Caldwell, 2010).

Einer der Gründe für diese Ablehnung dürfte die Tatsache sein, dass Google Scholar nicht nur Publikationen aus Fachzeitschriften, Monographien und Herausgeberwerken, sondern auch andere wissenschaftliche Dokumente (z.B. Qualifikationsarbeiten, Präsentationsfolien, sogenannte „graue Literatur“) beinhaltet, deren Qualität unterschiedlich und nicht gesichert ist. Verglichen mit dem Funktionsumfang von Fachdatenbanken fehlen zudem weitgehende

Filterfunktionen und ein kontrolliertes Vokabular (Thesaurus; vgl. Giustini & Kamel Boulos, 2013). Gleich nach dem Start von Google Scholar begann in wissenschaftlichen Kreisen eine Debatte darüber, welche Vor - und Nachteile Google Scholar gegenüber Fachdatenbanken bietet und ob Google Scholar für fachliche Recherchen überhaupt benutzt werden sollte. Im Folgenden werden die Vor - und Nachteile von Google Scholar skizziert. Es folgt ein kurzer Literaturüberblick über Studien, die Google Scholar mit Fachdatenbanken verglichen. Im empirischen Teil des Beitrags wird der Frage nachgegangen, ob Studierende der Psychologie mithilfe von Fachdatenbanken effektiver recherchieren als mit Google Scholar.

Vorteile von Google Scholar

Als wesentliche Vorteile von Google Scholar werden neben der freien Verfügbarkeit im WWW und der – zumindest in der Basissuche – intuitiven Bedienbarkeit die Volltext-indexierung und der Nachweis von grauer Literatur hervorgehoben (Jacsó, 2005; Giustini & Kamel Boulos, 2013). Google Scholar macht damit auch Texte, Vortragsfolien und ggf. neue Publikationen auffindbar, die in Fachdatenbanken (noch) nicht verzeichnet sind (Giles, 2005; Shultz, 2007). Eine Besonderheit ist die automatische Analyse von Zitationen. Dies ermöglicht es Google Scholar zum einen, die Treffer nach ihrer Zitationshäufigkeit zu ordnen und somit für den Benutzer möglicherweise relevante Dokumente besser sichtbar zu machen. Zum anderen ermöglicht diese Funktion, zitierende Dokumente zu finden, die nicht durch Eingabe des Suchbegriffs gefunden werden (Mayr & Walter, 2007). Weiterhin handelt es sich bei Google Scholar um eine interdisziplinäre Suchmaschine (Mayr & Walter, 2007), was insbesondere bei Recherchen im Schnittstellenbereich mehrerer Fachgebiete von Vorteil sein kann.

Nachteile von Google Scholar

Einer der Hauptkritikpunkte an Google Scholar betrifft die mangelnde Transparenz. Im Gegensatz zu Fachdatenbanken, die genaue Angaben über den Abdeckungsbereich und –grad sowie die Aktualisierungsintervalle ihrer Datenbank machen, fehlen solche Angaben bei Google Scholar völlig (e.g. Jacsó, 2005; Gray et al., 2012). Auf die Abdeckung kann nur anhand empirischer Studien geschlossen werden, in denen systematisch Suchanfragen durchgeführt werden (z.B. Mayr und Walter, 2007). Wie oben beschrieben ist die Qualität der in Google Scholar verzeichneten Publikationen unterschiedlich und muss in jedem Einzelfall überprüft werden. Ein weiterer wichtiger Kritikpunkt besteht im Fehlen eines kontrollierten Vokabulars und dem Nichtvorhandensein spezifischer Filterfunktionen, beispielsweise eines Filters für die wissenschaftliche Disziplin (Giustini & Kamel Boulos, 2013).

Vergleiche von Google Scholar mit Fachdatenbanken

Bald nach dem Start von Google Scholar wurden Studien veröffentlicht, in denen Recherchen mithilfe von Google Scholar und Fachdatenbanken verglichen wurden. Die meisten dieser Studien wurden im Bereich Medizin/Biologie durchgeführt, was vermutlich daran liegt, dass PubMed (eine der wichtigsten Fachdatenbanken in diesem Bereich) ebenfalls frei verfügbar ist, sodass die freie Verfügbarkeit in diesem Feld keinen Wettbewerbsvorteil von Google Scholar darstellt. Google Scholar wurde in diesen Studien überwiegend kritisch beurteilt. So kam Shultz (2007) zu der Schlussfolgerung, dass PubMed nicht durch Google Scholar ersetzt werden kann, da ein kontrolliertes Vokabular fehlt.

Allerdings sei Google Scholar durchaus für den Einstieg in eine umfassende Recherche geeignet, da die Suchoberfläche äußerst benutzerfreundlich gestaltet sei. Andere Studien argumentieren, dass Google Scholar bei Recherchen zu medizinischen Themen gegenüber PubMed unterlegen sei, aber durchaus als ergänzende Quellen genutzt werden könne. So verglichen Anders und Evans (2010) die Rechercheergebnisse aus Google Scholar und PubMed mit einer vorliegenden Literaturliste. Es zeigte sich, dass PubMed dank spezifischer Filterfunktionen wesentlich präzisere Ergebnisse lieferte als Google Scholar. Aktuellere Untersuchungen beurteilen Google Scholar positiver. So stellte eine Studie fest, dass mithilfe von Google Scholar gefundene Publikationen von Experten als relevanter beurteilt wurden als mithilfe von PubMed gefundene Publikationen (Nourbakhsh, Nugent, Wang, Cevik, & Nugent, 2012). Dennoch dominiert weiterhin die Auffassung, dass Recherchen nicht allein auf Google Scholar beruhen sollten. Gestützt wird diese bspw. durch eine Arbeit von Giustini und Kamel Boulos (2013), die versuchten, die Literaturrecherchen für einen vorliegenden Überblicksartikel mithilfe von Google Scholar zu replizieren. Von insgesamt 506 Publikationen, die in dem Überblicksartikel genannt worden waren, konnten 30 nicht mithilfe von Google Scholar gefunden werden. Das Auffinden der übrigen Publikationen erforderte mehrere Versuche und verschiedene Kombinationen von Suchbegriffen und Syntaxausdrücken und war damit wesentlich aufwändiger als eine Recherche in PubMed.

Fragestellung

In den oben vorgestellten Vergleichsstudien wurden zuvor definierte Recherchen sowohl mit Google Scholar als auch mit Fachdatenbanken durchgeführt und die Ergebnisse verglichen. In der Studie von Shultz (2007) beispielsweise wurden von Recherche-Experten zehn Recherchen zu verschiedenen Themen mit PubMed realisiert und anschließend mit Google Scholar wiederholt. Durch dieses standardisierte Vorgehen werden vergleichbare Ergebnisse geliefert. Allerdings wird dabei nicht berücksichtigt, dass weniger recherche-erfahrene Benutzer möglicherweise nicht so effektiv vorgehen wie die Recherche-Experten in den Studien, da sie die in Fachdatenbanken enthaltenen Funktionen unzureichend ausschöpfen. Fachdatenbanken sind Google Scholar vor allem dann überlegen, wenn ihre erweiterten Funktionen (Verknüpfung von Suchen mit Booleschen Operatoren, Filterfunktionen) benutzt werden (Walters, 2011). In der Realität werden diese erweiterten Funktionen von vielen Nutzern allerdings nicht verwendet (Weiland & Baier, 2010). Uns ist nur eine Studie bekannt, die untersucht, ob die Fachdatenbank PubMed auch dann bessere Ergebnisse liefert als Google Scholar, wenn die Probanden keine Recherche-Experten sind (Perryman, 2013). In der Studie zeigte sich, dass Nicht-Experten mit Google Scholar verglichen mit PubMed zwar thematisch relevantere, aber vor allem ältere Publikationen fanden.

Im vorliegenden Beitrag wird vor diesem Hintergrund der Frage nachgegangen, ob Bachelor-Studierende der Psychologie mithilfe von Fachdatenbanken effektiver recherchieren als mithilfe von Google Scholar. Hierfür wird auf Daten aus einer Evaluationsstudie für ein Informationskompetenztraining zurückgegriffen, die im Sommer 2013 an der Universität Trier durchgeführt wurde. Primäre Inhalte des Trainings, das am Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) konzipiert wurde, waren die Verwendung von Fachdatenbanken (in diesem Fall insbesondere PSYNDEX und PsycINFO) und die Verwendung der erweiterten Suche in Google Scholar. Es wurde Wert darauf gelegt, den Studierenden zu vermitteln, dass Google Scholar bei einfachen Suchanfragen durchaus gute Ergebnisse liefern kann, dass aber bei komplexen Suchanfragen auf Fachdatenbanken zurückgegriffen werden sollte; die Nutzung dieser Fachdatenbanken (einschließlich ihrer speziellen Funktionen, z.B. Verwendung von Filtern etc.) wurde durch

Lehrtexte und -videos vermittelt und in einer Reihe von praktischen Rechercheaufträgen eingeübt. Ziel der Studie war, ein auf Psychologiestudierende zugeschnittenes Trainingskonzept zu entwerfen und zu evaluieren, das auch von anderen Universitäten genutzt werden kann, um die Informationskompetenz von Studierenden zu fördern. Zur Evaluation des Trainings kamen neben einem standardisierten Informationskompetenztest im Multiple-Choice-Format (vgl. Leichner, Peter, Mayer, & Krampen, 2013) standardisierte fachliche Rechercheaufgaben zum Einsatz.

Im Folgenden werden zwei Hypothesen geprüft:

(1) Das durchgeführte Informationskompetenztraining verbessert die Recherchefähigkeiten der Teilnehmer.

(2) Studierende, die Fachdatenbanken für ihre Recherche verwenden, erzielen bessere Rechercheergebnisse als diejenigen, die Google Scholar verwenden.

Methode

An der Evaluationsstudie nahmen N = 67 Psychologiestudierende (Alter M = 21.67 Jahre [SD = 2.38]) aus dem zweiten (n = 34) und vierten (n = 33) Semester des Bachelorstudiengangs teil. Diese Stichprobe wurde in zwei Gruppen eingeteilt (n = 37 bzw. n = 30). Im Verlauf der Evaluationsstudie wurden drei Messzeitpunkte (t1 – t3) realisiert. Im Anschluss an Messzeitpunkt 1 nahm Gruppe 1 am zweiwöchigen Training teil, Gruppe 2 fungierte als Wartekontrollgruppe. Unmittelbar nach dem Training folgte t2, worauf Gruppe 2 am Training teilnahm. Der dritte Messzeitpunkt fand statt, nachdem Gruppe 2 das Training absolviert hatte.

Zu jedem Messzeitpunkt bearbeiteten die Teilnehmer den oben erwähnten Informationskompetenztest, der aufgrund eigener Überlegungen und vorliegender englischsprachiger Tests (z.B. Noe & Bishop, 2005; Project SAILS, 2013) entworfen wurde. Der Test bestand aus 35 Fragestellungen zu den Bereichen „Recherche nach Fachinformationen“ und „Bewertung von Fachinformationen“. Für jede korrekt angekreuzte bzw. ausgelassene Antwortoption wurden Teilpunkte vergeben. Eine Frage aus dem Bereich Recherche lautete beispielsweise:

Welche Option ist am sinnvollsten, wenn Literatur nicht in Ihrer (Universitäts-)Bibliothek zu finden ist?

- *Bei Bibliotheken in benachbarten Städten anfragen, im Zweifelsfall dort hinfahren.*
- *Die Literatur per Fernleihe bestellen.(korrekte Antwortoption)*
- *Es gibt keine Alternative zum Kauf des Buchs/der Zeitschrift.*

Zudem bearbeiteten die Probanden zu jedem Messzeitpunkt drei fachliche Rechercheaufgaben; hierfür stand ihnen ein PC mit Internetanschluss und Zugang zu den an der Universität Trier lizensierten Fachdatenbanken (via Ovid-SP) zur Verfügung. Um die Aufgaben soweit wie möglich zu standardisieren, wurde zunächst eine Aufgabentaxonomie (Leichner, Peter, Mayer, & Krampen, in press) mit drei unterschiedlich schwierigen Aufgabentypen entwickelt, auf deren Grundlage inhaltlich verschiedene Rechercheaufgaben mit gleicher Struktur erstellt wurden. Bei jedem Messzeitpunkt bearbeiteten die Probanden eine Aufgabe pro Typ. In den folgenden Analysen wird nur einer der drei Aufgabentypen berücksichtigt, da nur bei diesem Aufgabentyp substanzielle Vorteile (d.h. präzisere Ergebnisse) durch die Verwendung spezifischer Funktionalitäten von Fachdatenbanken

(Thesaurussuche, Nutzung von Filterfunktionen, Verknüpfung durch Boolesche Operatoren) zu erwarten sind. Eine der drei Aufgaben des verwendeten Typs lautet beispielsweise:

Gibt es nach (nicht einschließlich) 2003 publizierte Meta-Analysen, die die Wirksamkeit der sogenannten Kognitiven Verhaltenstherapie („cognitive behavior therapy") bei der Behandlung von Depression („Major Depression") untersuchen? Bitte geben Sie – wenn möglich – zwei entsprechende Publikationen an.

Bewusst wurden Aufgaben mit klar definierter möglicher Treffermenge verwendet, um eine ökonomische und zugleich objektive Auswertung der Ergebnisse sicherzustellen. Die für die Aufgabenbearbeitung zur Verfügung stehende Zeit war auf zehn Minuten begrenzt; die meisten Probanden benötigten allerdings weniger Zeit. Im Anschluss an die Bearbeitung einer jeden Aufgabe beantworteten die Probanden einige Fragen zu ihrem Vorgehen während der Recherche. Unter anderem wurde gefragt, welche Suchmaschine sie zuletzt verwendet hatten, um die angegebene Literatur zu finden.

Die Aufgaben wurden anhand eines vorher festgelegten standardisierten Auswertungsschlüssels bewertet. Dabei wurde jede der bis zu zwei dokumentierten Publikationen separat beurteilt, wobei Punkte für jedes in der Aufgabenbeschreibung genannte Kriterium vergeben wurden, die von der gefundenen Publikation erfüllt wurden. Im Falle der obigen Beispielaufgabe konnte die maximale Punktzahl erreicht werden, wenn (1) die angegeben Suchbegriffe in Titel, Abstract oder Schlagwörtern der gefundenen Publikation aufgeführt waren; (2) die gefundene Publikation nach 2003 erschienen war und (3) es sich um eine Meta-Analyse handelte. Wurden eines oder mehrere der Kriterien nicht erfüllt, wurden entsprechend weniger Punkte vergeben. Falls ein Proband nur eine passende Publikation gefunden hat, wurden für die zweite Publikation entsprechend keine Punkte vergeben. Die Punktwerte der beiden pro Messzeitpunkt gefundenen Publikationen wurden für die Auswertung addiert und auf einen Wertebereich von 0 bis 1 skaliert.

Ergebnisse

Die von den Probanden bei den drei Messzeitpunkten erreichten Punktwerte sind in Abbildung 1 dargestellt. Obwohl die Mehrzahl der Probanden die Aufgabenbearbeitung innerhalb der vorgegebenen Bearbeitungszeit abschloss, weisen die erzielten Punktwerte darauf hin, dass die erzielten „Treffer" die in den Aufgabenbeschreibungen genannten Kriterien vielfach nicht erfüllen. Die Aufgaben lassen sich zum ersten Messzeitpunkt in beiden Versuchsgruppen mit einem Mittelwert von $M = .50$ als im psychometrischen Sinne mittelschwer charakterisieren.

Um Hypothese 1 zu überprüfen, wurde eine zweifaktorielle Varianzanalyse (Gruppe x Messzeitpunkt) mit Messwiederholung auf dem zweiten Faktor berechnet. Es zeigt sich eine signifikante Interaktion zwischen der Gruppenzugehörigkeit und dem messwiederholten Faktor, also dem Verlauf der Punktwerte ($F[2,130] = 6.51$, $p < .01$).

Es ist deutlich zu erkennen, dass sich die Punktwerte im Verlauf des Trainings verbesserten, was durch die signifikante Interaktion bestätigt wird. Es gab keine Unterschiede zwischen den beiden Gruppen vor Beginn des Trainings ($t[65] = 0.90$, n.s.). Wie erwartet erreichte Gruppe 1 zu t2 höhere Werte, da diese Gruppe (im Gegensatz zu Gruppe 2) bereits am Training teilgenommen hatte ($t[65] = 3.82$, $p < .01$). Beim dritten Messzeitpunkt gab es erneut keine Unterschiede zwischen den beiden Gruppen ($t[65] = 1.24$, n.s.).

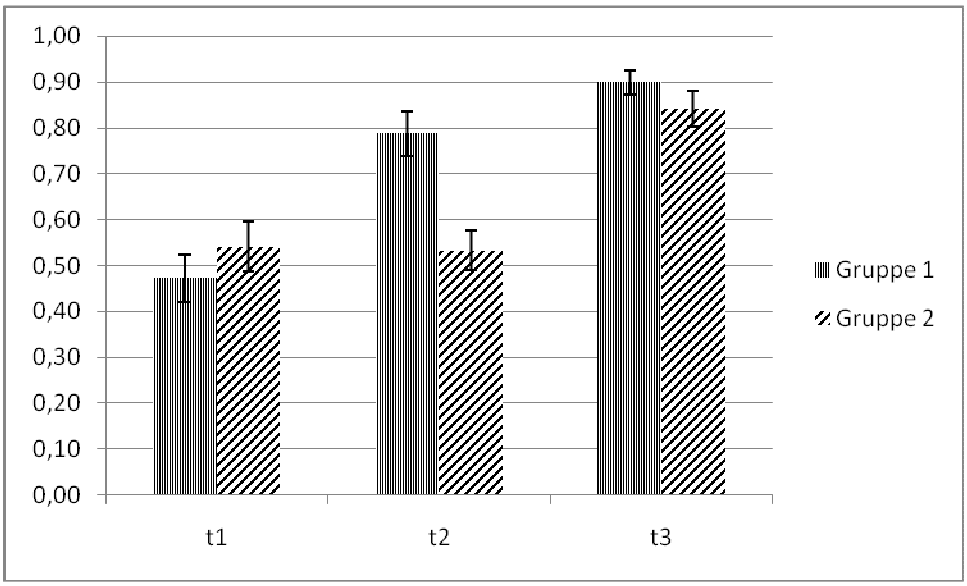
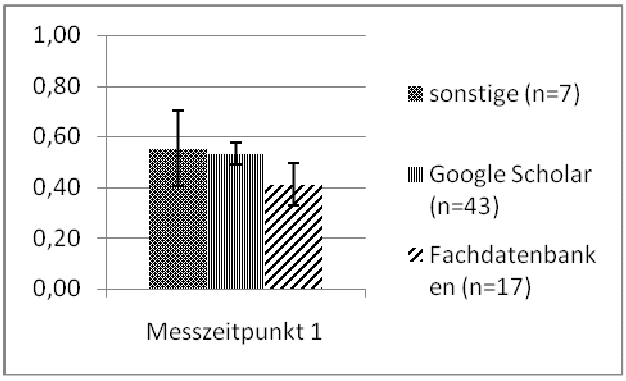
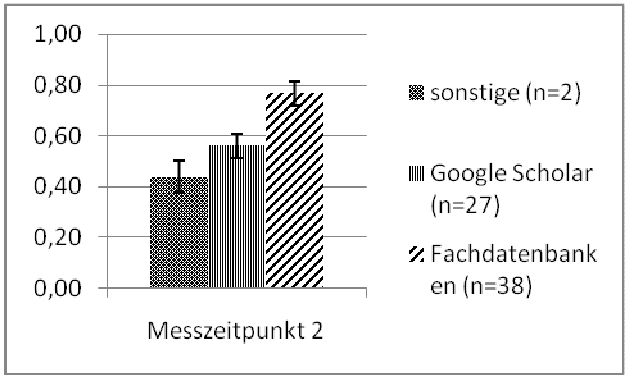
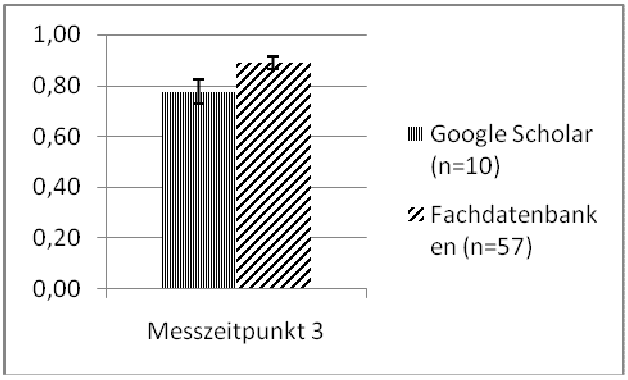


Abbildung 1: Mittelwerte und Standardfehler der Leistungen in den fachlichen Rechercheaufgaben in Abhängigkeit von Versuchsgruppe und Messzeitpunkt.

Zur Prüfung von Hypothese 2 wurde zu jedem der drei Messzeitpunkte untersucht, ob Probanden, die nach eigenen Angaben für ihre Recherche eine Fachdatenbank genutzt hatten, höhere Werte bei der Rechercheaufgabe erzielten als Probanden, die entweder Google Scholar oder sonstige Suchwerkzeuge (z.B. Google) verwendet hatten. Die Ergebnisse sind in den Abbildungen 2 bis 4 veranschaulicht.







Abbildungen 2 bis 4: Punktwerte und Standardfehler bei Verwendung der jeweiligen Suchmaschine.

In den folgenden Analysen wurden jeweils nur diejenigen Probandengruppen verglichen, die Google Scholar bzw. Fachdatenbanken genutzt hatten, d.h. die wenigen Probanden, die zu t1 bzw. t2 sonstige Ressourcen (z.B. Google) zur Aufgabenlösung verwendet hatten, wurden ausgeschlossen. Es zeigte sich, dass Fachdatenbank-Benutzer zu t1 (vor der Trainingsteilnahme) nicht erfolgreicher waren ($t[58] = 1.40$, n.s.). Zu t2 ($t[63] = 2.90$, $p < .01$) und t3 ($t[65] = 1.82$, $p < .05$) waren Fachdatenbank-Benutzer erfolgreicher, d.h., sie erreichten höhere Punktwerte bei der Rechercheaufgabe.

Diskussion

Bezüglich Hypothese 1 konnte gezeigt werden, dass Probanden nach der Trainingsteilnahme – gemessen an den erzielten Leistungen bei einer Rechercheaufgabe – effektiver recherchieren konnten. Bezüglich Hypothese 2, und damit der Hauptfragestellung der vorliegenden Arbeit, zeigte sich ein interessantes Muster. So waren Fachdatenbank-Benutzer bei den Messzeitpunkten 2 und 3 erfolgreicher als die Benutzer von Google Scholar. Bei t1 hingegen waren keine Unterschiede zwischen Fachdatenbank- und Google-Scholar-Nutzern zu beobachten, allerdings war das allgemeine Leistungsniveau zu diesem Zeitpunkt auch relativ niedrig. Offenbar kennen einzelne Studierende zwar bereits die relevanten Datenbanken ihres Fachs und rufen diese für ihre Literaturrecherchen auf. Es fehlt ihnen jedoch das Wissen über spezifische Funktionalitäten der Datenbanken bzw. die Fähigkeit, diese Funktionalitäten zu nutzen. Ohne dieses Wissen bzw. diese Fähigkeiten können Fachdatenbanken nicht effektiv verwendet werden – und ihre Vorteile kommen nicht zur Geltung. Dies entspricht Befunden anderer Studien, wonach Fachdatenbanken nur dann bessere Ergebnisse liefern als Google Scholar, wenn fortgeschrittene Recherchetechniken verwendet werden (Perryman, 2013; Walters, 2011).

Implikationen

Die vorliegenden Befunde stehen in Einklang mit den Ergebnissen anderer Studien, die einen Bedarf nach Informationskompetenzschulungen bei Studienanfängern feststellen (Smith, Given, Julien, Ouellette, & DeLong, 2013). Möglicherweise lassen sich die Benutzeroberflächen von Fachdatenbanken zudem benutzerfreundlicher gestalten, sodass auch Nicht-Experten die Vorteile von Fachdatenbanken nutzen können. Die in etlichen anderen Studien gezogene Schlussfolgerung, dass Fachdatenbanken nicht durch Google Scholar ersetzt werden können (z.B. Falagas, Pitsouni, Malietzis & Pappas, 2007; Anders & Evans, 2010), kann mit diesen Befunden dennoch prinzipiell bestätigt werden.

Einschränkungen

Zum einen ist darauf hinzuweisen, dass die Probanden zusätzlich befragt wurden, ob sie weitere Suchmaschinen verwendet und ob diese zum Ergebnis beigetragen hatten. Etwa die Hälfte der Probanden gab an, auch weitere Suchmaschinen verwendet zu haben; von diesen berichtete jedoch nur ungefähr ein Fünftel, dass die weiteren Suchmaschinen zum Ergebnis beigetragen hätten. Diese Daten sprechen gegen die Vermutung, dass die Ergebnisse dadurch verzerrt werden, dass einige Probanden zuvor weitere Suchmaschinen verwenden haben, diese Suchmaschinen von unserer Auswertung jedoch nicht erfasst werden.

Eine zweite Alternativerklärung ist, dass Fachdatenbank-Nutzer primär deshalb erfolgreicher waren, weil sie bessere Suchstrategien anwendeten und/oder ein höheres Niveau von Informationskompetenz aufwiesen. Bei der Auswertung der Daten von t2 zeigte sich, dass

32 der insgesamt 38 Fachdatenbank-Benutzer zuvor am Training teilgenommen hatten, sodass diese Alternativerklärung durchaus realistisch erscheint. Möglicherweise wären diese Probanden auch dann erfolgreicher gewesen, wenn sie Google Scholar verwendet hätten, da sie bessere Suchstrategien beherrschten. Gegen diese Erklärung spricht allerdings, dass die Fachdatenbank-Nutzer in dem zu Evaluationszwecken eingesetzten Informationskompetenztest keine signifikant höheren Punktwerte erreichten als die Google Scholar-Nutzer nachdem alle Probanden am Training teilgenommen hatten (t3).

Um diese beiden Alternativerklärungen jedoch vollständig auszuschließen, müsste man die Probanden im Rahmen einer experimentellen Studie instruieren, ausschließlich entweder Google Scholar oder Fachdatenbanken zu verwenden; dies könnte Thema einer Folgestudie sein.

Zuletzt sei darauf hingewiesen, dass in der vorliegenden Studie lediglich Aufgaben eines spezifischen Typs verwendet wurden. Daher verbietet sich die vorschnelle generalisierende Schlussfolgerung, dass Recherchen mit Google Scholar prinzipiell weniger erfolgreich sind. Wie die Literatur zeigt, gibt es durchaus Rechercheprobleme, die sich mit Google Scholar effektiv bearbeiten lassen. Die Vorteile von Fachdatenbanken (insbesondere die Möglichkeit, mithilfe eines kontrollierten Vokabulars zu recherchieren und die Ergebnisse mithilfe von Filterfunktionen einzuschränken) kommen nur bei hinreichend komplexen und spezifischen Aufgabenstellungen (wie bei dem hier verwendeten Aufgabentyp) zum Tragen (vgl. Shultz, 2007; Gray et al., 2012).

Schlussfolgerung

Die Ergebnisse zeigen zum einen, dass alle Probanden (auch diejenigen, die Google Scholar verwendeten) nach der Trainingsteilnahme höhere Punktwerte in der Rechercheaufgabe erreichten. Offenbar verbessern die während des Trainings vermittelten Suchstrategien auch die Nutzung von Google Scholar. Gleichwohl konnten Studierende, die Fachdatenbanken zur Recherche verwenden, im Anschluss an ein Informationskompetenztraining Recherche-aufgaben des verwendeten Typs effektiver lösen als Google Scholar-Verwender. Damit bestätigt dieser Befund vorliegende Publikationen (z.B. Anders & Evans, 2010; Giustini & Kamel Boulos, 2013), wonach Recherchen mithilfe von Fachdatenbanken präziser und effektiver gestaltet werden können.

Zudem zeigte sich, dass der Anteil der Probanden, die Google Scholar (oder sonstige Suchmaschinen) verwendeten, durch die Trainingsteilnahme von ungefähren 75 auf ungefähr 15 Prozent sank. Wie in der Einleitung erwähnt, zeichnet sich informationskompetentes Verhalten dadurch aus, dass je nach Rechercheproblem das angemessenste Suchwerkzeug gewählt werden kann. Bei den von uns gestellten Aufgaben kann davon ausgegangen werden, dass sie hinreichend komplex sind, um die Vorteile von Fachdatenbanken zur Geltung zu bringen. Insofern stellt dieses Ergebnis einen weiteren Beleg für die Wirksamkeit des Trainingsprogramms mit Blick auf die Förderung der Informationskompetenz dar.

Literatur

Anders, M. E. & Evans, D. P. (2010). Comparison of PubMed and Google Scholar literature searches. *Respiratory Care,* 55(5), 578–583.

Association of College and Research Libraries. (2010). *Psychology information literacy standards*. Retrieved from http://www.ala.org/acrl/standards/psych_info_lit

Falagas, M. E.; Pitsouni, E. I.; Malietzis, G. A. & Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal,* 22(2), 338–342. doi:10.1096/fj.07-9492LSF

Giles, J. (2005). Science in the web age: Start your engines. *Nature,* 438(7068), 554–555. doi:10.1038/438554a

Giustini, D. & Kamel Boulos, M. N. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, 5(2). doi:10.5210/ojphi.v5i2.4623

Gray, J. E.; Hamilton, M. C.; Hauser, A.; Janz, M. M.; Peters, J. P. & Taggart, F. (2012). Scholarish: Google Scholar and its value to the sciences. *Issues in Science and Technology Librarianship*, 70. Retrieved from http://www.istl.org/12-summer/article1.html

Hightower, C. & Caldwell, C. (2010). Shifting sands: Science researchers on Google Scholar, Web of Science, and PubMed, with implications for library collections budgets. *Issues in Science and Technology Librarianship*, 63.

Hochschulrektorenkonferenz. (2012). *Hochschule im digitalen Zeitalter: Informations-kompetenz neu begreifen - Prozesse anders steuern*. Retrieved from http://www.hrk.de/uploads/media/Entschliessung_Informationskompetenz_20112012.pdf

Jacsó, P. (2005). Google Scholar: The pros and the cons. *Online Information Review*, 29(2), 208–214. doi:10.1108/14684520510598066

Leichner, N.; Peter, J.; Mayer, A.-K. & Krampen, G. (2013). Assessing information literacy among German psychology students. Reference Services Review, 41(4), 660–674. doi:10.1108/RSR-11-2012-0076

Leichner, N.; Peter, J.; Mayer, A.-K., & Krampen, G. (in press). Assessing information literacy using information search tasks. *Journal of Information Literacy*.

Mayr, P., & Walter, A.-K. (2007). An exploratory study of Google Scholar. *Online Information Review*, 31(6), 814–830. doi:10.1108/14684520710841784

Noe, N. W. & Bishop, B. A. (2005). Assessing Auburn University library's Tiger Information Literacy Tutorial (TILT). *Reference Services Review*, 33(2), 173–187. doi:10.1108/00907320510597372

Nourbakhsh, E.; Nugent, R.; Wang, H.; Cevik, C. & Nugent, K. (2012). Medical literature searches: A comparison of PubMed and Google Scholar. *Health Information & Libraries Journal*, 29(3), 214–222. doi:10.1111/j.1471-1842.2012.00992.x

Perryman, C. (2013). For non-expert clinical searches, Google Scholar results are older with higher impact while PubMed results offer more breadth. *Evidence Based Library and Information Practice*, 8(2), 254–257. Retrieved from http://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/18535/15264

Project SAILS. (2013). *Project SAILS Information Literacy Assessment*. Retrieved from https://www.projectsails.org/

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association* : JMLA, 95(4), 442–445. doi:10.3163/1536-5050.95.4.442

Smith, J. K.; Given, L. M.; Julien, H.; Ouellette, D. & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for

undergraduate academic work. *Library & Information Science Research*, 35(2), 88–96. doi:10.1016/j.lisr.2012.12.001

Walters, W. H. (2011). Expert searches in Google Scholar and eight other databases. *portal: Libraries and the Academy*, 11(4), 971–1006.

Weiland, P. & Baier, C. (2010, April). *Simplify your Interface - Wie die Auswertung von User Logs das Design einer Suchoberfläche beeinflusst am Beispiel der Weiterentwicklung von PSYNDEX Direct*. Vortrag auf dem 25. Oberhofer Kolloquium, Barleben/Magdeburg. Retrieved from http://www.zpid.de/pub/research/weiland_baier_Simplify-your-Interface.pdf

Die Modellierung von Interdisziplinarität – Ontologiedesign für den Exzellenzcluster „Bild Wissen Gestaltung“

Christian Stein

Der Exzellenzcluster Bild Wissen Gestaltung[1] ist ein hochgradig interdisziplinäres Forschungsgroßprojekt mit über 25 Disziplinen in über 20 Basisprojekten an der Humboldt-Universität zu Berlin. Die grundlegende Idee dahinter ist es, unterschiedlichste wissenschaftliche Disziplinen in ungewöhnlichen und neuen Zusammenstellungen an Fragen unserer Zeit arbeiten zu lassen. Dabei kommen nicht nur Geistes-, Natur- und Technikwissenschaften zusammen, sondern erstmals gleichberechtigt auch Designer und Gestalter.

So arbeiten in einzelnen Basisprojekten beispielsweise Architekten, Kunsthistoriker, Informatiker, Chemiker, Kulturwissenschaftler und Interaction Designer zusammen an einer gemeinsamen Fragestellung. Mit über 200 Mitarbeitern ist damit eine Forschungskonfiguration geschaffen, die einzigartig ist und Gelegenheit bietet, das Zusammenarbeiten und die Kommunikationsstrukturen in solchen massiv interdisziplinären Konstellationen zu untersuchen. Darin liegt eines der Hauptinteressen des Clusters: Was ist Interdisziplinarität? Wo funktioniert sie tatsächlich? Und wo nicht? Wann ist sie wirklich hilfreich und zielführend und wann verkommt sie zu einem Schlagwort? Schließlich: Wie kann man das Gelingen der Kommunikation zwischen den Disziplinen sicherstellen bzw. verbessern?

Damit steht eine genuin ontologische bzw. terminologische Fragestellung im Zentrum dieses Forschungsgroßprojektes und gleichermaßen eine extreme Kommunikationssituation. Die Erfahrung der ersten anderthalb Jahre Clusterarbeit zeigt, dass Terminologiediskussionen einen außergewöhnlich großen Raum einnehmen. Dabei ergibt es sich sehr häufig, dass verwendete Benennungen von anderen Disziplinen aufgegriffen, diskutiert und mit disziplinspezifischen Verständnisweisen und Assoziationen angereichert werden. Dabei ist ein generelles Bemühen spürbar, in höheren Abstraktionsschichten wieder einen gemeinsamen, überdisziplinären Bedeutungskern zu finden, der die jeweiligen disziplinspezifischen Begriffe inkludiert.

So schwierig und teilweise abstrakt diese Bemühungen um die Terminologie auch erscheinen mögen, als so wertvoll haben sie sich dennoch für den interdisziplinären Dialog erwiesen. Dabei geht es tatsächlich weniger darum, ein eindeutiges Verständnis zu erzeugen, als die eigenen Denkstrukturen und Perspektiven produktiv aufzubrechen und in einem gemeinsamen Gesamtbild zu verorten. In vielen Fällen hat diese Herangehensweise nach der ersten Irritation tatsächlich zu einem besseren Verständnis des eigenen und fremden Denkens geführt. Selten jedoch konnte sich bisher auf eine spezifische Definition geeinigt werden, die alle Beteiligten vollständig unterstützen. Auch das kann als Terminologiearbeit verstanden werden – allerdings mit einer völlig anderen Zielsetzung als in der klassischen Terminologiearbeit. Dennoch bleibt es ein wichtiges Ziel, auch dieses Arbeiten an Terminologie zu formalisieren, zu ordnen und beschreibbar zu machen.

Diese und einige weitere Prozesse will der Cluster an sich selbst beobachten und Interdisziplinarität in all ihren Ausprägungen besser verstehen lernen. Der Aufgabe widmet sich eine ganze Research Area aus verschiedenen Blickwinkeln. Um alle Modellierungen und Beobachtungen kombiniert auswerten zu können, wurde eine Ontologie entwickelt, die die Informations- und Kommunikationsstrukturen des Clusters abbilden können soll.

[1] https://www.interdisciplinary-laboratory.hu-berlin.de/

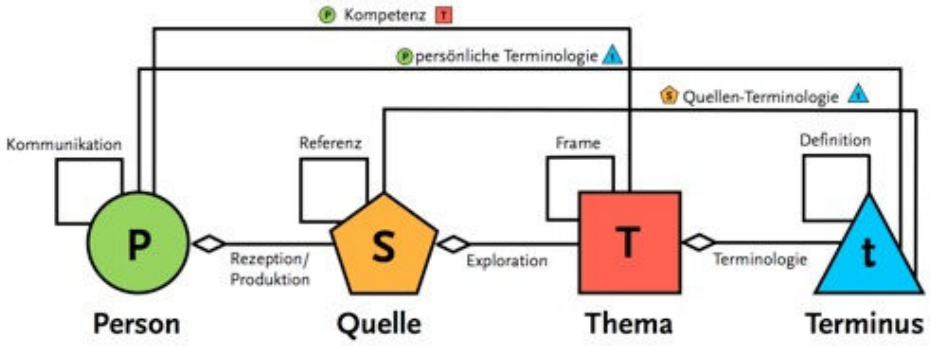


Abbildung 1: Entwurf Entitätenmodell zur Abbildung von interdisziplinären Forschungsprozessen im BWG-Cluster.

Das in Abbildung 1 dargestellte Modell zeigt nur die abstrakten Basisentitäten und die allgemeinen Relationen zwischen ihnen. Hier ist jedoch gut zu sehen, wie vielfältig vernetzt sich bereits auf dieser Ebene eine Forschungslandschaft präsentiert: Kommunikation findet zwischen Personen statt, diese haben Kompetenzen zu bestimmten Themen und sprechen bzw. verstehen eine persönliche Auswahl von Terminologie. Sie arbeiten mit Quellen, indem sie sie rezipieren oder produzieren. Quellen haben implizite oder explizite Referenzen auf andere Quellen, behandeln verschiedene abstrakte oder konkrete Themen (Exploration) und benutzen eine eigene Auswahl an Quellen-Terminologie, die sich in ihnen finden lässt. Themen wiederum können sich zu häufig zusammen auftretenden Themengruppen gruppieren (Frame) und haben eine charakteristische Terminologie. Jeder Terminus wiederum ist durch seine Relationen mit anderen, definierenden Termini definiert. Diese Struktur wird ergänzt durch raum- und zeitbezogene Kategorien und eine Vielzahl von Spezialisierungen. So ergibt sich eine Struktur die zur formalen Beschreibung von Forschungsaktivitäten dienen kann und auch dem einzelnen Forscher viel Information über die eigene Arbeitsweise zurückgibt.

Die Ontologie wird befüllt mit manuellen und automatischen Beobachtungen und dient als zentrale Datenbasis für alle Auswertungen. Dabei geht es um die sehr heterogenen Arbeitsweisen, Raumnutzung, Kommunikation, Organisation, Themen und Quellen sowie die verwendete Terminologie. Zur automatischen Aufzeichnung gehört u.a. eine Software-eigenentwicklung namens Diary, die in der Lage ist, Forschungsaktivitäten am Rechner aufzuzeichnen, zu kategorisieren, Inhalte zu bestimmen und die Aufzeichnungen an einen Triplestore zu übermitteln. Die Hoffnung ist es, neben der reinen Beobachtung auch hilfreiche Verbindungen und Statistiken in den Cluster zurückspiegeln zu können und so zu einer transparenteren Forschungslandschaft zu kommen, in der Bedarfe besser mit Kompetenzen vernetzt werden können. In diesem Vortrag wird die Konzeption der Ontologie, der Stand der Umsetzung verschiedener Maßnahmen sowie die interne Kommunikation und der Datenschutz vorgestellt und diskutiert werden.

Konzeption einer Architektur für flexible Suche und Informationsvernetzung in der Fahrzeugentwicklung

Markus Zoier, Manfred Rosenberger, Nadja Marko, Selver Softic, Christian Kaiser, Bernd Fachbach, Heimo Gursch, Mario Zechner, Roman Kern & Alexander Stocker

Fahrzeugentwicklung ist eine wissensintensive Tätigkeit, die optimale Zusammenarbeit über Disziplin-, Abteilungs- und Unternehmensgrenzen erfordert. Gerade deswegen ist der möglichst effektive Zugang zu Information und Wissen ein wesentlicher Erfolgsfaktor. Der Einsatz flexibler und intelligenter Technologie für die Suche und Vernetzung von Information für Entwickler kann ein möglicher Ansatz sein, wie sich der Zugang zu Informationen verbessern lässt. Im folgenden Beitrag werden Ergebnisse aus einem gemeinsam mit drei Automobilherstellern und einem großen Zulieferer gestarteten Forschungsprojekt vorgestellt, welches ein Informationscockpit für Fahrzeugentwickler zum Ziel hat. In rund 50 Gesprächen mit Fahrzeugentwicklern wurde der Status Quo der Informationsbeschaffung in erste Anforderungen für ein Informationscockpit bei den Projektpartnern erhoben. Auf Basis vorab definierter Projektziele sowie der erhobenen, erweiterten, abgestimmten und verfeinerten Anforderungen der Fahrzeugentwickler wurde eine erste Architektur erstellt, welche gemeinsam mit den Ergebnissen der Gespräche in diesem Beitrag präsentiert wird. Aufgrund stringenter Bedingungen hinsichtlich Geheimhaltung in der Fahrzeugentwicklung bei den Projektpartnern, kann nur ein Ausschnitt in diesem Beitrag im Detail vorgestellt werden.

1 Motivation

1.1 Informationsbeschaffung in der Fahrzeugentwicklung

Fahrzeugentwicklung zeichnet sich heute durch mehr Varianten und Derivate aus, als jemals zuvor, denn jede Nische zwischen Fahrzeugklassen wird sofort gefüllt. In der Folge drängen Hybrid- und Elektroantriebsvarianten sowie gänzlich neue Elektrofahrzeuge in das bestehende Produktportfolio (Bernasch und Fachbach, 2011). Zusätzlich sorgen wachsende Anforderungen an Fahrzeuge dafür, dass aufgrund explodierender Variantenvielfalt und steigender Produktkomplexität auch der Informationsbedarf in der Fahrzeugentwicklung stetig zunimmt.

Für das Informationsmanagement stellt diese Entwicklung eine große Herausforderung dar, denn Gesamt¬fahrzeugentwicklung und -optimierung macht eine intensive Zusammenarbeit über Disziplinen, Abteilungs- und Unternehmensgrenzen mehr denn je notwendig. Die rasche Verfügbarkeit relevanter Daten und Informationen ist heute einer der wesentlichsten Erfolgsfaktoren in der Fahrzeugentwicklung. Doch bisherige Methoden und Werkzeuge des Informations- und Datenmanagements sind derzeit nur bedingt erfolgreich (Dau und Konzag, 2010). Vor allem die virtuelle Entwicklung führt zu einer Vielzahl von Problemen und Herausforderungen im Datenmanagement (Geist und Vornholt, 2005). Um Ingenieure in ihrer jeweiligen Entwicklungsdisziplin bestmöglich zu unterstützen, sind dort historisch gewachsene abteilungsspezifische Insellösungen und Datencontainer entstanden. Viele Projekte mit dem Ziel der Konsolidierung und Zentralisierung, etwa in Datenbanken für

Product Lifecycle Management (PLM), scheiterten in der Vergangenheit an der mangelnden Akzeptenz der Nutzer.

Die Vermutung liegt nahe, dass es neuer flexiblerer Lösungen bedarf, um die Deckung des Informationsbedarfs der Entwickler zu befriedigen. Denn ein wesentlicher Enabler für effektive Zusammenarbeit zwischen Entwicklern liegt in der Schaffung eines optimalen Informationszugriffs auf Engineering-Daten, die in unzähligen Datenbanken oder in der Form „spezieller" Dokumente vorliegen. Dazu gehören Anforderungen, Stücklisten, Zeichnungen, Simulationsmodelle, Simulationsergebnisse, Testergebnisse, Meeting-Protokolle, Projektstatusdaten, Lessons Learned, uvm. Der Fahrzeugentwickler von heute muss sich intensiv mit der Identifikation von Wegen für die Informationsbeschaffung auseinandersetzen, was ihm je nach Erfahrung und Reputation im eigenen Unternehmen mehr oder weniger gut gelingen kann und wird.

1.2 Von der Internet-Suche zur Informationsvernetzung von Unternehmensinhalten

Im Internet ermöglichen heute einfach zu bedienende Suchmaschinen wie Google einen einfachen und umfassenden Zugang zu Informationen (Ellwein, 2002). Es liegt auf der Hand, dass Suchparadigmen und ansätze aus dem Web früher oder später in die Unternehmenslandschaft diffundieren. Vergleichbare Entwicklungen haben auch in anderen Bereichen bereits stattgefunden, etwa durch die Übernahme von Web 2.0-Anwendungen und -Technologien im Wissensmanagement (Stocker und Tochtermann, 2011).

In der wissenschaftlichen Community hat sich für Suchansätze im Kontext von Unternehmen der Begriff „Enterprise Search" etabliert. Am Markt finden sich heute zahlreiche Anbieter solcher Lösungen, wie beispielsweise Autonomy, Sinequa oder Intrafind. Doch bisher ist das von den Anwendern ersehnte „Google im Unternehmen" noch immer mehr Vision als Wirklichkeit (Lewandowski 2009). Denn die Suche in Unternehmensdaten stellt eine ebenso große Herausforderung für das Information Retrieval dar wie zuvor die Suche im Web (Hawking, 2004). Es gilt nun Suchansätze aus dem Web in die Unternehmenslandschaft zu holen und dabei die in den Unternehmen vorherrschenden Rahmenbedingungen hinschlich der Natur der Inhalte, des Verhaltens von Mitarbeitern und den wirtschaftlichen Rahmenbedingungen mit einzubeziehen (Mukherjee und Mao, 2004).

Vor diesem Hintergrund wurde Ende 2012 ein auf zweieinhalb Jahre angesetztes Forschungsprojekt gestartet. Gemeinsam mit drei Automobilherstellern und einem großen Zulieferer wird ein Konzept erarbeitet, das Engineering-Informationsquellen nutzt, den Zugriff auf diese ermöglicht und ohne tiefe physische Integration eine Suche und semantische Vernetzung von Informationen erlaubt. Als Projektergebnis wird eine prototypische Umsetzung eines Informationscockpits für den Zugriff auf dezentrale Informationen in der Fahrzeugentwicklung angestrebt. Dabei gilt es Ansätze zu finden, wie strukturierte und unstrukturierte Inhalte erschlossen und möglichst flexibel miteinander vernetzt werden können, um den Informationsbedarf von Fahrzeugentwicklern am Arbeitsplatz zu befriedigen. Diese Vernetzung muss sowohl systemgestützt, als auch interaktiv durch den Nutzer passieren können. Der entwickelte Ansatz muss die Suche nach Informationen möglichst nutzerorientiert, ansprechend und nachvollziehbar gestalten, damit er von Fahrzeugentwicklern akzeptiert wird.

Nach dieser Darstellung von Ausgangssituation und Motivation werden im nächsten Abschnitt die Grundlagen des Phänomens Suche in Unternehmensdaten erläutert. Es folgt eine Beschreibung der im Forschungsprojekt gewählten Methodik in Abschnitt drei. Abschnitt 4 präsentiert die Ergebnisse der rund 50 geführten Gespräche mit Personen aus der

Fahrzeugentwicklung in den Rollen Entwickler, Entscheider und IT und liefert Einblicke in die Architekturerstellung. In Abschnitt 5 werden die Ergebnisse dieses Beitrags diskutiert, während Abschnitt 6 mit einer Zusammenfassung und einem Ausblick auf weitere Forschungsarbeiten schließt.

2. Suche und Informationsvernetzung

2.1 Suche in Unternehmensdaten

Enterprise Search bzw. Suche in Unternehmensdaten ist ein Teilgebiet von Information Retrieval (Hawking, 2004) und bezeichnet die computergestützte inhaltsorientierte Suche mit Hilfe einer unternehmensinternen Suchmaschine, welche Inhalte mit Hilfe von speziellen Computerprogrammen, sog. Crawlern, indexiert. Die Suche wird nicht live auf den ursprünglichen Datenquellen durchgeführt, sondern auf dem zuvor erstellten Suchindex. Gefundene Dokumente werden im Kontext der Suchanfrage üblicherweise als Textauszug ergänzt um Metadaten angezeigt. Durch diese Vorschau soll sich schnell die Relevanz der Ergebnisse durch den Suchenden beurteilen lassen. Durch eine fortlaufende Indexierung der einzelnen Datenquellen wird die Aktualität der Resultate gewährleistet. Die Nutzung von Filtern, auch Facetten genannt, ermöglicht eine rasche Navigation durch die Trefferliste.

In der Literatur wird die Frage aufgeworfen, ob die akademische Information Research (IR) Community überhaupt genug über die in den Unternehmen vorliegenden Daten sowie den Such- und Informationsbedürfnissen ihrer Nutzer und deren Paradigma der Relevanzbewertung weiß (Hawking 2006). Obgleich sich Nutzer die Einfachheit der Google-Suche auch hinter der Firewall wünschen, zeichnen sich Internet-Suche und Intranet-Suche durch zahlreiche konzeptionelle Unterschiede aus:

- Bei der Internet Suche besteht der Suchraum vorwiegend aus verlinkten Web-Seiten und vergleichsweise einfach zu erschließenden Dokumenten in Standardformaten. Im Unternehmen sind Informationsressourcen heterogener, angefangen von unstrukturierten Dokumenten in speziellen Formaten bis hin zu strukturierten Inhalten in großen Datenbanken.
- Internet-Suchmaschinen wie Google bewerten die Relevanz von Webseiten anhand der auf sie zeigenden Hyperlinks [Page et al 1999]. Doch Daten und Dokumente sind in Unternehmen in der Regel überhaupt nicht verlinkt und daher bedarf es vollkommen anderer Ansätze zur Relevanzbewertung.
- Im Internet finden sich Informationen auf Web-Seiten, welche auch gefunden werden „wollen“ und daher speziell für Suchmaschinen optimiert wurden. In Intranets ist dies in den meisten Fällen nicht so, da Anwender in der Regel wenig Initiative ergreifen, um ihre Inhalte für Intranet-Suchmaschinen, etwa durch systematische Metadatenpflege, zu optimieren.
- Bei der Internet-Suche wird im öffentlichen Teil des Intranets gesucht, Zugriffsrechte sind daher irrelevant. Dokumente in Unternehmen werden oft durch komplexe Rollen- und Rechtestrukturen verwaltet, die Sicherstellung von Zugriffsbeschränkungen ist essentiell.
- Internet-Suchanfragen sind verglichen mit Intranet-Suchanfragen viel allgemeinerer Natur. Der Nutzer ist bei einer Internet-Suche in vielen Fällen mit einer Antwort zufrieden, wobei aber viele mögliche Lösungen existieren. Bei der Internet-Suche existiert oftmals nur das eine richtige Dokument, das rasch gefunden werden will.

2.2 Ansätze zur Informationsvernetzung

Gewissermaßen basiert Enterprise Search auch auf Informationsvernetzung. Denn bei Enterprise Search wird diese Informationsvernetzung über gemeinsame Eigenschaften von Dokumenten wie etwa gemeinsame Autoren durch Filterkategorien (Facetten) im Suchindex abgebildet. In der Wissenschaft wird Informationsvernetzung heute stark mit Forschung im Bereich von Semantic Web assoziiert, wobei vor allem die Forschung rund um ‚Linked Data' eine wichtige Rolle spielt.

Im Semantic Web (Berners Lee, 1998; Berners Lee, 2001) stellt ein Uniform Ressource Identifier (URI) eine Zeichenfolge dar, welche eine Ressource wie beispielsweise eine Person, eine Webseite oder einen Ort eindeutig identifiziert. Dabei ist es auch möglich, mehrere URIs für ein und dieselbe Ressource zu vergeben. Alle Informationen im Semantic Web werden maschinell verarbeitbar in der Form von Triples, also Statements bestehend aus Subjekt, Prädikat und Objekt abgebildet. Dabei beschreibt das Subjekt eine Ressource, das Prädikat die Eigenschaft dieser Ressource und das Objekt den Wert dieser Eigenschaft. Das Resource Description Framework (RDF) ist ein einheitliches Modell zur formalen Beschreibung von Informationen über Objekte. Alle Ressourcen werden einheitlich über URIs identifiziert. Das Ressource Description Framework Schema (RDFS) und die Web Ontology Language (OWL) stellen formale Sprachen dar, um die Bedeutung der verwendeten Vokabeln einheitlich zu spezifizieren.

Bei Linked Data handelt es sich um eine durch das World Wide Web Consortium (W3C) koordinierte technische Entwicklung. Linked Data ist ein durch Tim Berners-Lee veröffentlichtes Set an Prinzipien zur Publikation und Verlinkung strukturierter Daten im Web (Berners Lee, 2006). Linked Data nutzt bestehende W3C-Standards wie Uniform Ressource Identifier (URI) zur Identifikation von Ressourcen im Web, Resource Description Framework (RDF) zur Kodierung und Verlinkung von Daten und SPARQL Protocol and RDF Query Language (SPARQL) als graphbasierte Abfragesprache für RDF. Durch Linked Data werden die Daten am Web langfristig durch Maschinen interpretierbar gemacht und miteinander vernetzt. Das ermöglicht die Entwicklung von Softwaretools (oft als Semantic Web Agenten bezeichnet), die komplexe Informationsaufgaben für Menschen erledigen können.

Informationsvernetzung wird auch über Föderation von Datenbanken angestrebt und in dieser Domäne dann auch mit dem Begriff Informationsintegration versehen. Das Ziel der Informationsintegration besteht in der Schaffung einer zentralen, integrierenden Komponente mit einer einheitlichen Schnittstelle für Anwender und Anwendungen, um den Zugriff auf bestehende Informationssysteme zu erleichtern. (Vgl. Leser und Naumann, 2007). Die dabei entstehenden integrierten Informationssysteme stellen eine einheitliche Sicht auf die Datenquellen (und die darin enthaltenen Daten) zur Verfügung.

2.3 Nutzungshäufigkeit und Nutzerzufriedenheit von Enterprise Search

Eine Literaturrecherche hat aufgezeigt, dass es nur wenig empirische Arbeiten zu Nutzung und Nutzerzufriedenheit mit Suche in Unternehmensdaten gibt. In einer 2009 durchgeführten Untersuchung bei 140 Unternehmen wurde identifiziert, dass die Mehrzahl der Unternehmen Suchmaschinen heute zwar „verwendet", diese aber nicht die gesamten Informationen in Unternehmen erschließen. Ein zentraler Informationszugang fehlt zumeist und interne Suchmaschinen werden wesentlich weniger genutzt, als externe (Bahrs, 2009). In der Dissertation von Bertram (2011) wurden Ergebnisse zum Status Quo der unternehmensweiten Suche in österreichischen Großunternehmen aus Sicht von Wissensmanagement

vorgestellt. Dabei wurde herausgefunden, dass 70 Prozent der befragten Unternehmen noch keine Enterprise Search Lösung einsetzen.

Ein Blick in Arbeiten aus der Praxis zeigt folgendes Bild: Findwise (2012) führt mit dem „Enterprise Search and Findability Report“ regelmäßig eine praxisnahe Umfrage durch, um herauszufinden, wie Organisationen über die Einführung von Enterprise Search denken. Die Ergebnisse zeigen, dass es für die Mehrheit der befragten Unternehmen heute schwer ist, die relevanten Informationen zu finden. Gründe dafür liegen etwa in der schlechten Suchfunktionalität, der Inkonsistenz der Verschlagwortung von Inhalten, dem gänzlichen Fehlen adäquater Schlagworte, der schlechten Navigation oder der Unkenntnis, wo überhaupt nach Informationen gesucht werden soll. Fast die Hälfte der Anwender war überwiegend oder sehr unzufrieden mit ihrer Suchanwendung. Auch eine Studie von MindMetre (2012) kommt zu diesem Ergebnis, denn mehr als die Hälfte der dort befragten Anwender konnte die gesuchte Information mit der vorhandenen Lösung nicht finden.

Der Wissenschaft ist noch immer wenig über die speziellen Suchbedürfnisse bekannt, welche Nutzer in Unternehmen und vor allem Ingenieure in der Fahrzeugentwicklung hegen. Freund und Toms (2006) haben in einer Studie das Suchverhalten von Softwareentwicklern untersucht und dabei beispielsweise eine durchschnittliche Query-Länge von 4,38 Keywords identifiziert, wobei auffallend eine umfassende Nutzung spezieller Keywords vorgenommen wurde. Doch über die Nutzerzufriedenheit mit Enterprise Search finden sich lediglich Ergebnisse in den von Praktikern durchgeführten Umfragen.

2.4 Funktionsumfang und Anwendung in der Fahrzeugentwicklung

Ulbrich et al (2009) und Softic et al (2013) haben sich mit dem Vergleichen von Suchmaschinen-Technologien beschäftigt. In beiden Fällen ging es darum, anhand von Feature-Vergleichen den State of the Art von Suchmaschinen zu beschreiben, um die Auswahl von Anbietern zu erleichtern. Ulbrich et al (2009) untersuchen zusätzlich den Zugang der befragten Suchmaschinenanbieter zum Thema Wissenserschließung. Bahrs (2008) beschreibt eine Architektur für eine lernende Suchmaschine mit profil- und kontextorientiertem Ranking, welche sich besser an Umgebungen mit heterogenen Informationsquellen anpassen soll als bisherige. Ein zentrales Element stellt dabei cases based reasoning (CBR) dar, welche das Lösen von Problemen mit Hilfe von Erfahrungen aus ähnlichen Situationen ermöglichen soll.

Die Literaturrecherche konnte aufzeigen, dass es zwar vereinzelt Arbeiten zum Funktionsumfang am Markt erhältlicher Suchmaschinen gibt, jedoch wenig über Suchansätze in der Fahrzeugentwicklung bekannt ist. Wissenschaftliche Arbeiten von Blommquist et al (2007) und Maier et al (2003) diskutieren den Einsatz von Ontologien in der Automobilindustrie. Gehle (2000) beschreibt in seinem Beitrag zum IT-unterstützten Wissenstransfer einen mit „Marktplatz des Wissens“ bezeichneten Ansatz von BMW aus dem Jahr 1997 sowie dessen Architektur. Dieser Marktplatz soll Dreh- und Angelpunkt für den Informations- und Wissensaustausch sein und den Mitarbeitern einen zentralen Einstiegspunkt zu verschiedenen Informationsressourcen bieten. In diesem Zusammenhang wird auch eine Such- und Retrievalanwendung mit semantischen Netzen beschrieben. Dau und Konzag (2010) erläutern anhand eines Use Case bei BMW Grundlagen und Methoden für ein Informationssystem, das einem Fahrzeugentwickler Entwicklungsinformationen und Produktdaten aus der Konzeptentwicklung zugängig macht. Dabei werden Volltextsuche, semantische Suche, facettierte Suche und Visualisierung über Semantische Netze eingesetzt.

3 Methodik

Das Projektziel besteht, wie eingangs angeführt, in der Entwicklung eines Prototyps für flexible Suche und Informationsvernetzung in der Fahrzeugentwicklung. Dazu wurden gemeinsam mit den Projektpartnern zur Projektdefinition übergeordnet Projektziele definiert, Anforderungen erhoben und ein Konzept- und Architekturentwicklungsprozess gestartet. In diesem Beitrag wird lediglich die Methodik für die Konzeption der Architektur vorgestellt, welche in nachfolgender Abbildung dargestellt und dann im Detail beschrieben wird.

Im Forschungsprojekt wurde auf eine saubere Methodik der Architekturerstellung Wert gelegt und dazu ein toolgestützter Ansatz ausgewählt, der Transparenz und Nachvollziehbarkeit gestattet. Für die Modellierung wurde das Tool ARCWAY (arcway.com) ausgewählt, da es visuell ansprechende Darstellungen komplexer Zusammenhänge erlaubt und von einem durchgängigen Anforderungsmanagement über die Modellierung bis zur Projektkommunikation mit unterschiedlichen Stakeholdern unterstützt.

In einem ersten Schritt wurden von zwei Forschern Gespräche mit rund 50 auf die Projektpartner verteilte Personen in den Rollen Entwickler, Entscheider und IT-Mitarbeiter geführt:

- Den von den Projektpartnern ausgewählten Personen wurde durch den gesprächsführenden Forscher die Projektintention anhand eines gedruckten Projektsteckbriefs vorgestellt.
- In weiterer Folge wurden mit Hilfe eines Leitfadens Fragen zu Aufgaben und Verantwortung, Ablauf der Informationsbeschaffung und relevante Informationsquellen, besondere Herausforderungen in der Informationsbeschaffung und Vision/Idealszenario gestellt.
- Gesprächsprotokolle wurden vom gesprächsführenden Forscher gegengelesen, freigegeben und an die Gesprächspartner mit der Bitte um Durchsicht, Überarbeitung und Freigabe weitergeleitet, um die Gültigkeit des Inhalts sicherzustellen.

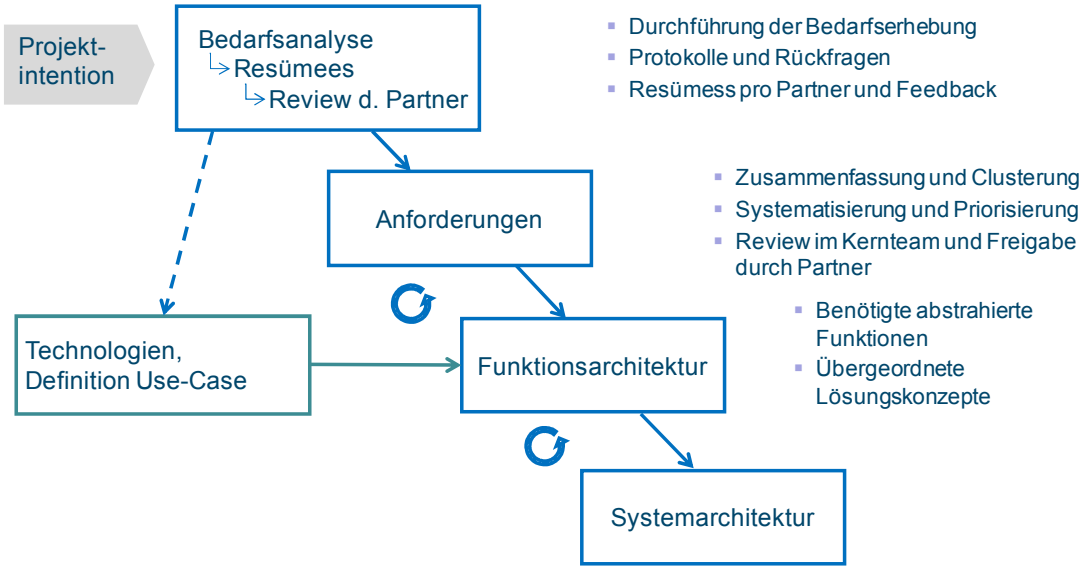


Abbildung 1: Vorgehensweise bei der Erstellung von Funktions- und Systemarchitektur.

In einem zweiten Schritt wurden die Ergebnisse aus den Gesprächen ausgewertet und in Stakeholder-Anforderungen überführt. Dazu wurde folgende Vorgehensweise gewählt.

- Die freigegebenen Gesprächsprotokolle wurden für jeden der vier Projektpartner auf der Basis von Gemeinsamkeiten und Unterschieden ausgewertet.
- Die wesentlichen Ergebnisse wurden je Projektpartner in einem Powerpoint-Foliensatz und einem Poster festgehalten.
- In einem Projektmeeting wurden die Ergebnisse mit den Projektpartnern diskutiert, um weitere Anforderungen an das zu entwickelte Informationscockpit zu generieren.
- Projektziele und Anforderungen wurden mit ARCWAY erfasst, wobei die rund 10 im Projektantrag definierten Projektziele die Spitze der Anforderungsmodellierung darstellen (Top-Level-Ziele).
- Die Anforderungen der Projektpartner wurden aus den Resümees der Interviews und der Diskussion am Projektmeeting je Projektpartner erfasst, mit diesem abgestimmt und freigegeben (Stakeholder-Anforderungen).
- Durch das wissenschaftliche Projektteam wurden die Anforderungen mit dem State of the Art zu Suche in Unternehmen ergänzt (erweiterte Anforderungen).
- In einem letzten Schritt wurden die Anforderungen zusammengefasst und nach Anwendungsfall oder Grob-Architektur kategorisiert und mit den Projektpartnern abgestimmt (konsolidierte Anforderungen).

Erst in einem dritten Schritt wurde die eigentliche Architekturentwicklung getrennt nach fachlicher und technischer Architektur gestartet. Dazu wurde folgende Vorgehensweise gewählt:

- Ein grobes Lösungskonzept wurde bereits im Rahmen der Projektanbahnung erarbeitet, da ohne dieses das Forschungsprojekt gar nicht zustande gekommen wäre.
- Die Prinzip-Architekturen von Enterprise Search Anbietern wurden auf Basis der zur Verfügung gestellten technischen White-Paper sowie einzelner Gespräche mit Vertretern in ARCWAY modelliert (Technologiebausteine).
- Gemeinsam mit den Projektpartnern wurden Use-Cases definiert, die eine Entwicklung und ein Testen des Funktionsmodells zulassen, welches möglichst unabhängig von der Implementierung sein soll (Funktionsarchitektur).
- Aufgrund des Wissens über die Funktionsarchitektur und der Technologiebausteine kann als Gesamt-Lösungsansatz die IT-technische Darstellung entstehen (Systemarchitektur).

4 Vorstellung der Ergebnisse

4.1 Bedarfsanalyse

Aus der Literatur ließen sich kaum Erkenntnisse gewinnen, wie heute Fahrzeugentwickler ihren Informationsbedarf decken und welche Anforderungen sie an eine Suche haben. Das deckt sich auch mit der Aussage von Hawking (2006), dass die Information Retrieval Community womöglich noch zu wenig über die in den Unternehmen vorliegenden Daten sowie über Such-, Informations- und Relevanzbewertungsmechanismen ihrer Nutzer weiß. Die im Projekt durchgeführten Gespräche mit Methodenentwicklern (12), Entwicklern (12), Entscheidern (9) und fachbereichsnahen IT-Mitarbeitern (13) lieferten tiefe Einblicke über kulturelle Aspekte bei der Informationsbeschaffung in der Fahrzeugentwicklung, vorhandene Informationsquellen und -typen, besondere Herausforderungen bei der Informationsbeschaffung und Visionen und Anforderungen an ein Informationscockpit.

Zusammengefasst kann aus der Perspektive Organisation und Kultur gesagt werden, dass heute bei den Projektpartnern das Need-to-Know-Prinzip bei der Informationsversorgung vorherrschend ist: Aktuell erhalten Ingenieure formal und IT-basiert nur Zugriff auf diejenigen Informationen, die zwingend benötigt werden. Alles was darüber hinaus für eine optimale Zusammenarbeit gebraucht wird, passiert über informelle Kommunikation und Netzwerke.

Dieser Zielkonflikt zwischen Unternehmens-IT-Strategie (pro: Datensicherheit und Einführung von Schutzklassen) und der Bedarf der Entwicklung nach einen besseren Zugang zu Informationen (pro: offenerer Zugang zu Information) ist erkannt und bewusst. Allerdings ist noch keine Lösung in Sicht. Auch das Verständnis der Notwendigkeit der Weiterverwendung von Engineering-Daten (z.B.: Simulationsmodelle, Input-Files) durch andere Fachbereiche speziell für Systemsimulation ist erst im Aufbau begriffen.

Als Datenquellen wurden in den Gesprächen fast ausschließlich Laufwerke genannt. Für Computer-aided Engineering (CAE) werden bei den Gesprächspartnern noch keine umfassenden Simulationsdatenmanagement-Systeme eingesetzt. In Einzelfällen sind lokale Datenbanken für CAE-Daten vorhanden. Zunehmend werden Wikis als Quelle für Informationen, Strukturen und Übersichten genutzt. Die Schaffung von Übersicht über Daten und die Steuerung wird oftmals mit Hilfe von Excel-Tabellen durchgeführt, welche beispielsweise in der Root von Ordnern liegen und die Inhalte beschreiben.

Informationstypen betreffen entsprechend dem Projektfokus schwerpunktmäßig Simulationsmodelle und Daten für die Bedatung dieser Modelle (z.B.: Input Decks). Als weitere relevante Dokumente wurden Produktdaten (z.B. Konfiguration und Ausstattung), Produktparameter (z.B. Getriebeabstufung), Simulations- und Messdaten, CAD-Daten und Projektdaten (Anforderungen, Status, Entscheidungen, Änderungen, Planung) genannt.

In der Informationsbeschaffung stehen heute persönliche Kommunikation und Nutzung des eigenen Netzwerks im Vordergrund. Wenn Prozessdefinitionen für die Informationsversorgung vorhanden sind, sind diese in erster Linie meilensteinorientiert. Abstimmungen und Synchronisationen finden zu Meilensteinen statt, zwischen den Meilensteinen gibt es wenig formale Prozesse, was die Informationsversorgung dort wesentlich erschwert. Generell finden sich im CAE-Bereich (Simulation) weniger Definitionen, als im CAD Bereich (Konstruktion). Das Ausmaß an definierten Prozessen variiert je nach Projektpartner.

Die Weiterverwendung von Berechnungsdaten ist eine neue Entwicklung und teilweise noch nicht etabliert. Auch die Verantwortung für die Dokumentation und Beschaffung von Daten für die Berechnung ist meist nicht hinreichend geklärt, was das Pull-Prinzip notwendig macht. Es ist nicht geklärt ob der Nutzer oder der Ersteller von Engineering-Daten für deren Verwendung verantwortlich ist. Diese mangelnde Klärung sorgt dafür, dass Daten oft nur widerwillig oder gar nicht zur Verfügung gestellt werden.

Kernfragestellungen betreffen aus Sicht der Fahrzeugentwickler die Recherche nach Informationen mit Fokus auf die frühe Phase, d.h. das Verschaffen eines ersten Überblicks und die Herstellung von Kontakten zu für bestimmte Entwicklungsthemen relevanten Personen, das Abonnement von Änderungen an Inhalten und letztendlich den eigentlichen Zugriff auf Daten über das Informationscockpit, was Vertrauen in die Relevanz und Gültigkeit der gefundenen Inhalte betrifft. Der Kontext einer gefundenen Information soll systemtechnisch erfragt werden können, um die Gültigkeit und Relevanz zu bewerten (z.B. von welcher Person wurde zu welchem Zeitpunkt ein gefundenes Simulationsmodell in welchem Fahrzeugentwicklungsprojekt verwendet).

4.2 Architekturentwicklung

Im Projektantrag wurden die folgenden acht Ziele (High-Level Anforderungen) definiert: Komplexität und Dynamik handhaben, interdisziplinäre Zusammenarbeit ermöglichen, effizienten Informationszugriff ermöglichen, flexible Einbindung von Informationsquellen ermöglichen, Vernetzung von Informationen ermöglichen, Kontext von Information erschließen, kontextgetriebenes Rollen- und Rechtekonzept erarbeiten, hohe User-Akzeptanz sicherstellen.

Für jeden Partner wurden separat Stakeholder-Anforderungen erhoben und in die folgenden Kategorien eingeteilt: Informationstyp, Informationsdarstellung, Personalisierung und Interaktion, Gesamtarchitektur, Technische Informationsquellen, Zugriffsrechte, Nicht-funktional und Randbedingungen. In Summe wurden rund 150 solcher Stakeholder-Anforderungen erhoben. Die Ergebnisse aus der Bedarfserhebung und den dort geführten Gesprächen liefern die Basis dazu. Rund 70 ergänzende Anforderungen wurden vom Projektteam mit dem Ziel definiert, die Lücken zwischen den Anforderungen der Stakeholder und den Zielen abzudecken. Diese umfassen die Bereiche Frontend und Benutzerinteraktion (Sucheingabe und Filterung, Informationsdarstellung, Personalisierung, Allgemein), Relevanzbewertung und Kontext, Gesamtarchitektur und Funktionen, Quellsysteme und Dokumente (Unstrukturierte Datenquellen, Strukturierte Quellsysteme, Dateitypen, Informationstypen) sowie Handlungsfelder und Randbedingungen.

Die Stakeholder-Anforderungen sowie die ergänzenden Anforderungen wurden zu rund 40 konsolidierten Anforderungen zusammengefasst. Konsolidierte Anforderungen wurden in folgende Kategorien geclustert: Frontend (Personalisierung, Interaktion Benutzer, Informationsdarstellung Benutzer, Interaktion Vernetzungsadministrator, Informationsdarstellung Vernetzungsadministrator), Business Logic, Technische Informationsquellen und -typen, nicht funktionale Anforderungen, IT-Anforderungen, sowie Handlungsfelder und Randbedingungen.

Das Konzept für flexible Suche und Informationsvernetzung in der Fahrzeugentwicklung wird sowohl als „fachliche Architektur“ mit Fokus auf das WAS, als auch als „technische Architektur“ mit Fokus auf das WIE modelliert. Sie ist einerseits durch Aufbaubilder (Pläne, statische Aspekte) und andererseits Ablaufbilder (Funktionsabläufe, dynamische Aspekte) dargestellt und wurde mit Hilfe der Fundamental Modeling Concepts (Knöpfel et al 2005) modelliert. Im Allgemeinen wird die Kompositionsstruktur als Blockdiagramm mit aktiven (Agenten als Rechtecke), passiven Komponenten (Informationsspeicher als runde Form) und Kommunikationskanälen (Linien und Kreise) genutzt. Im folgenden Beitrag kann nur ein Auszug der Architektur, die vier Detaillierungsstufen umfasst und in rund 20 Plänen abgebildet wird, wiedergegeben werden.

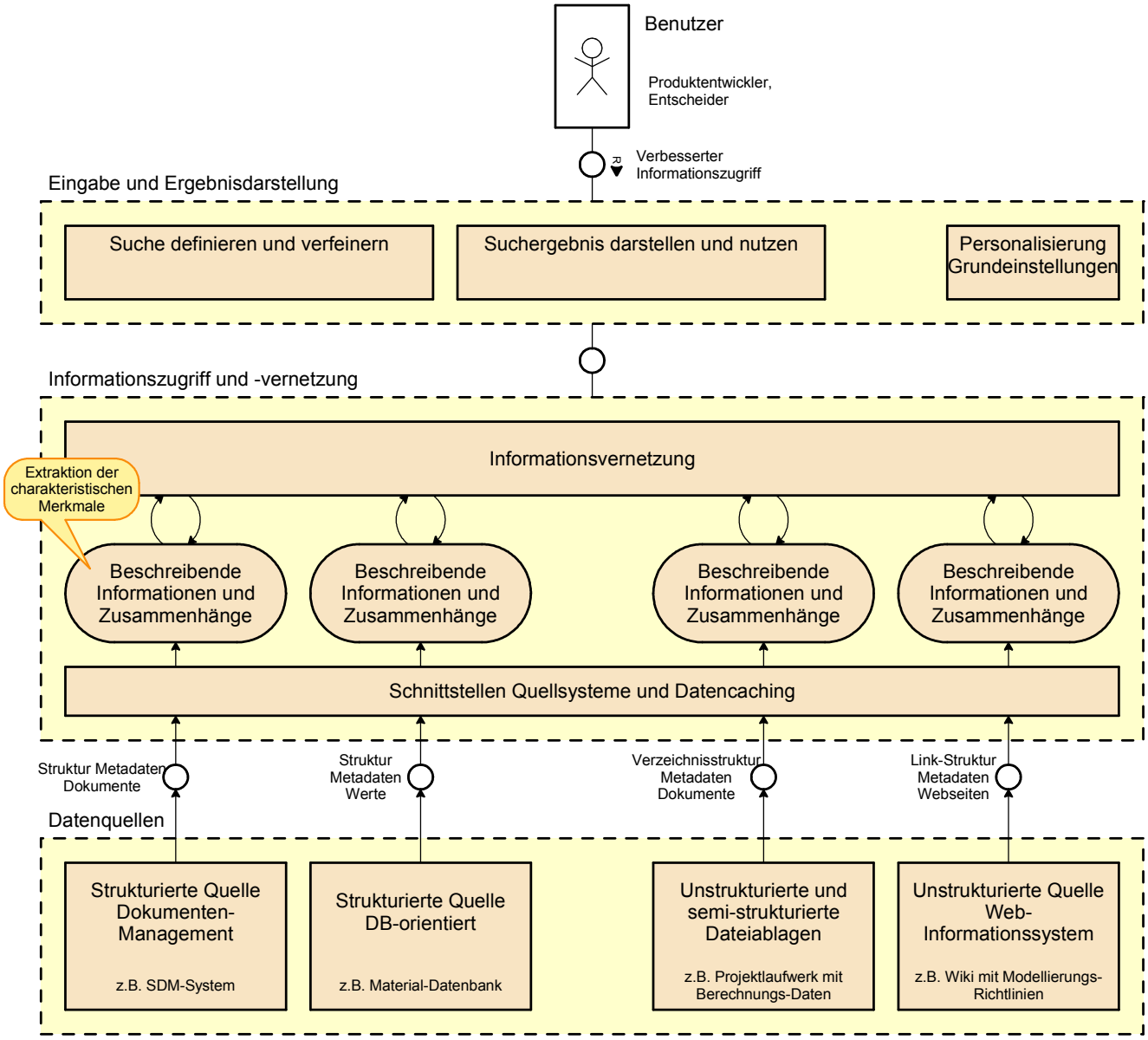


Abbildung 2: Grobkonzept der Architektur für flexible Suche und Informationsvernetzung (1. Detaillierungsstufe).

Die fachliche Architektur folgt einer klassischen „Drei Schichten Architektur“ betrieblicher Informationssysteme mit Benutzungsschnittstelle, Anwendungskern und Datenhaltung (Keller, 2002) und beschreibt die Funktionen Eingabe und Ergebnisdarstellung, Informationszugriff und vernetzung sowie Datenquellen. Aus den an das Informationscockpit durch den Administrator angebundenen Datenquellen mit strukturierten und unstrukturierten Daten werden Inhalte, beschreibende Informationen und Ergebnisse extrahiert und miteinander vernetzt. Der Benutzer kann über ein Frontend Suchanfragen formulieren und Ergebnisse kontextbezogen analysieren, filtern und Zusammenhänge eruieren.

Insgesamt werden vier unterschiedliche Datenquellen für Engineering-Daten unterschieden: Strukturierte Quellen mit Dokumenten im Dateisystem und dazugehörigen Metadaten in einer Datenbank (z.B.: Simulationsdatenmanagement-System), strukturierte Quellen als Datenbank mit Werten (z.B.: Materialdatenbank), unstrukturierte bzw. semi-strukturierte Dateiablagen mit Dokumenten (z.B.: Projektablage die mehr oder weniger einem festgelegten Schema folgt) und unstrukturierte bzw. semistrukturierte Web-Informationssysteme

(z.B.: Berechner-Wiki). Analysiert werden die Struktur der Inhaltsorganisation (z.B.: nach Projekten und Disziplinen geordneter Baum von Berechnungstasks), die Metadaten der Inhalte (z.B.: Bearbeiter, Fachbereich, Änderungsdatum, Freigabestatus, …), die Inhalte selbst (z.B.: Berechnungsmodelle, Input-Decks, Analyseergebnisse) und Werte (z.B.: Testergebnisse, Kennfelder).

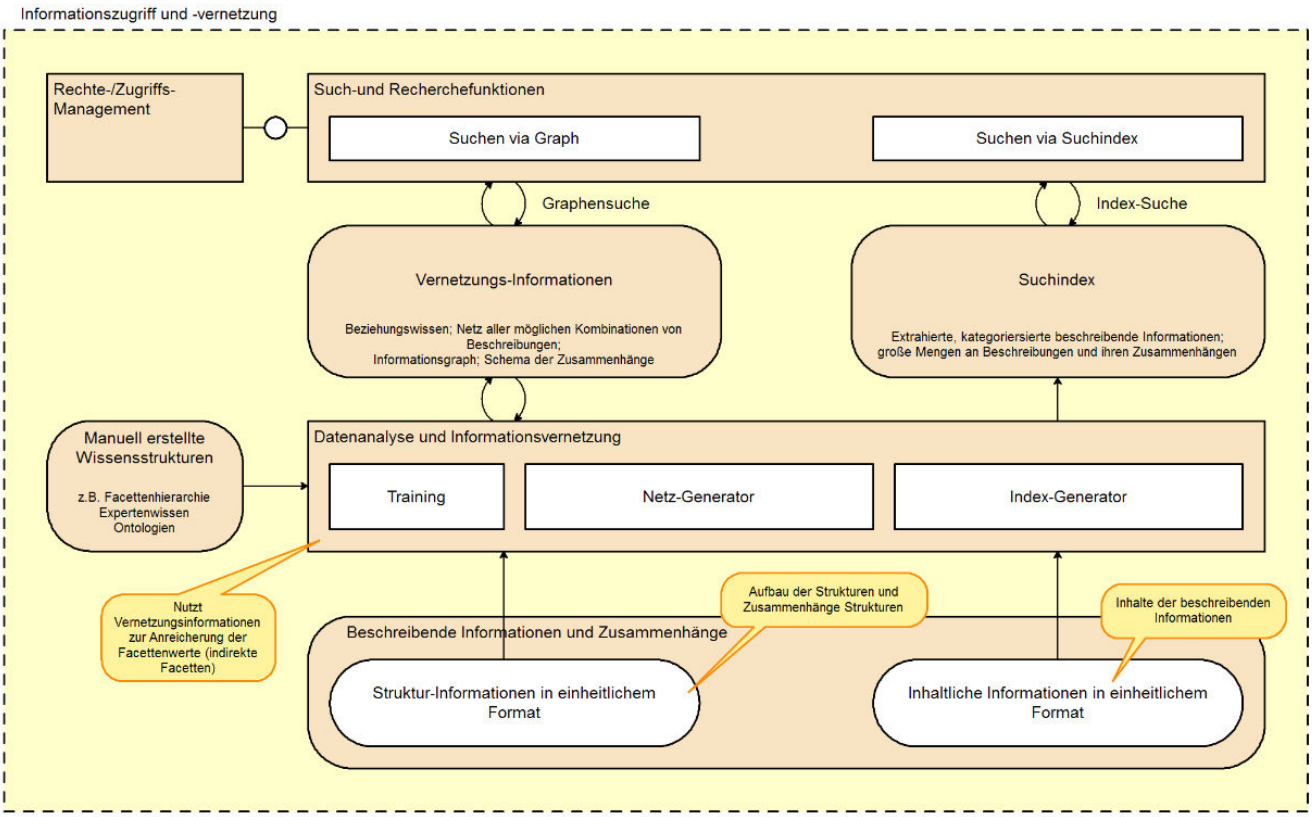


Abbildung 3: Layers „Informationszugriff und -vernetzung“ (2. Detaillierungsstufe).

Sobald die beschreibenden Informationen aus den unterschiedlichen Quellen extrahiert und in ein einheitliches Format gebracht wurden, werden sie durch Datenanalyse und Informationsvernetzung miteinander in Bezug gebracht und kategorisiert. Die Strukturinformationen der Quellsysteme werden, so gut es geht, dazu verwendet, um die Analysequalität zu verbessern. Die Analyse erfolgt durch Hinzuziehen von manuell- sowie systemtechnisch erstellten Wissensstrukturen (z.B. Organisationsstruktur, Fahrzeugprojektstruktur, Baugruppenstruktur, Stücklisten). Dabei werden moderne Ansätze wie Linked Data einbezogen.

Das Informationscockpit unterscheidet zwischen Strukturinformationen (bzw. Vernetzungsinformationen) und Inhaltsinformationen. Vernetzungsinformationen bestehen aus dem Mapping verschiedener Wissensstrukturen (z.B. Fahrzeugprojektstruktur, Bauteilstruktur, …) auf eine zentrale Systemstruktur. Zugeordnete und verknüpfte Informationen werden im Suchindex gespeichert, der aus unstrukturierten Inhalten generiert wird. Aus Sicht der Such- und Recherchefunktionen steht ein Metadaten-Netz mit direkt und indirekt verknüpften Informationen zur Verfügung. Sowohl eine Suche über einen Suchindex (d.h. über die Inhalte der Dokumente und Datenbanken), als auch über einen Graph (d.h. über das Beziehungswissen zwischen den Inhalten) wird unterstützt.

Neben der Möglichkeiten einer klassischen Enterprise Search wie Schlagwortsuche, Boolesche Suche und Facettierte Suche (Mukherjee und Mao, 2004) soll das Informationscockpit weitere Funktionen beinhalten. Dazu gehören wie die Suche nach relevanten Dokumenten auch die Suche nach zu den Suchergebnissen relevanten Datenbanken und die Ausführung vordefinierter Datenbankabfragen. Damit können etwa vordefinierten Datenbankabfragen auch als Suchergebnisse angezeigt und durch den suchenden Benutzer ausgeführt werden. Eine Kombination aus Suchmaschine und strukturierter Abfrage auf einer einfach zu nutzenden Oberfläche wird ein zentrales Ergebnis darstellen.

5. Diskussion der Ergebnisse

5.1 Begleitende Rahmenbedingungen

Die vorgestellte Architektur für flexible Suche und Vernetzung in der Fahrzeugentwicklung erlaubt die Entwicklung eines funktionalen Prototyps. Eine bei den Partnern betriebene Suchlösung, die auf diesem Prototyp aufbaut, kann nur dann nutzenstiftend sein, wenn bestimmte Rahmenbedingungen bei den Projektpartnern erfüllt werden. Für die Ausschöpfung des Potenzials des zu entwickelten Ansatzes für flexible Suche und Informationsvernetzung wurden anhand der Gespräche eine Reihe von Risiken identifiziert: Allgemein ist bei allen Projektpartnern sehr wenig Beschreibung zu Simulationsmodellen vorhanden, d.h. Eingangsinformation (z.B.: CAD-Stand), Versionsbeschreibung, Verwendungszweck und Intention, Gültigkeitsbereich und Schnittstellenbeschreibung sind kaum dokumentiert. In vielen Fällen liegt zu wenig Information für die Bewertung von Kontext und Zusammenhängen vor, welche durch das entwickelte System angezeigt werden kann. Generell sind Metadaten wenig oder nicht vorhanden und eine aktive Metadatenpflege wird kaum durchgeführt.

Das heute vorherrschende Need-to-Know-Prinzip widerspricht tendenziell dem Ansatz des Forschungsprojekts, welches einen offeneren Zugang zu Information vorsieht. Wer keinen Zugriff auf eine Information haben darf, der darf sie auch in einer Suchmaschine nicht finden. Im Regelfall werden Suchende mit geringen Zugriffsrechten nur sehr wenige Inhalte erschließen können und das Informationscockpit dann kaum als nützlich empfinden. Das Forschungsprojekt beinhaltet neben den technologischen Aspekten auch die Erarbeitung von begleitenden Handlungsempfehlungen. Darin sollen die organisatorischen, prozesstechnischen und kulturellen Auswirkungen untersucht, optimale Rahmenbedingungen oder Voraussetzungen definiert und Handlungsempfehlungen abgeleitet werden. Es bedarf eines Konzepts für die Verantwortlichkeit für Dokumentation und Beschaffung von Informationen, welches vor allen eine Win-Win-Situation zwischen Ersteller und Nutzer von Modellen ermöglichen soll. Auch der Einfluss der Unternehmenskultur auf die Potentialausschöpfung wird als enorm erkannt.

5.2 Suche und Vernetzung von Inhalten in Unternehmen

Obwohl mittlerweile zahlreiche Anbieter von Enterprise Search auf dem Markt sind, existiert noch kaum wissenschaftliche Literatur über Nutzung und Zufriedenheit. Die einzelnen über Literaturrecherche identifizierten wissenschaftlichen Studien (Bahrs, 2009; Bertram, 2011) zeigen an, dass die Nutzung und Verbreitung von Enterprise Search wohl doch geringer ist, als zu erwarten wäre, und der zentrale Zugang zu Information noch offen ist. Die beiden referenzierten Studien aus der Praxis (Findwise, 2012; MindMetre, 2011) zeigen eine sehr verhaltene Zufriedenheit der Anwender mit existierenden Lösungen.

Für die Forschung bedeutet das, die Charakteristik von Informationssuche und -vernetzung in Unternehmen, den konkreten Informationsbedarf der Anwender und deren Relevanzbewertungsmechanismen weiter zu untersuchen, um die Entwicklung maßgeschneiderter Lösungen zu erlauben. Die wissenschaftliche Literatur bietet derzeit lediglich allgemeine Erkenntnisse. Aus diesem Grunde wurden für die Architekturentwicklung im Projekt auch rund 50 Gespräche geführt. Ein Informationscockpit für Suche und flexible Informationsvernetzung in der Fahrzeugentwicklung muss viel stärker auf die Besonderheiten der Domäne eingehen, da generische Enterprise Search Lösungen nicht den gewünschten Erfolg bringen, wie auch die Analysegespräche aus der Praxis vermuten lassen. Hier liegen relevante Informationen nicht nur in Form von Dokumenten vor, sondern in unterschiedlichen Ausprägungen. Aus der optimalen Verknüpfung von strukturierten mit unstrukturierten Inhalten kann ein großes Potential für den Fahrzeugentwickler erschlossen werden.

6 Zusammenfassung und Ausblick

Die rasche Verfügbarkeit relevanter Daten und Informationen ist ein wesentlicher Erfolgsfaktor in der Fahrzeugentwicklung. Viele Projekte mit dem Ziel der Konsolidierung und Zentralisierung der Informationsversorgung über monolithische Systeme (z.B.: PLM) sind in der Vergangenheit in großen Unternehmen an der mangelnden Akzeptanz der Nutzer gescheitert. Heute sind flexible Lösungen zur Suche und Informationsvernetzung mehr denn je gefragt, um die historisch gewachsenen dezentralen Lösungen zu erschließen.

Im vorliegenden Beitrag wurde ausgehend von der Einführung in den State-of-the-Art zur Suche im Unternehmenskontext ein diesbezügliches Forschungsprojekt mit drei Automobilherstellern und einem großen Zulieferer vorgestellt. In rund 50 Gesprächen mit Fahrzeugentwicklern hat sich gezeigt, dass Informationsbeschaffung heute hauptsächlich über Gespräche und Networking stattfindet und in keinem der Fälle bereits ein umfassender IT-gestützter zentraler Zugang zu Information vorliegt. Diese Ergebnisse lieferten den Grundstein für ein umfangreiches Projekt zur Entwicklung eines Informationscockpits.

Für das Forschungsprojekt war eine Methodik notwendig, die alle Schritte von der Definition des Projekts, der Erhebung der Anforderungen bis hin zur Entwicklung der Architektur nachvollziehbar und transparent abbildet. Aus diesem Grund wurde ein strukturierter, toolgestützter Ansatz gewählt und ausführlich im Beitrag vorgestellt. Gemeinsam mit den Partnern wurden Anforderungen abgeleitet, verfeinert, diskutiert und in eine funktionale Architektur für flexible Suche und Informationsvernetzung in der Fahrzeugentwicklung überführt. Grundlage für diese Architektur stellte eine intensive Beschäftigung mit dem Themen Enterprise Search und Informationsvernetzung, dem State of the Art dazu sowie den Kapazitäten der diesbezüglichen Anbieter dar. Des Weiteren wurden die Bedürfnisse der zukünftigen Anwender integriert. Die vorgestellte Architektur soll keine Suchlösung für allgemeine Anwendungsfelder darstellen, sondern speziell den optimalen Informationszugriff in der Fahrzeugentwicklung erleichtern und mit den dortigen Nutzungs-Charakteristiken, dem Informationsbedarf und der vorhandenen Systemlandschaft umgehen können.

In weiterer Folge ist die prototypische Umsetzung geplant. Dazu liegt bereits ein Frontend-Demonstrator vor, der als zweites Instrument neben der Modellierung zur Kommunikation von Funktionsweise und Nützlichkeit bei den Partnern dient. Derzeit wird ein Systemlabor mit von den Partnern erhaltenen Testdaten aufgesetzt, um mehr über deren Eigenheiten zu erforschen.

Danksagung

Die Autoren danken dem „COMET K2 Forschungsförderungs-Programm“ des Österreichischen Bundesministeriums für Verkehr, Innovation und Technologie (BMVIT), des Österreichischen Bundesministeriums für Wirtschaft, Familie und Jugend (BMWFJ), der Österreichischen Forschungsförderungsgesellschaft mbH (FFG), des Landes Steiermark sowie der Steirischen Wirtschaftsförderung (SFG) für die finanzielle Unterstützung. Ebenfalls danken wir den unterstützenden Firmen und Projektpartnern sowie der Technischen Universität Graz.

Literatur

Bahrs, J. (2009): Enterprise Search – Suchmaschinen für Inhalte im Unternehmen. In: *Handbuch Internet-Suchmaschinen*, Heidelberg: AKA Verlag.

Bahrs, J.; Meuthrath, B.t; Peters, K. (2008): Selbstlernende Suchmaschine als zentraler Informationszugang bei heterogener Informationslandschaft. In: *Beherrschbare Systeme dank Informatik*, Proceedings 133, Band 1. Springer (Berlin), S. 365-371.

Berners Lee, T. (1998): *Semantic Web Roadmap*. www.w3.org/DesignIssues/Semantic.html.

Berners Lee, T.; Hendler, J., Lassila, O. (2001): The Semantic Web. In: *Scientific American*, May 2001, p. 29-37.

Berners Lee, T. (2006): Linked Data, http://www.w3.org/DesignIssues/LinkedData.html

Bernasch, J., Fachbach, B. (2011): Integration als Schlüssel für die virtuelle Fahrzeugentwicklung, *ATZextra*, April 2011, Volume 16, Issue 5, pp 18-23.

Betram, J. (2011): *Informationen verzweifelt gesucht – Enterprise Search in österreichischen Großunternehmen*. Dissertation, 2011.

Blomqvist, E.; Öhgren, A. (2008): Constructing an enterprise ontology for an automotive supplier. In: *Engineering Applications of Artificial Intelligence*, Volume 21, Issue 3, April 2008, Pages 386–397.

Dau, F.; Konzag, A. (2010): Management von Entwicklungsinformationen und Produktdaten als eine Grundlage zur Erhöhung der Qualität von Fahrzeugkonzepten. Text- und Data Mining für die Qualitätsanalyse in der Automobilindustrie. Gerhard Heyer, *Leipziger Beiträge zur Informatik* (Sonderheft), 2010.

Ellwein, C. (2002): *Suche im Internet für Industrie und Wissenschaft*. München: Oldenbourg Wissenschaftsverlag.

Findwise (2012): *Enterprise Search and Findability Survey 2012*, http://findwise.com/solutions/enterprise-search-and-findability-report

Gehle, M. (2000): IT-unterstützter Wissenstransfer in der internationalen Forschung & Entwicklung: Ein Praxisbericht über den „Marktplatz des Wissens“ der BMW AG, *Wirtschaftsinformatik*, Oktober 2000, Volume 42, pp S119-S123.

Geist, I.; Vornholt, S. (2005): Management of User-defined Meta Information for Virtual Product Development. In: *Proceedings of 9th International Conference on Database and Expert Systems Application*, 2005.

Hawking, D. (2004): Challenges in Enterprise Search. In: *Proceedings of the 15th Australasian database conference* - Volume 27, Pages 15-24.

Hawking, D. (2006): Enterprise Search — The New Frontier? *Advances in Information Retrieval. Lecture Notes in Computer Science* Volume 3936, 2006, p 12.

Keller, W. (2002): *Enterprise Application Integration*. Heidelberg: dpunkt Verlag, 2002, ISBN 3-89864-186-4

Knöpfel, A.; Gröne, B.; Tabeling, P. (2005): *Fundamental Modeling Concepts. Effective Communication of IT Systems*. Chichester: John Wiley & Sons Ltd.

Leser, U.; Naumann, F. (2007): *Informationsintegration, Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Heidelberg: dpunkt Verlag, 2007.

MindMetre (2011): *Mind the Enterprise Search Gap*: A Mind Metre Research Report sponsored by Smart Logic, 2011 http://www.smartlogic.com/home/knowledge-zone/white-papers/1600-mindmetre-research-report-sponsored-by-smartlogic

Mukherjee, R.; Mao, J. (2004): Enterprise Search: Tough Stuff. In: *Queue - Search Engines*, Volume 2 Issue 2, April 2004 Pages 36.

Page, L.; Brin, S.; Motwani, R.; Winograd, T. (1999): *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Softic, S.; Rosenberger, M.; Zoier, M.; Mondelos, K.; Pillinger, E. (2013): A Preliminary Short Survey of State of the Art Enterprise Search Engines for Future Work Place. In: *Proceedings of WEBIST 2013 – 9th International Conference on Web Information Systems and Technologies*, Aachen, 2013.

Stocker, A.; Tochtermann, K. (2011): *Wissenstransfer mit Wikis und Weblogs*. Fallstudien zum erfolgreichen Einsatz von Web 2.0 in Unternehmen. 2. Auflage. Gabler Verlag, 2011.

Ulbrich, A.; Kraker, P.; Luidolt, C. (2009): Ergebnisse einer Untersuchung von Anbietern von Suchmaschinen-Technologien. In: *Proceedings of Wissensmanagement*. 2009, 118-127.

Notizen

Von der Dokumentation zum organisationalen WissensPerformanceSystem

Klaus Mak, Christian Meurers, Johannes Göllner & Robert Woitsch

„Von der Dokumentation zum organisationalen WissensPerformanceSystem“ beschreibt die Entwicklung einer Dokumentationseinrichtung und ihren Hauptaufgaben in Zusammenhang mit den notwendigen Schritten zu einem operativen Unterstützungselement für ein Unternehmen. Die Steuerung durch Wissen als „Idee“ im Hintergrund leitete mehrere Jahre Entwicklungs- und Forschungstätigkeit im Bereich des Dokumentations- und Fachinformationswesens und des Wissensmanagements in der Zentraldokumentation („ZentDok“) der Landesverteidigungsakademie. Neben den durchzuführenden Arbeiten des praktischen Dokumentations- und Fachinformationswesens im gesamten Unternehmen, wurde auch ein F&E-Bereich kontinuierlich in der Organisation aufgebaut und etabliert. Die Ergebnisse dieses Beitrages entstanden somit aus mehr als vierzig Jahren praktischer Erfahrung in der Organisation, wie auch aus unterschiedlichsten Projekten mit Partnern aus der Industrie und der Forschung.[1] Der Beitrag beschäftigt sich zunächst mit dem organisatorischen Hintergrund und der zugrundeliegenden WM-Strategie im ÖBH. In weiterer Folge wird der konzeptuelle Hintergrund und der theoretische Ansatz sowie das entwickelte Referenzarchitekturmodell vorgestellt, bevor anhand eines durchgeführten Projektes zur Erstellung einer Forschungsbilanz für das ÖBH die praktische Umsetzung beschrieben wird.

1 Einleitung

Organisatorisch entwickelte sich die ZentDok bis 2013 zu einem Instrument für Unterstützungs- und Dienstleistungen sowie Anwendungsentwicklungen für das Österreichische Bundesheer (ÖBH) in den Bereichen: Dokumentation, Fachinformation, Wissensmanagement und Wissensentwicklung.

Diese Aufgaben sind für alle Unternehmensbereiche abzudecken und mit Erstellung von Konzepten, Bereitstellung von Open Source Content, Entwicklung von Verfahren und Methoden, sowie Werkzeugen zur Wissensentwicklung, aber auch bei der Entwicklung von Fähigkeiten der Mitarbeiter, sicherzustellen.

Ausgangspunkt dabei ist nicht nur ein pragmatischer Zugang zu den Begrifflichkeiten sondern parallel dazu die Entwicklung von anwendbaren Vorgangsweisen, Methoden und Werkzeugen, die in Pilotprojekten in unterschiedlichsten Unternehmensdomains entwickelt und getestet sowie bereits in Teilorganisationen eingeführt wurden.

Nachfolgend wird der Begriff Wissen als erforderlicher „Wert“ in allen Ausprägungsformen implizit oder explizit als eine zentrale Ressource definiert und durch heuristische Modelle für die jeweiligen Anforderungen in den Geschäftsprozessen analysiert, modelliert und dokumentiert.

Wissen ist die Voraussetzung für eine Handlung oder Nichthandlung in einem System oder einer Domain. Wissen sind somit „interpretierte Daten oder Informationen“, die in einem

[1] Die detaillierten Inhalte des Beitrags sind in der Schriftenreihe der Landesverteidigungsakademie publiziert worden und bei Bedarf verfügbar. (Woitsch, Mak & Göllner 2010; Woitsch et.al. 2010a)

Entscheidungsprozess gleichzeitig „Gewissheit“ schaffen müssen. Daten und Information sind in sich nicht zu Interpretation fähig.

Die Grenzen der Vollständigkeit und Richtigkeit der Informationen sind neben mangelhafter Interpretationsleistung und Zeitmangel bestimmend für die Höhe des Risikos, der Chancen und der (Un-)Sicherheit der (Nicht-)Entscheidung!

Die erforderliche „Interpretationsleistung“ oder „Relevanzprüfung“ muss von einer „Entscheidungsinstanz“ geleistet werden. Diese „Entscheidungsinstanz“ kann ein Mensch oder eine Maschine sein, wobei „implizites Wissen“ durch Fähigkeiten des Menschen/der Rolle überprüft werden und unter anderem aus Daten, Informationen, Dokumenten, Intuition, Erfahrung, Erkenntnissen usw. generiert worden sein kann.

„Explizites Wissen“ muss ebenfalls geprüft in das System gebracht werden und Content muss durch Context-Informationen mit Mehrwert zum „Dokument“ transformiert werden.

Das „Dokument“ ist damit das system- oder domainrelevante Wissensobjekt. Dabei kann es sich um Text, Ton-, Bild-, Multimedia-, u. a. -inhalte und deren Beziehungen und Muster (Systeme, Domains, Lagebilder, etc.…) handeln.

In der Wissensentwicklung stellen alle unternehmerischen Entscheidungsprozesse ähnliche Herausforderungen in unterschiedlichsten Konstellationen und Intensitäten für die Mitarbeiter der ZentDok dar. Abbildung 1 stellt dies generisch - ohne Anspruch auf Vollzähligkeit - dar.

Abbildung 1: Generischer Blick auf die Wissensentwicklung.

Ziel ist es, durch Fachinformation, Dokumentation, Wissensmanagement und Wissensentwicklung eine kontinuierliche Steigerung der Potentiale zur Erhöhung der Handlungsfähigkeit des gesamten Unternehmens zu erreichen. Der Rahmen der Fähigkeiten des

ursprünglichen „Dokumentars“ erweiterte sich in den letzten Jahren rasant und dieser entstand aus dem Bedarf der zu unterstützenden Unternehmensbereiche.

2 Voraussetzungen im Unternehmen

Alle Voraussetzungen zur Steuerung eines unternehmensweiten WM-Systems sind fundierte theoretische Grundlagen sowie ein generisches Rahmenwerk. Dies wird mit dem Konzept „Wissensmanagement im ÖBH“ (BMLVS 2008), einem operativen Querschnittskonzept, auch dezidiert ausgewiesen.

Dadurch ergibt sich die Bedeutung für alle organisatorischen Hauptgeschäftsfelder wie Einsatz, Lehre und Forschung einschließlich der Betriebsführung in allen Ebenen und Teilbereichen der Organisation.

Der Fokus liegt dabei auf den Einsatzerfordernissen, die zielgerichtet durch WM-Tätigkeiten unterstützt werden. Ein besonderes Augenmerk muss daher auf die Verknüpfungsmöglichkeiten zwischen allen entscheidenden Organisationsbereichen und deren Abhängigkeiten gelegt werden. Wissensmanagement als Enabler (vgl. Cuviello 2009) von „Interoperabilität“ nimmt dabei eine wesentliche einsatzrelevante Unterstützungsaufgabe wahr.

Die Wissensmanagement Vorgehensweise basiert auf dem BPMS Paradigma (vgl. Karagiannis 1995). Genau wie im BPMS Paradigma werden fünf Phasen für modellunterstütztes Wissensmanagement definiert und vier Anwendungsszenarien realisiert, die zur unterschiedlichen Zielerreichung beitragen.

(1) Die Wissensbasierte Produktstrategie ermöglicht die Steuerung des Wissensmanagementsystems aus einer strategischen Produktsicht mit Werkzeugen ähnlich der Balanced Scorecard oder der Wissensbilanz.
(2) Die Wissensbasierten Geschäftsprozesse ermöglichen die Ausrichtung des Wissensmanagements auf die Geschäftsprozesse hin, um diese direkt oder indirekt zu unterstützen. Dabei werden Werkzeuge wie Workflows, Process-Steppers oder Geschäftsregeln verwendet.
(3) Die Wissensbasierte Organisation ermöglicht die Ausrichtung des Wissensmanagements auf die Mitarbeiter und die Organisation hin. Dabei werden Werkzeuge wie Skill-Management, Skill-Profile und Ausbildungsszenarien eingesetzt.
(4) Die Wissensbasierte Infrastruktur verbindet Wissensprodukte mit den Bedarfsträgern, indem die Infrastruktur wie Internet, Datenbanken oder Dokumentenmanagement-Systeme mittels Wissensmanagementprozessen verbunden werden.

3 WM-Strategie

In Anlehnung an den Begriff der Strategie in einer Organisation kann die WM-Strategie wie folgt definiert werden:

Eine WM-Strategie verfolgt das Ziel, Wissensmanagement in einer Organisation dauerhaft lebensfähig und auf die Geschäftsziele ausgerichtet zu integrieren. Ziel ist es, bestehendes Wissen so lange wie möglich nutzbar zu machen, zu erhalten sowie rechtzeitig zukünftige Potentiale aufzubauen.

Dazu werden fünf Dimensionen als Grundgerüst für die Erstellung einer unternehmensbezogenen WM-Strategie angenommen. (vgl. PROMOTE 2001; vgl. Knowledge Research 2002)

Dimension 1: Wissensstrategie als Geschäftsstrategie (vgl. WIEM 2009)

Dieser Blickwinkel konzentriert sich auf das Bereitstellen von qualitätsgesichertem und zeitgerechtem Wissen in der täglichen Arbeit des Wissensarbeiters. Durch die Kopplung von Geschäftsprozessen – die ja die Unternehmensziele unmittelbar ermöglichen – und dem Wissensmanagement, wird der Output des Wissens direkt oder indirekt an den Ergebnissen des Geschäftsprozesses gemessen. Daher werden folgende WM-Aktivitäten besonders gefördert: das Erzeugen, Verteilen, Dokumentieren, Speichern, Erneuern und Nutzen von geschäftsprozessrelevantem Wissen. Dieser Blickwinkel betrachtet daher die Effektivität des Wissensmanagements.

Dimension 2: Immaterielles Vermögen (vgl. Nemetz 2006)

Dieser Blickwinkel legt das Schwergewicht auf das Wissen in Form von immateriellem Vermögen in einer Organisation. Das immaterielle Vermögen kann explizit in Form von Patenten, Technologien, Methoden, Beziehungen oder anderen strukturieren Wissensformen auftreten. Dadurch wird der Wert des Wissens im Unternehmen gesichert und ermöglicht eine Marktdifferenzierung. Daher werden folgende WM-Aktivitäten besonders gefördert: das Erkennen, Schützen, Dokumentieren, Speichern und Vermarkten von Wissen. Dieser Blickwinkel betrachtet daher die Sicherstellung des Wissens.

Dimension 3: Persönliches Wissensmanagement (vgl. PKM 2009)

Dieser Blickwinkel fokussiert auf den eigentlichen Wissensträger, den Mitarbeiter, der für die Generierung, Erneuerung und Nutzung des Wissens in den Anwendungsbereichen verantwortlich ist. Ziel ist die Entwicklung und Unterstützung von unternehmensrelevanten

Wissensträgern, indem Eigenverantwortung, Arbeitsklima, Weiterentwicklung sowie die individuelle Vernetzung gefördert werden. Dabei stehen folgende WM-Aktivitäten besonders im Mittelpunkt: das Fördern und die Weiterbildung von Mitarbeitern, Übertragen von Eigenverantwortung, geeignete Team- und Gruppenbildung, sowie das Fördern von Netzwerken. Dieser Blickwinkel betrachtet daher den Wissensträger und seine Entwicklung als zentrale Wertschöpfung.

Dimension 4: Innovationsstrategie

Dieser Blickwinkel betrachtet die ständige Weiterentwicklung eines Unternehmens durch die Erzeugung von neuem Wissen. Dabei werden eine Technologiefrüherkennung sowie eine Technologieführerschaft angestrebt, um diesen Technologievorsprung in weiterer Folge durch Differenzierung zu einem Wettbewerbsvorteil umzuwandeln. Daher werden folgenden WM-Aktivitäten besonders berücksichtigt: die Grundlagenforschung, die angewandte Forschung, die experimentelle Entwicklung, das ständige Prototyping sowie die Weiterentwicklung aufgrund von Lessons Learned oder ähnlicher Instrumente. Dieser Blickwinkel betrachtet vor allem die zukünftige Wettbewerbsfähigkeit des Unternehmens.

Dimension 5: Wissenstransfer

Dieser Blickwinkel beobachtet den Wissenstransfer sowie die Dokumentation und Archivierung von Wissen. Im Zentrum der Aktivitäten stehen die zielgerichtete Verteilung und die dafür notwendige Transformierung des Wissens in eine Form, die für Adressaten notwendig ist. Daher werden folgende WM-Aktivitäten besonders berücksichtigt: das Organisieren, Umstrukturieren, Dokumentieren, Verteilen, Übersetzen und Anwenden des Wissens. Dieser Blickwinkel betrachtet daher die Effizienz des Wissensmanagements.

In Abhängigkeit der Gesamtunternehmensstrategie, die zuerst expliziert in Vision, Mission und Strategie sowie weiterführender Strategiewerkzeuge abgebildet werden sollte, können

obige fünf Dimensionen entweder für sich alleine oder in unterschiedlich gewichteten Kombinationsmöglichkeiten verwendet werden, um eine für das Unternehmen relevante WM-Strategie zu definieren.

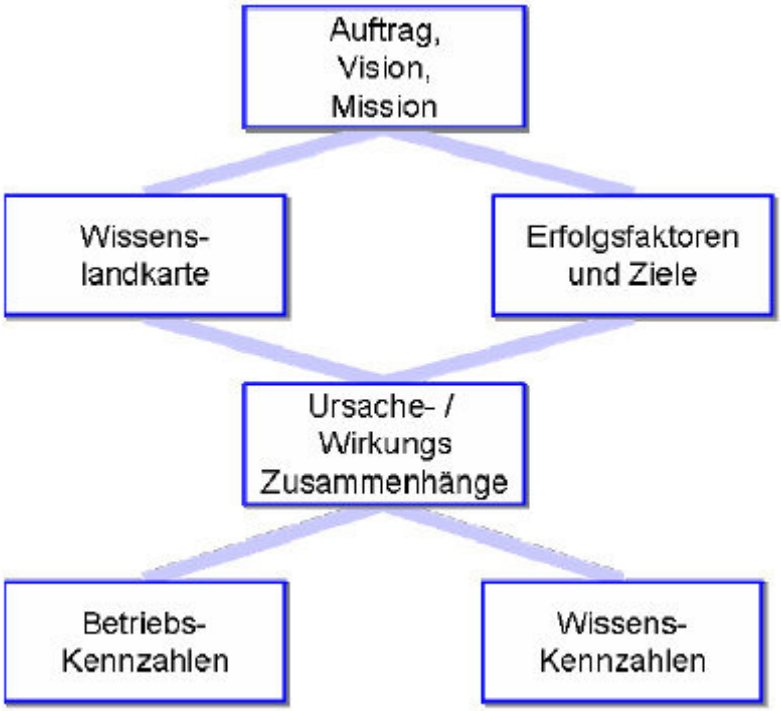


Abbildung 2: Kernelemente der WM-Strategie.

Strategisches Wissensmanagement ist daher ein Prozess, welcher

1) die Formulierung der Ziele,
2) deren Umsetzung sowie
3) laufende Überprüfung und Anpassung des WM-Systems liefert.

Es gilt dabei, sowohl externe Chancen und Risiken als auch interne Stärken und Schwächen zu berücksichtigen und Entscheidungen mit Zukunftswirkung zu treffen.

Kernfragen sind daher:

4) Welches Wissen ist für den heutigen Geschäftserfolg relevant?
5) Wie kann das Wissen für den zukünftigen Geschäftserfolg zur Verfügung gestellt werden?
6) Wie muss sich die Organisation weiterentwickeln, damit sie auch in Zukunft über das erforderliche Wissen verfügt, um lebensfähig und profitabel zu bleiben?

Um diese Fragen konkret und detailliert beantworten zu können, muss sich eine strategische Analyse regelmäßig mit dem Umfeld der Organisation auseinandersetzen, also insbesondere mit

a) den Bedürfnissen der Bedarfsträger
b) mit dem Umfeld, welches das Wissen für die Organisation maßgeblich bestimmt.

4 Konzeptueller Hintergrund

Das Konzept einer Wissensbilanz muss abgestimmt auf das vorhandene Betriebs-Organisations-Modell entwickelt werden. Alle in Anwendung stehenden betriebswirtschaftlichen Methoden und Instrumente müssen in der Erstellung der Wissensbilanz nicht nur berücksichtigt, sondern auch integriert werden.

Die Prozessqualität spielt dabei eine zentrale Rolle. Dadurch wird das Zusammenwirken von Auftrag, den geforderten Fähigkeiten der Organisation, ihrer organisatorischen und strukturellen Entwicklung, den Kosten sowie Qualitäten der Leistungserbringung über konkrete Ergebnisse abbildbar.

Aus der Mission, Vision und den strategischen Zielen sowie ergänzenden strategischen Dokumenten wie Leitbild, Konzept oder SOLL-Profil können konkrete Fähigkeiten der Organisation und Kompetenzen der Mitarbeiter samt den dazu notwendigen Produkten abgebildet werden.

Um konkret Daten aus der jeweiligen Organisation zu generieren, können alle zur Verfügung stehenden betriebswirtschaftlichen Methoden und Instrumente des Controllings und des Qualitätsmanagements angewendet werden.

Zum Beispiel:

- Prozessmanagement
- Balanced Scorecard
- Qualitätsmanagement
- Assessment Frameworks
- Kosten- und Leistungsrechnung
- Kontinuierlicher Verbesserungsprozess (KVP)
- Produktlandkarte
- Wissensprozesslandkarte

Als Darstellungsform der Organisationsumgebung in den vier Sichtweisen des Wissensbilanzrahmenwerkes ergibt sich die „Idee“ eines generischen Performance-Frameworks.

Eine Wissensbilanz ist ein Instrument zur gezielten Darstellung sowie Entwicklung des Intellektuellen Kapitals (IK) (Woitsch et.al. 2010b) einer Organisation. Sie zeigt die Zusammenhänge zwischen den organisationalen Zielen, den Geschäftsprozessen, dem Intellektuellen Kapital und dem Geschäftserfolg einer Organisation auf und beschreibt diese Elemente mittels Indikatoren (vgl. BMWA 2008).

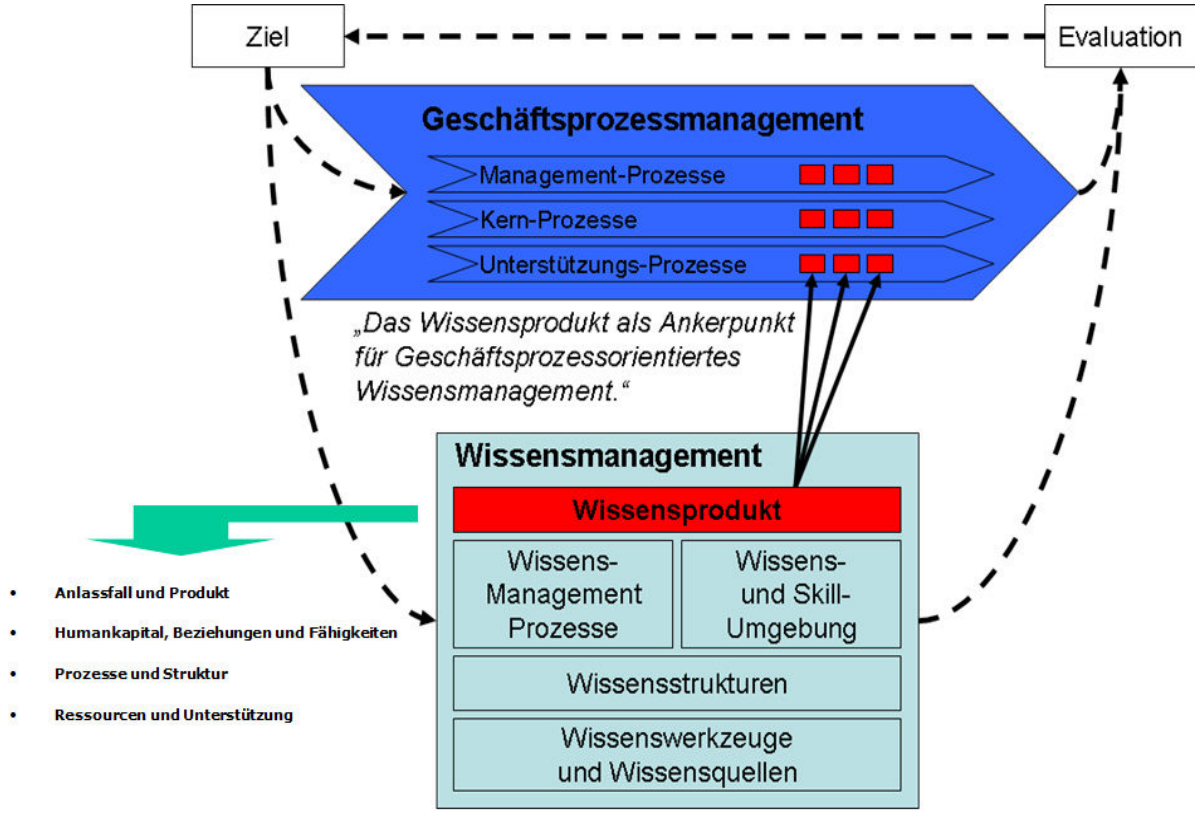


Abbildung 3: „Idee" des generischen Performance-Frameworks.

Mit der Wissensbilanz werden folgende direkte Nutzenkategorien erreicht (Bornemann & Reinhardt 2008):

1. Höhere Wettbewerbsfähigkeit und weniger Risiko,
2. Bessere Nutzung der immateriellen Vermögenswerte,
3. Klare strategische Ausrichtung.

Die Wissensbilanz zeigt einerseits die Zusammenhänge zwischen den organisationalen Zielen, den Prozessen, dem Intellektuellen Kapital und dem Erfolg von wissensbasierten Organisationen auf. Andererseits generiert die Wissensbilanz auch Kenngrößen für strategische Entscheidungen und Potentiale. Diese Indikatoren werden, besonders wenn sie sensible Informationen und Daten enthalten, vorrangig für interne Organisationszielgruppen verwendet.

4.1 Referenzmodell/-architektur

Das Referenzmodell basiert auf der Analyse von existierenden Wissensbilanzen, dem Balanced Scorecard Ansatz sowie vorhandenen Instrumenten der Betriebsorganisation.

Dieses Referenzmodell unterscheidet vier Perspektiven:

1. Anlassfall und Produkt Perspektive: Ziele, Kennzahlen, Vorgaben und Maßnahmen in Bezug auf Anlassfälle, Produkte sowie das wahrgenommene Ergebnis

2. Prozesse und Struktur Perspektive: Ziele, Kennzahlen, Vorgaben und Maßnahmen in Bezug auf Kernprozesse, qualitätsrelevante Prozesse sowie führungs- und qualitätsrelevante Prozesse.
3. Humankapital, Beziehungen und Fähigkeiten-Perspektive: Ziele, Kennzahlen, Vorgaben und Maßnahmen in Bezug auf Personen (Humankapital), sowie Fähigkeiten und Beziehungen (Beziehungskapital).

Ressourcen und Unterstützungs-Perspektive: Ziele, Kennzahlen, Vorgaben und Maßnahmen in Bezug auf Budget, Infrastruktur (Strukturkapital), Material und Gerät, sowie Information und Kommunikation. (vgl. HESIG 2008)

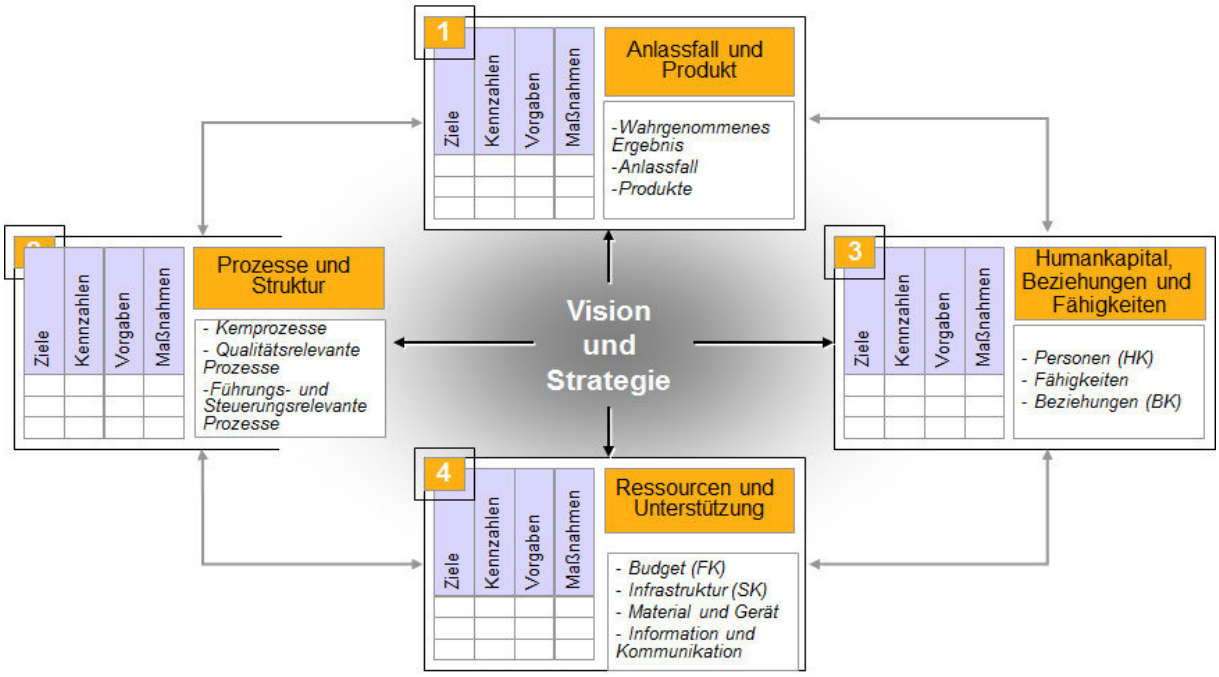


Abbildung 4: PROMOTE®-Sichtweisen einer Wissensbilanz.

Zur verfahrenstechnischen und somit nachvollziehbaren und qualitätsgesicherten Umsetzung von Wissensmanagement haben sich modellorientierte Ansätze als vorteilhaft erwiesen.

Ziel dieses Abschnittes ist es die Modellierung von Wissensmanagement vorzustellen und das PROMOTE® Referenzmodell (Mak, Göllner, Meurers & Utz 2010; Hinkelmann, Karagiannis & Telesko 2002) anzubieten. Nachdem die Strategie des Wissensmanagements definiert worden ist und die Veränderung des derzeitigen Systems vorgegeben ist, wird das WM-System geplant.

Das Wissensmanagement hat den Geschäftsprozess direkt oder indirekt zu unterstützen. Um ein eindeutiges und in der Praxis handhabbares Werkzeug für die Kopplung von Geschäftsprozessen und Wissensmanagement zur Verfügung zu haben, wird hier der Begriff „Wissensprodukt“ eingeführt.

Wissensprodukte sind die Ergebnisse aus dem Wissensmanagement, die im Geschäftsprozess verwendet werden. Das hier vorgestellte Referenzmodell beschreibt, wie Wissensprodukte hergestellt werden.

Die Wissensprodukte werden vom WM-System aus an das Unternehmen angeboten. Sie werden in Informationsprodukte, Beratungsprodukte und Anwendungsprodukte unterteilt.

Sie werden zum Triggern von Prozessen, bei der Durchführung von Aktivitäten oder beim Treffen von Entscheidungen von den Wissensarbeitern genutzt.

Informationsprodukte stellen explizite Informationen zur Verfügung, die von einem Mitarbeiter verarbeitet wird. Darunter fallen Handlungsanleitungen, Dokumente, Bücher, Richtlinien, Protokolle, schriftlich festgehaltene Regeln, Präsentationen, Filme, Datenbankabfragen oder ähnliches.

Beratungsprodukte stellen implizites Wissen zur Verfügung und können daher nur von einem Mitarbeiter zu einem anderen Mitarbeiter weitergegeben werden. Darunter fallen Auskünfte, Beratungen, Moderationen, das Abhalten von Kursen, periodische Besprechungen oder ähnliches.

Anwendungsprodukte stellen explizites Wissen zur Verfügung, die von Software-programmen verarbeitet werden. Darunter fallen Entscheidungsunterstützungssysteme, Workflowsysteme, automatische Textanalysesysteme, Suchmaschinen oder Ähnliches.

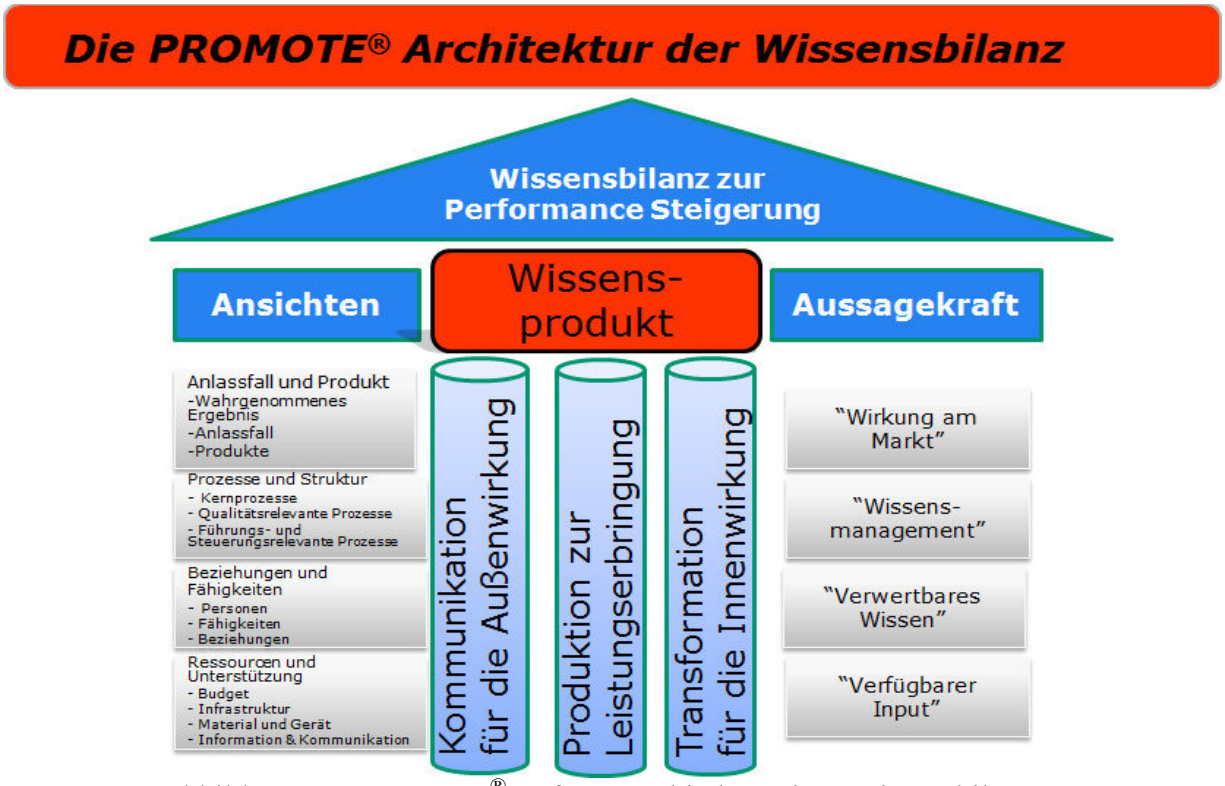


Abbildung 5: PROMOTE® Referenzarchitektur einer Wissensbilanz.

Das WM-System baut im Wesentlichen auf den drei Säulen

a) Kommunikation,

b) Produktion und

c) Transformation

auf.

Die Kommunikation bezieht sich dabei auf die Außenwirkung am Markt, Produktion bezieht sich auf die Leistungserbringung und Transformation auf die Innenwirkung. Diese drei Säulen werden in vier unterschiedlichen Sichtweisen (Perspektiven) betrachtet.

Ausgangspunkt für die erste Sichtweise ist das Produkt selbst und die Analyse seiner Wirkung am Markt. Die zweite Sichtweise sind Prozesse und Strukturen, somit wird das Wissensmanagement der Organisation analysiert. Die dritte Perspektive bilden Humankapital, Beziehungen und Fähigkeiten der Organisation, somit wird das verwertbare Wissen analysiert. Die vierte Sichtweise sind Ressourcen und Unterstützungen wie Kapital, Infrastruktur, Material und Gerät sowie Information und Kommunikation, somit wird der verfügbare Input analysiert. Anhaltspunkte für das Referenzmodell liefern die so genannten Wissenscodes aus dem EU-Projekt MATURE (MATURE 2010). Codes sind Beschreibungen von Wissenstätigkeiten, die der Referenzarchitektur für das Wissensmanagementsystem zugeordnet werden können, wobei kein Anspruch auf Vollzähligkeit besteht. Sie bieten Anhaltspunkte in der Referenzarchitektur um Wissensarbeit beschreiben zu können und haben somit beispielhaften Charakter.

Die Wissensarbeit in Kommunikation, Produktion und Transformation in der Produktebene ist abhängig vom jeweiligen Unternehmensziel und individuell anzupassen.

4.2 Roadmap zur Wissensbilanz

Beschreibung der Ausgangslage

Bevor mit der Erstellung der Wissensbilanz begonnen werden kann, muss im ersten Schritt die Ausgangssituation beschrieben werden. Dabei gilt es die Wissensorganisation zu analysieren und die für die Steuerung relevanten Teile der Organisation zu spezifizieren. Die Abbildung mittels Modellen ermöglicht nicht nur komplexe Zusammenhänge durch die Modellsprache einfach und graphisch darzustellen, sondern unterstützt auch die Analyse einer Wissensorganisation durch ein formales und strukturiertes Vorgehen. Es muss zuerst ein Überblick über das vorhandene Wissensmanagement gegeben werden, um anschließend im Detail verschiedene Wissensaspekte genauer analysieren zu können.

Spezifikation der Ziele

Zur Ableitung der Ziele werden zuerst die Zielgruppen der Wissensbilanz erhoben. Pro Zielgruppe werden der Zweck der Wissensbilanz, die Qualität der Berichte, sowie der Berichtszyklus festgelegt. Eine Tabelle ermöglicht einen Überblick über die Zielgruppen, den Zweck der Wissensbilanz für die jeweilige Zielgruppe sowie den Berichts- und Steuerzyklus. In einem weiteren Schritt werden die Einflussfakturen der Instrumente auf die Wissensbilanz identifiziert.

Erkennen der Wirkungszusammenhänge

Nachdem die kritischen Erfolgsfaktoren erarbeitet und zu Wissenszielen zusammengefasst wurden, erfolgt eine Analyse der Ursachen-Wirkungszusammenhänge. Die Ziele können in übergeordnete (strategische) und untergeordnete (operative) Ziele aufgeteilt werden. Ein untergeordnetes (operative) Ziel dient als Zwischenschritt zu einem übergeordneten (strategischen) Ziel. Die Aufteilung in übergeordnete und untergeordnete Ziele erfolgt daher aufgrund des zeitlichen Horizonts des Zieles, die direkte oder indirekte Auswirkung auf das Wissensprodukt oder die Anzahl der beeinflussenden Erfolgsfaktoren. Danach werden die Abhängigkeiten der Ziele basierend auf der Einschätzung von Experten definiert.

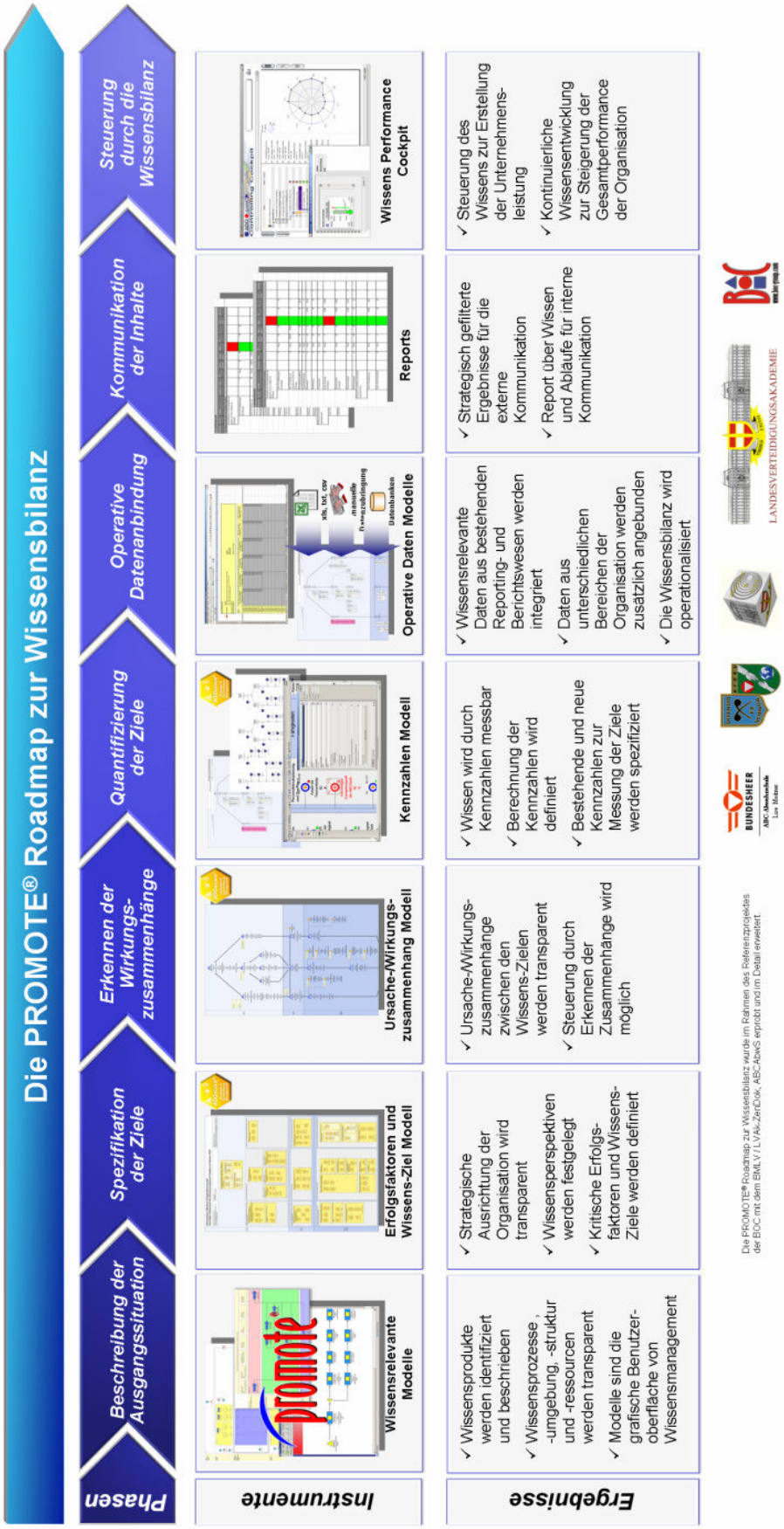


Abbildung 2: RoadMap zur Erstellung der Wissensbilanz.

Quantifizierung der Ziele

Effektive Unternehmensdaten werden in Kennzahlen hinterlegt bzw. an Datenquellen angebunden. Kennzahlen beinhalten die aktuellen Werte, sowie Vergleichswerte (Soll-Werte) und Toleranzgrenzen. Die Plausibilität jeder Kennzahl wird durch kritisches Hinterfragen sichergestellt. Jede Kennzahl wird durch den Bezugszeitraum, Maßeinheit, Begrenzungsart, Toleranzgrenzen, ihrer Funktion sowie der Art des Periodenwertes beschrieben. Großteils werden die Kennzahlen von bestehenden Systemen bereitgestellt, sodass eine plausible Integration bestehender Sichtweisen mit neuen Kennzahlen erarbeitet werden kann.

Operative Datenanbindung

Die Verknüpfung von Daten mit der Wissensbilanz erfolgt durch die Anbindung externer Datenquellen wie Datenbanken, DataWarehouse oder manuellen Dateneingaben mittels Excel an die jeweiligen Kennzahlen. Die verwendeten komplexeren Kennzahlen werden erst im Werkzeug auf Basis der verknüpften elementaren Daten berechnet. Diese werden zur Simulation der Kennzahlen vor der eigentlichen Cockpit-Generierung herangezogen, um die Plausibilität auf Indikatoren-Ebene überprüfen zu können.

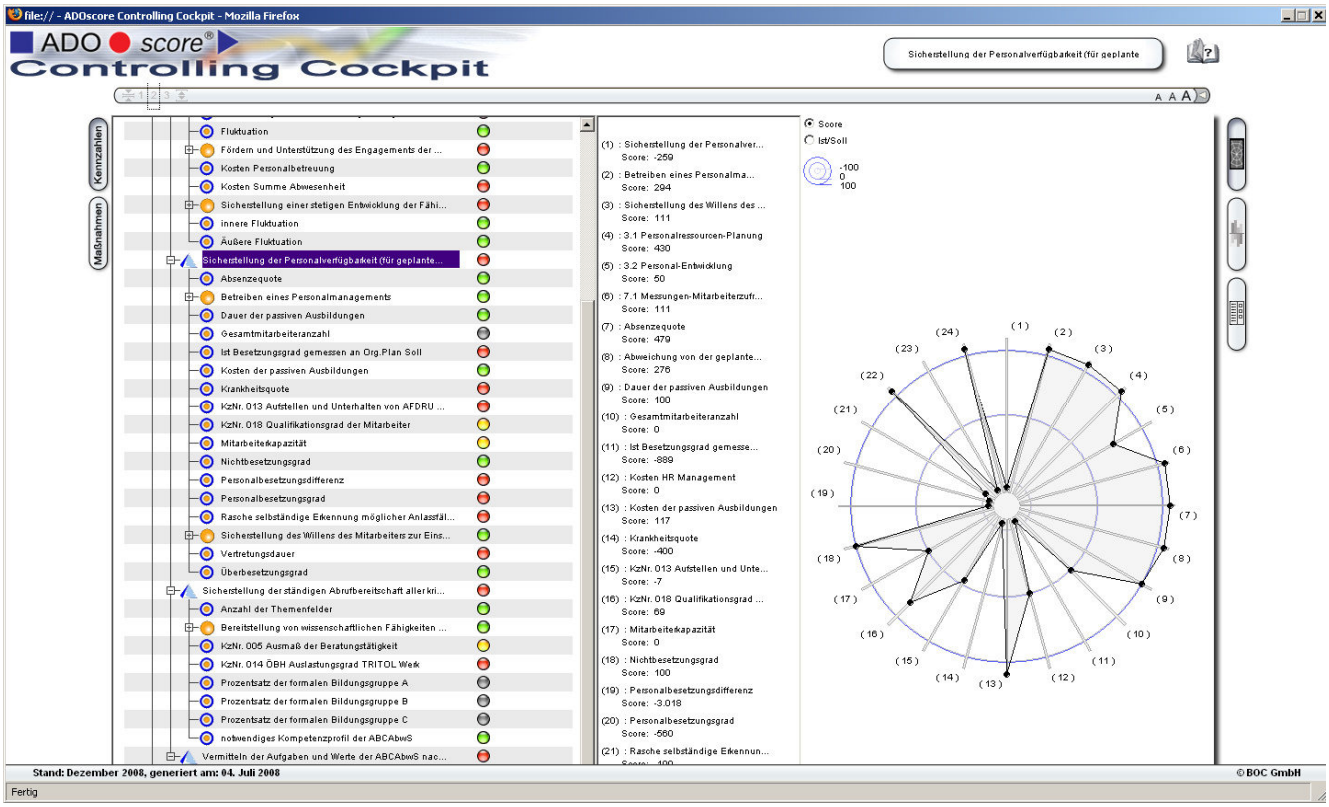


Abbildung 7: Performance Cockpit.

Kommunikation der Inhalte

Nach der Modellierung der Wissensbilanz, der Anbindung der Daten sowie der Simulation können das Wissens-Performance-Cockpit sowie die Wissensbilanz-Berichte generiert werden. Das eingerichtete „Wissens-Performance-Cockpit“ bietet weitreichende Controlling- und Analyse-Funktionen. Die Inhalte können anhand vielfältiger Darstellungs- und Filteroptionen aufbereitet werden. Zur Aufbereitung der Werte, der Ziele sowie der Kennzahlen, kann dabei das „traffic-light-coding“ verwendet werden. Durch die Signalwirkung

der Ampelfarben können positive und negative Entwicklungen frühzeitig erkannt und entsprechende Maßnahmen gesetzt werden. Über einen Navigationsbaum wird die Auswahl der gewünschten Perspektiven, Ziele oder Kennzahlen ermöglicht.

Steuerung der Wissensbilanz

Die Berichtsformen des Performance-Cockpits einer Wissensbilanz listen die Kennzahlen, ihre Zugehörigkeit zu den Sichten (Perspektiven), sowie die Wirkung auf Zielvorgaben auf. Neben einer Ampeldarstellung kann der Trend, sowie ein normierter Zielerreichungsgrad – Score – ausgerechnet und neben den wichtigsten Kennzahlenwerten dargestellt werden. Ein Performance-Cockpit ermöglicht ebenfalls umfangreiches recherchieren in den Detail-beschreibungen sowie in der Kennzahl-Zielabhängigkeit. Somit können Schwachstellen frühzeitig erkannt und die Auswirkungen auf die Ziele rechtzeitig beeinflusst werden.

5 Umsetzung des Modells im Rahmen des Aufbaus eines Forschungsmanagement systems

Auf Basis der oben vorgestellten entwickelten Referenzarchitektur wurde den konzeptionellen und strategischen Vorgaben sowie dem Konzept „Forschung im ÖBH“ und dem operativen Querschnittskonzept „WM im ÖBH“ folgend im Rahmen eines internen Projektes die Wissensbilanz für das Forschungsmanagementsystem (FMS) im ÖBH erstellt (BMLVS 2010). Die dabei benötigten Ressourcen, Fähigkeiten, Prozesse, Strukturen für den Anlassfall „Forschungsbedarf“ wurden abgebildet und mit den verfügbaren „Forschungs-ressourcen“ verglichen. Dieses Monitoring gab einen Überblick über die Verfügbarkeit der Forschungsmanagement- und Forschungsdurchführungskompetenz in einer bestimmten Zeitperiode.

Zunächst wurde ein Übersichtsmodell in Form einer Wissenslandkarte generiert, in der die „Inputgeber“ und „Konsumenten“ der Forschung im ÖBH mit ihren Schnittstellen zum FMS abgegrenzt und die steuerungsrelevanten Teile der Organisation abgebildet wurden. Der Vision/Mission sowie den aus den strategischen Vorgaben für Forschung im ÖBH folgenden Strategie und Zielen wurde eine Forschungsprozesslandkarte entwickelt, die alle Kern- und Unterstützungsprozesse sowohl in Bereich der Steuerung und Management aber auch bei den forschungsdurchführenden Stellen abbildet.

Zur Spezifikation der Ziele wurden zuerst die Zielgruppen der Wissensbilanz erhoben. Pro Zielgruppe konnte dann der Zweck der Wissensbilanz, die Qualität der Berichte, sowie der Berichtszyklus festgelegt werden. Nach der Spezifikation der Ziele wurden die kritischen Erfolgsfaktoren identifiziert. Die Erfolgsfaktoren wurden im Zuge von Workshop-Sitzungen gesammelt und in weiterer Folge auf Plausibilität überprüft und angepasst. Bekannte Erfolgsfaktoren aus Referenzmodellen wurden dabei ergänzend zur Verfügung gestellt und integriert. Die Erfolgsfaktoren wurden den Perspektiven der Wissensbilanz (Anlassfall und Produkt- Perspektive, Prozesse und Struktur-Perspektive, Humankapital, Beziehungen und Fähigkeiten – Perspektive, Ressourcen und Unterstützungs-Perspektive), sowie den Säulen der Wissensbilanz (Kommunikation, Produktion, Transformation), zugeordnet. Durch diese Zuordnung ergab sich eine erste Gruppierung der Erfolgsfaktoren. Ähnliche Faktoren wurden zu strategischen Zielen aggregiert.

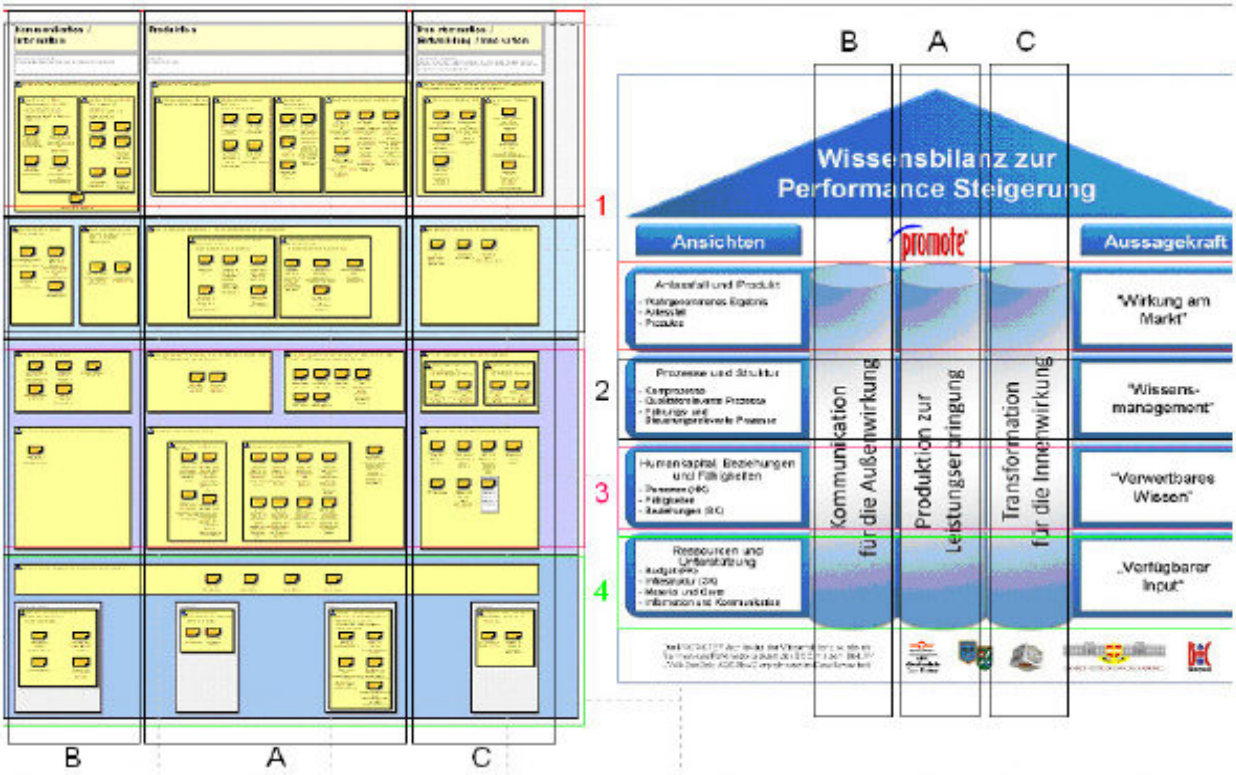


Abbildung 8: Erfolgsfaktoren und deren Zuordnung zu den Säulen der Wissensbilanz.

Auf Basis der erarbeiteten Ergebnisse wurden Ursachen- und Wirkungszusammenhänge identifiziert und dargestellt und den strategischen Zielen operative Ziele zugeordnet. Die Erreichung eines Ziels wird dabei durch eine oder mehrere Kennzahlen gemessen. Anhand der Abhängigkeiten von Zielen sowie Kennzahlen, konnten Einflüsse zwischen den Faktoren aufgezeigt werden. Abbildung 9 zeigt einen Auszug aus dem Ursache- Wirkungsmodell.

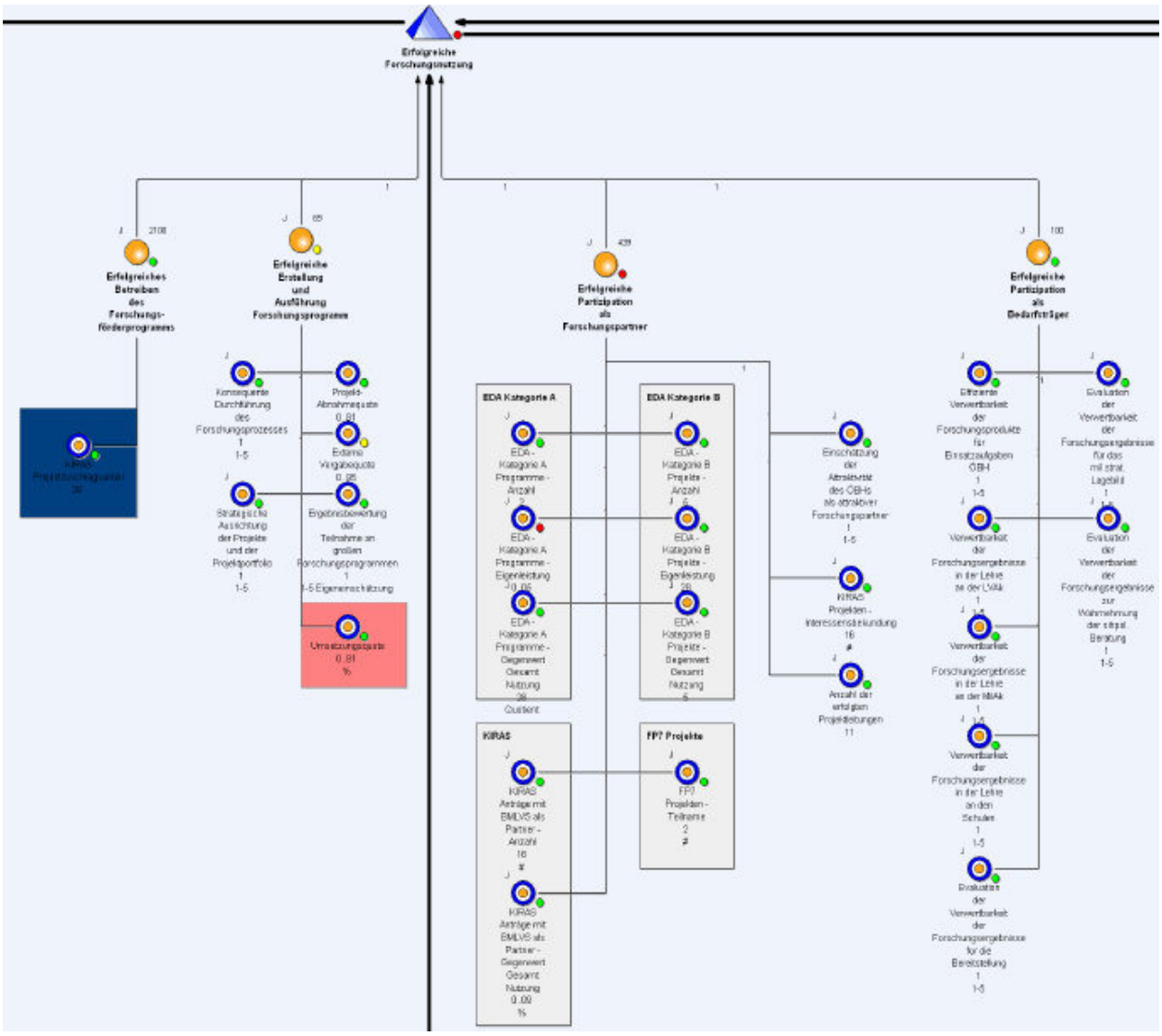


Abbildung 9: Auszug aus dem Ursache-Wirkungsmodell.

Gemäß der Roadmap wurden die Ziele quantifiziert und mit Kennzahlen hinterlegt. Die Kennzahlen wurden mit effektiven Unternehmensdaten hinterlegt bzw. an Datenquellen angebunden. Kennzahlen beinhalten die aktuellen Werte, sowie Vergleichswerte (Soll-Werte) und Toleranzgrenzen. Die Plausibilität jeder Kennzahl wurde durch Bereitstellung und Analyse von Daten sowie kritisches Hinterfragen sichergestellt.

Name	Bezugszeit-raum	Messgröße	Tolleranz-art	Begrenzungs-art	Toleranz grün / gelb	Toleranz gelb / rot	FDS / FMS
Abnahmequote Forschungsprojekte	Jahr	%	absolut	von unten	0.2	0.28	
Anpassungsmöglichkeit des FDS an den tatsächlich erkannten Bec	Jahr	1-5	absolut	von oben	2	3	
Anzahl der am FMS teilnehmenden Dienststellen	Halbjahr	%	absolut	von unten	0	0.2	
Anzahl der an FDS teilnehmenden Dienststellen	Halbjahr	%	absolut	von unten	0	0.2	
Anzahl der erfolgten Projektleitungen	Jahr		absolut	von unten	0	4	
Anzahl der Informationsveranstaltungen für Externe	Jahr	#	absolut	von unten	0	0	
Anzahl der Informationsveranstaltungen für Interne	Jahr	#	absolut	von unten	0	0	
Anzahl der Kommunikationsprodukte	Quartal	#	absolut	von unten	0	9	
EDA - Kategorie A Programme - Anzahl	Jahr	#	absolut	von unten	0	1	
EDA - Kategorie A Programme - Eigenleistung	Jahr	%	absolut	beidseitig	0	0.01	
EDA - Kategorie A Programme - Gegenwert Gesamt Nutzung	Jahr	Quotient	absolut	von unten	13	20	
EDA - Kategorie B Projekte - Anzahl	Jahr	#	absolut	von unten	0	1	
EDA - Kategorie B Projekte - Eigenleistung	Jahr	€	absolut	von unten	0	0	
EDA - Kategorie B Projekte - Gegenwert Gesamt Nutzung	Jahr		absolut	von unten	0	0.06	
Effiziente Verwertbarkeit der Forschungsprodukte für Einsatzaufga	Jahr	1-5	absolut	von oben	2	3	
Einschätzung der Attraktivität des ÖBHs als attraktiver Forschungs	Jahr	1-5	absolut	von oben	1	3	
Erfüllung aller Kennzeichen des funktionierenden Systems	Jahr	1-5	absolut	von oben	2	3	
Ergebnisbewertung der Teilnahme an großen Forschungsprogrami	Jahr	1-5	absolut	von oben	1	3	
Evaluation der Mitwirkung an int. Forschungsprogramme	Jahr	1-5	absolut	von oben	2	4	
Evaluation der Mitwirkung an int. und nationalen. Forschungsgremi	Jahr	1-5	absolut	von oben	2	3	
Evaluation der Verwertbarkeit der Forschungsergebnisse für das m	Jahr	1-5	absolut	von oben	2	3	

Abbildung 10: Kennzahlen (Auszug).

Jede Kennzahl, sowohl quantitativ als auch qualitativ, wird durch den Bezugszeitraum, Maßeinheit, Begrenzungsart, Toleranzgrenzen, ihrer Funktion sowie der Art des Periodenwertes beschrieben. Großteils waren die Kennzahlen von bestehenden Systemen bereitgestellt, sodass eine plausible Integration bestehender Sichtweisen mit neuen Kennzahlen erarbeitet werden konnte.

Nach der Operativen Datenanbindung erfolgte die Kommunikation der Inhalte sowie die Steuerung durch die Wissensbilanz durch die Implementierung eines Wissens-Performance-Cockpits, das weitreichende Controlling- und Analyse-Funktionen bietet. Die Inhalte, Werte, Ziele und Kennzahlen können anhand vielfältiger Darstellungs- und Filteroptionen aufbereitet werden. Dazu wird beispielsweise das „traffic-light-coding“ verwendet, das über die Darstellung von Ampelfarben und der damit verbundenen Signalwirkung positive und negative Entwicklungen frühzeitig erkennen lässt, sodass entsprechende Maßnahmen gesetzt werden können.

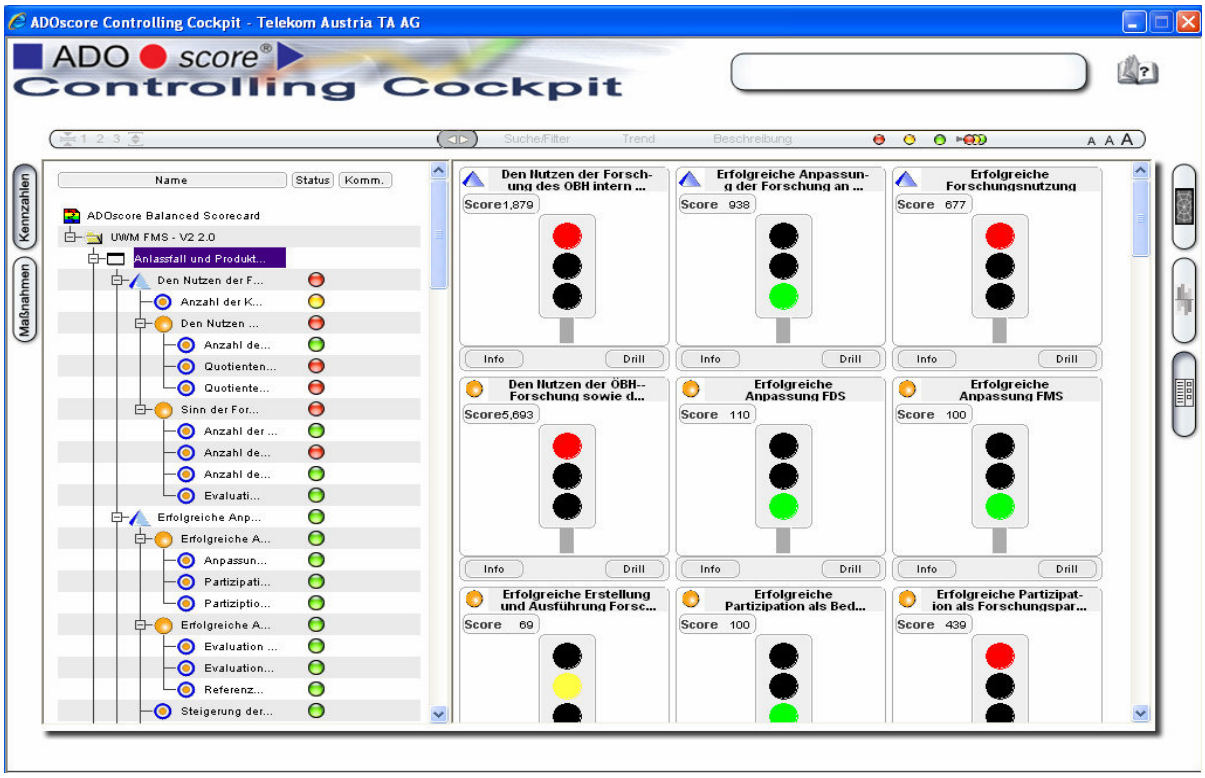


Abbildung 11: Ampeldarstellung im Wissens-Performance-Cockpits.

Der Navigationsbaum auf der linken Seite des Cockpits ermöglicht die Auswahl der gewünschten Perspektiven, Ziele oder Kennzahlen. Die Berichtsformen im Performance-Cockpit der Wissensbilanz listen die Kennzahlen, ihre Zugehörigkeit zu den Sichten (Perspektiven), sowie die Wirkung auf Zielvorgaben auf. Neben der Ampeldarstellung wird der Trend, sowie ein normierter Zielerreichungsgrad – Score – ausgerechnet und neben den wichtigsten Kennzahlenwerten dargestellt. Das Performance-Cockpit ermöglicht umfangreiches recherchieren in den Detailbeschreibungen sowie in der Kennzahl-Zielabhängigkeit. Somit können Schwachstellen frühzeitig erkannt und die Auswirkungen auf Ziele rechtzeitig beeinflusst werden.

Kennzahlenübersicht												
Pers-pektive	Strat. Ziel	Kennzahl	Verantwortlich	Einheit	BZR	Toleranz grün/gelb	Toleranz gelb/rot	Status	Soll	Ist	Ist Vor-periode	Kommentar
Anlassfall und Produkt Perspektive												
	Sinn der Forschung für das Unternehmen BH transparent, nachvollziehaber und Nutzbringend zu kommunizieren (OZ)											
		Anzahl der Informationsveranstaltungen für Interne		#	Jahr	0..00	0..00		1	2	2	
		Evaluation des Mehrwertes für beteiligte OrgEinheiten		1-5	Jahr	1..00	3..00		1	1	1	
		Anzahl der am FMS teilnehmenden Dienststellen		%	Halbjahr	0..00	0..20		1	0..60	0..59	
		Anzahl der an FDS teilnehmenden Dienststellen		%	Halbjahr	0..00	0..20		0..80	0..80	0..79	
	Den Nutzen der Forschung des ÖBH intern und extern zu kommunizieren											
		Anzahl der Kommunikationsprodukte		#	Quartal	0..00	9..00		15	8	7	
	Den Nutzen der ÖBH-Forschung sowie die ÖBH-Leistungsfähigkeit als Forschungspartner und Forschungsbedarfsträger zu kommunizieren (OZ)											
		Anzahl der Informationsveranstaltungen		#	Jahr	0..00	0..00		1	2	2	

Abbildung 12: Kennzahlenbericht.

6 Abschließende Betrachtung und Zusammenfassung

Im Zuge des vorgestellten Projektes als Anwendungsbeispiel konnten die Referenzarchitektur und die RoadMap zur Wissensbilanz validiert werden und der theoretische Ansatz praktisch umgesetzt werden. Im Zuge des Projektes traten aber auch klar die neuralgischen Punkte bzw. die kritischen Phasen des Referenzmodells zu Tage. So könnte sich die Abstraktion bzw. die Reduktion von Zielen und in weiterer Folge das Fehlen von eventuell wichtigen Ursache-Wirkungszusammenhängen für die Wissensbilanz als kritisch erweisen. Daher erfolgte die Umsetzung der einzelnen Schritte zur Erstellung einer Wissensbilanz bzw. eine Wissens-Performance-Cockpits teilweise in mehreren Iterationsschritten, in denen immer wieder Plausibilitätsprüfungen und Validierungen insbesondere von Zielen und Kennzahlen durchgeführt wurden. Die Kennzahlen können in jedem Segment des Ursache-Wirkungsmodells quantitativ oder qualitativ sein. Eine genaue Beschreibung des Kontextes der Steuerungsfunktionalität ist aber dezidiert erforderlich, um für die Führungspersonen gleiche Vorraussetzung in der Steuerung zu garantieren. Dies stellt über den gesamten Erstellungsprozess der Wissensbilanz eine besondere Herausforderung, da hier die Grenzen des Systems sichtbar werden.

Eine weitere Problematik ist die Anbindung von Kennzahlen an Datenquellen. Für das vorgestellte Projekt waren alle relevanten Unternehmensdaten in verschiedenen Ausprägungen vorhanden und konnten in das System integriert werden. Diese Datenanbindung könnte allerdings in anderen Anwendungsbereichen nicht immer in dieser Form sichergestellt werden, man denke hier beispielsweise an klassifizierte Informationen, die aus dem simplen Grund der Unternehmenspolitik nicht in ein solches System eingebunden werden dürfen. Außerdem ist immer auf eine Validierung der den Kennzahlen hinterlegten Werte zu achten, da das Wissens-Performance-System sonst falsche Ergebnisse liefern kann.

Abschließend konnte mit der praktischen Anwendung des vorgestellten theoretischen Ansatzes nicht nur dessen Tauglichkeit, sondern auch der Nutzen einer Wissensbilanz für eine Organisation unterstrichen werden. Die Management-Ebene erhält damit ein Instrument, das Vorgänge, Ressourcen und Prozesse transparent und nachvollziehbar macht und

gleichzeitig die Grundlagen für ein Forschungscontrolling zur Betriebssteuerung und -planung, für die Entwicklung und Umsetzung der Organisationsfähigkeiten, zur Verbesserung der Planungssicherheit sowie der Entwicklung von effizienten und zielgerichteten Strategien in allen abgebildeten Bereichen schafft.

Durch dieses Rahmenwerk, ausgehend von bisherigen Arbeiten wie Studien und WM-Projekten, wurden fundierte Grundlagen geschaffen, um einen prototypischen Demonstrator „Wissensbilanz“ für das ÖBH oder für ausgesuchte Organisationseinheiten zu entwickeln. Damit ist die Möglichkeit zur Schaffung von transparenten Abläufen als Grundlage zur Steuerung in den Bereichen Einsatz, Lehre und Forschung u.a. als Grundlage für Bildungs- und Forschungscontrolling zur Betriebssteuerung und -planung, für die Entwicklung und Umsetzung der Organisationsfähigkeiten und zur Verbesserung der Planungssicherheit, gewährleistet.

Begleitend erfolgt eine Standardisierung der Fach-Terminologien und -Prozesse sowie Anpassung an die in Entstehung befindliche Wissensmanagement-Terminologie des ÖBH. Der Aufbau von Kompetenzen im Umgang mit modernsten Managementinstrumenten und

-methoden für Mitarbeiter des ÖBH sowie für Mitarbeiter in WM-Projekten, stellt eine wesentliche Wertsteigerung für die Organisation dar. Dieses Wissen ist Grundlage für die Weiterentwicklung des Wissensmanagements des ÖBH. Damit verbunden ist eine Steigerung der Reaktionsfähigkeit und Lernfähigkeit und somit eine wesentliche Erhöhung der gesamten Einsatzbereitschaft.

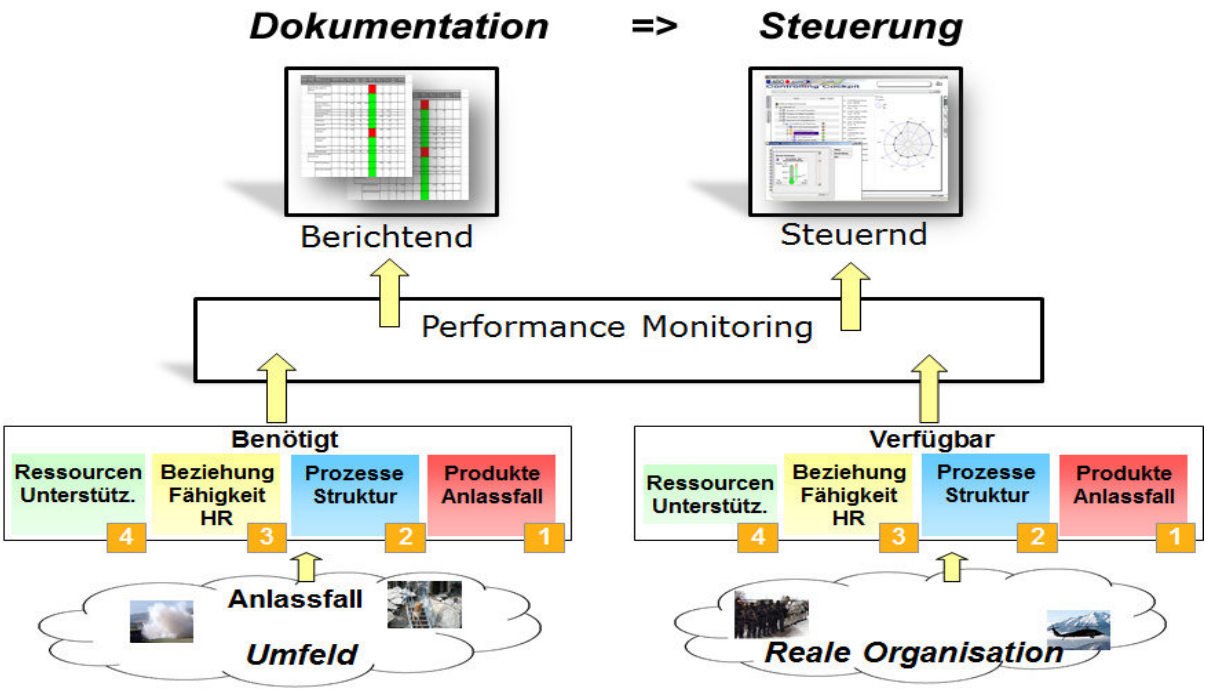


Abbildung 13: Von der Dokumentation zur Steuerung.

Die Dokumentation von Ursache- und Wirkungszusammenhängen von Komponenten einer Einsatzorganisation mittels vorgestellter Methodik kann umgesetzt werden. Die Transparenz aller Ressourcen der Organisation, insbesondere die des intellektuellen Wertes der Mitarbeiter, kann für alle Bereiche der Organisation beispielhaft abgebildet werden.

Die Kommunikation des Unternehmens über eine dynamische Performance-Darstellung könnte ebenso beispielgebend sein wie eine fundierte Organisations- und Personalentwick-

lung, basierend auf einer integrativen und historisch gewachsenen Bertachtungsweise der Organisation.

Für das Management ergibt sich außerdem die Möglichkeit des frühzeitigen Erkennens von Wissensbedarf, der durch eine zielgerichtete Mitarbeiterentwicklungsstrategie effizient, rechtzeitig und nachvollziehbar abgedeckt werden kann. Für Kommandanten und Leiter ergeben sich Perspektiven für transparente Entscheidungseinflüsse auf die jeweiligen Führungsbereiche.

Literatur

BMLVS (2008), *Wissensmanagement im ÖBH*, Bundesministerium für Landesverteidigung und Sport, Wien, 2008

BMLVS (2010), Mak, Hofmeister, Göllner, Woitsch, *WM-Projekt Forschungsmanagementsystem (FMS) – ÖBH Modell: „Die Forschungsbilanz ÖBH“,* Wien, 2010

BMWA (2008), *Leitfaden Wissensbilanz – Made in Germany.* Bundesministerium für Wirtschaft und Arbeit in Zusammenarbeit mit dem Arbeitskreis Wissensbilanz, URL: http://www.akwissensbilanz.org/Infoservice/Infomaterial/WB-Leitfaden_2.0.pdf [13.11.2013]

Bornemann & Reinhardt (2008), *Handbuch Wissensbilanz-Umsetzung und Fallstudien*, ESVVerlag, Berlin, S. 13

Cuviello (2009), *Army Knowledge Management - The Interoperability Enabler,* LTG, Army Chief Information Officer, US Army, Vortrag [27.03.2009]

Hesig (2008), Wissensbilanz – Made in Europe, *Wissensmanagement – Das Magazin für Führungskräfte*, Heft 4/2008

Hinkelmann, Karagiannis & Telesko (2002), PROMOTE - Methodologie und Werkzeug zum geschäftsprozessorientierten Wissensmanagement. In: *Geschäftsprozessorientiertes Wissensmanagement*, Springer-Verlag

Karagiannis (1995) *BPMS: Business Process Management Systems: Concepts, Methods and Technologies*, SIGOIS Special Issue, SIGOIS Bulletin 10-13, 1995

Knowledge Research 2002, Verfügbar unter: http://www.knowledgeresearch.com/strategies.htm [01.11.2002]

Mak, Göllner, Meurers & Utz (2010), *PROMOTE® in the Austrian Armed Forces: The Knowledge Management Processes in the Department of Central Documentation and Information Service/National Defence Academy,* 17th EKAW 2010 - Knowledge Engineering and Knowledge Management – Workshop OKM-Open Knowledge Model at the Gulbenkian Foundation, Lissabon, Portugal

MATURE (2010), MATURE Projekt, “*D1.1 Results of the Ethnographic Study and Conceptual Knowledge Maturing Model*”, URL: http://mature-ip.eu/files/deliverables/D1.1_Ethnographic_Studies_Knowledge_Maturing_Model.pdf [25.10.2010]

Nemetz (2006), A Meta-Model for Intellectual Capital Reporting In: Reimer, Karagiannis, (Hrsg.): *Proceedings of the 6th International Conference on Practical Aspects of Knowledge Management*. Berlin: Springer.

PKM (2009), *1st Workshop on Personal Knowledge Management (PKM)*, Verfügbar unter: http://www.wm-konferenz2009.org/workshops/PKM.php, [12.12.2009]

PROMOTE (2001), (IST-1999-11658), *Deliverable 2.1 Introduction of a knowledge management strategy, setting evaluation criteria*, PROMOTE-Consortium, Wien, 2001

WIEM (2009), *WIEM 2009, Messen, Bewerten und Benchmarken des wirtschaftlichen Erfolgs von Wissensmanagement*, Verfügbar unter: http://www.wm-konferenz2009.org/workshops/WIEM2009.php, [10.12.2009]

Woitsch, Mak & Göllner (2010), *Grundlagen zum Wissensmanagement im ÖBH – Teil 1: Ein WM-Rahmenwerk aus Sicht praktischer Anwendungen*, Wien, 2010

Woitsch, Mak & Göllner (2010a), *Grundlagen zum Wissensmanagement im ÖBH – Teil 2: Wissensbilanz als Steuerungsinstrument im ÖBH: Ein Evaluierungs-Rahmenwerk aus der Sicht praktischer Anwendungen*, Wien, 2010

Woitsch, Mak & Göllner (2010b), *Intellectual Capital Management using Knowledge Scorecards: The Austrian Defence Academy Showcase*, DEXA 2010-EGOVIS ´10-International Conference on Electronic Government and the Information Systems Perspective at the University of Deusto, Bilbao, Spanien

Social Media als Informationsquelle in der Unterrichtsvorbereitung: Das Beispiel einer berufsbezogenen Online-Community

Angela Vorndran, Silvana Keutz, Daniela Olenik, Aida Raimbekova & Marc Rittberger

Kollaborative Wissensgenerierung und Informationsaustausch werden durch zahlreiche im Internet angebotene Plattformen unterstützt. Auch im beruflichen Arbeitsfeld können diese Formen des Wissensaustauschs relevant sein und für viele verschiedene Informationsbedarfe eine Informationsquelle mit expertengeneriertem Wissen und praxisrelevanter Information darstellen z.B. in Form einer berufsspezifischen Online-Community. Webbasierte Portale dienen auch in traditionell stärker individuell geprägten Arbeitsbereichen wie der Unterrichtsvorbereitung von Lehrerinnen und Lehrern als geeignete Informationsquellen, die von der in dieser Suchumgebung einzigartigen Mischung aus handlungsrelevanter Information, fachkundig erteilten Empfehlungen und Erfahrungswissen profitieren. In dieser Studie werden bereitgestellte Unterrichtsentwürfe, Unterrichtsmaterialien und Diskussionsforen in den Fächern Biologie und Geschichte in einem von Lehrpersonen selbst organisierten Onlineportal inhaltlich analysiert. Auch wird untersucht, für welche berufsbezogenen Informationsbedarfe und Nutzungsweisen diese Art eines kollaborativ organisierten webbasierten Informationsangebotes geeignet ist.

1 Einführung

Laut einer Studie von Michel (2008) nutzen 90 Prozent der Lehrerinnen und Lehrer das Internet zur Unterrichtsvorbereitung. Eine entsprechend hohe Anzahl von Angeboten ist im Laufe der letzten Jahre entstanden, die diese Zielgruppe bedienen. Von Lehrpersonen selbst organisierte Portale spielen ebenso eine Rolle wie von öffentlichen Institutionen erstellte Angebote und solche von Schulbuchverlagen oder Privatpersonen. Frühere Forschung zeigt, dass im Lehrerberuf traditionell eine stark individualisierte Arbeitsweise vorherrscht und Kooperation auch speziell in der Unterrichtsvorbereitung nur selten stattfindet (vgl. Heinen/Blees 2011:112, Tebrügge 2001:68). Die Teilnahme an onlinebasierten community-orientierten Angeboten etabliert sich dementsprechend auch erst langsam. Breiter et al. ermitteln zehn Prozent befragter Lehrpersonen, die mindestens gelegentlich Materialien auf einer Plattform zur Verfügung stellen, und sieben Prozent, die sich dort austauschen (Breiter et al. 2010:238). Eine Untersuchung unter Lehrpersonen aus dem Jahr 2011 beschreibt allerdings eine Nutzung von Online-Netzwerken von 54 Prozent der befragten mathematisch-naturwissenschaftlichen Fachlehrer (Initiative D21 2011). Im Gegensatz dazu ergeben eine Anzahl älterer Untersuchungen eine hohe Wichtigkeit der Kolleginnen und Kollegen für die Informationsversorgung (vgl. Mundt et al. 2006, Savolainen 1995), so dass sich diese Präferenz bei weiterer Etablierung entsprechender Angebote auch im Internet abzeichnen könnte. Bei der Nutzung von Information aus webbasierten Quellen ist für Lehrende wichtig, dass es sich um redaktionell erstelltes Material handelt (Michel 2008:12), überprüfbare Qualitätskriterien, ein Mehrwert zu konventionellem Material und beständige Lokalisierbarkeit (Breiter et al. 2010:81) gegeben sind.

Im Mittelpunkt dieser Untersuchung steht das Portal *4teachers*, das von Lehrerinnen und Lehrern selbst organisiert und betrieben wird. Es gehört zu den meist genutzten schul-

bezogenen Onlineportalen in Deutschland (vgl. Hron/Neudert 2006:121, Schindler/Botte 2007:41, Initiative D21 2011:17).

Das Portal *4teachers* verfügt über verschiedene interaktiv ausgerichtete Angebote, die von einer hohen Teilnehmerzahl frequentiert werden (über 980.000 Mitglieder). Beispiele dafür sind ein umfangreicher Forenbereich mit nahezu 29.000 Diskussionen und über 410.000 Beiträgen, ein Tauschbereich von Unterrichtsmaterialien mit über 57.000 Materialien, die zusätzlich über 100.000 Kommentare verzeichnen und u. a. einen Chat und ein Wiki (www.4teachers.de, 7.2.2014).

In dieser Untersuchung soll das Informationsverhalten der Nutzerinnen und Nutzer des Portals bei der Suche nach unterrichtsrelevanter Information näher beleuchtet werden. Hierzu zählen insbesondere Aspekte des Informationsbedarfs, des Suchverhaltens und der Verwendung von Information (vgl. Pettigrew et al. 2001:44). Es soll ein genauerer Blick darauf geworfen werden, welche Fragestellungen in Onlineangeboten mit kollaborativem Charakter auftreten und wie die speziellen Charakteristiken eines solchen Angebots im Hinblick auf die Suchumgebung und die Art der angebotenen Information genutzt werden. Weiterhin sollen Bezüge zum beruflichen Aufgabenfeld der Unterrichtsvorbereitung von Lehrenden gezogen werden, da frühere Untersuchungen gezeigt haben, dass der berufliche Kontext für die Informationssuche eine wichtige Rolle spielt, (vgl. Leckie et al. 1996, Byström/Hansen 2005).

2 Methoden

Die methodische Herangehensweise umfasst eine qualitative Inhaltsanalyse von Forendiskussionen, Unterrichtsentwürfen und Unterrichtsmaterialien sowie den dazu verfassten Kommentaren. Die Inhaltsanalyse orientiert sich an im Vorfeld erarbeiteten oben zusammengefassten Forschungsfragen zu Informationsverhalten und Besonderheiten der Suchumgebung. Dieser problemorientierte Ansatz wird im weiteren Vorgehen durch bei der Arbeit am Textkorpus auftretende Aspekte ergänzt. Angewendet wird demnach eine Kombination aus problemorientierter und textbasierter Inhaltsanalyse (vgl. Krippendorff 2004: 340ff.).

Zunächst wird ein grobes Kategorienschema deduktiv erstellt und im Laufe der Analyse durch weitere Kategorien, die aus dem Text induktiv entwickelt werden, ergänzt. Weiterhin können bestehende Kategorien konkretisiert und um weitere Unterkategorien erweitert werden (vgl. Kuckartz 2007:60ff.). In der Textanalyse werden somit Kategorien entwickelt, die das Datenmaterial im Hinblick auf die gewählten Forschungsfragen umfassend und systematisch abbilden.

Als Grundlage für diese Untersuchung wird eine Auswahl an Datenmaterial aus den im Portal *4teachers* bereitgestellten Inhalten getroffen. Aus den Fächern, die ausreichende Datengrundlagen bieten und verschiedene Fachbereiche abdecken, wurden Biologie und Geschichte randomisiert ausgewählt. Eine Untersuchung von Tebrügge (2001) zeigt, dass Unterschiede in der Informationsnutzung verschiedener Fachlehrer bestehen, so dass auch hier diese Aspekte vergleichend berücksichtigt werden sollen. Beispielsweise spielen für das Fach Mathematik Schulbücher eine große Rolle, während in Chemie und Deutsch eigene Materialsammlungen den höchsten Stellenwert einnehmen. Weiterhin werden in allen drei Fächern sowohl Publikationen von Schulbuchverlagen als auch andere Publikationen häufig genutzt. Materialaustausch unter Kollegen spielt eine untergeordnete Rolle, webbasierte Quellen werden in der Untersuchung nicht genannt.

Für die qualitative Inhaltsanalyse werden je 30 Unterrichtsentwürfe, 30 Unterrichtsmaterialien und 20 Threads der Diskussionsforen pro Fach zugrunde gelegt. Die Auswahl des zu untersuchenden Datenmaterials erfolgt basierend auf für die einzelnen Materialbereiche entwickelten Samplingkriterien. Unterrichtsentwürfe und Unterrichtsmaterialien werden jeweils anteilig aus den bestehenden thematischen Unterordnern der Fächer nach Aktualität entnommen und unter der Voraussetzung einbezogen, dass sie mindestens einen Kommentar erhalten haben, um dadurch eine größere Datengrundlage für die angestrebten Erkenntnisse zu erhalten und einen Anhaltspunkt für eine tatsächlich erfolgte Nutzung des Materials zu haben. Für die Auswahl zu analysierender Forendiskussionen werden Threads nach Aktualität sowie mit der Einschränkung, dass sie drei bis zwanzig Beiträge enthalten, einbezogen.

3 Ergebnisse

3.1 Unterrichtsentwürfe

Unterrichtsentwürfe sind schriftliche Planungen für den Ablauf einer Schulstunde mit detaillierten Unterrichtsverlaufsplänen und Erläuterungen der verwendeten Methoden. Teilweise enthalten diese Entwürfe zusätzliche Materialien, etwa Arbeitsblätter oder Tafelbilder. Es wird davon ausgegangen, dass solche Materialien von Lehrerinnen und Lehrern selbst erstellt werden, die durch vorhandene Schulbücher nicht abgedeckt werden, nach denen aber ein Bedarf besteht (vgl. Bromme/Hömberg 1981:72, Tebrügge 2001:141).

Die meisten Unterrichtsentwürfe (knapp 80 %) beinhalten die zu bearbeitenden Arbeitsmaterialien wie Arbeitsblätter, Folien oder Texte. Insgesamt werden in Biologie (B) 59, in Geschichte (G) 55 Arbeitsmaterialien den je 30 Unterrichtsentwürfen beigefügt. Das Unterrichtsmaterial, das am häufigsten enthalten ist, ist das Arbeitsblatt (B 25 %, G 27 %) (vgl. Abb. 1). Häufig enthalten die Entwürfe auch Vorlagen für Tafelbilder (B 24 %, G 27 %). Die Kategorie „Sonstiges“ ist bei den Unterrichtsentwürfen für Biologiestunden weniger häufig belegt (14 %) als bei den Entwürfen für Geschichtsstunden (20 %). Diese Kategorie beinhaltet z.B. Hausaufgaben, Arbeitsaufträge, Karten, Tests oder Bewertungsbögen.

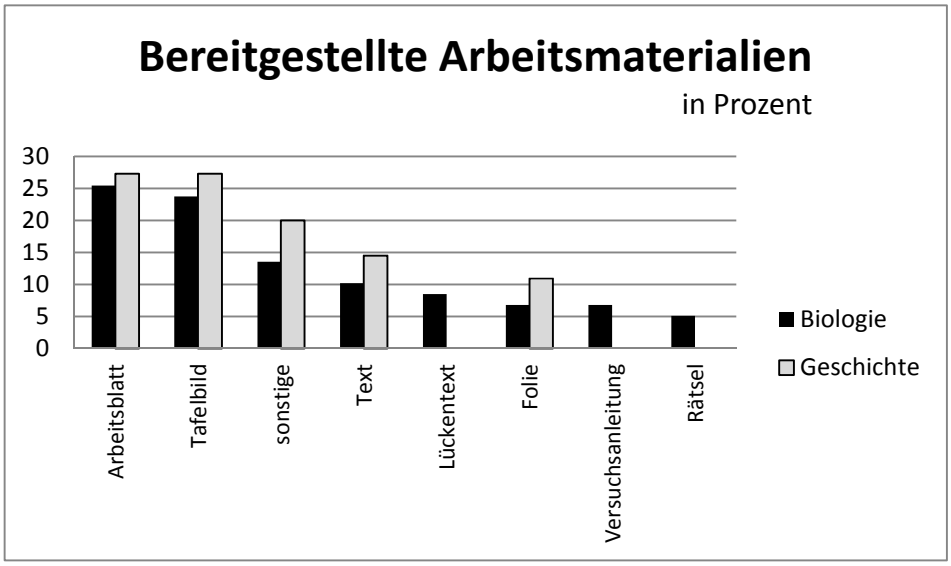


Abbildung 1: In Unterrichtsentwürfen für Biologie und Geschichte enthaltene Arbeitsmaterialien.

Die Mehrzahl der Unterrichtsentwürfe enthält weiterhin Angaben zum Medieneinsatz innerhalb der geplanten Schulstunde (G 97 %, B 93 %). Insgesamt dominiert der Einsatz der traditionellen Tafel (25 %) und von Overheadfolien (24 %), gefolgt von Arbeitsblättern (19 %) (vgl. Abb. 2). Diese Aufteilung zeigt sich auch bei Betrachtung der einzelnen Fächer. Am häufigsten werden im Fach Geschichte Overheadfolien (26 %) eingesetzt, gefolgt von der Tafel (24 %) und dem Arbeitsblatt (19 %). In mehr als 75 % der Entwürfe werden Tafel und Overheadfolien eingesetzt. Im Fach Biologie stehen die Tafel (21 %) aber auch die Kategorie „Sonstiges“ (21 %) an erster Stelle. Die Kategorie „Sonstiges“ umfasst z.B. Medien wie Filme, Hörbeispiele, Bilder aber auch unterschiedliche Materialien oder Hilfsmittel zur Durchführung von Versuchen (etwa Mikroskope, Petrischalen usw.).

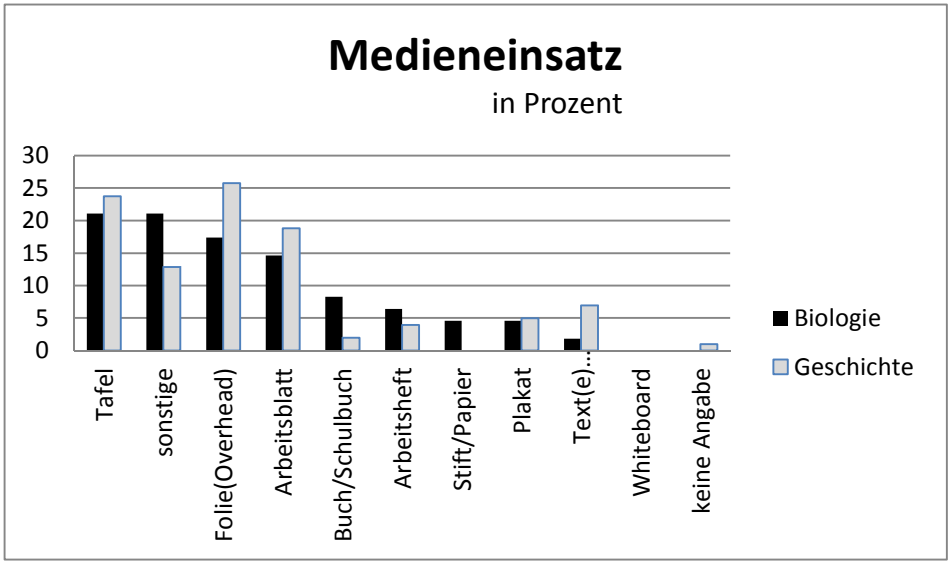


Abbildung 2: In Unterrichtsentwürfen für Biologie und Geschichte eingesetzte Medien.

Für beide Fächer liegen keine Entwürfe vor, in denen der Einsatz des Whiteboards angegeben wird. Beim Medium „Buch“ gibt es einen deutlichen Unterschied zwischen den Fächern zugunsten von Biologie (8 %, G 2 %). Bei dem Medium „Text“ findet sich eine umgekehrte Verteilung (B 2 %, G 7 %).

Die Unterrichtsentwürfe beider Fächer erhalten durchschnittlich drei Kommentare, die mehrheitlich positiv ausfallen (69 %). 13 Prozent der Kommentare enthalten negative Beurteilungen, 17 Prozent sind als neutral zu bezeichnen.

Die positiven Bewertungen im Fach Biologie enthalten zur Hälfte ein allgemeines Lob für den Unterrichtsentwurf. 31 Prozent loben den Aufbau der Schulstunde bzw. die verwendete Methode. 19 Prozent erwähnen das enthaltene Arbeitsmaterial positiv. Negativ bewerten die Lehrkräfte Entwürfe, die kein Arbeitsmaterial enthalten oder bei denen die Quellenangaben fehlen. 30 Prozent der abgegebenen Kommentare für die Unterrichtsentwürfe im Fach Biologie enthalten keine weiteren auswertbaren Bemerkungen.

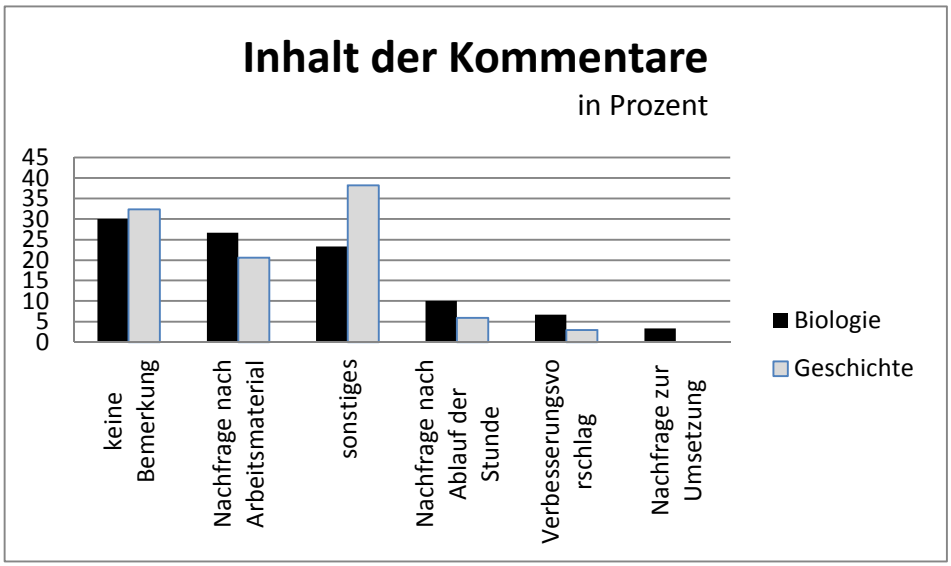


Abbildung 3: Inhalt der Kommentare zu Unterrichtsentwürfen in Biologie und Geschichte.

Für beide Fächer kann festgestellt werden, dass häufig Arbeitsmaterial nachgefragt wird (26 % B, 21 % G) (vgl. Abb. 3). Seltener wird nach dem Ablauf der Stunde gefragt (10 % B, 6 % G) oder Verbesserungsvorschläge gemacht (7 % B, 3 % G). Die Kategorie „Sonstiges" ist häufig vertreten, hierunter fallen die unterschiedlichsten Äußerungen, z.B. Nachfragen nach der Bewertung oder Anmerkungen, dass die Stunde so ähnlich durchgeführt wurde oder der Entwurf auch für andere Klassenstufen, als die angegebenen, geeignet sei. Aber auch negative Bemerkungen sind in dieser Kategorie enthalten, z.B. wenn das Fehlen des Unterrichtsverlaufsplans bemängelt wird.

3.2. Unterrichtsmaterialien

Unter Unterrichtsmaterialien werden sämtliche Unterlagen verstanden, die im Unterricht zum Einsatz kommen können, z.B. Arbeitsblätter, Klausuren oder Texte. Unterrichtsmaterialien, die auf dem Portal *4teachers* angeboten werden, können als Lernmittel direkt an die Schüler weitergegeben oder von den Lehrpersonen ihren Anforderungen angepasst werden.

Die Auswertung hinsichtlich des Medientyps des bereitgestellten Unterrichtsmaterials zeigt ein Überwiegen von Arbeitsblättern mit Fragen (34 %), gefolgt von sonstigen Materialien (wie z.B. Mindmap oder Steckbrief) (26 %) und Lückentexten (14 %) (vgl. Abb. 4). Im Fach Biologie werden häufiger auch Rätsel (10 %) und Kartenspiele (10 %) bereitgestellt, selten Puzzles (4 %), Schaubilder (3 %), Einstiegstexte (3 %) und Klausuren (3 %). Im Fach Geschichte sind dagegen die Medientypen Rätsel, Kartenspiel und Puzzle überhaupt nicht vertreten, Lückentexte hingegen zu 12 %, ebenso Einstiegstexte (11 %) und Schaubilder (6 %). Der große Anteil an nicht klassifizierbaren Materialien von 20 Prozent in Biologie und 31 Prozent in Geschichte zeigt die große Diversität der angebotenen Unterrichtsmaterialien und lässt auf Kreativität und methodischen Einfallsreichtum der Lehrkräfte, die diese Materialien zur Verfügung stellen, schließen. Im Vergleich mit Unterrichtsmaterialien, die in den Unterrichtsentwürfen integriert sind zeigt sich, dass in diesen das Tafelbild eine deutlich größere Bedeutung einnimmt, während bei den alleine angebotenen Materialien eher Material zum selbständigen Arbeiten angeboten wird.

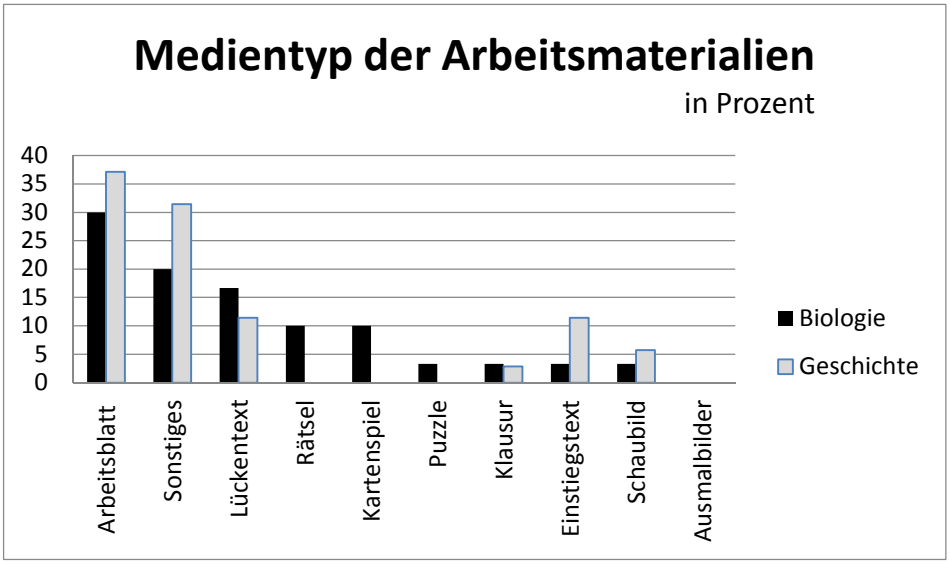


Abbildung 4: Medientypen der Arbeitsmaterialien in Biologie und Geschichte.

Aus der Art der Unterrichtsmaterialien können Rückschlüsse über wahrscheinliche Sozialformen des Unterrichts bei Einsatz des Materials gezogen werden. Hierbei, zeigt sich in den Sozialformen des Unterrichts, dass eigenständiges Arbeiten der Schülerinnen und Schüler wie Einzelarbeit (80 %) oder Gruppenarbeit (14 %) stark dominiert und nur ein kleinerer Teil Arbeit im Klassenverband (9 %) thematisiert. Die Auswertung der in den Unterrichtsmaterialien beinhalteten unterrichtlichen Handlungsmuster, die sich in diesem Fall hauptsächlich an den Aktivitäten der Schülerinnen und Schüler orientieren, zeigen einen Großteil an Übungen (51 %), Rezeption von Texten oder anderen Informationen (24 %) und in geringerem Umfang Spiele (14 %) und Aktivitäten wie Rollenspiele oder eigenständig durchgeführte Experimente (12 %).

Die untersuchten Unterrichtsmaterialien verzeichnen durchschnittlich zwei Kommentare. Die Kommentare enthalten in den meisten Fällen positive Bewertungen des Materials (70 % B, 74 % G), zu einem kleineren Teil auch neutrale Bewertungen (21 % B, 26 % G). Negative Kommentare treten in vier Prozent auf und sind auf das Fach Biologie beschränkt.

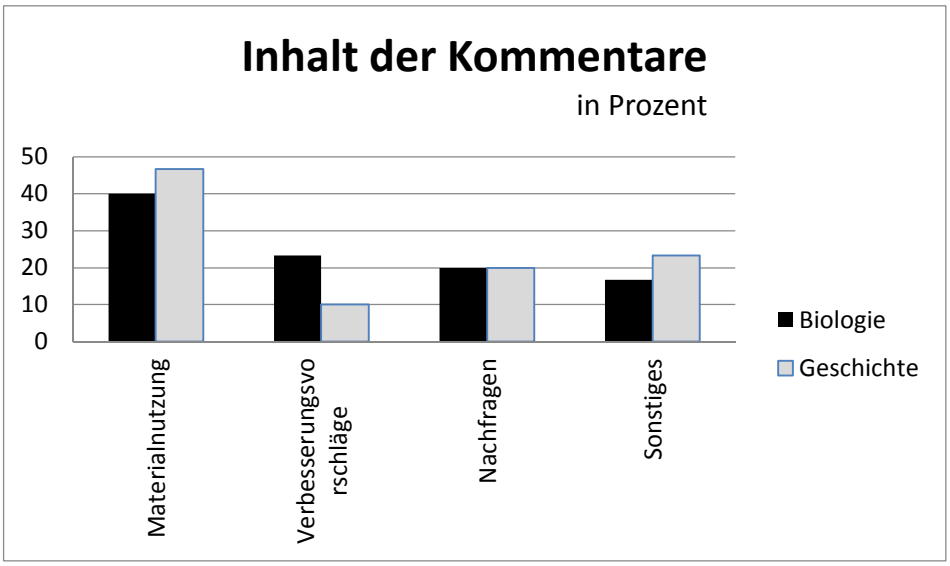


Abbildung 5: Inhalte der Kommentare zu Unterrichtsmaterialien in Biologie und Geschichte.

Inhaltliche Schwerpunkte der Kommentare, die mithilfe von Kategorien erfasst werden zeigen, dass Kommentare mit Angaben zur Materialnutzung den größten Anteil stellen (40 % B, 47 % G). Einerseits ist hier der Einsatzzweck der Information im Unterricht relevant, z. B. als Vorlage zur Unterrichtsgestaltung, Unterstützung für schwache Schüler oder als Hintergrundinformation. In anderen Fällen wird mehr Wert auf die Funktion des Informationseinsatzes im Unterricht gelegt, z. B. als Festigung, Prüfungsvorbereitung oder Zusammenfassung. Auch der Zeitpunkt des Einsatzes im Verlauf einer Unterrichtsstunde kann als dominanter Aspekt der Nutzungsabsicht zutage treten, z. B. bei einem Einsatz der Information als Stundeneinstieg oder im Rahmen einer Freiarbeitsphase für die Schülerinnen und Schüler.

Eine weitere Kategorie der Kommentarinhalte stellen Verbesserungsvorschläge dar, die vor allem im Fach Biologie einen großen Anteil einnehmen (23 %, G 10 %) (vgl. Abb. 5). Sonstige Bemerkungen (B 17 %, G 23 %) beziehen sich beispielsweise auf technische Schwierigkeiten oder Danksagungen. Weiterhin stellen Nachfragen zum Material ein Fünftel der Kommentare in beiden Fächern. Aus der Auswertung dieser Kommentare wird deutlich, dass Nutzer auf die Beigabe von Lösungen zu den angebotenen Materialien großen Wert legen, da diese in den meisten Fällen nachgefragt werden. Größtenteils werden diese Erwartungen auch erfüllt, denn in beiden Fächern beträgt der Lösungsanteil über 80 Prozent.

3.3 Forendiskussionen

Die Daten der Forendiskussionen bieten einen reichhaltigen Bestand an auswertbaren Informationen, die das Informationsverhalten von Nutzerinnen und Nutzern des Portals betreffen. In den Anfragen der Threaderöffner finden sich zum Teil sehr ausführliche Angaben darüber, welche inhaltlichen Schwerpunkte die Anfrage hat, welche Motivation für die Informationssuche den Anstoß gibt und welche Anforderungen an gesuchte Information gestellt werden. Die inhaltliche Analyse der Forenkommunikation eröffnet somit die Gelegenheit, sowohl über inhaltliche Problemstellungen des beruflichen Kontextes von Lehrpersonen, als auch über die konkreten Einflussfaktoren des Arbeitsumfeldes und die Besonderheiten der Kommunikation in der speziellen Suchumgebung eines Diskussionsforums Erkenntnisse zu erlangen. Die je zwanzig untersuchten Threads der beiden Fächerforen enthalten 115 Beiträge in Biologie und 134 Beiträge in Geschichte.

3.3.1 Anfragen

In den Anfragen der Forenkommunikation werden Angaben zu den Informationsbedarfen der Teilnehmer erhoben, die sich didaktischen Arbeitsbereichen zuordnen lassen, die in der pädagogischen Forschung von Klafki als Binnengliederung der Didaktik definiert werden (1970:151f., 2007:92f.). Die Auswertung der Forenanfragen zeigt, dass vor allem methodische Fragestellungen in beiden Fächern stark in Forendiskussionen thematisiert werden (58 %) (vgl. Abb. 6). Inhalte und Themen des Unterrichts werden vor allem im Fach Geschichte gesucht (40 %, B 15 %), während Fragen zur Mediennutzung stärker für das Fach Biologie relevant sind (15 %, G 5 %). Unterrichtsziele und Themen zu Kontroll-, Beurteilungs- und Sanktionsmaßnahmen werden in den untersuchten Threads nicht angesprochen.

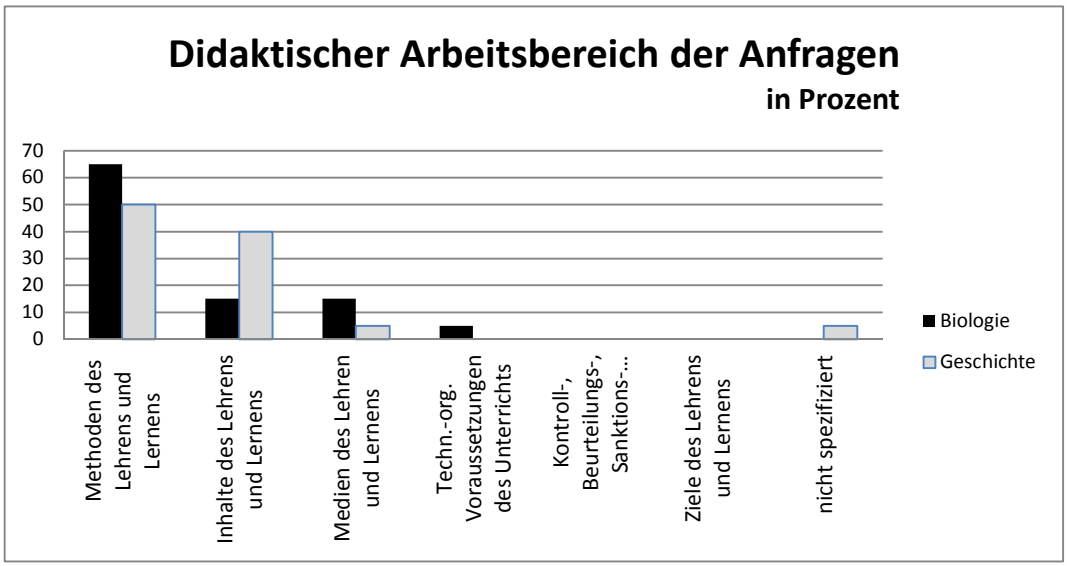


Abbildung 6: Didaktischer Arbeitsbereich der Forenanfragen in Biologie und Geschichte.

Die Anfragen in der Forenkommunikation beinhalten außerdem häufig Hinweise auf die Präferenzen der Fragenden hinsichtlich eines Medientyps, der für die Fragestellung geeignet ist (vgl. Abb. 7). In beiden Fächern ist bemerkenswert, dass zum größten Teil (56 %) eigene Erfahrungen im Sinne von Kommentaren zu eigenen Vorschlägen, Ideen für beschriebene Fragestellungen oder Beurteilungen einer Situation erfragt werden. Vor allem im Fach Biologie werden auch weitere konkrete Anforderungen an gesuchte Medientypen gestellt, wie nach Arbeitsblättern (10 %) für die selbständige Arbeit der Schülerinnen und Schüler, Filmen (10 %) und Anleitungen für Versuche (20 %), die in einem naturwissenschaftlichen Fach erwartungsgemäß stärker vertreten sind als in Geschichte. Im Fach Geschichte wird neben der Präferenz für eigene Erfahrungen der Forenteilnehmer eine weniger starke Spezifizierung nach gesuchten Medientypen vorgenommen.

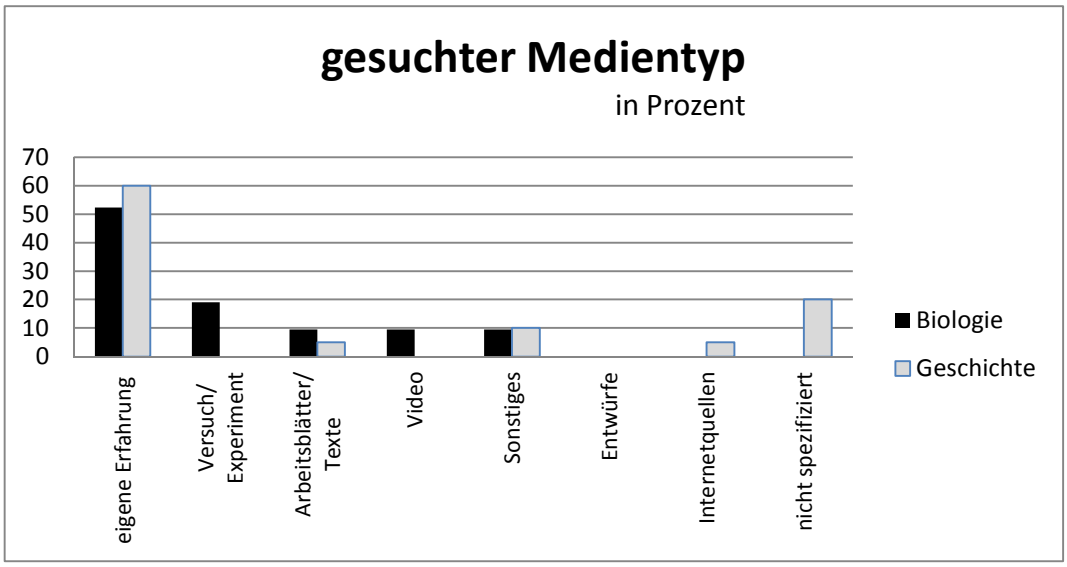


Abbildung. 7: Gesuchte Medientypen der Forenanfragen in Biologie und Geschichte.

Ein weiteres Element in den Anfragen im Diskussionsforum ist der Umfang der Angaben zu Hintergrund und Inhalt der Anfrage. Beide Fächer weisen mit 80 bzw. 86 Angaben in 20

Anfragen ähnliche Anzahlen auf. Ebenfalls in beiden Fächern werden am häufigsten Angaben zum Thema der gesuchten Information gemacht (20 %, vgl. Abb. 8). Auch Angaben zum Grund der Suche (14 %) und die gesuchte Materialart (17 %) werden häufig genannt. In Geschichte werden weiterhin stärker strukturelle Aspekte wie der Jahrgang (17 %, B 13 %) und die Schulform (10 %, B 0 %) in der die Information eingesetzt werden soll integriert, während in Biologie eher die eigenen Hintergründe des Suchenden beschrieben werden, wie z. B. der bisherige Suchverlauf 8 %, G 2 %) und eigene Vorkenntnisse (5 %, G 1 %). Generell werden wenige Angaben zu den thematischen Vorkenntnissen der Suchenden und zum Zeitrahmen der geplanten Nutzung (B 6 %, G 3 %) gemacht.

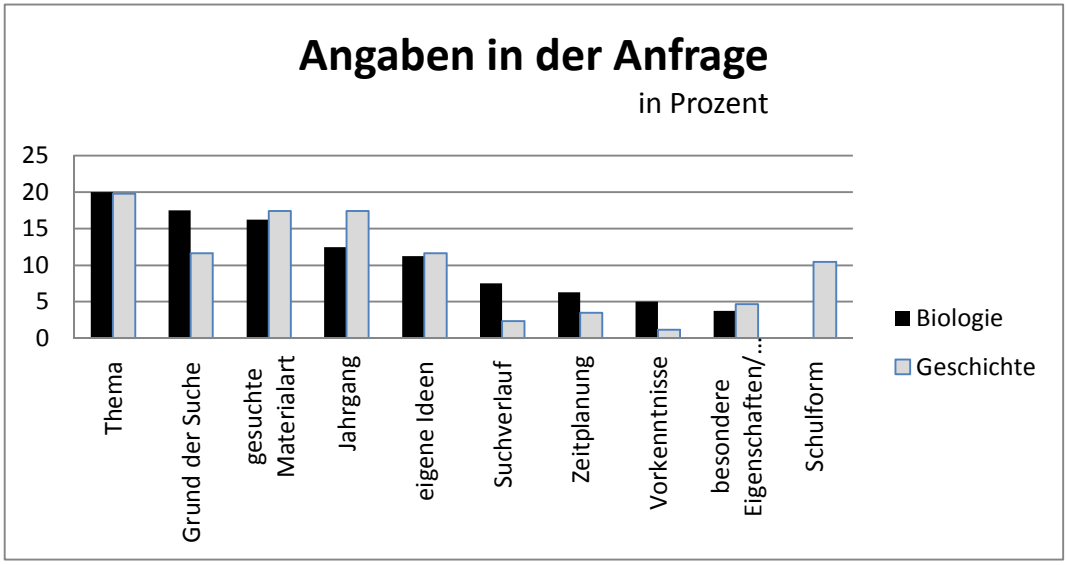


Abbildung 8: Angaben in den Forenanfragen in Biologie und Geschichte.

In einer geringen Zahl von Fällen werden auch weitere inhaltliche Anforderungen an die gesuchte Information gestellt, wie z.B. die Eignung zur differenzierten Bearbeitung durch Schülerinnen und Schüler mit unterschiedlichen Kompetenzstufen. Es zeigt sich, dass die Möglichkeit in einem Forenbeitrag durch eine unbeschränkte natürlichsprachliche Eingabe umfangreichere Informationen in eine Anfrage einzubringen von den Forenteilnehmerinnen und -teilnehmern mit einer durchschnittlichen Nennung von vier Hintergrundangaben pro Anfrage rege genutzt wird.

Die Motivation, die zur Eingabe einer Anfrage im Diskussionsforum führt, wird nach ihrem Ursprung in extrinsischen und intrinsischen Faktoren unterschieden. Insgesamt kann bei 47 Prozent der Anfragen eine extrinsische Motivation festgestellt werden, die zu etwas mehr als der Hälfte auf Anforderungen aus Studium oder Lehrerausbildung zurückzuführen ist.

Intrinsische Motivation, die auf dem eigenen Streben nach Information zur besseren Vorbereitung von Unterricht basiert, tritt in verstärktem Ausmaß im Fach Geschichte auf, wo 45 Prozent der Anfragen diesen Beweggründen zugeordnet werden, während in Biologie lediglich 20 Prozent der Anfragen diese Motivation zugrunde liegt. In 22 Prozent der Anfragen kann keine Angabe zum Ursprung der Motivation ermittelt werden.

Zum allergrößten Anteil (85 %) wird in den Angaben eine geplante Nutzung in direktem Kontext zur Anfrage angestrebt. Das heißt, die gesuchte Information soll in einem konkreten Szenario im Unterricht eingesetzt werden und wird zu diesem Zweck erfragt. Eine konzeptionelle Nutzung für einen noch nicht bestimmten späteren Einsatzzweck wird in fünf Prozent der Anfragen ermittelt.

3.3.2 Antworten

Eine Analyse der Art der Antwortbeiträge trägt zu einem besseren Verständnis der Kommunikationsstrukturen innerhalb des Forums bei. Ein Vorteil der Informationssuche in einem Forum kann eben in dieser Interaktion zwischen verschiedenen Akteuren liegen, die unterschiedliche Sichtweisen, Erfahrungen und Kenntnisse einbringen können. Die Beiträge beinhalten zu einem Großteil (50 %) direkte Antworten auf die Anfragen, die als Empfehlungen für an anderen Orten vorhandene Information durch Links oder Literaturtipps (26 %) und Antworten in den Beiträgen, die sich auf die eigene Erfahrung (5 %) oder konkrete Ideen der Antwortenden stützen (18 %), klassifiziert werden (vgl. Abb. 9).

Weiterhin ist in einigen Fällen ein Klärungsprozess zu einzelnen Elementen der Anfrage oder der Antworten zu beobachten, der in der Forenkommunikation im Gegensatz zu anderen Informationsangeboten möglich ist und zu einem größeren Nutzen der Antworten führen kann. Im Fach Geschichte liegt der Anteil der Antworten um drei Prozentpunkte niedriger als in Biologie. Es ist ein größerer Anteil an inhaltlicher Diskussion zu beobachten mit deutlich mehr kritischen Beiträgen als in der Biologie (G 4 %, B 1 %) und einer ebenfalls höheren Zahl an Kommentaren (G 15 %, B 13 %).

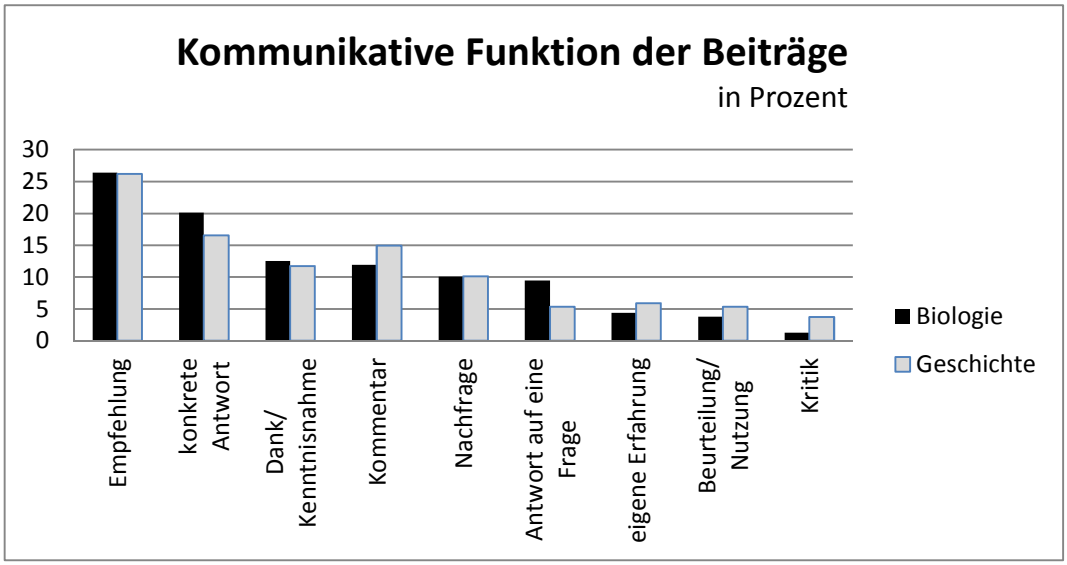


Abbildung 9: Kommunikative Funktionen der Antwortbeiträge in Biologie und Geschichte.

Auch eine Auswertung der Art der Antworten zeigt, dass im Fach Geschichte in größerem Ausmaß interaktive Kommunikation unter den Beitragenden stattfindet. Ein größerer Anteil an Beiträgen mit erklärenden Antworten tritt hier auf. Antworten, die nur Links enthalten sind zu sieben Prozent vertreten, in der Biologie zu 18 Prozent. Dagegen enthalten in Geschichte 24 Prozent der Antworten Links mit weiteren Anmerkungen, während nur 21 Prozent in Biologie dies aufweisen. Inhaltliche oder andere Nachfragen in den Diskussionsthreads kommen zum größten Teil von antwortenden Teilnehmern (77 %). Im Fach Geschichte sind dies sogar 84 Prozent der Nachfragen. Insgesamt verzeichnen beide Fächer in zehn Prozent der Beiträge eine Nachfrage. Dies zeigt, dass die Kommunikation der Forenteilnehmer andere Beiträge zur Kenntnis nimmt und auch hinterfragt oder weitere Erklärungen erbittet. Aus der regen Nutzung dieser Möglichkeit ist zu schließen, dass eine interaktive Suchumgebung bei der Experten eines Faches kollaborativ zusammenarbeiten für die hier geäußerten Anfragen von Vorteil ist.

Die empfohlene Information in den Antworten der Forendiskussionen wird hinsichtlich ihrer Medienarten und der jeweiligen Quellen ausgewertet. Im Fach Biologie werden insgesamt 88 Empfehlungen in den untersuchten 20 Threads gegeben, in Geschichte 118. Dies entspricht einer Verteilung von 0,76 Empfehlungen pro Beitrag in Biologie und 0,88 in Geschichte. Ein Großteil der empfohlenen Information liegt in Form von Websites (30 %) oder Berichten eigener Erfahrung (27 %) vor (vgl. Abb. 10), weiterhin werden zahlreiche Videos empfohlen (21 %). Auch unterrichtsspezifische Medienarten wie Arbeitsblätter (4 %) oder Unterrichtsentwürfe (9 %) spielen eine Rolle. Unterschiede in den beiden untersuchten Fächern zeigen sich in der Zahl an Empfehlungen für Websites (B 34 %, G 25 %), Printmaterialien (B 2 %, G 7 %) und Videos (B 25 %, G 15 %). Auch der Anteil an Empfehlungen, die auf eigener Erfahrung basieren liegt in Biologie deutlich höher als in Geschichte (B 35 %, G 18 %). Eine Zusammenfassung von Versuchs- und Bastelanleitungen in einer Kategorie führt hier zu einem häufigeren Auftreten im Fach Geschichte.

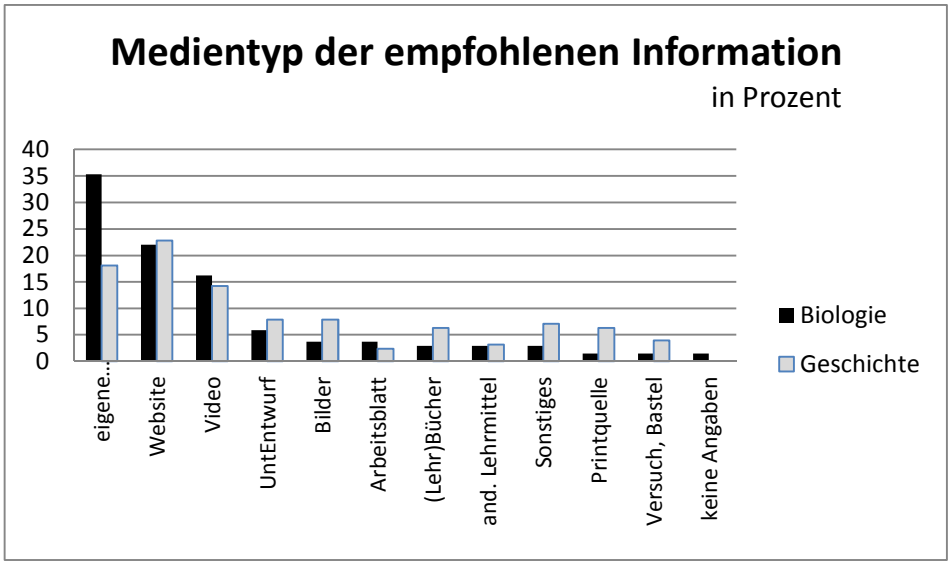


Abbildung 10: Medientyp der empfohlenen Information in Forenbeiträgen für die Fächer Biologie und Geschichte.

Die Auswertung online- und erfahrungsbasierter Quellen der empfohlenen Information ergibt eine Dominanz von allgemeinen Internetquellen (30 %) und eigenen Erfahrungen (26 %) (vgl. Abb. 11). Websites öffentlicher Institutionen werden zu 14 Prozent und Websites mit der spezifizierten Zielgruppe der Lehrerinnen und Lehrer zu neun Prozent empfohlen. Verweise auf eigene Informationen des Portals *4teachers* sind in 20 Prozent der Empfehlungen zu finden. Unterschiede zwischen den Fächern Biologie und Geschichte treten hier in größerem Umfang auf. Beispielsweise in Form einer klaren Präferenz im Fach Biologie für Erfahrungsberichte (37 %, G 16 %). Im Fach Geschichte werden allgemeine Websites (33 %, B 27 %), für Lehrpersonen konzipierte Websites (12 %, B 7 %) und Information aus dem Portal *4teachers* (26 %, B 13 %) häufiger empfohlen. In Bezug auf Offlinemedien werden keine weiteren Auswertungen zu den Quellen vorgenommen.

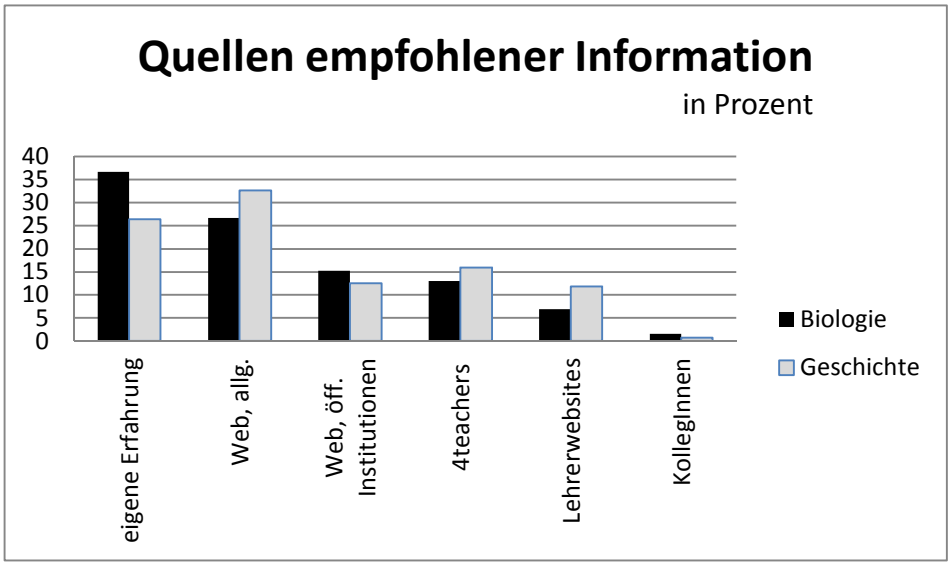


Abbildung 11: Quellen empfohlener Information in Forenbeiträgen für die Fächer Biologie und Geschichte.

Die Zufriedenheit der Fragesteller mit der empfohlenen Information wird in 59 Prozent der Threads positiv beurteilt In einem Drittel der Threads liegt keine Angabe zur Zufriedenheit vor.

4 Diskussion

Die Untersuchung der ausgewählten Daten der Plattform 4teachers kann in verschiedenen Hinsichten Erkenntnisse zum Informationsverhalten dieser Berufsgruppe in dieser speziellen Informationsumgebung liefern. Die Auswertung der ausgewählten Forenbeiträge zeigt, dass die prägenden Arbeitsbereiche für die Suche nach Information „Methoden des Lehrens und Lernens“ (58 %) und „Inhalte des Lehrens und Lernens“ (28 %) (vgl. Klafki 1970, 2007) sind.

Zu methodischen Fragestellungen werden in der Forenkommunikation vor allem eigene Erfahrungen (69 %) oder Versuchsbeschreibungen (15 %) gesucht. Empfehlungen zu diesen Anfragen beinhalten eine Vielzahl an verschiedenen empfohlenen Medientypen unter denen Erfahrungen (38 %) und allgemeine Webseiten (18 %) den größten Anteil bilden. Quellen für genannte Information sind häufig das Portal 4teachers selbst (34 %) oder allgemeine Webseiten (31 %). Auch Kommentare zu Unterrichtsmaterial thematisieren oft methodische Aspekte z.B. zur inhaltlichen, funktionalen oder zeitlichen Integration des Materials in den Unterricht.

In Anfragen, die den Arbeitsbereich Unterrichtsinhalte betreffen, werden deutlich weniger eigene Erfahrungen nachgefragt (23 %), stattdessen stehen konkrete Materialarten wie Arbeitsblätter oder Videos (jeweils 15 %) im Fokus. Allerdings wird auch in 23 Prozent der Anfragen keine Spezifikation hinsichtlich des Medientyps gemacht. Auch hier zeigen die Empfehlungen eine große Bandbreite an Medientypen, bei denen Webseiten und Videos (beide 26 %) dominieren. Unter Unterrichtsmaterialien werden vor allem Unterrichtsentwürfe (9 %) und (Lehr)bücher (8 %) empfohlen. Auch in den Kommunikationsstrukturen lassen sich Unterschiede je nach Arbeitsbereich feststellen. So werden für Unterrichtsmethoden mehr direkte Antworten im Beitrag gegeben (25 % im Vgl. zu 10 %), während zu Inhalten mehr Empfehlungen auf andere Quellen gegeben werden (43 % im Vgl. zu 21 %).

Weiterhin sind aus der Art der angebotenen Unterrichtsmaterialien Präferenzen für bestimmte Inhalte abzulesen, da die Erstellung dieser in einem wahrgenommenen Bedarf gründet. Nach Hochladen auf der Plattform 4teachers wird das Material auch genutzt und größtenteils positiv bewertet (ca. 70 %). Zu den Materialien in Unterrichtsentwürfen gehören häufig explizit für den Unterricht entworfene Formate wie Arbeitsblätter (26 %) oder Tafelbilder (25 %). Bei alleinstehenden Unterrichtsmaterialien spielen auch unterhaltenden Formate wie Rätsel oder Spiele mit elf Prozent eine größere Rolle. Sowohl die Materialien als auch die vorherrschenden Arten der Mediennutzung, wie Tafel (19 %), Overheadfolien (18 %) und Arbeitsblätter (14 %), weisen dabei eher auf einen Einsatz von Methoden und Inhalten nach etablierten Vorgehensweisen der Kombination von Lehrervortrag und Einzelarbeit der Schülerinnen und Schüler hin. Allerdings ist in allen Unterrichtsinhalten ein hoher Anteil an als ‚Sonstiges' klassifizierten Materialien zu bemerken, was für eine große Vielfalt an weiterhin relevanten Medientypen spricht.

Insgesamt sind deutliche Unterschiede im Informationsverhalten bei der Suche nach Information für verschiedene Arbeitsbereiche zu beobachten. Die Informationsumgebung der Online-Community bietet in beiden Bereichen Vorteile, da Information für methodische Fragestellungen offensichtlich leichter durch Erfahrungsaustausch erlangt wird als durch Publikationen. Im Arbeitsbereich Inhalte sind konkrete Materialien relevant, die hier durch Empfehlungen und Hinweise zum Einsatz angereichert werden können und so einen konkreteren und individualisierteren Handlungsbezug im beruflichen Kontext aufweisen. Dies wird beispielsweise auch darin deutlich, dass in Kommentaren explizit nachgefragt wird, dass Arbeitsmaterial handlungsrelevant aufbereitet und beispielsweise mit Lösungen versehen ist. Die Beobachtung, dass Information für die instrumentelle Nutzung in konkret geplanten Szenarien in großer Zahl benötigt wird (85 % aller Forenanfragen), unterstützt diesen Eindruck.

Die Tatsache, dass in einem Großteil der Forenanfragen eigene Erfahrungen oder Einschätzungen als gewünschter Antworttyp auftreten zeigt, dass der interaktive Austausch in kollaborativer Zusammenarbeit von Experten des Feldes, die im Portal stattfindet, ein Auszeichnungsmerkmal dieser Suchumgebung ist, das von den Teilnehmerinnen und Teilnehmern explizit als Vorteil gesehen und genutzt wird. Eine große Anzahl an Experten in Form von Fachkollegen ist verfügbar, die über einen reichhaltigen Erfahrungsschatz und weit gefächerte Kenntnisse über geeignete Informationsquellen und methodische Vorgehensweisen verfügt, die in anderer Weise nur schwer zu erhalten sind. Die Voraussetzungen dieser Suchumgebung ermöglichen außerdem eine häufig wahrgenommene umfangreiche Darstellung der Informationsbedarfe und ihrer Hintergründe mit ca. vier zusätzlichen Angaben pro Anfrage. Durch die Interaktion im Forum werden oft in Zusammenarbeit von Fragenden und Antwortenden Lösungen für die geschilderten Problemsituationen erarbeitet. Durch Nachfragen, die 14 Prozent der Forenbeiträge ausmachen, können Klärungs- und Austauschprozesse angestoßen werden. Im Gegensatz zu anderen Suchumgebungen können Antworten gefunden werden, die neue Aspekte beinhalten ohne dass diese durch Suchbegriffe durch den Fragenden vorgegeben werden müssen. Auch in den Kommentaren zu Unterrichtsentwürfen und Unterrichtsmaterialien wird durch Verbesserungsvorschläge (5 bzw. 8 %) und Nachfragen zu methodischen Aspekten (10 %) ein Diskurs über Fragestellungen der Unterrichtsvorbereitung und methodischer Vorgehensweisen geführt. Hierbei wird deutlich, dass in den Portalaktivitäten nicht nur das Teilen von Information zu beobachten ist, sondern durch Kommentare und in Forendiskussionen gemeinsame Idee zur Unterrichtsgestaltung und eventuell Weiterentwicklungen des Materials angestoßen werden. Im Austausch spielen auch stärker methodische Fragestellungen eine Rolle, die bei reinem Materialaustausch nicht thematisiert werden können.

5 Schlusswort

Die Untersuchung konnte zeigen, dass die kollaborative und interaktive Suchumgebung eines Onlineportals für die beruflichen Fragestellungen bei der Unterrichtsplanung besonders relevant sein können. Stark praxisrelevante und methodische Fragestellungen, die vermehrt auftreten, profitieren besonders von der angebotenen informellen von Experten der eigenen Zielgruppe generierten Information. In dieser Funktion können communityorientierte Informationsdienste eine geeignete Ergänzung zu formaler Informationsvermittlung durch Fachinformationsdienste sein. Im Weiteren wäre eine tiefergehende Untersuchung einer größeren Datenstichprobe, die durch die eine Auswertung von Webanalysedaten ergänzt werden könnte, interessant.

Literatur

Breiter, A., Welling, S., & Stolpmann, B. E. (2010). *Medienkompetenz in der Schule. Integration von Medien in den weiterführenden Schulen in Nordrhein-Westfalen* (Vol. 64). Berlin: Vistas.

Bromme, R., & Hömberg, E. (1981). *Die andere Hälfte des Arbeitstages. Interviews mit Mathematiklehrern über alltägliche Unterrichtsvorbereitung* (Vol. 25). Bielefeld: Institut der Didaktik der Mathematik der Universität Bielefeld.

Byström, K. & Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology*, 56(10), 1050–1061.

Heinen, R., & Blees, I. (2011). Social Bookmarking als Werkzeug für die Kooperation von Lehrkräften : Das Projekt Edutags für den Deutschen Bildungsserver. In *J. Griesbaum, T. Mandl, C. Womser Hacker (Hrsg.), Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011),* 111–122). Boizenburg: Verlag Werner Hülsbusch.

Hron, A., & Neudert, S. (2006). *Projekt EVA-LO Online-Befragung. Befragung zur Nutzung und Bewertung des Internetportals "Lehrer-Online."* Tübingen: IWM - Institut für Wissensmedien.

Initiative D21, (Hg.) (2011). *Bildungsstudie: Digitale Medien in der Schule.* Eine Sonderstudie Im Rahmen Des (N)ONLINER Atlas 2011. http://www.initiatived21.de/wp-content/uploads/2011/05/NOA_Bildungsstudie_140211.pdf, [12.2.2014].

Klafki, W. (1970). *Funkkolleg Erziehungswissenschaft: eine Einführung in drei Bänden; eine Vorlesungsreihe der Philipps-Universität Marburg in Zusammenarbeit mit dem Hessischen Rundfunk, dem Saarländischen Rundfunk, dem Süddeutschen Rundfunk und dem Südwestfunk (Quadriga)*. Frankfurt am Main: Fischer-Taschenbuch-Verl.

Klafki, W. (2007). *Neue Studien zur Bildungstheorie und Didaktik: zeitgemaße Allgemeinbildung und kritisch-konstruktive Didaktik* (6., neu ausgest. Aufl.). Weinheim: Beltz.

Krippendorff, K. (2004). *Content analysis an introduction to its methodology* (2nd ed.). Thousand Oaks [u.a.]: Sage Publ.

Kuckartz, U. (2007). *Einführung in die computergestützte Analyse qualitativer Daten* (2., aktualisierte und erw. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Leckie, G.J., Pettigrew, K.E. & Sylvain, C. (1996). Modeling the Information Seeking of Professionals: A General Model Derived from Research on Engineers, Health Care Professionals, and Lawyers. *Library Quarterly*, 66(2), 161-193.

Michel, L. P. (2008). *Digitale Schule – wie Lehrer Angebote im Internet nutzen*. Tübingen: MMB-Institut für Medien- und Kompetenzforschung. http://www.dlr.de/pt/Portaldata/45/Resources/dokumente/bildungsforschung/MMB_Veroeffentlichung_Lehrer_Online_ 20080505_final.pdf [12.11.2013].

Mundt, M., Stockert, K., & Yellesetty, L. (2006). The Information Behavior of Teachers. University of Washington Information School, Seattle. http://katestockert.com/LIS510_Information_Behavior_of_Teachers.pdf [13.4.2010]

Pettigrew, K.E., Fidel, R. & Bruce, H. (2001). Conceptual Frameworks in Information Behavior. *Annual Review of Information Science and Technology*, 35, 43–78.

Savolainen, R. (1995). Everyday life information seeking: Approaching information seeking in the context of "way of life." *Library & Information Science Research*, *17*(3), 259–294.

Schindler, C., & Botte, A. (2007). *Evaluation des Bildungsservers Hessen*. Frankfurt: Deutsches Institut für Internationale Pädagogische Forschung.

Tebrügge, A. (2001). *Unterrichtsplanung zwischen didaktischen Ansprüchen und alltäglicher Berufsanforderung. Eine empirische Studie zum Planungshandeln von Lehrerinnen und Lehrern in den Fächern Deutsch, Mathematik und Chemie*. Frankfurt a.M.: Lang.

Notizen

The Tweet and the City: Comparing Twitter Activities in Informational World Cities

Thorsten Förster, Lennart Lamerz, Agnes Mainka & Isabella Peters

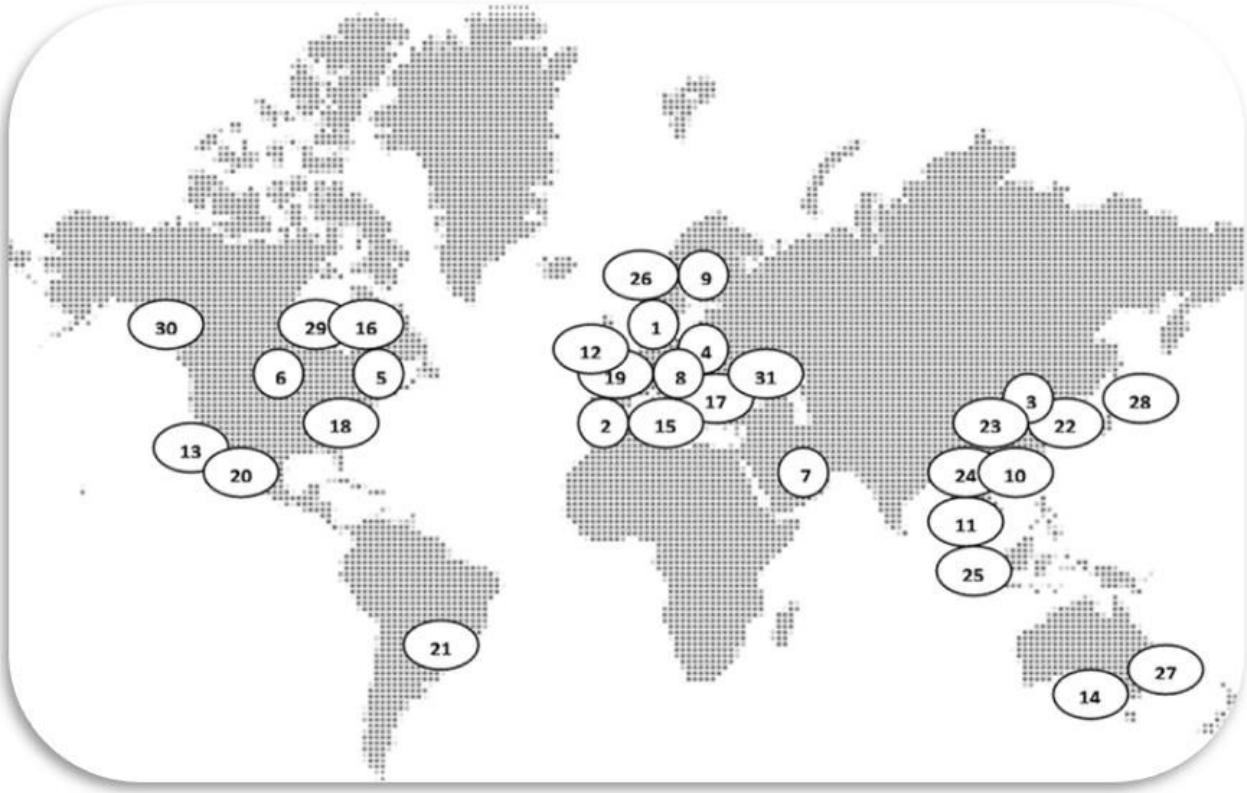
This paper informetrically monitors Twitter activities that are related to 31 Informational World Cities. It is a big data analysis of 18 million tweets that have been downloaded via Twitter's Search API (content-based approach) and Twitter's Streaming API (location-based approach). The Tweets have been filtered either by search terms (i. e. the city's name) or geo-locations (coordinates of a city). The analysis was made by mainly using quantitative statistic methods endorsed by several qualitative investigations. It shows that tweet activity related to Informational World Cities varies from city to city. A city's area or its size of population does not necessarily affect these activities. Factors like the penetration rate of smart phones, number of tourists etc. influences the amount of tweets that are produced in or about a city. Topics are mostly event-driven or related to sports and politics. City names are popular in spam tweets and they are often chained to draw the attention to messages which are not city-related at all (e. g., religious comments). The paper presents an approach for quantitatively analysing tweeting behaviour in Information World Cities to prospectively find distinct indicators of how Twitter activities in Informational World Cities can be classified and how they vary between the different cities.

Introduction

Social media has become a communication tool which is used by millions of people in their everyday live. From the beginning of the internet until today there has been a huge shift in using mass media. People, who traditionally consume mass media products like newspapers, radio, TV or even the internet, now are prosumers (Toffler, 1980), who consume and produce content for the masses (Ritzer and Jurgenson 2010). According to Bruns (2008) we are not talking about users who "use" and producers who "produce" anymore but rather about produsage. This is manifested in social media tools, like the microblogging platform Twitter, where it is not possible to separate users and producers anymore as almost every active user produces content, e. g. when commenting and retweeting other users' tweets (Java et al., 2007). As Informational World Cities are hubs in a global network (Castells, 1989; 1994) it could be assumed that their citizens, companies and others (e. g. tourists) who are located in one of those cities produce a high amount of social media content. Furthermore, people from other places can mention those cities and "talk" about city-focused topics. Therefore we analyse the tweets which mention a city and/or are produced in an Informational World City in terms of quantity, discussed topics, and relation to city-specific properties like size or population. City-related research has already been carried out for different World Cities which visualised tweet creation times and investigated user behaviour (Rios and Lin, 2013), compared the spread of topics on Twitter with that in traditional media (Zhao et al., 2011) or analysed Twitter streams for crisis communication in the 2011 South East Queensland floods (Bruns et al., 2012). The latter work showed that Twitter quickly links people and improves communication among them resulting in a strong connectedness of users. Accordingly, Twitter forms a network which is on the one hand developed in urban

spaces and on the other hand expanded in digital spaces by social communities (Hartley et al., 2012) both driven by a shared topic.

Research on Informational World Cities is currently a very active field to which the authors heavily contribute (amongst others:; Peters et al., 2013; Mainka et al., 2013a). Therefore it is based on the definition of Informational World Cities given by Mainka et al. (2013b). They are defined by two conditions: First, a city has to be called World City (according to Friedmann (1995), Taylor (2004), or Sassen (2001)) in the literature, and second, the city should be also referred to as digital (Yigitcanlar and Han, 2010), smart (Shapiro, 2006; Hollands, 2008), knowledge (Ergazakis et al., 2004), or creative (Landry, 2000; Florida, 2005) city; at least one of the secondly mentioned preconditions must be fulfilled. Advices for 31 cities to be recognized as Informational World Cities were found in the literature (Orszullok et al., 2012). These cities also reflect global centres distributed over the world (Figure 1).



1. Amsterdam	8. Frankfurt	14. Melbourne	20. San Francisco	26. Stockholm
2. Barcelona	9. Helsinki	15. Milan	21. Sao Paulo	27. Sydney
3. Beijing	10. Hong Kong	16. Montreal	22. Shenzhen	28. Tokyo
4. Berlin	11. Kuala Lumpur	17. Munich	23. Seoul	29. Toronto
5. Boston	12. London	18. New York	24. Shanghai	30. Vancouver
6. Chicago	13. Los Angeles	19. Paris	25. Singapore	31. Vienna
7. Dubai				

Figure 1: The Informational World Cities (Source: Mainka et al., 2013b, 298).

As previously mentioned Twitter can be a communication tool to connect people. Therefore it is interesting to investigate the behaviour of users tweeting about or tweeting in one of the 31 Informational World Cities. Unique or shared characteristics in these cities could be

compared to other cities. Since Twitter allows indexing of tweets with geo-coordinates (i.e. the location from where the user sent the tweet) there are several studies which analyse these geo-tags downloaded via the Twitter API. Takhteyev et al. (2012) examined the impact of geographic distance, national boundaries, language or frequency of air travel and analysed the formation of Twitter ties. However, Weidemann (2013) showed that only 6 percent (about 30 million) of Twitter-users broadcast their GPS-information. Instead of using geo-codes you can also follow a content-based approach to determine the location of users to estimate geographical centres (Cheng et al., 2010). While guided by following research questions we aim at contributing to the ongoing research and examine tweets related to the 31 Informational World Cities:

RQ1 How many tweets mention the city (content-based approach)?

RQ2 How many tweets are sent from within the city (geo-tag approach)?

RQ3 What relation does tweet activity have to population or area?

RQ4 Where do users originally come from when tweeting about or from within a city?

RQ5 Do users only tweet about the city where they are actually located?

RQ6 What are the main topics in tweets related to Informational World Cities?

Methods

To examine the tweet characteristics and the relationship between Twitter-usage and the predicate of being an Informational World City, we extracted tweets from Twitter, which have one or both of the following criteria:

1. One or more city names where mentioned in the tweet, with or without hashtag.
2. A tweet was tagged with geo-information from one of the defined squares representing the cities' areas.

At least one of these criteria is necessary to establish a valid connection between a tweet and a city. For downloading tweets, we used both the Search API and the Streaming API provided by Twitter. By using the Search API tweets were collected which contained one of the names of the Informational World Cities. In order to retrieve a corpus as exhaustive as possible, all cities were searched in the respective national language and in the nine most spoken languages using the Latin alphabet: Spanish, English, Portuguese, German, French, Italian, Polish, Romania and Dutch (Lewis, 2009). This is advisable since the APIs often have problems with non-Latin alphabets (cf. various discussions in Twitters developer fora). To overcome this Twitter-specific problem, the queries containing non-Latin symbols, like the query for Hong Kong, were divided into two queries:

1. *"Hong Kong" OR Hongkong OR #Hongkong*
2. *Hēunggóng OR #Hēunggóng OR Xiānggǎng OR #Xiānggǎng OR 香港 OR #香港*

This method assured that queries containing only Latin symbols were not affected by this problem. For example the query for Vienna has no non-Latin symbols so all search terms could be searched through one query (*Wien OR #Wien OR Viena OR #Viena OR Vienna OR #Vienna OR Vienne OR #Vienne OR Wenen OR #Wenen OR Wiedeń OR #Wiedeń*).

The restriction to Latin alphabet-based languages also allowed us to read and translate languages more easily. Considering that in countries without Latin alphabet-based languages many tweets are in English anyway – according to Semiocast (2011) about 40 percent of all tweets are written in English – the amount of tweets which were not collected although they contain one of the cities names should be within acceptable limits. In addition, some popular abbreviations like *#la* for Los Angeles and *#nyc* for New York City were included in queries, if a well-known abbreviation existed. Other city-specific nicknames like *#bigapple* or attractions like *#eiffeltower* have not been taken into account yet but might be included in further research.

All tweets were collected which matched our criteria and were published on Twitter during the 20 and 27 June 2013. Unfortunately, there were some technical issues with our data collecting automata running for each city, so we had to collect the tweets for Los Angeles, Sydney, Shenzhen, Munich and Singapore again between 19 July and 26 July 2013. It is obvious that a short time like a week is not sufficient to get fully robust data. For example there could have been, and it's very likely that there were, special events in some cities while in other cities special events might have been a week earlier or later. Those events could influence the amount of tweets produced in a city. Another influencing factor is the seasonal climate. The 31 cities are placed all over the world what makes it very likely that the seasonality in tourism influenced the main corpus of tweets (e.g. natural phenomena or institutional/social phenomena; Butler, 2001). Research in this area is always affected by those factors and should be considered during data interpretation. Altogether we downloaded 17,775,290 unique tweets which fall into two different corpora. The first contains all tweets found through the geo-location and the second contains all tweets found through the cities' names. Using the Streaming API tweets were retrieved which had a valid geo-location that matched with one of the rectangles we chose to represent each city's geo-location (see Figure 2). The sizes of the rectangles are based on the official sizes of the city areas in case an official declaration is provided. Otherwise we used the Google Maps definition of the city area. Table 1 shows the coordinates for all cities.

Figure 2: The rectangle for Milan as an example.

City	Geo coordinates
Amsterdam	4.76,52.26,5.00,52.41
Barcelona	2.07,41.28,2.27,41.47
Beijing	116.10,39.68,116.67,40.12
Berlin	13.05,52.33,13.72,52.66
Boston	-71.19,42.22,-70.92,42.39
Chicago	-87.96,41.62,-87.39,42.07
Dubai	55.02,24.92,55.55,25.33
Frankfurt	8.47,50.02,8.8,50.23
Helsinki	24.83,60.1,25.37,60.29
Hong Kong	113.84,22.15,114.44,22.56
Kuala Lumpur	101.52,2.88,101.76,3.24
London	-0.56,51.26,0.28,51.68
Los Angeles	-118.42,33.72,-117.86,34.12
Melbourne	144.80,-37.87,145.11,-37.75
Milan	9.06,45.4,9.34,45.55
Montréal	-73.97,45.41,-73.47,45.70
Munich	11.33,48.06,11.71,48.25
New York City	-74,40,-73,41
Paris	2.08,48.65,2.63,49.04
San Francisco	-122.75,36.8,-121.75,37.8
São Paulo	-46.80,-23.78,-46.36,-23.39
Seoul	126.82,37.46,127.13,37.64
Shanghai	121.10,30.97,121.80,31.51
Shenzhen	113.76,22.44,114.6,22.86
Singapur	103.58,1.19,104.08,1.45
Stockholm	17.77,59.23,18.2,59.44
Sydney	150.15,-34.08,151.34,-33.55
Tokyo	139.36,35.41,139.92,35.82
Toronto	-79.63,43.58,-79.11,43.85
Vancouver	-123.26,49.19,-123.02,49.31
Vienna	16.18,48.08,16.55,48.31

Table 1: The coordinates for every city defining the rectangles we used to find geo-located tweets.

The first two values define the left lower corner of the rectangle the second two define the right upper corner (Figure 2). The use of rectangles to define locations is forced by the Twitter-API itself and supports searching for tweets published within a defined geographic area. Contrary to the use of the Search API this search is limited to the tweets which have a valid geo-location i.e. that the user has explicitly allowed for sending his geo-coordinates to Twitter along with the tweet he publishes.

In order to enhance the comparability, the data can be relativized by using the size of the employed rectangles or by using the residential population of each city. The population of a city could only be determined through official statistics. Unfortunately, the standards for

surveying the statistics differ from country to country and influence the results of our analysis. Eurostat mentions in its Reference Guide for the European Regional and Urban Statistics (2010) that the "'Administrative City' does not always yield comparable spatial units". Paris for example has an official resident population from only about 2.2 million, while London has an official resident population from over 8 million since Paris counts only the population of the inner core, while London also counts the population of the urban area called Greater London. Similar problems occur with statistics about tourism. However, statistics like residential population and tourism help determining the reasons why the extent of tweeting-activities in Informational World Cities is different.

The relational database model used to save the data is shown in Figure 3. All queries are contained in the table *query*. In a first step, all data from the APIs were saved in the table *json_data* as JSON-Strings. These strings contain not only the tweet-text itself but all additional information Twitter saves like the geo-location from where the tweet was sent, the name of the user who wrote the tweet, a unique numeric identifier for the tweet, etc. The collected data were split up into tables like *tweet, url, user, hashtag*, etc. In order to preserve the information which tweet has which author and which hashtags etc. the tables are connected through relational tables like *tweet_hashtag* which link a tweet to all hashtags included in that tweet. This approach is similar to the simple model used in the architecture for Twitter data collection described by Oussalah, Bhat, Challis and Schnier (2013).

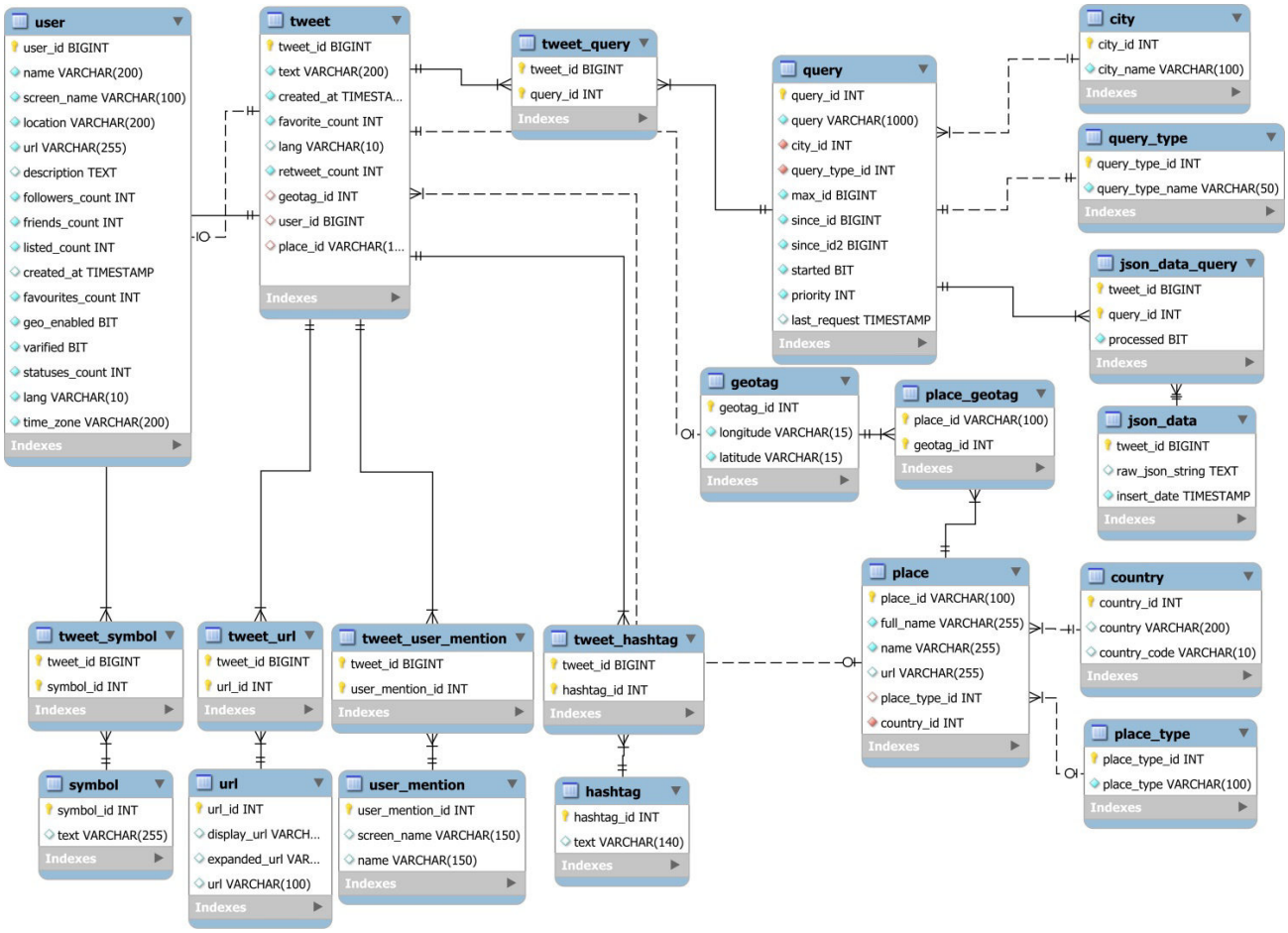


Figure 3: The database model.

Data description

The database contains over 18 million tweets that were found using the different queries for the two APIs mentioned before and therefore has duplicates. For example, when using Twitter's Search API, a tweet is found via the search term "New York" when it contains the string "New York City". If this tweet with the unique ID provided by Twitter does not exist in the database, a new record is stored and the query that matched will be connected with it. If the tweet is found again because it additionally contains *#paris*, only the connection between the query for Paris and the tweet will be stored in the database and the record, containing the tweet data, will not be stored twice. In addition it is possible that this tweet has geo-information assigned and its coordinates are located within the rectangle of Milan (see Figure 2). So this tweet is found a third time but this time via the Streaming API and a new connection between tweet and query will be stored. Hence, there is a difference between counting how often a tweet was found using any of the APIs and counting the unique tweets stored in the database. Figure 4 shows the number of tweets the corpus contains segmented into two categories with three subcategories each. The subcategory *Search API* represents all tweets that are only found using Twitter's Search API. So if our example tweet was found using the term "New York" and a second time using the term "Paris" it would fall into this category. As a result it increases the number of unique tweets found only via the Search API by one and the number of all tweets containing duplicates by two because it has been found using the queries for New York City and Paris. But since our example tweet has been found twice using the Search API and once using the Streaming API we have only one unique tweet but three hits for the category of all tweets containing duplicates that have been found in both APIs. As Figure 4 shows there is a difference of 423,169 tweets between unique and duplicate tweets. The values of the unique and duplicate tweets found by the Streaming API are identical, because a tweet cannot be tagged with two different geo-locations. Comparing the two query methods used to create the corpus it is obvious that more tweets were found using search terms than using geo-locations which is due to users not specifying their locations. About 56 % of all (not unique) tweets were found using the Search API, 42 % using the Streaming API and 2 % using both.

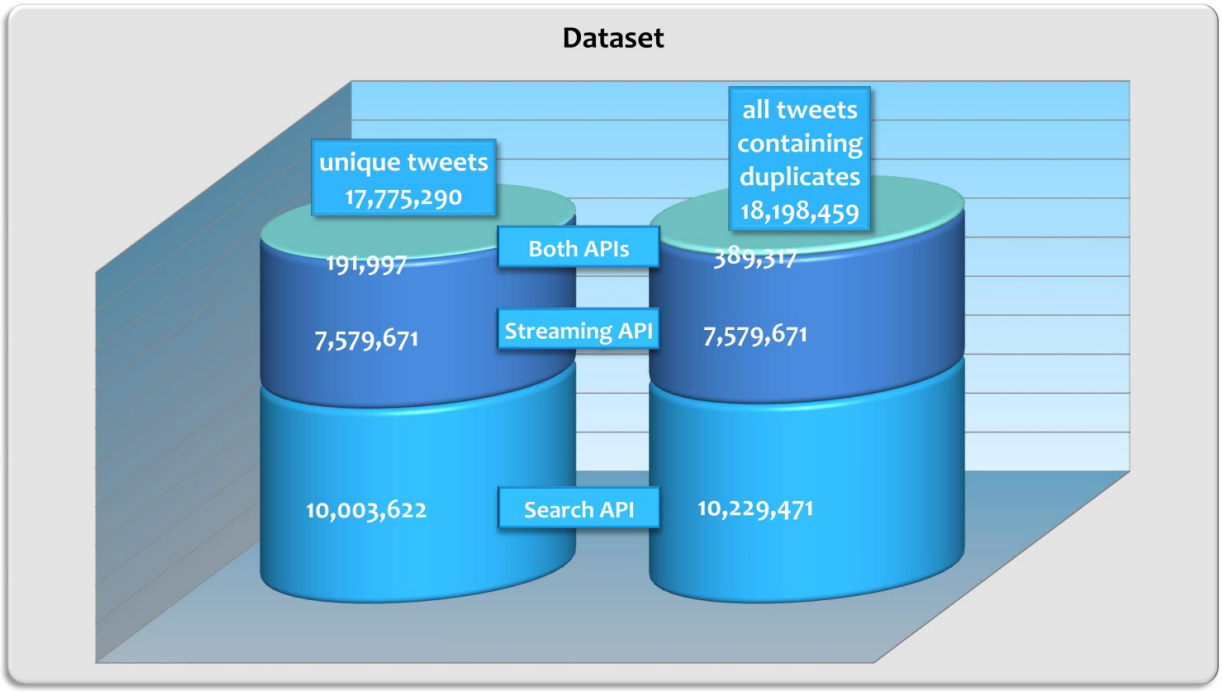


Figure 4: The database contains tweets fount via Twitter's Streaming API and Search API.

Results

In this section we visualise and interpret the results of our analysis by dividing it into sub-sections that relate to the research questions we asked in our introduction.

Location of users and tweet content

We begin our analyses by comparing the number of users who tweeted to the actual number of unique tweets stored in the database. If we count the unique users, we arrive at a number of 4,911,318 unique users that tweeted tweets matching either a search term or a location query. That would mean that every user in our corpus created 3.6 city-related tweets on average. Figure 5 shows the top 25 users in terms of number of tweets created in the corpus. Typically we find a skewed distribution with few users producing an exceptional large amount of tweets. A detailed investigation of the tweets produced by the users revealed that some of the accounts are not real users but rather spam accounts used to advertise. A typical tweet for user *@JulienLanger* is:

"#fashion #marketing Concrete Chunks from Patio (90065 GLASSELL PARK EAGLE ROCK) http://t.co/7LP8dogXAK #socialmedia #movies".

Even the link is not working (anymore). In the meantime this account has probably been identified as a spam account and closed from Twitter. Our corpus contains tweets of users that do not exist anymore since we only have a static snapshot of tweets published in the period described before.

Not all user accounts in the top 25 are spam accounts. The user on the second place is the official account of an FC Barcelona fan club which uses terms like *Barcelona* very often. Also almost every tweet of this user is a retweet of tweets produced by other fans. This is substantiated by the study of Weller and Bruns (2013) who figured out that sport clubs (in that case German) use Twitter as a marketing and fan communication tool. In addition there are accounts like *@q8_b7*, *@bah_sa*, *@ba7_lions*, *@m7_galia* or *@h_llah* that are originally situated in the United Arab Emirates which post religious messages and use hashtags from other cities (i.e. *#NewYork*, *#paris* etc.). Those accounts may try to reach a larger audience for the messages that should be transported. A sample check revealed that some accounts only tweet about the city or region where they are situated (i.e. *@Toronto_Follow*, *@CRECanada*) or they are the accounts of local radio stations or news agencies (i.e. *@soundloop*, *@JapanNewsTwo4*). In Figure 5 we can see the location of the user besides the user name (when the location was specified) and the city the user tweets about most besides the number of tweets. That does not mean that the specific user only tweets about this city. For example *@JapanNewsTwo4* created tweets about all 31 cities we were looking for. An example tweet of this user looks like this:

#world#london#paris#moscow#milano#tokyo#mumbai#munich#madrid#barcelona#spain#germany#dubai#indonesia#malaysia#kualalumpur#milano#italy#qa...

This again looks like a spam tweet, but this account has not been closed yet. Comparing the specified location in a user's profile and the city the user tweets about we can see that in most cases these two values match. For example *@tebakanlagu* specified Stockholm as his location which is also the city he tweets about most. Some users only give the country where they are located like *@CRECanada* who is tweeting about Vancouver. But the most common case is that a user does not indicate any location in his profile. So we have to rely on what users specify in their profile without knowing whether the information is correct.

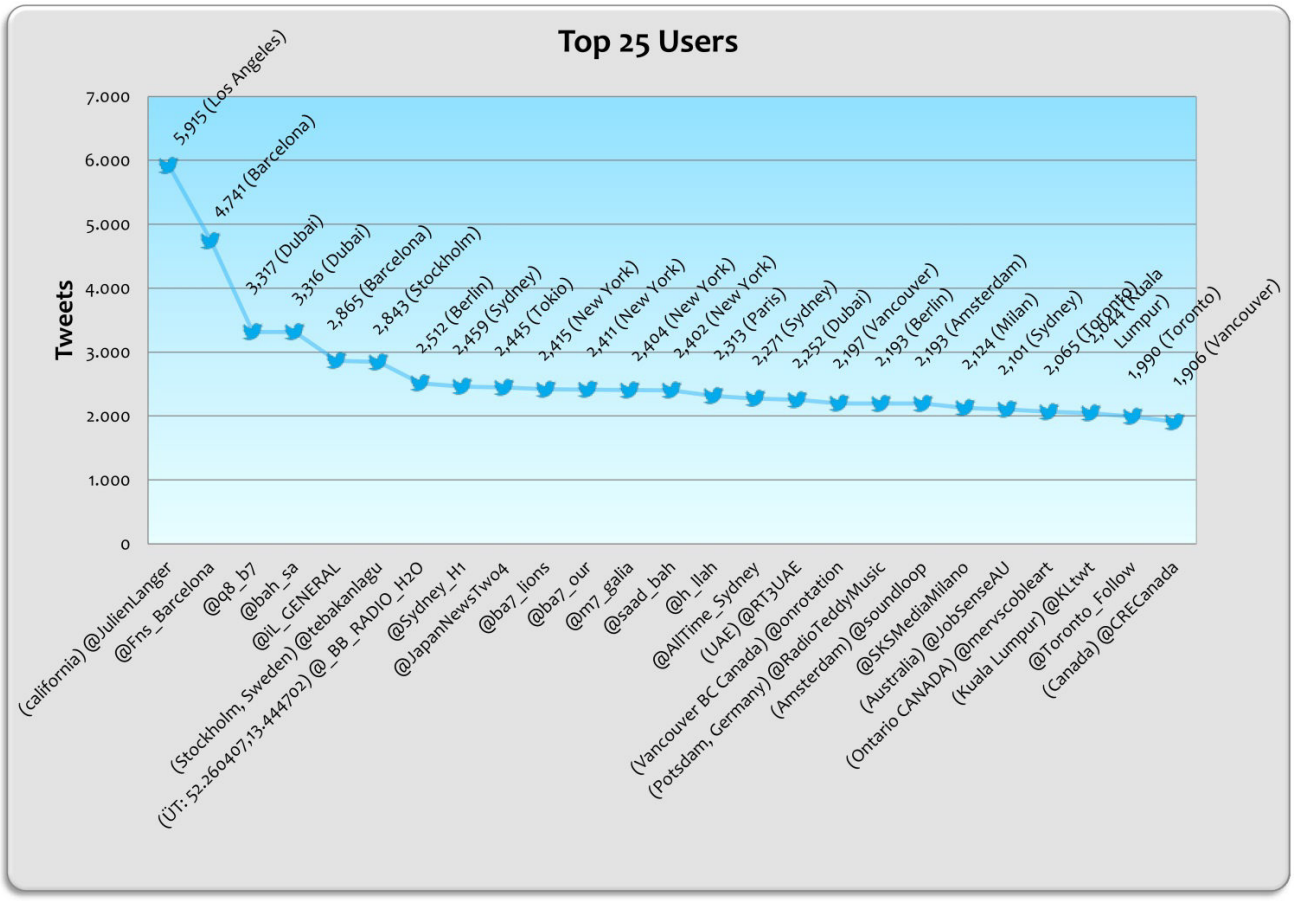


Figure 5: Top 25 users with locations and their top-mentioned cities.

Extent to which tweets mention or are sent from within a city

Table 2 shows the absolute numbers of tweets for every city. It is divided into two groups: Tweets found by searching for the city's name (collected through search terms) and tweets found by searching for a geo-location. As one can see, Paris has the highest number of location-collected tweets, followed by São Paulo, Kuala Lumpur and New York City. Regarding the tweets collected through search terms, Tokyo is in lead, followed by New York City, London and Paris. It seems, that Tokyo profits from its great number of inhabitants (ca. 13 million) which is the same for New York and London (both over 8 million inhabitants).

City	Search Term	Location
Tokyo	1260790	477929
New York City	836934	785768
London - Greater London	816654	763315
Paris	774241	1419325
Boston	613278	156601
Chicago	559100	416579
Barcelona	515876	80188
Sydney	483648	41976
Milan	436341	65039
Seoul	406454	91890
Dubai	373268	110391
Hong Kong	348802	42622
Munich	331425	8910
Los Angeles	311464	429305
Toronto	311044	163759
São Paulo	272015	1029477
Singapore	252153	115957
Amsterdam	200679	50599
Berlin	191511	49797
San Francisco	165562	333234
Shanghai	152716	9454
Melbourne	148130	37107
Vancouver	142887	29562
Beijing	120620	9543
Kuala Lumpur	111942	898722
Vienna	98066	20878
Montréal	94718	68730
Frankfurt	38737	12038
Stockholm	37461	44323
Helsinki	17585	8363
Shenzhen	2690	287

Table 2: The amounts of tweets found for every city.

Since there are other cities with high residential population and much less tweets in which they are mentioned, for example São Paulo or the Chinese cities, there must be other influencing factors. First, the relatively small amount of tweets that are connected to the Chinese cities is due to the unavailability of Twitter and the extended usage of the Chinese Twitter-clone Weibo (Eichstädt and Wei, 2013, 108). Second, popular people, like Paris Hilton, with many followers might increase the amount of tweets mentioning Paris (770,000 tweets contain Paris), although this assumption could not be proved here. Indeed, there are only 16,805 tweets (2.17 %) which were found in our data set due to containing Paris and *"hilton"*. Thus, the impact of false hits for Paris is rather small, possibly because tweets get only "connected" to a person by using the "@" message, i.e. *@parishilton*. But tweets with @-terms were not collected by the implemented bots. Anyway homonymy remains problematically having in mind that not only persons could be named like cities but that different cities can bear the same name (e.g. Sydney, MT, USA and Sydney, AUS). This issue has not been focussed for the underlying quantitative analysis of this paper. Third, tourists could be a factor that influences the statistics shown in Table 2. The four top-mentioned cities are all popular destinations for urban tourism. New York City (over 10 million), London (over 15 million), and Paris (over 8 million) belong to the cities in our set which had the most international visitors in 2011 (Euromonitor, 2013). Tokyo had only about 2.7 million international visitors, but it has had always a strong domestic tourism (430 million domestic visitors in 2008).

There are some immense differences between the amounts of tweets collected through searching for geo-locations and collected through search terms. Especially São Paulo and Kuala Lumpur have a very high number of tweets sent from within the city and a rather small amount of tweets containing the city's name. On the contrary, the three German cities, Berlin, Frankfurt and Munich have only few tweets with geo-locations. When trying to explain this phenomenon we can only rely on hypotheses. Possible impact could be based in different levels of awareness of data privacy protection in different countries – e.g. in Germany private data on the internet is a highly sensitive subject which is discussed quite often in the general public (see Unisys (2013) for general privacy concerns in Germany regarding the internet). In fact, the least tweets with a geo-location were sent from Munich. Only 2.61 % of all tweets that are linked with Munich in both corpora are coming from the corpus that contains the tweets with geo-locations. Another influencing factor could be the penetration rate of smartphones in the different countries. Obviously, it rather makes sense to send tweets with geo-locations when the user is moving and uses a mobile device like a smartphone for tweeting and informing followers where he is. According to Rowinski (2013) there are about 70 million smartphone users in Brazil and only 32 million in Germany. However, the rate of smartphone users in France (33 million) is nearly the same as in Germany but the most tweets with geo-locations have been sent from Paris. One could speculate again that tourism has an impact on these statistics. To fully understand these characteristics a deeper subject analysis of the tweets as well as a user survey is needed.

The impacts of events hosted in a city during the period of data collection can be particularly shown by the extensive use of hashtags combined with city names. In the data for Amsterdam *#rihanna* is the hashtag appearing most often which is most likely because the pop singer Rihanna had a concert on June 23rd in Amsterdam. Another popular hashtag is *#ziggo* referring to the location the concert took place.

The relation between tweet activity and size of population and area

In order to get more comparable data the absolute number of tweets found through the search for geo-locations are relativized by the size of the rectangle used to match the geo-tagged tweets per city in square kilometres. The results are shown in Table 3. Looking at the results the first cites are nearly the same, except for London replacing New York City, but in different order: Kuala Lumpur before Paris before London before São Paulo. After that there is a great gap to Boston (356 Tweets per km²). Altogether, places with a high population density like Barcelona and cities with small city centres like Vancouver profit from this relativization. The results show that the size is not alone a factor that explains the different amounts of tweets we retrieved for the various cities. The top five cities in Table 3 are all neither very small (the rectangle for Paris has over 2,000 km²) nor are they very large (the rectangle for Boston has only 460 km²) but the numbers differ extremely and range from 1,084 tweets to less than seven, if the Chinese cities would be ignored. If the size of the areas were a leading factor for the amount of geo-located tweets per city, the difference between these numbers would have been much smaller. We could also ask for the relation between residential population statistics and our statistics about tweets found through city names. The amount of tweets found through search terms per one hundred inhabitants (Table 3) shows a quite different ranking as in Table 2. With 993 tweets per 1,000 inhabitants Boston is on top of the ranking. Second ranked Paris has not only a far smaller value but also, as discussed before, profits from only about 2.2 million official inhabitants living in the city's inner core. If this ranking is compared to the absolute ranking (Table 1), one can see that the top three cities in absolute numbers (Tokyo, New York, London) are not the top cities anymore,

because of their large population. Other cities with large residential populations, like Seoul or São Paulo, do not to profit since they are less often mentioned in tweets. Vice versa cities with small populations, like Helsinki or Frankfurt, are not automatically high listed in the ranking.

City (Size of the rectangle in km², residential population)	Tweets found through geo location per km²	Tweets found through search terms per 1000 inhabitants
Kuala Lumpur (864, 1.887.674)	1084	59
Paris (2.145, 2.257.981)	679	343
London - Greater London (1.176, 7.400.000)	677	110
São Paulo (1.716, 11.253.503)	615	24
Boston (459, 617.594)	356	993
Barcelona (380, 1.615.448)	220	319
Tokyo (2.296, 13.189.000)	215	93
Los Angeles (2.240, 3.863.839)	192	104
Chicago (2.565, 2.695.598)	170	207
Seoul (558, 10.442.426)	170	39
Milan (420, 1.367.733)	161	36
Amsterdam (360, 820.500)	147	245
Toronto (1.404, 2.744.000)	120	113
Vancouver (288, 2.313.328)	106	62
Melbourne (372, 4.169.103)	104	319
Singapore (1.300, 3.818.200)	90	101
New York City (10.000, 8.336.697)	81	100
Dubai (2.173, 1.870.000)	53	197
Stockholm (903, 863.100)	51	43
Montréal (1.450, 1.613.260)	49	59
San Francisco (10.000, 825.111)	34	201
Vienna (851, 1.757.353)	25	56
Berlin (2.211, 3.292.400)	23	58
Frankfurt (693, 682.854)	18	54
Hong Kong (2.460, 7.173.900)	18	48
Munich (722, 1.447.614)	14	90
Helsinki (1.026, 603.968)	9	29
Sydney (6.307, 4.605.992)	7	84
Beijing (2.508, 17.400.000)	4	7
Shanghai (3.780, 23.470.000)	3	6
Shenzhen (3.528, 10.470.000)	0	0

Table 1: The amount of tweets denominated through the size of the defined areas respectively the residential population[1]

The relationship between geo-locations and city mentions

The collected data allows counting both how many tweets were sent from a location in a city and how many mention this city's name in the tweet-text as well. Table 4 shows the overlap of the two sets ("found through search term" and "found through geo-location") for each city. Paris has surprisingly few tweets (only 7,531) which were sent from Paris and contain one or more of the search terms belonging to Paris.

As Table 4 shows, only about 0.34% of all tweets belonging to Paris. The highest percentage of tweets in the overlapping area of the two sets relative to the overall amount of tweets belonging to the city, are given for Frankfurt (2.8%), Singapore (2.7%) and Helsinki (2.5%). All three cities are rarely mentioned on Twitter. In the ranking in which every tweet is assigned to the city it comes from or contains the city's name, Helsinki is on rank 30, Frankfurt is on 29 and Singapore is on 18 of 31 cities. Altogether, the relative numbers are very low. It seems that people who send their location to Twitter do not waste space in the tweet-text to describe in which city they are.

[1] The residential populations were retrieved from official websites. The numbers are as up-to-date as possible. Most represent the population number for 2010 till 2012.

City	Overlap search term and geo location	Number of all tweets linked to the city
Tokyo	20412	1733253
São Paulo	15179	1312107
New York City	13370	1629775
Kuala Lumpur	11626	1036942
London - Greater London	11465	1601618
Chicago	10388	984149
Singapore	9767	359492
Dubai	7639	480861
Paris	7531	2223166
Los Angeles	6680	735237
San Francisco	6596	501950
Toronto	6247	472797
Boston	5662	771075
Amsterdam	4680	248763
Barcelona	4539	594960
Berlin	4484	238965
Melbourne	3248	500877
Hong Kong	2848	390324
Seoul	2556	498739
Montréal	2476	162745
Sydney	2443	524421
Vancouver	1921	171364
Milan	1670	185238
Munich	1483	340001
Frankfurt	1403	49898
Shanghai	1221	161260
Vienna	1128	118399
Beijing	953	129631
Stockholm	890	82199
Helsinki	644	25717
Shenzhen	49	4077

Table 4: Number of all tweets per city found through search term and geo-location.

The main topics of tweets represented by hashtags

To determine the main topics of the tweets that were collected, using either the Search API or the Streaming API, we investigated the hashtags in the tweets related to every city. Hashtags represent a method how users can index their tweets. We filtered out hashtags which are built from city names since they were already used to find those tweets and would automatically be the hashtags with the highest frequency in our corpus. As shown in Figure 6 the hashtags are ranked according to their frequency of occurrence which is also visualised by font size. For some cities the main topics in the collection period were specific events like concerts as it is shown in the hashtag cloud for Amsterdam:

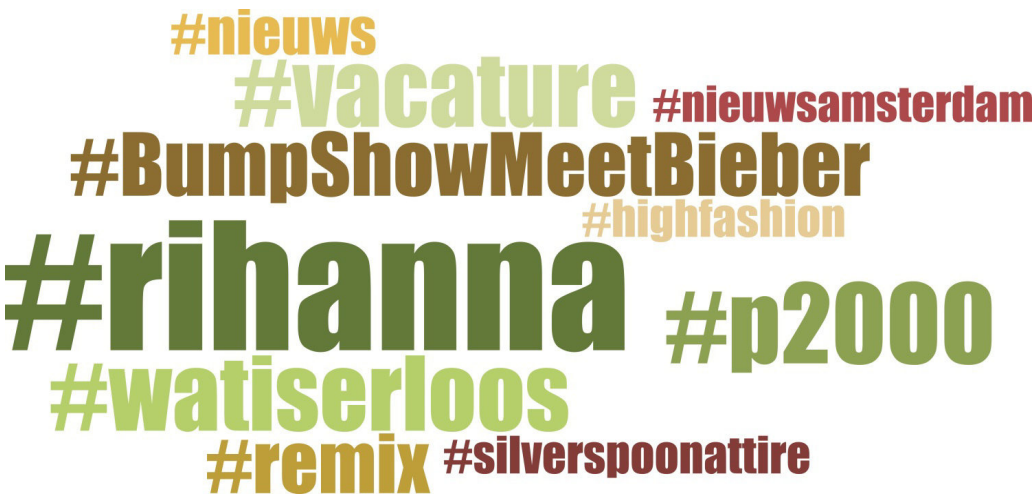


Figure 6: Tag cloud for Amsterdam.

The hashtag *#rihanna* has the highest frequency (5,192) in the Amsterdam related tweets. There are also very common hashtags like *#nieuws* ("news") that occur in the top list of hashtag for the other cities and in other languages as well. The hashtag *#BumpShowMeetBieber* is very popular in different variants since a radio show has been promoting it (see Figure 7). It occurs in the top ten lists of Amsterdam, Frankfurt, Melbourne and Vienna. The top ten list of hashtags for Frankfurt am Main (Figure 7) is even dominated by this hashtag:

Figure 7: Tag cloud for Frankfurt am Main.

Here, another trend can be detected. Many hashtags are related to sports, specifically sport clubs or sport events. The hashtag*#Eintracht* relates to the soccer club Eintracht Frankfurt and the hashtag *#Bundesliga* relates to the highest German soccer league. Also other sport related hashtags are very popular (i.e. *#nba*, *#nhl, #nfl, #blackhaws*or *#sports*). This hashtag cloud relates more Frankfurt-specific hashtags like *#Börse* which is the German Stock Exchange that is located in Frankfurt am Main and *#Blockupy* which relates to the anti-capitalistic movement protesting against it.

The hashtag analysis also revealed that the tweets of some cities contain hashtags about other cities. A perfect example is the top ten hashtag list of Paris, where seven hashtags relate to other cities and three of them relate to New York City (Figure 8):

Figure 8: Tag cloud for Paris.

Further investigations of tweet-texts and hashtags showed that some people use those hashtags, when they plan to visit the city. Paris, London and New York City are often mentioned together in tweets concerned with fashion. There are also many spam tweets that simply chain hashtags of different cities to reach a greater audience with their advertisements:

"KING D ROYALTY http://t.co/mqI6TZt3ST … #ATLANTA #ATL #MIAMI #LIKE #RETWEET #SHARE #DJS #DJ #LONDON #PARIS #NYC #NY #DC #VA #PA #NEWYORK 36"

This tweet promotes a music video on YouTube by using that many hashtags to reach as many users as possible. The account has already been closed by Twitter. Such tweets show the limitations of content-based tweet research and have to be filtered out for further investigations, because they do not represent the average user and may distort some quantitative statistics by unrealistically increasing the received values from the dataset. Of course not every tweet containing more than one hashtag about a city (i.e. #nyc and #paris) can be classified as a spam tweet. Therefore a simple automatic filter would cause a high false positive rate. So to achieve an adequate result an algorithm based on machine learning techniques has to be used (cf. Miller et al., 2014). Because of its complexity this will be part of the next research steps related to this topic.

Discussion

Our first two research questions referred to the quantitative amounts of tweets in our data set. Table 2 visualizes the great differences between Informational World Cities. Looking at the different numbers of tweets produced about or in a city we tried to determine some influencing factors, which are the size of the residential population or tourism, the penetration rate of mobile devices, or the importance of privacy protection in the public view. It could be shown that the factors have not equally influenced the cities. Regarding the third research question the analysis of our dataset revealed that some cities like Tokyo, London, New York with both high numbers in population and strong tourism equals high amount of tweets (see Table 2). On the other hand, cities like Seoul or the Chinese cities

perform completely different, and although Kuala Lumpur or Chicago are far smaller they have a relatively high number of tweets. There is also a difference between the amounts of tweets found through search terms and found through geo-location for some cities. The penetration rate of mobile devices and the importance of privacy protection are possible explanations for this discrepancy, but further answers can only be given by a content analysis of the tweets.

As a first result (RQ4 & RQ5) the data shows a skewed distribution where only few users produce a large amount of tweets, but many users only tweet occasionally. This phenomenon is well-known from other social media tools and Web 2.0 platforms (Stock, 2008). Also, the analyses revealed problems in locating the origin of a user to investigate whether he produces tweets about the places where he lives, because every user can specify what he wants in his profile. In addition, spam accounts can distort the statistics and have to be filtered out in future research to provide more significant data.

Regarding RQ6 we analysed hashtags in our dataset. The most frequent hashtags across cities are often related to specific events, political circumstances, sport or fan clubs, or viral promotion campaigns started during our investigation period. Although hashtags representing the searched city's name had to be deleted from our analysis, we found some indicators of connectedness, i.e. several cities often mentioned together in tweets. For example the tag cloud for Paris (Figure 8) contains several other city names like New York or London.

When using Twitter in scientific analyses we encountered several pitfalls. First of all we had to rely on data Twitter provides. For example, there can be technical problems on's side which could not even be noticed (Weller & Bruns, 2013). In addition, server problems on the client side can cause data loss as mentioned before. Collected tweets can only be a snapshot from the whole data volume. Because of the high dynamics of social media networks the snapshots, taken at different moments in time, may lead to different results during analyses. Therefore, it is difficult to formulate generalised assertions and prognoses or to repeat the study. But in retrospection a comprehensive investigation of this particular snapshot can be made and conclusions on what factors influencing these dynamics can be drawn.

Conclusion and future research

As a first approach we analysed our dataset in a pure quantitative way in order to test the suitability of tweets for Informational World City research. The results already provide some interesting facts and thought-provoking impulses for both fields of research, Twitter analysis and Informational World Cities. However, there still is a bunch of open research questions for future work which should include the revision of the corpus and delete spam. The study will also benefit from a thorough content-based analysis which will support drawing more general conclusions. Another field that has to be analysed is the conversations that may even happen between users from different cities. In addition, a detailed analysis of given geo-information in tweets and/or user profiles may reveal clusters of regions that are producing many tweets about a specific city. Here we hope to find further connections between the 31 Informational World Cities. To come to adequate results in this field it is mandatory to create a detailed analysis for every city including all influencing factors and key figures (i.e. followers, retweets, favourites etc.) being used for social media content. By this we hope to find distinct indicators of how Twitter activities in Informational World Cities can be classified and how they vary between the different cities.

References

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and Beyond. From Production to Produsage.* New York: Peter Lang.

Bruns, A., Burgess, J., Crawford, K., & Shaw, F. (2012). .*#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods.* Brisbane: ARC Centre of Excellence for Creative Industries and Innovation.

Butler, R. (2001). Seasonality in tourism: issues and implications. *Seasonality in tourism*, pp. 5-22.

Castells. (1994). European cities, the informational society, and the global economy. *New Left Review, 204*, pp. 18-32.

Castells, M. (1989). *The Informational City. Information Technology, Economic Restructuring, and the Urban-Regional Process.* Oxford, UK: Basil Blackwell.

Cheng, Z., Caverlee, J.& Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759-768.

Eichstädt, B.& Wei, C. (2013). Im Social Web der Mitte. In R. Leinemann (Ed.), *Social Media* (pp. 107-110). Berlin Heidelberg: Springer.

Euromonitor. (2013). *Top 100 cities destination ranking*. Retrieved August 23, 2013, from http://blog.euromonitor.com/2013/01/top-100-cities-destination-ranking.html

Eurostat. (2010). European Regional and Urban Statistics. Luxembourg: Publications Office of the European Union.

Hartley, J., Potts, J., MacDonald, T., Erkunt, C.& Kufleitner, C. (2012). *CCI-CCI Creative City Index 2012.* Queensland University of Technology: ARC Centre of Excellence for Creative Industries and Innovation.

Java, A., Song, X., Finin, T. & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis at ACM SIGKDD, San Jose, CA* (S. 56-65). New York: ACM.

Lewis, M. P. (2009). *Ethnologue: Languages of the World, Sixteenth edition. Dallas, Texas: SIL International.* (M. P. Lewis, Ed.) Retrieved from http://www.ethnologue.com/16

Mainka, A.& Khveshchanka, S. (2012). Digital Libraries as Knowledge Hubs in Informational Cities. *Libraries in the Digital Age (LIDA) Proceedings Vol. 12, 18 - 22 June 2012.* Zadar, Croatia: University of Zadar.

Mainka, A., Fietkiewicz, K., Kosior, A., Pyka, S.& Stock, W. G. (2013a). Maturity and usability of e-government in informational world cities. In F. E, & C. W, *Proceedings of the 13th European Conference on e-Government. University of Insubria Varese, Italy, 13-14 June 2013* (pp. 292-300). Reading, UK: Academic Conferences and Publishing International (ACPI).

Mainka, A., Hartman, S., Orszullok, L., Peters, I., Stallmann, A. & Stock, W. G. (2013b). Public Libraries in Informational World Cities. *Libri, 63*(4), 295-319.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W.&Wang, H. A. (2014). Twitter spammer detection using data stream clustering. *Information Sciences 260,* 64-73.

Oussalah, M., Bhat, F., Challis, K.& Schnier, T. (2013). A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems (37)*, pp. 105-120.

Peters, I., Hartmann, S.& Mainka, A. (2013). Social Media Use and Outreach of Selected Public Libraries in Informational World Cities. *Proceedings of the Second Association for Information Science and Technology ASIS&T European Workshop* (pp. 79-93). Åbo/Turku, Finland: Skrifter utgivna av Informationsvetenskap vid Åbo Akademi Vol. 2. Åbo: Åbo Akademi University.

Rios, M. & Lin, J. (2013). Visualizing the "Pulse" of World Cities on Twitter. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Ritzer, G. & Jurgenson, N. (2010). Production, Consumption, Prosumption: The nature of capitalism in the age of the digital "prosumer". *Journal of Consumer Culture*, pp. 13-36.

Rowinski. (2013, May 29). *Think Mobile Is Big Now? Here's Proof That It's Just Getting Started.* Retrieved August 23, 2013, from http://readwrite.com/2013/05/29/huge-potential-only-15-of-global-internet-traffic-is-mobile#awesm=~ofKwdMIxymYk5E

Semiocast. (2011, November 24). *Arabic highest growth on Twitter.* http://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter [23.08.2013]

Stock, W.& Stock, M. (2008). *Wissensrepräsentation.* München: Oldenbourg Verlag.

Takhteyev, Y., Gruzd, A.& Wellman, B. (2012). Geography of Twitter networks. *Social Networks 34*, pp. 73-81.

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

Unisys. (2013). *Unisys Security Index*. http://www.unisyssecurityindex.com/usi/germany [23.08.2013]

Weidemann, C. (2013). Social Media Location Intelligence: The Next Privacy Battle - An ArcGIS add-in and Analysis of Geospatial Data Collected from Twitter.com. *International Journal of Geoinformatics, 9(2)*, pp. 21-27.

Weller, K.& Bruns, A. (2013). Das Spiel dauert 140 Zeichen: Wie deutsche Fußballvereine Twitter für Marketing und Fan-Kommunikation entdecken. In J. Griesbaum, B. Heuwing, J. Ruppenhofer & K. Werner (Eds.), *HiER 2013. Proceedings des 8. Hildesheimer Evaluierungs- und Retrievalworkshop* (pp. 119-131). Hildesheim: Universität Hildesheim.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, Ee-Peng, . . . Li, X. (2011). Comparing twitter and traditional media using topic models. *ECIR'11 Proceedings of the 33rd European conference on Advances in information retrieval*, pp. 338-349.

Enterprise 2.0 in Bibliotheken und Informationseinrichtungen – Neue Anforderungen an Wissensmanagement in Zeiten des Social Web

Sarah Baldin & Beate Renate Somorowsky

Der Beitrag befasst sich mit dem Modell Enterprise 2.0 und seiner Adaption in Bibliotheken und Informationseinrichtungen. Das Augenmerk liegt auf innerbetrieblichen Rahmenbedingungen für einen Erfolg versprechenden Einsatz von Web 2.0 für Wissensgenerierung und Wissensmanagement. Derartige Bedingungen werden vordergründig in Personalmanagement und Organisationsentwicklung lokalisiert. Der Begriff Enterprise 2.0 wird als ein umfassender konzeptueller Gesamtansatz vorgestellt, welcher die Prinzipien des Social Web auf zentrale Bereiche einer Organisation überträgt. Enterprise 2.0 geht somit essentiell über die rein technische Einführung von Social-Web-Anwendungen in Bibliotheken und Informationseinrichtungen hinaus und stellt eine weit reichende, soziotechnische und soziokulturelle Innovation und Veränderung dar.

Vorbemerkung

Der Aufsatz basiert auf der Masterthesis „Enterprise 2.0 in Bibliotheken − neue Anforderungen an Personalmanagement und Organisationsentwicklung in Zeiten des Social Web“.[1] Diese Thematik wurde bislang in der Fachliteratur nicht hinlänglich beleuchtet.[2] Die Aspekte Business Information Quality sowie organisationale Wissensgenerierung werden in der Masterarbeit insbesondere im Kapitel „Dimension Organisation und Strukturen“ untersucht. Die nachfolgenden Ausführungen stellen zentrale Ergebnisse dieses Kapitels in komprimierter Form dar. Hierbei wird der Fokus der Masterthesis, der sich auf die potentiellen Auswirkungen auf Personalmanagement und Organisationsentwicklung in Bibliotheken und Informationseinrichtungen richtet, wiedergegeben.

Einführung

Das Konzept Enterprise 2.0 (E 2.0) ist seit einigen Jahren in Wissenschaft und Fachliteratur, aber auch in der wirtschaftlichen Praxis angekommen – für das Bibliotheks- und Informationswesen lässt sich dies noch nicht feststellen. Dabei ist das soziale Web (Social Web) inzwischen sowohl in den privaten Bereich als auch in den beruflichen Kontext vorgedrungen. Offene, egalitäre, interaktive Anwendungen − wie Wikipedia, Facebook oder Bewertungsportale −sind aus dem Alltag kaum mehr wegzudenken. An diesem so genannten Mitmachweb kann jeder Anwender bzw. Teilnehmer, unabhängig von Ort und Zeit, freiwillig und flexibel mitwirken. Durch selbst erstellte Inhalte kann er vom passiven Informationskonsumenten nun auch zum aktiven Informationsproduzenten werden. Während sich hierdurch eine neuartige Kultur des Umgangs mit Information und Kommunikation entwickelt, verändert sich zeitgleich der Stellenwert der Wissensteilung. Vor diesem Hintergrund scheint der geflügelte Satz „Wissen ist Macht“ eine neue Bedeutung zu erhalten.

[1] Die Masterthesis wurde von den Autorinnen in Kooperation im Wintersemester 2012/13 im Rahmen des Studiums Master in Library and Information Science (MALIS) an der Fachhochschule Köln verfasst.
[2] Stand der Analyse: 31. Januar 2013 (Abschluss der o.g. Masterthesis).

Unternehmen der privaten Wirtschaft erkennen das Potenzial des Mitmachwebs und versuchen zunehmend, dessen Wesensmerkmale und Strukturen für sich nutzbar zu machen und in ihre Unternehmenskultur, ihre Arbeitsabläufe sowie an die Zusammenarbeit mit Kunden und Stakeholdern zu adaptieren. Im Mittelpunkt steht die Wissensteilung und -produzierung durch die Beschäftigten. Denn in der Situation, „[…] wo im Web schon heute ‚user generated content' eine Vorreiterrolle einnimmt, gewinnt in Unternehmen analog ‚employee generated content' verstärkt an Bedeutung" (Stocker; Tochtermann 2010, S. V). Hiermit hält ein neues Konzept Einzug in die Arbeitswelt: Enterprise 2.0.

Die Autorinnen stellen die These auf, dass kollaborative Instrumente des Wissensmanagements und der Kommunikationsgestaltung, wie sie übergreifend mit Social Web beschrieben werden, die Betriebsorganisation und -kultur nachhaltig verändern und umgekehrt nur dann effektiv funktionieren können, wenn Führungsstil, Kreativitätsbedingungen und betriebliche Hierarchie diese neuen Formen des vernetzten Arbeitens entsprechend unterstützen.

Begriff Enterprise 2.0

E 2.0 bezeichnet die durch das Social Web bedingten innerbetrieblichen, organisatorischen und strukturellen Veränderungen. Grundlegende Prinzipien des Web 2.0 − wie offene Kommunikation und demokratische Partizipation, kollaboratives Arbeiten und hierarchieübergreifende Teilnahme unterschiedlichster Protagonisten − werden auf zentrale Bereiche von Organisationen übertragen. Dies erfolgt zur Unterstützung der institutionsweiten Zusammenarbeit, des Wissens- und Innovationsmanagement sowie der Innen- und Außenkommunikation. Freier, funktionsunabhängiger Austausch von Informationen und Wissen bildet dafür Grundlage sowie Voraussetzung (vgl. Howaldt; Beerheide 2010, S. 362). Der moderne Unternehmenstyp im Sinne von Enterprise 2.0 verfolgt die Konvergenz sowohl technologischer als auch sozialer und kultureller Aspekte. Dabei ist E 2.0 ausdrücklich nicht auf die Implementierung bestimmter Technologien beschränkt (ebd.).

Trotz der Tatsache, dass E 2.0 ein relativ junger Begriff ist, wird er bereits mit diversen, teils auch in der Fachliteratur ungenügend reflektierten Synonymen belegt: Social Enterprise, Social Business, New Economy (vgl. Klotz 2010, S. 1), Enterprise Social Software (vgl. Howaldt; Beerheide 2010, S 361) oder Enterprise Web 2.0 (ebd. und Koch; Richter 2009, S. 15). Zusätzlich problematisch wird die Begriffsvermengung durch die im englischsprachigen Raum gängige Verwendung der Bezeichnungen Social Enterprise und Social Business, die zugleich für gemeinnützige Unternehmen stehen.

In Verbindung mit E 2.0 wird in Wissenschaft und Wirtschaft angestrebt, die Charakteristika und Innovationspotenziale von Social Web im Sinne eines umfassenden, interdisziplinären und mehrdimensionalen Managementkonzeptes zu betrachten und betriebswirtschaftlich zu nutzen. Dabei wird die Position vertreten, dass der Einsatz von Web-2.0-Anwendungen nicht isoliert betrachtet werden darf: Die tragenden, soziotechnisch und soziokulturell bedingten Attribute von Web 2.0 − wie Enthierarchisierung und Vernetzung − müssen mindestens auf gleicher Augenhöhe wie Software und Tools beachtet werden.

Begriffsgenese oder Enterprise 2.0 als ganzheitliches Wissensmanagement

McAfee (2006), Namensgeber des Begriffes, beschreibt in seinem Aufsatz „Enterprise 2.0: The dawn of emergent collaboration" das Konzept ausdrücklich als eine Entwicklung, welche sich auf die mittels Social Software vernetzte, interne Kommunikation und Zusammenarbeit in einem Unternehmen bezieht. Zugleich betont er den Unterschied zum Web 2.0,

welches *nicht* die im unternehmerischen Kontext institutionalisierten Aktivitäten im Internet meint: „These new digital platforms for generating, sharing and refining information are already popular on the Internet, where they ’re collectively labeled ‚Web 2.0' technologies. I use the term „Enterprise 2.0“ to focus only on those platforms that companies can buy or build in order to make visible the practices and outputs of their knowledge workers.” (McAfee 2006, S. 23).

Daneben problematisiert McAfee einen Änderungsbedarf in der Führungs- und Organisationskultur (vgl. ebd., S. 26-28), wie offene Kultur, an Bedürfnissen der Anwender statt an formalen Prozessen orientiertes Veränderungsmanagement sowie aktive Unterstützung seitens der Unternehmensführung. Die Mitarbeiter – in McAfees zitiertem Basisaufsatz als Wissensarbeiter bezeichnet − spielen in seinem Enterprise-2.0-Modell eine wichtige, da inhaltsgebende und gestalterische Rolle: Die Technologie verändert sich und entwickelt sich weiter mit dem aktiven Gebrauch durch die Wissensarbeiter.[3]

Im Kontext ebenfalls erwähnenswertes Begriffsverständnis geht auf das Plädoyer der Gesellschaft für Wissensmanagement e.V. (GfWM) zurück, E 2.0 ganzheitlich zu begreifen (dieser Absatz vgl. Wissensmanagement und Enterprise 2.0 2011, S. 2-3.). Dabei wird explizit zwischen den Zuständen „Enterprise 2.0 einzuführen" und „Enterprise 2.0 zu sein" unterschieden. Die GfWM befürwortet den zweiten Zustand, weil der erste meist nur die Einführung von Technologie beinhaltet. Aus ihrer Sicht bezieht sich E 2.0 zu dem nicht ausschließlich auf Unternehmen, sondern auf alle Arten von Organisationen. Damit wird ein entscheidender Impuls in die Begriffskonzeption von E 2.0 eingebracht.

Enterprise 2.0 in Bibliotheken und Informationseinrichtungen?

Für den tatsächlichen Grad der flächendeckenden Nutzung von Social Web in deutschen Bibliotheken und Informationseinrichtungen liegen, nach Einschätzung der Autorinnen, keine repräsentativen Erhebungen vor.[4] Vatter (2012) weist darauf hin, dass „die heutige Bibliothekswelt […] sich nach wie vor in zwei Lager [spaltet]: diejenigen, die bereits Fuß im Social Web gefasst haben und diejenigen, die sich beharrlich dagegen sträuben“ (Vatter 2012, S. 278). Davon unberührt bleiben die Fragen, inwiefern E 2.0 auf Bibliotheken und Informationseinrichtungen übertragbar ist und wie sich dabei die Auswirkungen auf die innerbetriebliche Informationsqualität und organisationale Wissensgenerierung beschreiben lassen. Kann letztendlich der Einsatz von Web 2.0 und seinen Weiterentwicklungen einen qualitativen Mehrwert erzielen?

In Verbindung mit der Implementierung von Web 2.0 in Bibliotheken wird in Fachliteratur und Praxis häufig die Formulierung Bibliothek 2.0 bzw. Library 2.0 gebraucht. Eine kongruente Definition ist aber in der Fachliteratur nicht eindeutig auszumachen, wie etwa Danowski (2008) anmerkt. Seine eigene Definition „Bibliothek 2.0 ist das Anwenden und Umsetzen der Prinzipien und Techniken des Web 2.0 im Kontext der Bibliothek“ weist er selbst als „[…] nicht unbedingt befriedigend“ aus (Danowski 2008, S. 2), da hier ein unbestimmter Term mithilfe eines weiteren unbestimmten Terms erläutert wird. Sowohl Bibliothek 2.0 als auch Library 2.0 werden darüber hinaus i. d. R. auf öffentliche Biblio-

[3] Dies kommt bei McAfee (2006) an diversen Stellen zum Ausdruck und wird durch Aussagen wie diese impliziert: „One of the most surprising aspects of Enterprise 2.0 technologies is that even though they’re almost completely amorphous and egalitarian, they appear to spread most quickly when there’s some initial structure and hierarchy” (McAfee 2006, S. 27).

[4] Dies wurde durch Recherchen der Autorinnen zum Stichtag 26.11.2012 ermittelt und ihnen auf Anfrage durch Vertreter der Berufsverbände BIB und BID schriftlich bestätigt.

theken bezogen, zudem uneinheitlich gebraucht und kontrovers diskutiert (vgl. z.B. Houghton 2010). Kaden (2009) konstatiert ferner, dass die Bezeichnung Library 2.0 von geringer begrifflicher Substanz und rhetorisch sei (vgl. Kaden 2009, S. 44 und 29). Daneben würde sie „[…] fast durchgängig anwendungsbezogen und als Schlagwort verwendet“ (ebd., S. 25). Insgesamt stellt Kaden das Fehlen eines wissenschaftlichen Diskurses zum Thema und einer kritischen Auseinandersetzung in der Bibliothekswissenschaft fest. In der Praxis zielt Bibliothek 2.0 auf ein externes Publikum, nämlich die Bibliotheksnutzer. Social Networks, Blogs, RSS-Feeds, Auskunftschats usw. werden schwerpunktmäßig für die Informationsversorgung bzw. Kommunikation mit Nutzern eingesetzt und nicht für interne Prozesse. Dieser nach außen zielende Fokus muss bedacht werden, um zu erkennen, inwiefern E 2.0 in seiner Veränderung der internen Prozesse (und, so gesehen, als Folge dieser auch verbesserter externer Dienstleistungen) über das Konzept Bibliothek 2.0 hinausgeht.

Die mit einer Social-Web-Einführung verbundenen wesentlichen Herausforderungen für Bibliotheken und Informationseinrichtungen gehen daher aus Sicht der Autorinnen über die bloße Einführung einzelner Anwendungen merklich hinaus und liegen demnach nicht vordergründig in technischem Know-how der Beschäftigten oder redaktioneller Auswahl der zu kommunizierenden Beiträge. Im Hinblick auf Mehrwerte und langfristigen Erfolg von Web 2.0 in Bibliotheken und Informationseinrichtungen erscheint es notwendig, eine sich an ihren Spezifika orientierende Organisationsstruktur und -kultur, Personalführung und -motivation zu erreichen. Dadurch sollen die Partizipation der Beschäftigten erhöht sowie die Weitergabe und Generierung von Wissen gewährleistet werden. Die Hauptanforderungen dabei werden von den Verfasserinnen auf den Gebieten des Personalmanagements und der Organisationsentwicklung identifiziert. Eine solche erweiterte, soziotechnisch und soziokulturell verankerte Problematik, die durch den Begriff Bibliothek 2.0 − soweit eruiert − in dieser Ausprägung nicht abgedeckt wird, findet sich im Begriffskonzept E 2.0 wieder.

Wissensmanagement in Enterprise 2.0

Im Enterprise-2.0-Ansatz stellt Wissensmanagement einen zentralen Aspekt dar (vgl. Wissensmanagement und Enterprise 2.0 2011, S. 2-3). Die Arbeitsabläufe in E 2.0 erfolgen webbasiert und sind an eine gemeinsame Informations- und Wissensbasis gekoppelt. Sie sind stets und für alle transparent und können situationsgerecht angepasst werden. Die verschiedenen Social-Software-Tools − wie Wikis, RSS-Feeds, Bookmarks − werden in die Arbeitsdurchführung unmittelbar eingebunden (vgl. Zukünftige Qualifikationserfordernisse 2010, S. 30-31).

Aus dem technischen Blickwinkel betrachtet, bieten die diversen Web-2.0-Anwendungen in besonderem Maße geeignete Plattformen zur gemeinschaftlichen und transparenten Teilung, Dokumentation, Strukturierung und (Weiter-)Entwicklung von Informationen und Wissen. So ermöglichen bspw. fachliche, berufsbezogene Online-Netzwerke, dass Mitarbeiter problembezogen und lösungsorientiert miteinander kommunizieren, Erfahrungen und Wissen ort- und zeitunabhängig austauschen und somit voneinander lernen können. Zudem können sich Mitarbeiter jederzeit und flexibel mit ihrem Informations- und Wissensstand in die verschiedenen Arbeitsabläufe einbringen und dadurch zur Qualitätssteigerung des gesamten Geschäftsprozesses beitragen. Die Prozesse öffnen sich so „[…] für die Erschließung von Mitarbeiterpotenzialen und Synergien, aber auch auf Erschließung und Nutzung externen Know-hows“. (Zukünftige Qualifikationserfordernisse 2010, S. 47-48). Die Qualität der Standardisierung in den Abläufen kann insgesamt erhöht werden.

In E 2.0 sind die Organisationsstrukturen bewusst flach konzipiert. Hierdurch sollen eine möglichst schnittstellenarme Organisationsstruktur, unmittelbares Zusammenwirken sowie reibungslose und verlustarme Information und Kommunikation, Wissensteilung und -generierung erfolgen (vgl. Integriertes Personalmanagement 2012, S. 33). Kürzere Entscheidungsinstanzen und schlankere Prozesse werden in Wissenschaft als weitere Vorteile flacher Hierarchien diskutiert. So sollen zügig bedarfsgerechte und bewegliche Reaktionen und Neustrukturierungen ermöglicht werden, ohne jedes Mal aufwändige Anpassungen der Organisation vornehmen zu müssen. Außerdem sind Hierarchien „[…] für Wissensarbeiter vollkommen ungeeignet, weil Wissen nicht hierarchisch organisiert ist, sondern situationsmäßig entweder relevant oder irrelevant ist“ (Leppert 2011, S. 6).

Für den tatsächlichen, praktischen Nutzen und Mehrwert von E 2.0 ist auf organisationsstruktureller Ebene eine Durchdringung der Arbeiten und Abläufe durch das Social Web also essentiell (vgl. Jäger; Petry 2012, S. 31), auch wenn E 2.0 nicht auf den Einsatz einer bestimmten Technik eingegrenzt oder fixiert ist (vgl. Howaldt; Beerheide 2010, S. 362). Das technische System bzw. die Datenablage sind so gestaltet, dass sich darin Teams, Projektgruppen, Netzwerke und Communities jederzeit und bedarfsgerecht selbst organisieren und miteinander kollaborieren können. Anhand webbasierter Zusammenarbeit können in E 2.0 Wissen und Innovation letztendlich hierarchiefrei und multidirektional entstehen sowie vernetzt, verzahnt und in die alltägliche Aufgabenbewältigung integriert werden. Die gesamte Wissensbasis der Organisation kann sichtbarer werden, das kollektive Wissen wird aktiviert und die Schwarmintelligenz gefördert.

Siemann (2012) hebt die Bedeutung derart strukturierter, offen zugänglicher Plattformen in E 2.0 im direkten Vergleich zu den informellen Netzwerken und den sogenannten Flurabsprachen in einer Organisation hervor. Um in dieser Hinsicht Informations- und Interaktionsdefiziten zu begegnen, werden in Enterprise-2.0-Betrieben die bisherige Kommunikation und Arbeitsabläufe durch Wikis, Blogs, RSS-Feeds, Microblogging u. ä. ergänzt und ersetzt (Siemann 2012, S. 10.11). Den informellen Organisationsstrukturen sowie dem sogenannten Bottom-up-Ansatz wird damit in E 2.0 Rechnung getragen.

Als zentrales Element einer Enterprise-2.0-Organisation werden in Wissenschaft und Wirtschaftspraxis die Art und der Umfang der Selbstorganisation hervorgehoben. Dabei handeln einzelne Organisationseinheiten, z.B. Sachgebiete oder Fachreferate, selbstbestimmt und ergebnisverantwortlich. Stecher (2012) argumentiert, dass dies ökonomischerweise einen Verzicht auf zentrale Vorgabe und Kontrolle bedeutet und die Stärkung dezentraler Verantwortung fördert: „Mitarbeiter werden zur autonomen Bildung von Teams und einer weitestgehend selbst bestimmten Arbeitsteilung befähigt, was den zentralen Koordinationsaufwand auf ein Minimum reduziert“(Stecher 2012, S. 61-62). Dabei wird davon abgesehen, einzelne Arbeitsergebnisse oder bestimme Lösungswege zu fordern. Stattdessen werden seitens des Managements Angebote gemacht und Voraussetzungen geschaffen, welche die Selbstorganisation ermöglichen (vgl. Graf; Glatz-Götz 2011, S. 137).

Vor dem Hintergrund der autonomen Strukturen und selbstgelenkten Organisation werden − als zentrale Aufgaben des Managements − Information und Kommunikation, Austausch und Vernetzung betrachtet (vgl. ebd., S. 137). Ohne Erfüllung dieser Anforderungen erscheint gezielte Kooperation von Menschen als Wissensträger in E 2.0 nicht effizient durchführbar (vgl. Stecher 2012, S. 67). Eine Enterprise-2.0-Organisation orientiert sich an der Entwicklung der Strukturen von Web 2.0 im Sinne eines Perpetual Beta (vgl. ebd.). Ein solches, sich selbst entwickelndes System stellt nach Ansicht der Verfasserinnen die Grundlage für den organisatorischen Wandel in einer Enterprise-2.0-Bibliothek dar. Für eine zielgerichtete Führung einer derart komplexen, modernen Organisation, wie sie E 2.0 darstellt, ist eine

Veränderung bzw. Anpassung der bestehenden Strukturen in Bibliotheken jedoch auch ohne gezielte Fortbildungs- und Förderungsmaßnahmen kaum vorstellbar.

Das (theoretische) Nutzenpotenzial wird insbesondere bei der Reduktion der Koordinationskosten durch das gemeinschaftliche und selbstorganisierte Arbeiten vermutet. Stecher (2012) fragt, „inwieweit aus betriebswirtschaftlicher Sicht durch die Integration von Enterprise-2.0-Technologien in Verbindung mit einer auf deren Nutzung abgestimmten Aufbau- und Ablauforganisation, Effektivitäts- und Effizienzgewinne in unterschiedlichen Bereichen erzielt werden können“ (Stecher 2012, S. 80). Als zentrale ökonomische Kenngrößen erkennt Stecher etwaige Kostensenkungen und Zeiteinsparungen einerseits sowie Qualitätssteigerungen im Informations- und Wissensfluss andererseits (vgl. ebd., S. 80). Dies kann sich bspw. bei der Informationssuche, -beschaffung sowie -weitergabe ergeben sowie beim Vermeiden von Doppelarbeit und Fehlern aus Desinformation, innerhalb und außerhalb der Institution (vgl. ebd., S. 80-81). Eine verbesserte Kontexteinbindung der Mitarbeiter stärkt die Sozialisation und den Wissenstransfer. Der optimierte Wissensaustausch kann zur Innovationssteigerung der Mitarbeiter und Organisation positiv beitragen, wohingegen Hierarchien als innovationshemmend angesehen werden.

Enterprise 2.0 und neue Anforderungen an Wissensmanagement in Zeiten des Social Web

Wird in einer Organisation ein gezielter Umgang mit Wissen als Ressource angestrebt, gilt das Wissensmanagement als ein wichtiger Aspekt der Führung (vgl. Flicker; Paul 2013, S. 132). Im vorliegenden Beitrag steht die Frage im Mittelpunkt, wie man Wissen im Kontext von E 2.0 managen muss, um daraus Wertschöpfung für Bibliotheken und Informationseinrichtungen zu erreichen.

Die Wechselwirkungen zwischen ganzheitlichem Personalmanagement und systemischer Organisationsentwicklung im Sinne dieses Beitragslassen eine Interpretation von Bibliotheken als komplexe Lernende Organisationen zu. Da die Wissensteilung ein differenzierter und sensitiver Vorgang an sich ist, hängt er insbesondere in Lernenden Organisationen von bestimmten Kontextfaktoren ab, wie Regulierung, Motivation und sozialem Feedback. Dabei sieht Reinisch (2007) in Verbindung mit Wikipedia eines der wesentlichen Prinzipien des Wissensmanagements erfüllt. Dieses besagt, dass Wissen sich vermehrt, wenn es mit anderen geteilt wird (vgl. Reinisch 2007, S. 149).

Im Rahmen von Wissensmanagement wird der Umgang mit Wissen gezielt geplant, organisiert und geleitet. Die Aufgabe eines ganzheitlichen Wissensmanagements ist es, das in einer Organisation existierende Wissen bestmöglich zu erhalten, zu nutzen, zu pflegen und mit weiterem, neuem Wissen anzureichern, um damit „[…] aktuell und zukünftig erfolgreiche Produkte, Prozesse und Geschäftsfelder vertreiben, abwickeln und bedienen zu können“ (Reinisch 2007, S. 151).Aus der Festlegung von Wissenszielen in Abstimmung mit den Zielsetzungen und Strategieneiner Organisation soll letztendlich das Geschäftsergebnis optimiert werden. Wissen wird als wertsteigernde Ressource betrachtet.

Laut Leppert (2011) ist in E 2.0 ein besonderes, neues Verständnis für die Ressource Wissen und das Wissensmanagement obligatorisch. Demnach sollte Wissen als Flussbetrachtet werden und das Wissensmanagement sich „[…] auf die Gestaltung von Kontexten, Beziehungen und Wissenskultur durch ein wissensorientiertes Management […]“ fokussieren (Leppert 2011, S.4). In ähnlicher Konnotation argumentiert Reinisch (2007), indem er beim Thema Wissensmanagement 2.0 betont, dass bspw. Wikipedia deswegen erfolgreicher als aufwendig gefilterte und gesteuerte IT-Systeme sei, weil die Plattform so geschaffen ist,

„[…] dass es die Quellen des Wissens zum Sprudeln bringt und die richtigen Impulse für den Fluss des Wissens gibt“ (Reinisch 2007, S. 148). Demnach ist der Mensch die Quelle des Wissens selbst. Wissen basiert auf Daten und Informationen, ist aber − im Gegensatz zu diesen – immer an Personen gebunden (vgl. Probst; Raub; Romhardt 2003, S. 22; Flicker; Paul 2013, S. 136). Das individuelle Wissen eines Menschen als Wissensträger ist, so gesehen, ein abstraktes Konstrukt und kann durch Kommunikation zum Ausdruck gebracht, ausgetauscht und reflektiert und weiterentwickelt werden. Die Nützlichkeit und der Wert des Wissens ergeben sich aus seiner kontextbezogenen Anwendung. Auch das in einem selbst verborgen liegende Wissen kommt erst entsprechend zur Geltung, wenn man es situationsgerecht und ggf. praxisbezogenen einzusetzen und weiterzugeben vermag.

Social Software wird von Wissensmanagementexperten deswegen als besonders interessant bewertet, weil hier die Elemente Interaktivität, Austausch und Verlinkung sowie das soziale Miteinander explizit im Fokus stehen. Günther (2010) stuft das Potenzial von Social Software für den Transfer und die Generierung von (neuem) Wissen als „[…] am höchsten“ ein (vgl. Günther 2010, S. 16). Das verborgene, implizite Wissen (engl.: tacit knowledge)[5] kann nun mithilfe von Social Software, bis zu einem gewissen Grad, freigesetzt und damit auch für andere verfügbar gemacht werden, was die Hauptziele eines systematischen Wissensmanagements sind (vgl. Reinisch 2007, S. 150).

Gerade wegen der vorhandenen und von Fachleuten als passend eingeschätzten technischen Web-2.0-Möglichkeiten werden Personalmanagement und Organisationsentwicklung mehr denn je vor besondere Anforderungen gestellt – schließlich sind mit Social Software praktikable und im privaten Bereich bereits intensiv genutzte Werkzeuge zur partizipativen Informations- und Wissensteilung verfügbar. Die Schaffung adäquater Rahmenbedingungen, in denen besonders das verborgene, unbewusste Wissen zum Vorschein kommen und systematisch genutzt werden kann, stellt die Führungskräfte in einer Enterprise-2.0-Organisation vor wachsende und teilweise neue Anforderungen. Das betrifft insbesondere die Mitarbeitermotivation zur Partizipation und Wissensteilung. „You can’t manage knowledge. Knowledge is between two ears, and only between two ears“ (Drucker zitiert nach Hofmann 2010, S. 60), schrieb Drucker (1999) angesichts der Problematik, Menschen als Wissensträger zum Wissensaustausch zu mobilisieren. Um dies erreichen zu können, müssen Wissensarbeiter vernetzt werden − sowohl in technischer als auch in sozialer, zwischenmenschlicher Hinsicht (Hofmann 2010, S. 54). Wissensmanagement in E 2.0 ist daher in erster Linie eine Vernetzungs- und Kommunikationsaufgabe (Flicker; Paul 2013, S. 132). Im Vier-Ebenen-Modell von Davenport und Prusak (1998), welches die Stellen in Organisationen beschreibt, in denen das Wissensmanagement angesiedelt ist, werden Bibliothekare explizit als die Wissensarbeiter bezeichnet (vgl. Davenport; Prusak zitiert nach Lehner 2012, S. 292).

Für Stecher (2012) steht in E 2.0 der Wissensaustausch stellvertretend für das gesamte Wissensmanagement (Stecher 2012, S. 62). Offenbar kann ein weit reichendes Networking das individuelle und kollektive Wissen effektiver und effizienter nutzbar machen. Hierbei erschöpft sich Wissensmanagement nicht darin, das gesamte Wissen der Organisation in Datenbanken zu sammeln, zu speichern und zugänglich zu machen, sondern versucht vielmehr aktiv, die Informationsträger und -nachfrager situativ optimal in Kontakt zu bringen. Der Austausch und die Zusammenarbeit können so verbessert werden. Die Transparenz des explizit und implizit vorhandenen Wissens kann gesteigert werden. Die Kommunikation mittels Social Software hilft zudem, eine gemeinsame Perspektive auf die Organisationsziele zu gewinnen. Durch die Integration verschiedener Sichtweisen in der Bibliothek bzw. Informationseinrichtung kann die Qualität der Zielsetzungen gemeinsam gesteigert werden.

[5] Zur Definition des Begriffes tacit knowledge siehe Seidler-de Alwis; Habermann 2004, S. 375-377.

Leppert (2011) weist daraufhin, dass schnelle Interaktion innerhalb großer Teams mittels Social Software deutlich die Innovationsdynamik steigert, denn „[…] je besser die Vernetzung und die Anzahl der losen Bindungen ist, desto höher ist die Wahrscheinlichkeit für neue Lösungsmuster“ (Leppert 2011, S. 9).

Schütt (2009) sieht Wissensmanagement im Zusammenhang mit E 2.0 in einem „[…] dritten Frühling […]“ (vgl. Schütt 2009, S. 38). In diesem konzentriert sich das Thema ihm zufolge auf den Aspekt Innovationsmanagement, womit es sich allerdings „[…] noch entfernt von einer wirklich ganzheitlichen Betrachtung der Optimierung von Wissensarbeit […]“ befindet (vgl. ebd., S. 38).

Fazit

Im Fokus des Wissensmanagements in einer Enterprise-2.0-Bibliothek stehen gemeinschaftliche und hierarchieübergreifende Wissensgenerierung und -austausch. Dabei muss aus Sicht der Autorinnen, den Wechselwirkungen zwischen der Förderung von Autonomie und Selbstorganisation einerseits und der Unterstützung von Zusammenarbeit und Teambildung andererseits spezielle Beachtung geschenkt werden. Dennoch bleibt für eine Führungskraft auch in einer Enterprise-2.0-Bibliothek die spannende und anspruchsvolle Aufgabe bestehen, mit Wissensmanagement einen realen Mehrwert für die Institution und ihre Nutzer zu schaffen. Dies unterstreichen Viljakainen und Müller-Eberstein (2012) wie folgt:„Was genau können Sie tun, damit die Erfahrungen an der Schnittstelle zum Kunden in immer neue Dienstleistungen umgesetzt werden?“ (Viljakainen; Müller-Eberstein 2012, S. 95).

E 2.0 bietet auch für Bibliotheken und Informationseinrichtungen ein umfassenderes als bisher angewendetes Einsatzszenario von Social Web. Die Prinzipien des Web 2.0, die gemeinsam mit seinen Anwendungen unmittelbar in die internen Abläufe und Strukturen einer Bibliothek nützlich integriert werden, bieten weitgehende Potenziale und Modernisierungschancen. Die Voraussetzungen dafür, aber vielmehr die Folgen einer solchen Integration, sind für die gesamte Organisationsentwicklung tief greifend. Sie erfordern in jedem Fall eine veränderte Organisationsstruktur und Kommunikationskultur sowie ein anderes Verständnis von Führung und Personalmanagement.

Im Sinne von Effizienzsteigerung und Wertschöpfung reicht der Nutzwert von E 2.0 über die technisch orientierten Web-2.0-Tools hinaus und verspricht, in den Bereichen des Geschäftsprozess-, Wissens-, Innovations- und Projektmanagements den größten Mehrwert zu entfalten. Hierfür müssen jedoch die genannten Voraussetzungen erfüllt oder aber erst geschaffen werden, inklusive formaler und rechtlicher Determinanten.

E 2.0 hat, wie das Wissensmanagement, die Lernende Organisation als übergeordnetes Ziel. Folgt man der Definition der Lernenden Organisation als „eine Organisation mit der Fähigkeit, Wissen zu entwickeln, zu erwerben und zu (ver-)teilen“ (Wissensmanagement und Enterprise 2.0 2011, S. 2 und Vollmaar; Scheerer 2012, S. 10) trifft das aus Sicht der Autorinnen in zweifacher Hinsicht auf Bibliotheken und Informationseinrichtungen zu: intern und extern. Die Transformation zu einer Enterprise-2.0-Bibliothek kann demnach gelingen, wenn organisationales Lernen und eine Anpassung an die sich durch das Web 2.0 verändernde Umwelt als gemeinsames, innerbetriebliches Ziel verinnerlicht werden, um hierdurch den bibliothekarischen Auftrag noch besser und nutzergerechter erfüllen zu können. Für Bibliotheken und Informationseinrichtungen, die sich für den Schritt zum Enterprise 2.0 entscheiden, können sich vielseitige Vorteile ergeben. Dazu gehören eine veränderte Arbeits- und Informationskultur mit einer im Idealfall verbesserten Arbeitsatmosphäre. Organisationsinterne Abläufe können flexibilisiert und verschlankt werden. Durch ein angepasstes

Wissensmanagement können letztlich auch die Serviceleistungen für die Nutzeroptimiert werden.

Wollen Bibliotheken von den Möglichkeiten des Web 2.0 profitieren, haben sie substanziell die gleichen Optionen wie Unternehmen der freien Wirtschaft – die oft gegebenen Besonderheiten des öffentlichen Dienstes sind hierbei jedoch als Hürden nicht zu unterschätzen. Die vergleichsweise stark hierarchisch geprägte Struktur des öffentlichen Dienstes ist insbesondere ernst zu nehmen, mit weiteren damit verbundenen Fragestellungen. Dazu gehört etwa die Frage nach einer Neuregelung der tariflichen Eingruppierung und anzupassenden Entlohnungsmodalitäten, wenn sich das Arbeiten in E 2.0 in wechselnden Projekten und Positionen (je nach Kenntnisstand) sehr flexibel und durchlässig gestaltet. Eine separate Untersuchung, welche die Vereinbarkeit des öffentlichen Dienstes in seinen gegenwärtigen und zukünftigen Strukturen mit E 2.0 prüft und dabei auch die Aspekte des Arbeits- und Verwaltungsrechts berücksichtigt, erscheint erforderlich.

Literatur

Danowski, P. (2008). Bibliothek 2.0. Was verbirgt sich dahinter? In: *Büchereiperspektiven,* 4, S. 2-5. http://www.publikationen.bvoe.at/perspektiven/bp4_08/s2-5.pdf [17.03.2014].

Flicker, A.; Paul, T. M. (2013). Wissen sichern in Organisationen am Beispiel der Stadtbücherei Würzburg. In: Degkwitz, A. (Hrsg.): *Personal- und Organisationsentwicklung in Bibliotheken* (S. 131-148). Berlin: De Gruyter Saur.

Glatz, H.; Graf-Götz, F. (2011). *Handbuch Organisation gestalten. Für Praktiker aus Profit- und Non-Profit-Unternehmen, Trainer und Berater* (2. Aufl.). Weinheim, Basel: Beltz.

Günther, J. (2010). *Wissensmanagement 2.0. Erfolgsfaktoren für das Wissensmanagement mit Social Software. Eine empirische Studie zu organisatorischen und motivationalen Erfolgsfaktoren für den Einsatz von Social Software in Unternehmen*. Stuttgart: Fraunhofer-Verl.

Hofmann, J. (2010). Enterprise 2.0: Das Wissen der Mitarbeiter mobilisieren. Wissensmanagement als Vernetzungs- und Kommunikationsaufgabe. In: Eberspächer, J.; Holtel, S. (Hrsg.): *Enterprise 2.0. Unternehmen zwischen Hierarchie und Selbstorganisation* (S. 53-61). Berlin, Heidelberg: Springer.

Houghton, S. (2010). CIL2010: Organization 2.0. In: *Librarian in Black*. http://librarianinblack.net/librarianinblack/2010/04/org2dot0.html [17.03.2014].

Howaldt, J.; Beerheide, E. (2010). Innovationsmanagement im Enterprise 2.0.Auf dem Weg zu einem neuen Innovationsparadigma? In: Howaldt, J. (Hrsg.): *Soziale Innovation* (S. 355-370). Wiesbaden: VS-Verlag.

Integriertes Personalmanagement (2012). *Integriertes Personalmanagement in der Praxis. Prozesse und professionelle Standards*. Deutsche Gesellschaft für Personalführung e.V. (Hrsg.). Bielefeld: Bertelsmann.

Jäger, W.; Petry, T. (2012). Enterprise 2.0. Herausforderung für Personal, Organisation und Führung. In: Jäger, W.; Petry, T. (Hrsg.): *Enterprise 2.0. Die digitale Revolution der Unternehmenskultur. Warum Personalmanager jetzt gefordert sind* (S. 17-35). Köln: Luchterhand.

Kaden, B. (2009). *Library 2.0 und die Wissenschaftskommunikation*. Berlin: Simon.

Klotz, U. (2010). Schöne neue Arbeitswelt 2.0? In: Eberspächer, J.; Holtel, S. (Hrsg.): *Enterprise 2.0. Unternehmen zwischen Hierarchie und Selbstorganisation* (S. 1-16). Berlin, Heidelberg: Springer.

Koch, M.; Richter, A. (2009). *Enterprise 2.0. Planung, Einführung und erfolgreicher Einsatz von Social Software in Unternehmen*. (2., aktual. und erw. Aufl.). München: Oldenbourg.

Lehner, F. (2012). *Wissensmanagement. Grundlagen, Methoden und technische Unterstützung*. (4., aktual. und erw. Aufl.). München: Hanser, 2012.

Leppert, A. (2011). *SWR 2.0. Ein Pilotprojekt. Mit Social Software in Richtung Enterprise 2.0*. Potsdam: Fachhochschule Potsdam, Institut für Information und Dokumentation.

McAfee, A. P. (2006). Enterprise 2.0. The dawn of emergent collaboration. In: *MIT Sloan Management Review*, 3, S. 20-29. http://www.wikiservice.at/upload/ChristopheDucamp/McAfeeEntrepriseDeux.pdf [17.03.2014].

Probst, G.; Raub, S.; Romhardt, K. (2003). *Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen.* (4., überarb. Aufl.). Wiesbaden: Gabler.

Reinisch, F. (2007). *Die Köpfe sind das Kapital. Wissen bilanzieren und erfolgreichnutzen*. Heidelberg: Redlinie Wirtschaft.

Schütt, P. (2009). Führen im Zeitalter des Informationsüberflusses. *Wissensmanagement: das Magazin für Führungskräfte* 11(8), S. 34-36.

Seidler-de Alwis, R.; Hartmann, E. (2004). The significance of tacit knowledge on company's innovation capability. In: Bekavac, B.; Herget, J.; Rittberger, M. (Hrsg.): *Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004), Chur, 6.-8. Oktober 2004*, S. 373-439. Konstanz: UVK Verlagsgesellschaft mbH.

Siemann, C. (2012). Heißluft 2.0 oder Mehrwert? *Personalwirtschaft*, 6, S.10-12.

Stecher, M. A. (2012). *Enterprise 2.0. Sozio-technische Neuausrichtung von Unternehmen*.(Zugl. Dissertation Ludwig-Maximilians-Universität). München, Hamburg: Kovac.

Stocker, A.; Tochtermann, K. (2010). *Wissenstransfer mit Wikis und Weblogs. Fallstudien zum erfolgreichen Einsatz von Web 2.0 in Unternehmen*. Wiesbaden: Gabler.

Wissensmanagement und Enterprise 2.0 (2011). *Wissensmanagement und Enterprise 2.0. Positionspapier der Gesellschaft für Wissensmanagement e.V.* (Version 1.1). GfWM. http://de.slideshare.net/gfwm/gfwm-positionspapier-wissensmanagement-und-enterprise-20.

Vatter, A.(2012). Tipps für Einstieg ins Social Web. Klare Strategie schützt vor bösen Überraschungen / schnelle Reaktion wichtig. In: *BuB,* 64(4), S. 278-279.

Viljakainen, P.; Müller-Eberstein, M. (2012). *Digital Cowboys. So führen Sie die Generation PlayStation.* Weinheim: Wiley-VCH-Verlag.

Vollmaar, G.; Scheerer, B. (2012). Enterprise 2.0. Mehr als „nur" Social Software im Unternehmen? In: *Wissensmanagement* 1, S. 10-13. http://www.wissen-kommunizieren.de/download/Wima_Enterprise%202.0_01-2012.pdf [17.03.2014].

Zukünftige Qualifikationserfordernisse bei beruflichen Tätigkeiten auf mittlerer Qualifikationsebene aufgrund der Auswirkungen von Web 2.0 (web2skills). Untersuchung im Auftrag des BMBF. Abschlussbericht (2010). Schmidt-Hertha, B. et al. (Hrsg.). http://www.frequenz.net/uploads/tx_freqprojerg/Abschlussbericht_Web_2.0_final.pdf [17.03.2014].

From ESE to EDM and Beyond: How Europeana Provides Access to its Cultural Heritage Objects

Steffen Hennicke, Evelyn Dröge, Violeta Trkulja & Julia Iwanowa

The paper aims to describe the motivation for moving from a flat tree-based data model towards a more specific graph-based data model for storing collected metadata records that were represented with various data formats within Europeana. The metadata provided to Europeana describes objects from different domains and was stored in various formats before. The difference between the first model Europeana has used for representing data, the *Europeana Semantic Elements* (ESE), and the improved second model, the *Europeana Data Model* (EDM), is explained. The focus here lays on the EDM as this model replaces the first one due to its higher degree of flexibility and expressivity. The paper concludes with a brief look at the specialisation of the EDM for the manuscript domain made by the *Digitised Manuscripts to Europeana* (DM2E) project and the contextualisation of the mapped data.

1 Introduction

The *Europeana Data Model* (EDM) was designed in order to insert rich metadata about *Cultural Heritage Objects* (CHOs) into *Europeana*. Europeana, the European digital library[1], is by most users seen as a portal (Concordia, Gradmann & Siebinga, 2010) which gives access to more than 30 million library, archive, museum and audio-visual objects from 36 countries[2]. With the development of the EDM, Europeana can be much more than just a portal. The *Linked Open Data* (LOD) approach facilitates querying the Europeana content and unlocks the metadata on a massive scale (Isaac & Haslhofer, 2013). Linked (Open) Data is data that is made available on the Web, accessible for both, human users and applications, is linked to other data and dereferencable via stable identifiers (Berners-Lee, 2006; Heath & Bizer, 2011). Ideally, LOD is described in the representation language *Resource Description Framework* (RDF) (Manola & Miller, 2004) which can be processed by its standard query language *SPARQL* (Buil-Aranda, Corby et al., 2013). This allows not only to offer search functionality for Europeana content as provided by its website but also to reuse the produced RDF data. In this way, not only the Europeana portal but also other applications can be built on CHOs from institutions all over Europe which are modelled in EDM.

This paper gives an overview on metadata interoperability in Europeana (section 2), how the EDM works in comparison to the former model ESE (section 3), an introduction in RDF and which additional value RDF actually provides (section 4), the EDM representation in OWL (section 5), how the EDM can be specialised (section 6) and enriched by linking to other resources (section 7).

[1] Europeana– Homepage: http://europeana.eu [21.11.2013].
[2] Europeana Professional – Content: http://www.pro.europeana.eu/web/guest/content [29.11.2013].

2 Metadata interoperability in Europeana

Europeana was initiated in 2005 by the European Commission with the vision to establish a digital library that is a single, direct and multilingual access point to the European cultural heritage and that ensures a common access to Europe's libraries, archives and museums. Since November 2008, a first version of Europeana is online which is continuously refined. 27 projects are currently involved into Europeana either by providing content or technology – or both[3]. Beyond this, numerous Task Forces deal with issues across the field of Europeana's development and respond to strategic questions and key opportunities[4]. In December 2013, Europeana has started a Metadata Quality Task Force[5] aiming to improve the quality of metadata found in Europeana. As the number of records in Europeana increased in the last years it became more difficult to find a specific record. This could be changed by improving the metadata quality in each single record.

The major challenge Europeana is facing, apart from the improvement of data quality, is the integration of many different metadata standards from diverse sources which are used to collect syntactically and semantically heterogeneous data about cultural heritage. Existing metadata schemas in the cultural heritage domain are for example Web schemas like Dublin Core[6], cataloguing schemas like MARC or MODS for libraries and EAD for archives and harmonisation schemas like CIDOC-CRM and FRBR (Hyvönen, 2012).

Metadata is the key for efficient management of resources in information systems, whereas metadata interoperability is a prerequisite for uniform access to resources in multiple autonomous and heterogeneous information systems (Haslhofer & Klas, 2010; Duval, Hodgin et. al., 2002). The interoperability and exchange of metadata is facilitated by the "mapping of [...] elements, semantics, and syntax from one metadata scheme to those of another [and this, in turn] allows metadata created by one community to be used by another group that employs a different metadata standard" (NISO, 2004:11).

Europeana attempted to solve this issue of metadata interoperability during the prototyping phase by establishing the first used model, the ESE. This approach proved to be a simple and robust solution but also showed some major drawbacks which are described later. The discovered deficiencies guided to the further improvement of the schema and the first draft of the EDM. This new model is based on principles and technologies of Linked Data and the Semantic Web which are "a new promising approach to address" (Hyvönen, 2012:6) the problems of metadata interoperability within the domain of cultural heritage.

The functionality and the benefits of the EDM become obvious by taking a look at the data representation based on EDM compared to the data representation based on the metadata model ESE.

3 Replacing ESE with EDM

The ESE has been used since the first Europeana portal prototype in November 2008 (Europeana Semantic Elements Specification and Guidelines, 2013). Isaac & Haslhofer (2013:293) describe the ESE as follows: "[T]he Europeana Semantic Elements (ESE) XML

[3] Content and/or technology providing projects: http://pro.europeana.eu/web/guest/projects [26.11.2013].
[4] Europeana Task Forces. Europeana Pro website: http://pro.europeana.eu/web/guest/network/task-forces [26.11.2013].
[5] New Metadata Quality Task Force. Europeana Pro blog: http://pro.europeana.eu/pro-blog/-/blogs/1970052 [26.11.2013].
[6] Dublin Core (DC) namespace: http://purl.org/dc/elements/1.1/ [04.12.2013].

Schema [...] is essentially a flat record structure that uses the Dublin Core Element Set with some Europeana extensions". The schema represents the lowest common denominator between the different metadata standards which are present in the information space of Europeana (Europeana Data Model Primer, 2013:4).
The ESE approach to the problem of metadata interoperability is to force rich metadata records provided to Europeana into one common and very simple metadata schema. *Example 1* shows an excerpt of a fictional ESE record snippet about the painting of Mona Lisa that illustrates this principle.

```
<record>
      <dc:creator>Leonardo da Vinci (Maler)</dc:creator>
      <dc:title>La Joconde - Gioconda & Mona Lisa</dc:title>
      <europeana:dataProvider>Bildarchiv Foto Marburg
      </europeana:dataProvider>
</record>
```

Example 1: Fictional ESE record snippet about the painting of Mona Lisa.

The basic assumption behind ESE is that if everyone uses the same representation language, there ought to be no interoperability problems. This representation language needs to be simple and generic enough so that anyone can use and understand it, independent from the kind of institution, domain and the representation language that was initially used.

A problem arises if not sufficient semantics are represented in such a common language and if a data provider aims at going beyond the limits of it in terms of expressiveness or granularity. Therefore, disadvantages of the ESE are that there are not many elements for describing an object, that it is not intended that these can be specialised and that it is not possible to model complex (hierarchical) objects, for example a monograph and its subparts. In EDM, this can be done by creating links to objects that are part of the described CHO or of which the described CHO is part of and thus creating hierarchical objects or by creating links to objects that follow the described CHO in a sequence. Additionally, creating relations to other objects is not easy in ESE as data is stored in form of literals and not of resources which can be further described. The RDF-based[7] EDM offers those functions, which are a core component of the formal semantics behind RDF. One of the main motivations for the development of the RDF framework was to achieve "interworking among applications: combining data from several applications to arrive at new information" (W3C, 2004). Therefore, it is easier to build an application for creating relations to other objects based on EDM.

After the launch of the ESE-based prototype of the Europeana portal, another challenge arose regarding the provenance of the metadata. It was not possible to include additional information weather a value applies to the original cultural heritage object or to its digital representation and it was difficult to manage the ingestion of multiple records for the same item, containing potentially contradictory statements about it. The Europeana prototype had therefore clearly shown from the very beginning that the ESE is not an ideal solution for the ambitions of Europeana.

The EDM is the logical improvement of the first schema. While the ESE reduces different data standards used by different domains and data providers to one common denominator, the EDM reversed this approach by being developed as a flexible and extensible semantic

[7] The W3C concept of the graph data model: http://www.w3.org/TR/rdf-concepts/#section-data-model [25.11.2013].

layer, which can accommodate community- or domain-specific standards via specialisations. This does not imply that the EDM core classes are not very general as well, but that the model allows including contextual and more specialised resources. In this way, the EDM changes the way Europeana deals with metadata gathered from data providers and aggregators. It provides more expressivity and flexibility to distinguish "between the intellectual and technical creation that is submitted by a provider (a bundle of resources about an object curated by the provider), the object this structure is about, and the digital representations of this object, which can be accessed over the web" (Europeana Data Model Primer, 2013:5). The higher expressivity of the (unspecialised) EDM is primarily due to additionally established contextual classes for agents, time spans, places or concepts. Additionally, the EDM offers a lot more properties that can be used to represent the provided data and its aim is to integrate and connect data in order to enable search functionality among data sets[8].By contextualising objects with external information resources, rich additional context can be provided.

In June 2013, the first EDM data was ingested into Europeana[9]. The EDM replaces the former data representation schema ESE since then successively. Providers still can submit their data in ESE and Europeana will manually convert the data to EDM, although the converted EDM data will be in most cases less rich than original EDM data.

4 Building a new data model based on RDF

The first important choice for Europeana's new data model was the selection of an appropriate representation language for the model itself. The W3C standard RDF has been chosen to replace the former XML format of ESE. RDF offers a way of modelling objects by enabling statements about objects in form of triples. In the case of information about CHOs, this may look as follows:

```
ex:MonaLisaPaintingex:wasPaintedByex:LeonardoDaVinci .
ex:MonaLisaPaintingex:hasLocationex:Louvre .
```

Example 2: Exemplary triple statements about the Mona Lisa painting in RDF. "ex" stands for an example namespace of the resources.

Example 2 illustrates that RDF statements can be compared to simple sentences consisting of a subject, a predicate and an object: "The Mona Lisa painting was painted by Leonardo da Vinci" and "The Mona Lisa painting is located in the Louvre" are translations in natural language of the RDF triples. Information expressed in this way has a graph structure (as opposed to a tree structure in the case of XML) as every part of the RDF triple can be a resource which can be further described by another statement. A resource can be everything as long as it is represented on the Web. In the case of the Mona Lisa painting example above, this could be a photograph of the painting, a video documentation about the life of Leonardo da Vinci or the website of the Louvre where the painting is exhibited but also be a more abstract thing like the concept of the oil painting technique or the era of the Renaissance. With the use of resources, it is possible to make statements on the Web about any entity in the world no matter if it is a native Web entity, part of the physical world or just

[8] For additional considerations of the functional potential of the EDM cf. Gradmann (2010).
[9] Moving to new Europeana Data Model. Europeana Pro blog: http://pro.europeana.eu/web/guest/pro-blog/blogs/1836825 [26.11.2013].

an abstract concept. Coming back to the Mona Lisa example, the resource *ex:LeonardoDaVinci* could be specified in this way and by saying that Leonardo da Vinci lived during the Renaissance (see *example 3*).

```
ex:MonaLisaPaintingex:wasPaintedByex:LeonardoDaVinci .
ex:LeonardoDaVinciex:livedDuringex:Renaissance .
```

Example 3: An additional triple statement about the da Vinci resource in RDF.

Resources are represented with URIs (Uniform Resource Identifiers)which are the stable identifiers and part of the Linked Data Rules. Berners-Lee, Fielding & Masinter (2005:3) describe the function of URIs as follows: "A Uniform Resource Identifier (URI) provides a simple and extensible means for identifying a resource". They have a specific build-up: "The URI syntax is organized hierarchically, with components listed in order of decreasing significance from left to right" (Berners-Lee, Fielding & Masinter, 2005:10). HTTP URIs, which are mostly known from the Web, are one example for URIs. URIs as part of RDF statements are based on HTTP and consist typically of the namespace URI of the used vocabulary and an identifier or the name of the schema resource. A Europeana URI for an individual described by the EDM is defined as follows:

`http://data.europeana.eu/data/item/08501/E83B976C1E5BDFDB78F600B5206AD1FE300F141E`

All RDF resources described by the EDM are dereferenceable via the EDM namespace for individuals `http://data.europeana.eu/data/`[10] followed by a local identifier like `item/08501/E83B976C1E5BDFDB78F600B5206AD1FE300F141E`. Combined, they form the URI for a resource (in this case: "La Joconde – Gioconda & Mona Lisa"). The described resources are dereferenceable by their URIs regarding the Linked Data principles (Heath & Bizer, 2011:22). An HTTP GET request against the above mentioned URI leads either to the HTML representation of the resource for human users (if the request is coming from an usual browser) or it will return the raw RDF data for automated processing[11]. Instead of using resources, one can also use literals in the object position of an RDF triple which cannot be further described (apart from language tags or data type definitions).

In summary, RDF-based modelling offers much more flexibility for an open and distributed publishing as well as the reuse of data collections.

5 EDM's OWL representation

Ontologies as knowledge organisation systems are built on the RDF graph structure. An ontologyis used to store and retrieve data. A common definition of ontology is the one of "a specification of a conceptualization" (Gruber, 1995:908). Ontologies are made up of classes (concepts) and properties, which build the ontology schema, as well as of instances (individuals), the actual data. Logical elements or rules complete the possibilities that the construction of ontologies offer. They are suitable to represent complex structures in detail

[10] Within the EDM, the schema (classes and properties, see the section "EDM's OWL representation" for details) and the individuals (the provided data) are distinguished by using different name spaces. The data namespace is shown above, the schema namespace is http://www.europeana.eu/schemas/edm/.

[11] Examples and documentation of the technical details can be found on the Europeana Pro website: http://pro.europeana.eu/tech-details [26.11.2013].

within a specific domain or across domains in case of upper ontologies. Ontologies can be seen as one specific view on a domain and therefore there can be several ontologies that may describe the same domain in a different way.

A shared data model is defined in order to provide a specification which describes how to represent a specific object. In the case of the EDM, this object is above all a Cultural Heritage Object. The EDM is modelled with the *Web Ontology Language* (OWL) (Hitzler, Krötzsch et al., 2012)[12]. The individual OWL resources of the EDM cannot be accessed via their respective URIs yet.

An important aspect with regard to functionality performed on the data is that not only a human reader can interpret this type of information but also an application. The EDM gives the opportunity to indicate that Da Vinci is a person with the use of the class *edm:Agent*. In the following, it is described how this is done and which other concrete classes and properties are offered by the EDM.

5.1 A closer look: Classes and properties in EDM

One of the main differences between both models is that the EDM makes use of classes. Classes in the EDM are above all used to distinguish between the CHO, its metadata aggregation (including information on the provider), agents and Web resources, places, time spans, physical things and concepts. The model reuses many different vocabularies like *OAI-ORE*[13], Dublin Core, *SKOS*[14] and *FOAF*[15]. It centres around the *edm:ProvidedCHO* class, the class for the described object. The EDM makes an extensive use of properties which are especially used to describe the provided object and the provider of the object. The complete class build-up of the EDM can be seen in *figure 1*.

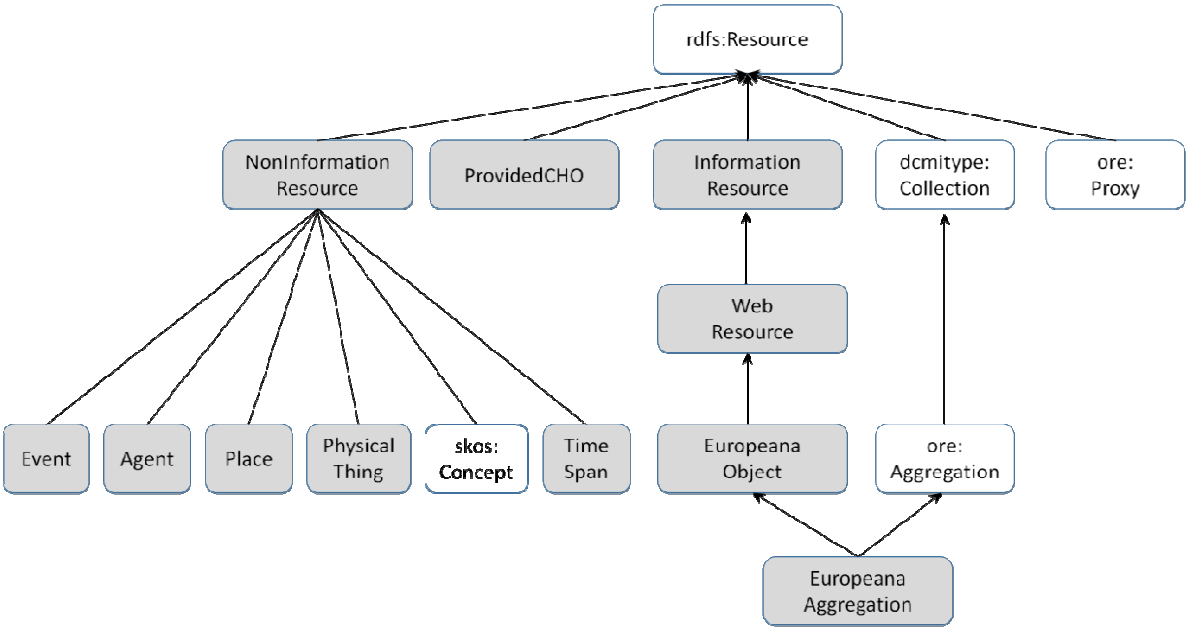


Figure 1: EDM class build-up (taken from Definition of the Europeana Data Model (2013:6).

[12] The latest OWL representation of the EDM on Europeana Labs: http://europeanalabs.eu/browser/europeana/trunk/corelib/corelib-solr-definitions/src/main/resources/eu/rdf/edm.owl [25.11.2013].
[13] OAI-ORE namespace: http://www.openarchives.org/ore/terms/ [04.12.2013].
[14] SKOS (Simple Knowledge Organization System) namespace: http://www.w3.org/2004/02/skos/core# [04.12.2013].
[15] FOAF (Friend of a Friend ontology) namespace: http://xmlns.com/foaf/0.1/ [04.12.2013].

The EDM has three core classes: *edm:ProvidedCHO*, *ore:Aggregation* and *edm:WebResource*. A closer look at these main classes makes clear that the most important resource is the one which represents the described object. This class is called *edm:ProvidedCHO*. In the presented example, the Mona Lisa painting is the described object. All the descriptive information about the painting like its creator, title and current location are stored at the class *edm:ProvidedCHO* (see *figure 2*).

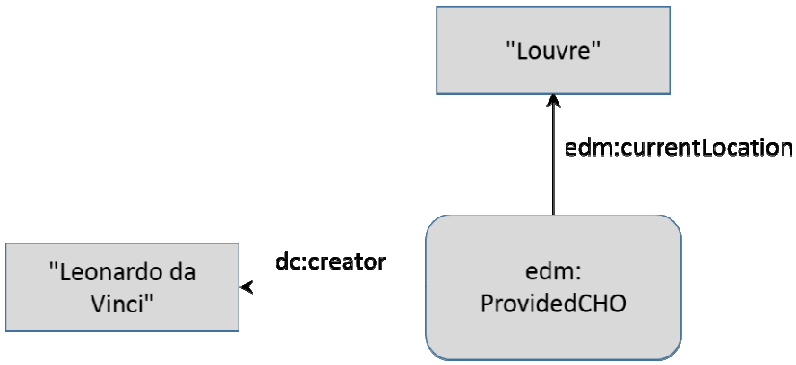


Figure 2: Example of the class edm:ProvidedCHO: The CHO is created by Leonardo da Vinci and its current physical location is the Louvre.

One speciality of the EDM is the explicit representation of the metadata record itself. Every time a data provider provides a description about an object in form of a metadata record to Europeana, a second resource is created for this metadata record. The class for this resource is *ore:Aggregation*[16]. If several objects are described in one metadata record, the information stored in the record is repeated, thus leading to exactly one *ore:Aggregation* class for each *edm:ProvidedCHO* class. *ore:Aggregation* is always connected to *edm:ProvidedCHO* via the property *edm:aggregatedCHO*.

ore:Aggregation holds together (or as the name already indicates: it aggregates) all information pieces which can be identified in a metadata record for one physical object including the metadata record itself, the described object and various kinds of views on the object. The class carries information about the provided metadata record itself, like the creator of the metadata record, its creation time and the times of updates, rights to the metadata or an image of the CHO, the provider who delivered the record to Europeana or the person who aggregated the data from the data provider and forwarded it to Europeana. In a heterogeneous context like Europeana this is very important, because you want to keep track of who provided what and when, in more general terms of provenance.

Attached to the *ore:Aggregation* is the class *edm:WebResource*. Web resources can be any kind of Web views of the objects. In the presented example case, this could be thumbnails showing the Mona Lisa painting, the HTML landing page of the Louvre website or a video showing the Mona Lisa painting. The property which links such Web resources to the aggregation is called *edm:hasView*. Similarly to *edm:ProvidedCHO* and *ore:Aggregation*, Web views can also be created as resources and thereby have explicit statements about them (see *figure 3*).

[16] "ore" is the prefix of the OAI-ORE namespace.

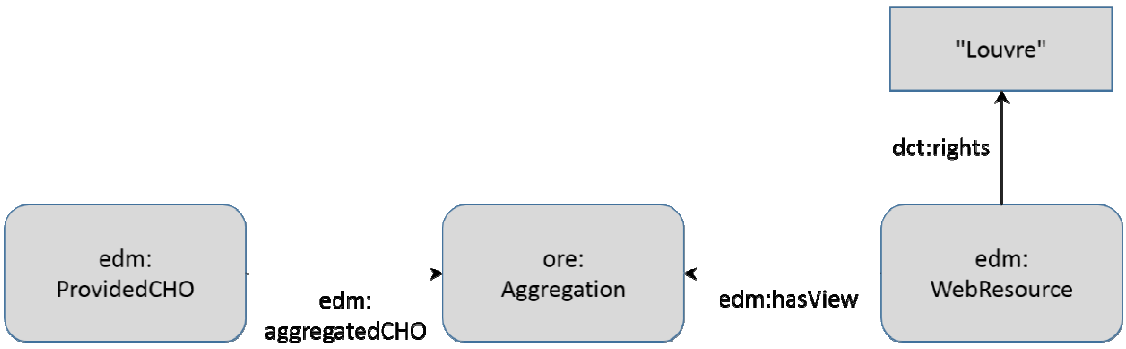


Figure 3: Example of the class relations within the Europeana Data Model

5.2 The upper layer of the EDM: Proxies

The institution of the class*ore:Proxy* is a special case which adds much to the flexibility of the model. The proxy is used to keep track of provenance of all statements about an object made by a specific provider. This is important for cases where there is more than one provider which creates statements about one and the same object. When a proxy about the object is created, all the descriptive metadata which was mapped to the CHO will then be mapped to the proxy. With this approach, it is possible to keep two different descriptions apart. Descriptions about one object may be contradictory. The EDM follows the open world assumption which is typical for RDF or OWL: Anyone can say anything about anything (Klyne& Carroll, 2004) including different statements on the same object that does not have to be consistently.

The ESE snippet from *example 1* is modelled as follows in EDM:

```
edm:aggregation/provider/08501/E83B976C1E5BDFDB78F600B5206AD1F
E300F141E
      aore:Aggregation ;
      edm:aggregatedCHOedm:item/
      08501/E83B976C1E5BDFDB78F600B5206AD1FE300F141E ;
      edm:dataProvider "Bildarchiv Foto Marburg" .
edm:item/08501/E83B976C1E5BDFDB78F600B5206AD1FE300F141E
      a edm:ProvidedCHO ;
      dc:creator "Leonardo da Vinci" ;
      dc:title "La Joconde – Gioconda& Mona Lisa" .
```

Example 4: The fictional ESE representation from example 1 modelled in EDM.

Compared to the ESE snippet, a difference between the metadata regarding the provider and the metadata of the actual CHO is made by using the classes *edm:ProvidedCHO* and *ore:Aggregation*.

But what about the properties which are used to make statements on the relation between resources? The EDM requires only a few mandatory properties like the kind of licence of the metadata, the title or description and the language of the object but does also add some new generic properties like *edm:hasView*,*edm:isRelatedTo* and reuses *owl:sameAs* for such modelling issues.

6 Specialising the EDM: The DM2E model

Taking a look on given classes and properties of the EDM, one might come to the conclusion that the EDM does not differ much from the ESE with respect to the genericity of the classes and properties and just extends the range of generic elements to describe a CHO. It may be true that the EDM is similar to the ESE in terms of class and property genericity because it is meant to function as a harmonisation schema for all different metadata standards represented in Europeana. A clear advantage compared to the former ESE is that the graph approach of RDF towards modelling information about resources provides an easy and flexible way of modelling and connecting complex data structures, e.g. as specialisations to the generic EDM. An example is the property *dc:creator* which is part of the EDM. It could be specialised by defining a more accurate sub-property like *ex:writer* or *ex:painter*. That allows us to say that Leonardo da Vinci is the painter (and not simply the creator) of the "Mona Lisa" which itself is a painting (and not simply a CHO).Example 5 shows how this is done by using a specialisation of the EDM.

```
dm2edata:aggregation/louvre/MonaLisa
      aore:Aggregation ;
      edm:aggregatedCHO dm2edata:item/louvre/MonaLisa ;
      edm:provider dm2edata:agent/louvre/Louvre .
dm2edata:item/louvre/MonaLisa
      a edm:ProvidedCHO ;
      dm2e:painter dm2edata:agent/louvre/LeonardoDaVinci ;
      dc:title "Portrait of the Mona Lisa"@en .
dm2edata:agent/louvre/louvre
      afoaf:Organization ;
      skos:prefLabel "Louvre"@en .
dm2edata:agent/louvre/LeonardoDaVinci
      afoaf:Person ;
      skos:prefLabel "Leonardo da Vinci"@en .
```

Example 5: Examples 1 and 4 modelled with the EDM specialisation made by DM2E (DM2E model): Da Vinci is not only creator but more specifically the painter of the Mona Lisa.

That is how the EDM enables the interoperability of heterogeneous metadata: It provides a generic level of description as an integration layer but allows at the same time more specialised layers of description beneath. The specialisation of the EDM which is used in *example 5* was created within the *Digitised Manuscripts to Europeana* (DM2E) project[17]. The specification of the DM2E model (Dröge, Iwanowa et al., 2013) facilitates semantically correct mappings on manuscripts as those from the *Wittgenstein Archive* or the *Deutsches Textarchiv*[18] that were previously not represented in Europeana. The aim of the model is to extend and enrich the EDM to enable a richer representation and a better semantic contextualisation for the provided manuscripts. The core EDM elements will not be changed, but extended by new resources like the classes *bibo:Book*[19], *bibo:Journal* or *dm2e:Manuscript*. These new classes are reused from other vocabularies whenever this is possible, analogue to the reusing practices in the EDM.Likewise the EDM, the DM2E model

[17] The DM2E project website: http://dm2e.eu [26.11.2013].
[18] The Wittgenstein Archives http://www.wittgensteinsource.org, Deutsches Textarchiv http://www.deutsches-textarchiv.de [26.11.2013].
[19] Bibo specification: http://bibliontology.com/specification [28.11.2013].

is more focussed on properties which are for example used to build more specific relations between the creator and the CHO (see the *dm2e:painter* property from *example 5*) or manuscript-related properties like *dm2e:incipit*which is used to indicate the opening words of a manuscript. The model does also make recommendations on data types (like the use of *xsd:datetime*), introduces specific URI build-ups for the most important contextual classes in the model, recommends which resources should be provided and encourages the provider to use as often as possible resources instead of literals in order to create rich RDF data. The latest stable model version can be accessed via the DM2E model schema namespace[20] and more detailed model specifications as well as mapping recommendations can be found in the DM2E wiki[21].At present, different data collections were mapped to the current revision of the DM2E model. These mappings will now be analyzed and will help to refine the model which was created in a bottom-up approach. The mapping output will be on the one hand rich EDM metadata that will be delivered to Europeana and on the other hand a Linked Open Data access point where users or applications can access the data, reuse it and thus build a continuously evolving semantic graph.

7 Linking to other resources

One of the most important merits of using the RDF approach for the EDM is the possibility to create semantic networks. One can easily assert any connection between any two resources and therefore it is possible not only to connect statements within or about one object, but also between different objects and any other resource entity which is described in RDF.

The most prominent example for a semantic network is the "Linked Open Data Cloud" (LOD Cloud)[22]where hundreds of free data sources are available summing up to trillions of triples, described in RDF and therefore any resource is connectable to any other RDF resource. Examples of data sources in the LOD Cloud are vocabularies about persons like *VIAF*[23], about places like *Geonames*[24] or about any kind of knowledge like *DBpedia*[25], which is the RDF representation of Wikipedia.

Related to the provided example in this paper, one could simply point to one of those resources in the LOD Cloud and thereby utilise, i.e. contextualise the provided object, all the knowledge which is attached to this resource. For example, a link to the entry about Leonardo da Vinci in VIAF could be created that would lead to additional information about Da Vinci which does not exist in the original record, like his dates of birth and death or the places he has worked at. This resource about Da Vinci in VIAF could be connected to other resources in other vocabularies. Currently, metadata records in Europeana are not strongly interlinked but this will change in the future – not least by projects like DM2E that are working on contextualising data.

[20] DM2E v 1.0 namespace: http://onto.dm2e.eu/schemas/dm2e/1.0/ [04.12.2013].
[21] DM2E wiki: http://wiki.dm2e.eu/Main_Page [04.03.2014].
[22] Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net [04.12.2013].
[23] Virtual International Authority File (VIAF) Data Source.http://viaf.org/viaf/data [04.12.2013].
[24] GeoNames Ontology http://www.geonames.org/ontology/documentation.html [21.11.2013].
[25] DBpedia: http://wiki.dbpedia.org/About [21.11.2013].

8 Conclusion

The creation of new functionalities enabled by the EDM allows users to formulate new queries, like "Give me all paintings which depict a specific person" or "Give me all paintings which are located in Paris and sort them by museum". Europeana is not yet able to precede such search queries among those data sets. In order to be able to provide the functionalities the EDM offers, first of all, Europeana has to store the data in triple stores and has to offer all data as RDF data via a Linked Open Data access point (the Europeana API[26] offers all EDM data, but not in RDF). However, it is important to keep in mind that the EDM is a data model for representing information in a certain way. It is another task to build new functionalities or applications on top of this data, if Europeana continues the technical development in the Linked Data direction.

The created specialisations of the EDM, developed by many other Europeana projects, cannot be displayed on Europeana yet and the presented DM2E model is still work in progress. The upcoming mappings to the model will show if the model has to be refined or if it meets the requirements. Basically, the model is able to describe the manuscripts provided to the DM2E project by its partners and is in some aspects very detailed (watermark, incipit etc.), whereas other aspects might still be specialised for manuscripts from other institutions which create metadata on manuscripts.

By the end of the project in January 2015, the DM2E data will be available in RDF under the CC0 1.0 license and can be processed via an SPARQL endpoint.

Acknowledgements

The authors are staff members of the DM2E research project at Humboldt-Universität zu Berlin which is funded by the European Commission under the "ICT Policy Support Programmes". We would like to thank all colleagues of DM2E and the Europeana Foundation for their helpful comments, as well as the European Commission for their support.

References

Berners-Lee, T. (2006).Linked Data - Design Issues.W3C Website. http://www.w3.org/DesignIssues/LinkedData [21.11.2013].

Berners-Lee, T., Fielding, R., &Masinter, L. (2005). Uniform Resource Identifier (URI): Generic Syntax. http://tools.ietf.org/html/rfc3986 [21.11.2013].

Buil-Aranda, C., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Harris, S., Hawke, S. et al. (2013). SPARQL 1.1 Overview.W3C Recommendation.W3C Website. http://www.w3.org/TR/sparql11-overview [21.11.2013].

Concordia, C., Gradmann, S., &Siebinga, S. (2010). Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. In *IFLA Journal*, 36(1):61-69.

Definition of the Europeana Data Model v5.2.4 (2013). Europeana Professional Website. http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5 [06.12.2013].

[26] Europeana API: http://pro.europeana.eu/api [04.12.2013].

Dröge E., Iwanowa, I., Hennicke, S., & Eckert, K. (2013). DM2E Model V1.0 Specification.Europeana Professional Website. http://pro.europeana.eu/documents/1044284/0/DM2E+Model+V+1.0+Specification [05.11.2013].

Duval, E., Hodgins, W., Sutton, S., Weibel, S.L. (2002). Metadata Principles and Practicalities.In *D-Lib Magazine*, 8 (4).D-Lib Website. http://www.dlib.org/dlib/april02/weibel/04weibel.html [26.11.2013].

Europeana Data Model Mapping Guidelines v2.0 (2013).Europeana Professional Website. http://pro.europeana.eu/documents/900548/60777b88-35ed-4bae-8248-19c3696b81fb [06.12.2013].

Europeana Data Model Primer, v14/07/2013 (2013). Europeana Professional Website. http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5 [06.12.2013].

Europeana Semantic Elements Specification and Guidelines v17/07/2013 (2013).Europeana Professional Website. http://pro.europeana.eu/documents/900548/2eee7beb-b9d8-4532-a089-8e8d6df38ce7 [06.12.2013].

Gradmann, S. (2010). Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana *(Europeana White Paper No. 1)*. Europeana Professional Website. http://pro.europeana.eu/c/document_library/get_file?uuid=cb417911-1ee0-473b-8840-bd7c6e9c93ae&groupId=10602 [04.12.2013].

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. In *Int. J. Human-Computer Studies*, 43 (5-6): 907-928.

Haslhofer, B., & Klas, W. (2010).A survey of techniques for achieving metadata interoperability. In *ACM Computing Surveys*, 42 (2):1–37.

Heath, T., & Bizer, C. (2011).*Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology (Bd. 1). Morgan & Claypool.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). OWL 2 Web Ontology Language. Primer (Second Edition). W3C Website. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/ [25.11.2013].

Hyvönen, E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology (Bd. 3). Morgan & Claypool.

Isaac, A., & Haslhofer, B. (2013). Europeana Linked Open Data - data.europeana.eu. Semantic Web 4 (3):291-297.

Klyne, G., & Carroll, J.J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Website. http://www.w3.org/TR/rdf-concepts/ [25.11.2013].

Manola, F., & Miller, E. (2004). RDF Primer. W3C Recommendation. W3C Website. http://www.w3.org/TR/rdf-primer [21.11.2013].

NISO (2004).Understanding metadata.National Information Standards Organization (NISO). http://www.niso.org/publications/press/UnderstandingMetadata.pdf [26.11.2013].

Fachinformationssysteme als Basis der Wissenschaftsforschung. Verwendungspotentiale infrastruktureller Dienste für ein indikatorenbasiertes Forschungsmonitoring am Beispiel der Bildungsforschung

Ute Sondergeld & Andreas Kempf

Der Beitrag geht der Frage nach, inwieweit sich Fachinformationssysteme für indikatorenbasierte Monitoringdienste eignen. Vor dem Hintergrund der zunehmenden Beobachtung und Bewertung von Forschungsaktivitäten werden am Beispiel der Bildungsforschung die gegenseitigen Abhängigkeiten zwischen Indikatorenbildung, der Beschaffenheit von Datengrundlagen und den Charakteristika des Forschungsfeldes aufgezeigt. Auf Basis von Daten des Projekts „Monitoring Bildungsforschung“ werden Befunde zu den Indikatoren Drittmittel, Förderquellen, Forschungskooperation, Mehrautorschaft und Peer Review-Veröffentlichungen präsentiert und hinsichtlich ihrer Aussagekraft diskutiert. Hieraus werden Implikationen für den Aufbau und die Verwendung infrastruktureller Dienste zur Beobachtung von Forschungsaktivitäten abgeleitet.

1 Einführung

Der Informationsbedarf zu Forschungsaktivitäten wächst kontinuierlich. Waren Hochschulen anfangs lediglich dazu angehalten, ihre Forschungsberichte so abzufassen, dass sie eine gewisse Vergleichbarkeit ermöglichten (WR, 1980), legte der Wissenschaftsrat als wichtigstes wissenschaftspolitisches Beratungsgremium in Deutschland Anfang letzten Jahres mit dem Ziel einer Standardisierung und Qualitätssicherung ein sogenanntes *Kerndatenset Forschung* vor (WR, 2013). In nahezu allen Forschungseinrichtungen gehören die Ermittlung und Bereitstellung von Daten zu Forschungsaktivitäten mittlerweile zum festen Bestandteil des eigenen Aufgabenportfolios. Die Verwendungsformen dieser Daten sind vielfältig. Neben institutionellen Bewertungsverfahren, wie etwa Akkreditierungen und Hochschulrankings, gehören in- und externe Steuerungsprozesse ebenso dazu wie die Berichterstattung an Mittelgeber, die amtliche Statistik sowie die Selbstbeschreibung der Wissenschaft. Die Auswertung dieser Daten erfolgt üblicherweise durch Indikatoren, die in Informationssystemen bereitgestellte Daten verdichten und deren Aussagekraft sowohl von der Beschaffenheit und Qualität der zugrundeliegenden Daten als auch von den Charakteristika des Untersuchungsgegenstandes abhängig ist. Zu Beginn in erster Linie für die Beschreibung von Wachstumsverläufen in der Wissenschaft verwendet, wurden sie im Zuge des erhöhten wissenschaftspolitischen Steuerungsbedarfs in den 1970er Jahren zunehmend als Leistungsindikatoren eingesetzt (Hornbostel, 1997). Dabei werden Indikatoren auch auf Datenbanken angewendet, die ursprünglich nicht für diesen Verwendungszusammenhang konzipiert wurden. Hieraus kann sich ein Spannungsverhältnis ergeben (Keiner, 2006, Dees & Rittberger 2009), das sowohl auf Seiten der Datenproduzenten als auch der -anwender Fragen nach der Eignung von Daten und Indikatoren, der Notwendigkeit ihrer Anpassung und nach ihrer Reichweite bzw. Aussagekraft aufwirft.

Vor diesem Hintergrund geht der vorliegende Beitrag am Beispiel des interdisziplinären Wissenschaftsfeldes der Bildungsforschung der Frage nach, inwiefern sich infrastrukturelle Dienste für eine indikatorenbasierte Beschreibung von Forschungstätigkeiten eignen. Anhand der in der Wissenschaftsforschung und -evaluation gebräuchlichen und wichtigen

Standardindikatoren Finanzierung, Kooperation und Peer Review-Status von Publikationen werden zeitliche Entwicklungen der Bildungsforschung aufgezeigt und die Aussagekraft dieser Indikatoren diskutiert. Grundlage hierfür bildet das Projekt "Monitoring Bildungsforschung (MoBi)"[1], das mit einem Schwerpunkt auf den Kerndisziplinen Erziehungswissenschaft, Soziologie und Psychologie eine indikatorenbasierte Analyse des Feldes der Bildungsforschung auf Basis von Forschungsprojekten und der aus ihnen hervorgegangenen Publikationen ebenso zum Ziel hat wie die Konzipierung eines kontinuierlichen Dienstes zur Beobachtung des Forschungsfeldes.

2 Forschungsfeld und Forschungsindikatoren

Insbesondere durch die Ergebnisse internationaler Vergleichsstudien wie PISA sind Fragen des Bildungserfolgs und seiner Voraussetzungen ins Zentrum des gesellschaftlichen und bildungspolitischen Interesses gerückt. Um wissenschaftlich fundierte Erkenntnisse über die Gelingensbedingungen von Bildung zu gewinnen, haben Forschungsförderer spezifische Förderprogramme zur Stärkung der Bildungsforschung ergriffen (Buchhaas-Birkholz, 2010). Die Einwerbung von Forschungsgeldern spielt für Forschungseinrichtungen zunehmend eine wichtige Rolle, da die Forschungsförderung seit den 1990er Jahren zunehmend über Drittmittel bei einem gleichzeitigen Rückgang der Grundfinanzierung von Forschungseinrichtungen gesteuert ist und deren erfolgreiche Akquise als Leistungsindikator in Evaluationen eingeht (Hinze, 2010). Die zunehmende Einwerbung von Drittmitteln bei anderen als den traditionellen Forschungsförderern ist eine Auswirkung dieser Entwicklung (Schmoch, 2008).

Ein weiteres forschungspolitisches Ziel bildet die Stärkung kooperativer Forschung. In Reaktion auf den Bericht der Internationale Kommission zur Systemevaluation der Deutschen Forschungsgemeinschaft und der Max-Planck-Gesellschaft (Krull, 1999), in dem die Segmentierung des Forschungssystems als eines der großen Probleme des deutschen Wissenschaftssystems identifiziert wurde, haben Wissenschaftsrat und die am Pakt für Forschung und Innovation sowie der Exzellenzinitiative beteiligten Körperschaften eine Reihe von Instrumenten zur Förderung personenbezogener und institutioneller Kooperation sowie thematisch fokussierter Konsortien entwickelt (WR, 2000, 2013), die sich auch in der Bildungsforschung niederschlagen (BMBF, 2007). Entsprechend dem Bedeutungszuwachs von Kooperation auf der Ebene von Forschungsprojekten wurde in einer Vielzahl von Studien auf den Zuwachs kollaborativen Publizierens hingewiesen (exemplarisch Parthey, 2006), der auch für die Sozialwissenschaften bestätigt wurde (Wuchty, Jones & Uzzi, 2007). Dabei wirken innerhalb der Bildungsforschung unterschiedliche disziplinäre Publikationstraditionen: während in Erziehungswissenschaft und Soziologie der Einzelautorschaft höhere Bedeutung zukommt, spielt die Gruppenautorschaft in der Psychologie durch ihre Nähe zu den Naturwissenschaften eine größere Rolle (Hicks, 2005; Piro, Asknes & Rørstad, 2013).

Solche disziplinspezifischen Publikationsmuster zeigen sich auch in der Präferenz verschiedener Publikationstypen und in der Veröffentlichung von Beiträgen in begutachteten Zeitschriften. Während sich die Psychologie generell durch eine Bevorzugung von Zeitschriftenpublikationen und Veröffentlichungen in Zeitschriften mit Peer Review

[1] Das Projekt mit dem offiziellen Titel „Entwicklung und Veränderungsdynamik eines heterogenen sozialwissenschaftlichen Feldes am Beispiel der Bildungsforschung“ wird von der Leibniz-Gemeinschaft im Rahmen des SAW-Verfahrens von 05/2011-07/2014 gefördert und kooperativ von folgenden Einrichtungen durchgeführt: GESIS – Leibniz-Institut für Sozialwissenschaften; Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID); Institut für Forschungsinformation und Qualitätssicherung (iFQ); Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

auszeichnet (Mutz, 2012; Hicks, 2005), sind die Publikationsmuster von Soziologie und Erziehungswissenschaft von einer breiteren Streuung gekennzeichnet: hier spielen Sammelwerksbeiträge und Monographien eine bedeutende Rolle (Dees, 2008; Krause, Riege, Stahl & Zens, 2009; WR, 2008).

Aus den vorangegangenen Überlegungen lassen sich folgende Hypothesen ableiten:

1. Drittmittelfinanzierte Forschung nimmt im Zeitverlauf zu; Drittmittel fließen zunehmend auch von EU und Stiftungen.
2. Kooperationsprojekte nehmen im Zeitverlauf zu. Ebenso steigt der Anteil an Mehrautorschaften, die in der Psychologie einen höheren Anteil als in den beiden anderen Disziplinen ausmachen.
3. Psychologische Publikationen werden häufiger in Zeitschriften sowie in Zeitschriften mit Peer Review-Verfahren veröffentlicht. Generell steigt der Anteil begutachteter Zeitschriftenbeiträge über die Zeit.

Bei der Anwendung von Indikatoren zur Beschreibung und Evaluation von Forschung gilt es, sich bewusst zu machen, dass diese keine objektiven Wahrheiten abbilden. Ihre Konstruktion geht mit bestimmten Wert- und analytischen Entscheidungen einher, die beeinflussen, was von einem Indikator erfasst wird (Weingart & Winterhager, 1984). Diese betreffen u.a. die Eigenschaften und die Beschaffenheit des Untersuchungsgegenstandes. So gehen aus Befunden quantitativer Analysen Varianzen sowohl zwischen als auch zum Teil innerhalb der Disziplinen hervor (Hornbostel, 2001). Block, Hornbostel und Neidhardt (1992) zeigen anhand des Drittmittelindikators, dass die Üblichkeit von Drittmittelforschung in den jeweiligen Disziplinen entscheidenden Einfluss auf die Anwendbarkeit und Bewertung eines Indikators hat. Die Varianz innerhalb von Disziplinen beschreibt Hornbostel (2001) am Beispiel der disziplinären Eigenheit der Erziehungswissenschaft: Diese bestehe neben geisteswissenschaftlichen und sozialwissenschaftlich-empirischen Wissenschaftstraditionen auch aus einem auf praktisch relevante Dienstleistungen spezialisierten Teil. Eine binnendisziplinäre Teilung in eine naturwissenschaftlich-experimentelle und eine geisteswissenschaftliche Ausrichtung, aufgrund derer keine eindeutige Zuordnung zu Natur- oder Geisteswissenschaften vorgenommen werden kann, gilt auch für die Psychologie (Jüttemann, 2006). Solche Unschärfen und Heterogenitäten haben Auswirkungen auf die indikatorengestützte Darstellung einer Disziplin, deren Aussagekraft umso weniger präzise erscheint, je breiter das Spektrum eines Faches ist (Hornbostel, 2001). Für ein interdisziplinäres Forschungsfeld wie die Bildungsforschung, dessen beteiligte Kerndisziplinen Erziehungswissenschaft, Soziologie und Psychologie von derart verschiedenartigen fachdisziplinären und intradisziplinären Forschungs- und Kommunikationstraditionen geprägt sind, gilt dies in besonderer Weise.

In der praktischen Anwendung von Indikatoren werden nicht nur aus mehreren verschieden gewichteten Kennzahlen zusammengesetzte Gesamtindikatoren genutzt, die Konzepte wie z.B. Forschungsqualität im Forschungsrating Soziologie abbilden sollen (WR, 2006). Weit verbreitet ist auch die Verwendung weniger komplex gebildeter Indikatoren, die Sachverhalte in absoluten oder relationalen Maßzahlen beschreiben (Meyer, 2004). So misst etwa einer der im ZPID-Monitor zur Internationalisierung der Psychologie genutzten Indikatoren den Anglisierungsgrad mithilfe des prozentualen Anteils englischsprachiger Veröffentlichungen an den in PSYNDEX insgesamt erfassten Fachpublikationen (Schui, Hoffmann & Krampen, 2013). Die Bindung an vorliegende Metadatenstrukturen führte auch in MoBi zur Nutzung solcher einfachen Indikatoren.

3 Methodischer Ansatz und Operationalisierung

3.1 Datengewinnung und Auswertungsmethodik

Aus der Datenbank SOFIS (Sozialwissenschaftliches Forschungsinformationssystem)[2] wurden alle in den Jahren 1995 bis 2009 abgeschlossenen Forschungsprojekte ausgewählt, in denen bildungsrelevante Themen untersucht wurden und an denen mindestens ein Forschungsinstitut mit Sitz in Deutschland beteiligt war. SOFIS enthält systematisch erfasste und nach einem einheitlichen Schema erschlossene Forschungsprojekte verschiedener sozialwissenschaftlicher Fachdisziplinen sowie weiterer interdisziplinärer Gebiete der Sozialwissenschaften und unterscheidet sich damit von Projektdatenbanken anderer Anbieter, die ein auf einzelne Körperschaften bezogenes oder ein regional begrenztes Sammelprofil haben. Die Auswahl der bildungswissenschaftlichen Projekte basierte auf der klassifikatorischen Inhaltserschließung der Projekte und erfolgte teils automatisch für die Disziplinbereiche Erziehungswissenschaft, Medienpädagogik und Bildungssoziologie, teils durch die intellektuelle Kontrolle von Projekten weiterer relevanter Bereiche wie z.B. Soziologie, Psychologie, Demografie, Berufsforschung. Zusätzlich wurden relevante Projekte anhand bildungsspezfischer Schlagwörter identifiziert. Dadurch ergab sich ein Korpus von 9119 Projekten, der durch den Einbezug von 20 bis zum Jahr 2009 abgeschlossener Projekte aus dem *Rahmenprogramm zur Förderung empirischen Bildungsforschung* des BMBF[3] auf insgesamt 9139 Forschungsprojekte erweitert wurde. Die Auswertung der Projekte erfolgte über Kreuztabellen mittels der χ^2-Statistik, das Alphafehler-Niveau konnte aufgrund des Stichprobenumfangs auf α=.01 festgelegt werden. Die Stärke des Zusammenhangs zweier Merkmale ist mit Cramers V (CrV) angegeben, welches als Effektstärkenmaß interpretierbar ist.

Zur Bildung des Publikationskorpus wurde für die geplanten Varianzanalysen aus dem Projektset nach Definition der gewünschten Teststärke von .90 für einen mittleren Effekt (Faul, Erdfelder, Buchner & Lang, 2009) eine Stichprobe von 270 Projekten zufällig ausgewählt. Als unabhängige Variablen wurden die *disziplinäre Zugehörigkeit* (Soziologie, Erziehungswissenschaften und Psychologie), *Finanzierungsart* (Eigenprojekt, staatliche Finanzierung und wissenschaftsautonome Finanzierung) sowie zwei weit auseinanderliegende *Zeiträume* (T1=1995-1997, T2=2006-2008) festgelegt. Im Autopsieverfahren wurde für jedes Projekt das gesamte Publikationsoutput recherchiert und nach einem einheitlichen, neu konzipierten Metadatenschema erfasst. Recherchiert wurden Publikationen von Projektbeteiligten, die im Zeitraum von Projektbeginn bis zu drei Jahre nach Projektende publiziert wurden. Aufgenommen wurden die Publikationstypen Zeitschriften- und Sammelwerksbeiträge, Bücher (Monographien, Sammelwerke, selbständige Forschungsberichte, Qualifizierungsarbeiten) sowie Abstracts aus Konferenzbänden. Auch Graue Literatur wurde erfasst. Von den insgesamt 1925 erfassten Publikationen aus 270 Forschungsprojekten waren 56,96% in den drei disziplinspezifischen Fachdatenbanken (SOLIS, FIS Bildung und PSYNDEX) der Projektpartner nachgewiesen. Die Metadaten aller anderen Publikationen wurden nachträglich erhoben. Bei einem Alphafehler-Niveau von α=.05 wurden mehrfaktorielle Varianzanalysen durchgeführt und Effektstärken (η^2) berechnet, wobei Post-hoc-Unterschiede nach Tukey-HSD berechnet wurden. Zur Validierung der Varianzanalyse wurden zusätzliche nonparametrische statistische Tests (Kruskal-Wallis-Test für unabhängige Stichproben) für jeden geprüften Haupteffekt berechnet. Die jeweiligen Testwerte bestätigen die varianzanalytischen Ergebnisse und lassen keine Verzerrung durch schief verteilte Outputwerte vermuten. Da Varianzanalysen

[2] http://www.gesis.org/unser-angebot/recherchieren/sofis/
[3] http://www.empirische-bildungsforschung-bmbf.de/

mit aggregierten Daten auf Projektebene nicht für alle Fragestellungen geeignet waren, wurden solche als kategoriale Daten zu den Publikationen aller Projekte in Kreuztabellen mithilfe der χ^2-Statistik ausgewertet.

3.2 Operationalisierung

3.2.1 Finanzierung

Auf Basis der Angaben im SOFIS-Erhebungsbogen und der Kennzeichnung einer Institution als finanzierende Einrichtung wurden die Projekte als eigenfinanziert und drittmittelfinanziert klassifiziert. Nur für Drittmittelprojekte wurden anhand von Informationen zu den finanzierenden Institutionen die wichtigen Fördergruppen EU, Bund, Länder, DFG und Stiftungen identifiziert. Bei der Analyse des Drittmittelindikators wird üblicherweise das Drittmittelvolumen einbezogen und am wissenschaftlichen Personal relativiert; eine solche Differenzierung war aufgrund fehlender oder nicht verlässlicher Inputdaten jedoch nicht möglich.

3.2.2 Kooperation

Projekte Als Kooperationsprojekt wurde jedes Projekt klassifiziert, das von mehr als einer Forschungseinrichtung durchgeführt wurde. Abweichend von der üblichen Berechnungsweise zur Ermittlung institutioneller Forschungskooperation wurde aufgrund fehlender Informationen nicht zwischen standortübergreifenden und standortinternen Kooperationen (z.B. zwischen verschiedenen Fachbereichen einer Universität) differenziert. Diejenigen Projekte, die der wissenschaftlichen Qualifizierung dienen und nicht im Rahmen eines größeren Projektes durchgeführt wurden, wurden bei der Berechnung ausgeschlossen.

Publikationen Zur Ermittlung von Mehrautorschaften wurde die durchschnittliche Anzahl der an einer Publikation beteiligten Personen gemessen. Angaben zur institutionellen Zugehörigkeit von Autoren standen nicht zur Verfügung und konnten nicht nacherhoben werden. Um einen Einfluss der Projektgröße auszuschließen, wurde als Schätzwert die Größe der Forschergruppe anhand der Anzahl unterschiedlicher Autoren, die im Rahmen eines Projektes publizierten, einbezogen. Allerdings konnten keine Informationen zum Umfang der Mitarbeit einzelner Personen berücksichtigt werden. Die in SOFIS nur ungenau vorhandenen Informationen über die Anzahl von Projektbeteiligten wurden hier nicht übernommen, weil die beteiligten Personen durch das Autopsieverfahren zuverlässiger bestimmt werden konnten.

3.2.3 Peer-Review

Für alle Zeitschriften wurde der Begutachtungsstatus möglichst zum jeweiligen Veröffentlichungszeitpunkt nachträglich unter Zuhilfenahme der Datenbank DEPOT[4], der Zeitschriftenlisten von PSYNDEX[5] und SOLIS[6] sowie durch Recherchen auf Verlagshomepages erhoben. Dabei wurde lediglich unterschieden, ob eine Zeitschrift ein Begutachtungsverfahren durchführt oder nicht; eine Differenzierung nach Verfahren wurde nicht vorgenommen. Es wurden auch solche Zeitschriften als begutachtet klassifiziert, die von

[4] http://www.fachportal-paedagogik.de/depot/zeitschriften.html
[5] http://www.zpid.de/index.php?wahl=products&uwahl=frei&uuwahl=journals_z
[6] www.wiso-net.de/download/SOLI.pdf

einem großen, möglichst international besetzten Herausgebergremium betreut sind. Die Befunde werden als Anteile von Artikeln aus Peer-Review-Zeitschriften an der Gesamtzahl der Zeitschriftenartikel eines Projektes angegeben. Da 90 Projekte keine Zeitschriften-veröffentlichungen nachweisen konnten, reduzierte sich die Anzahl der Projekte auf N=180.

4 Ergebnisse

4.1 Forschungsfinanzierung

Von den 9139 Forschungsprojekten werden insgesamt etwa zwei Drittel über Drittmittel finanziert (64,7%); bei 17,5% aller Projekte handelt es sich um Eigenprojekte, für 17,8% liegen keine Angaben vor. Im Zeitverlauf ist eine signifikante Zunahme drittmittel-geförderter Projekte und eine signifikante Abnahme der Eigenprojekte beobachtbar (Gesamtwerte: $\chi^2_{(.01;\ 8;\ N=9139)} = 127{,}11$; CrV= .08).

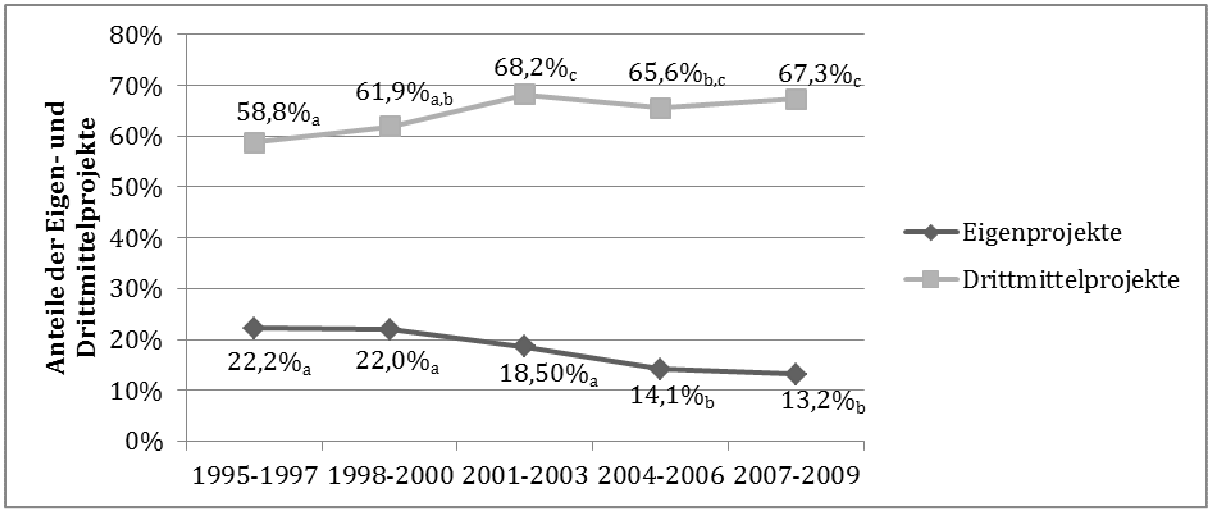


Abbildung 1: Anteile verschiedener Förderarten im Zeitverlauf.

Anmerkung. Die tiefgestellten Buchstaben hinter den Prozentwerten entsprechen jeweils homogenen Gruppen bei post-hoc-Tests innerhalb jeder Zeile.

Die fünf großen Fördergruppen finanzieren zusammen 67,3 Prozent (3980) aller 5915 Drittmittelprojekte. Im Zeitverlauf steigen die Förderanteile der EU vom ersten zum zweiten Zeitblock und verbleiben im weiteren Verlauf auf ähnlichem Niveau ($\chi^2_{(.01;\ 4;\ N=3980)} = 29{,}23$; CrV= .07), während der Förderanteil des Bundes abgesehen von einem Ausreißer im 2. Zeitblock über die Jahre konstant bleibt ($\chi^2_{(.01;\ 4;\ N=3980)} = 34{,}47$; CrV= .08). Die Förderanteile der Länder sinken ab 2004 ($\chi^2_{(.01;\ 4;\ N=3980)} = 55{,}00$; CrV= .10), gleichzeitig steigen ab dieser Zeit die Anteile der Stiftungen ($\chi^2_{(.01;\ 4;\ N=3980)} = 35{,}28$; CrV= .08). Für die DFG zeigen sich im Zeitverlauf keine signifikanten Änderungen. Die Anteile der Finanzierergruppen gleichen sich im Zeitverlauf zunehmend an.

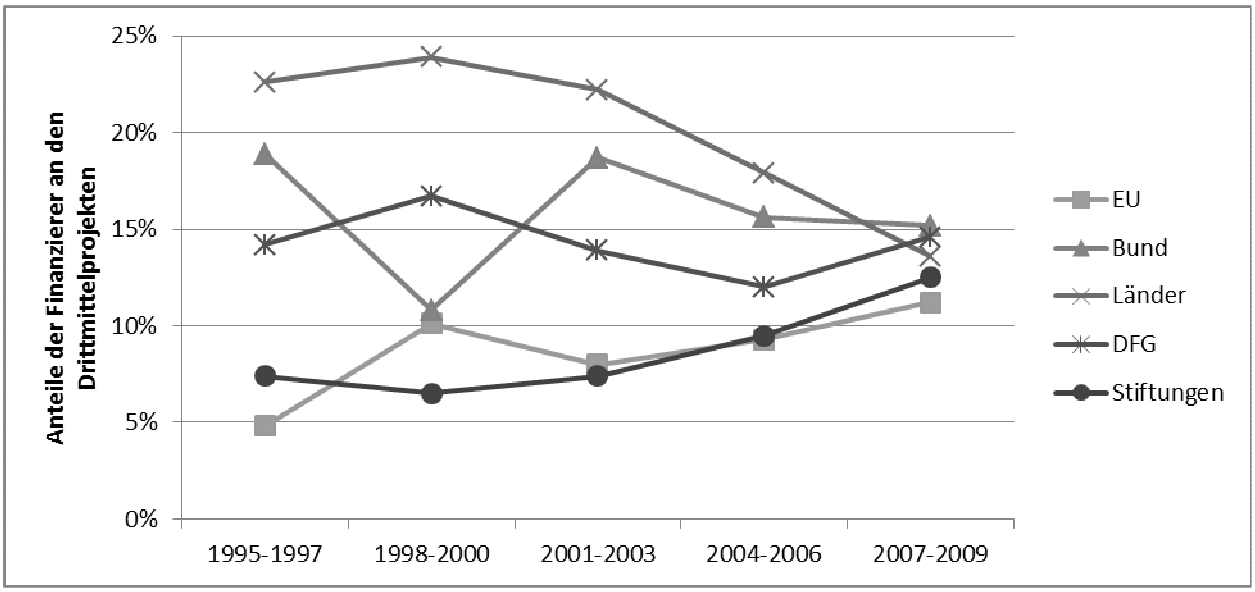


Abbildung 2: Anteile großer Geldgeber an der Drittmittelforschung im Zeitverlauf.

4.2 Kooperation

Projekte: 15,7 Prozent aller 7260 in die Analyse einbezogenen Forschungsprojekte werden in Kooperation durchgeführt. Über die Zeit hinweg steigt der Anteil der Kooperationsprojekte signifikant an ($\chi^2_{(.01;\ 4;\ N=7260)}$ = 119,45; CrV= .13). Dabei finden in der drittmittelgeförderten Forschung anteilig mehr Kooperationen statt als bei Eigenprojekten (17,4% vs. 9,9%; 15% keine Angabe ($\chi^2_{(.01;\ 1;\ N=6761)}$ = 48,56; CrV= .09).

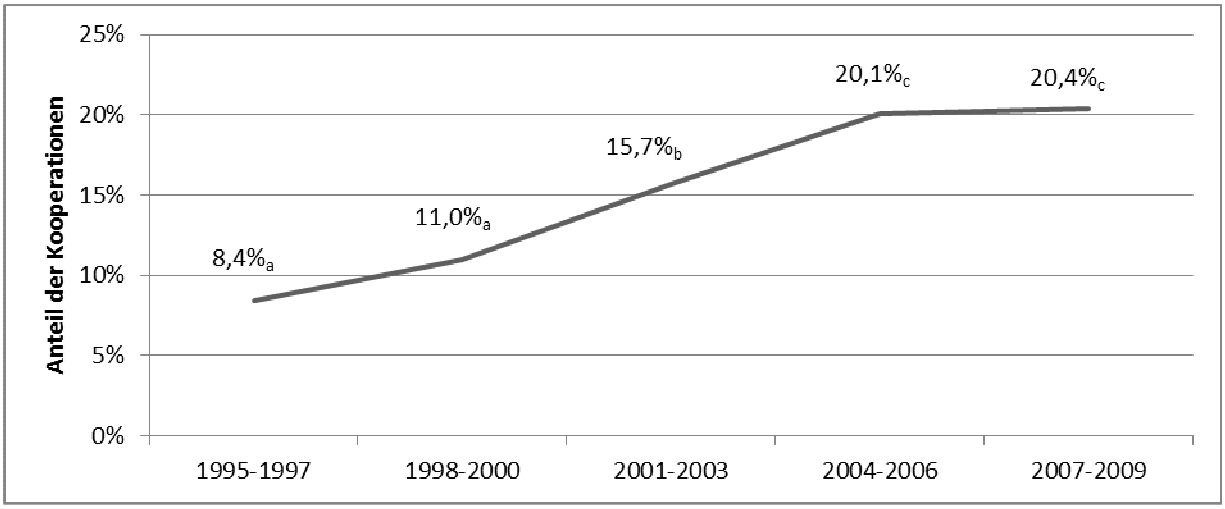


Abbildung 3: Anteil der Kooperationsprojekte im Zeitverlauf.

Anmerkung. Die tiefgestellten Buchstaben hinter den Prozentwerten entsprechen jeweils homogenen Gruppen im post-hoc-Test.

Publikationen: Während die Anzahl der an einem Projekt beteiligten Personen über die Zeit nicht ansteigt (keine Tabelle), zeigt sich in der Analyse der Mehrautorschaften ein signifikanter zeitlicher Effekt: Die durchschnittliche Anzahl der an einer Publikation beteiligten Personen erhöht sich von Zeitblock 1 zu Zeitblock 2 ($F_{(1,252)}$=10,83; p<.01;

η^2=.04). Hier zeigt sich auch ein Haupteffekt der Disziplin: Psychologische Projekte haben mit durchschnittlich 2,47 mehr beteiligte Autoren pro Publikation als die anderen Disziplinen ($F_{(2,252)}$=13,52; p<.01; η^2=.10).

	1995-1997	**2006-2008**	**Gesamt**
Soziologie	1,65	1,83	1,74 (SD=0,83)
Erziehungswissenschaft	1,51	2,17	1,84 (SD=1,01)
Psychologie	2,27	2,86	2,47 (SD=1,36)
alle Disziplinen	1,81 (SD=0,89)	2,23 (SD=1,30)	2,02

Tab. 1: Durchschnittliche Anzahl von Autoren pro Publikation in den verschiedenen Disziplinen im Zeitverlauf (SD=Standardabweichung)

4.3 Peer Review

Psychologische Projekte publizieren mit durchschnittlich 36,2 Prozent häufiger Zeitschriftenbeiträge als erziehungswissenschaftliche (28,9%) und soziologische Projekte (24,7%). Im disziplinären Vergleich sind die Anteile an Zeitschriftenbeiträgen mit Peer Review-Verfahren signifikant verschieden: während sich Erziehungswissenschaft und Soziologie mit 48,1 und 42,4 Prozent nicht voneinander unterscheiden, ist der Anteil von Zeitschriftenbeiträgen mit Peer Review-Verfahren an allen Zeitschriftenbeiträgen in der Psychologie mit 75,1 Prozent deutlich höher ($F_{(2,162)}$=10,58; p<.01; η^2=.12). Über die Zeit ist eine Zunahme von begutachteten Publikationen insgesamt jedoch nicht auszumachen; es zeigt sich lediglich ein statistischer Trend in der Psychologie, für die der Anteil von begutachteten Publikationen im Zeitverlauf von 62,0 auf 88,7 Prozent ($F_{(2,162)}$=2,94; p=.06; η^2=.04) ansteigt, während Zuwächse von 47,9 auf 49 Prozent in der Soziologie und von 41,6 auf 43,1 Prozent in der Erziehungswissenschaft nicht signifikant sind.

5 Diskussion

5.1 Charakteristika des Forschungsfeldes

Drittmittel: Die Befunde weisen darauf hin, dass die neuen Elemente der Wissenschaftssteuerung auch in der Bildungsforschung wirken. Neben Stiftungen spielt insbesondere die EU, deren Bedeutungszunahme für die Forschungsförderung auch Schmoch (2008) unterstreicht, eine wachsende Rolle. Die sich im Zeitverlauf zeigende Nivellierung der Förderaktivitäten verschiedener Geldgeber in der Bildungsforschung könnte auf eine zunehmende Dynamisierung des Feldes der Forschungsförderung verweisen. Obwohl die Indikatoren nicht in ihrer üblichen Weise berechnet werden konnten, lassen sie Aussagen über Strukturen der Forschungsförderung zu.

Kooperation: Der Anteil an kooperativ durchgeführten Forschungsprojekten innerhalb der Bildungsforschung steigt an; der höhere Anteil an Kooperationen in drittmittelgeförderter

Forschung könnte auf einen entsprechenden Schwerpunkt der Förderpolitik hinweisen. Der Anteil internationaler Kooperationen, die ebenfalls ein wichtiges Förderziel sind (Buchhaas-Birkholz, 2005; BMBF, 2007; WR, 2013), konnte nicht ermittelt werden, da Projektpartner aus dem Ausland in der zugrundeliegenden Datenbasis nicht systematisch erfasst werden.

Bei gleichbleibender Anzahl von Mitarbeitern pro Projekt nimmt die Anzahl an Autoren pro Publikation im Zeitverlauf deutlich zu, was für eine Bedeutungszunahme von Mehrautorschaften sprechen könnte. Dass diese Entwicklung in der Psychologie besonders ausgeprägt ist, kann als Hinweis auf ihre stärkere Orientierung an den Naturwissenschaften gewertet werden, innerhalb derer das kooperative Verfassen von Publikationen einen hohen Stellenwert einnimmt (Wuchty, Jones & Uzzy, 2007). Auf die Bedeutung individueller Autorschaft in der Erziehungswissenschaft und Soziologie weist die geringere Anzahl an Mehrautorschaften hin. In bibliometrischen Analysen werden oftmals nur diejenigen Publikationen als kooperativ bewertet, deren Autoren aus verschiedenen Forschungseinrichtungen stammen. Diese Methode der Indikatorenbildung scheint für die Ermittlung des Publikationsoutputs von Forschergruppen, die zum überwiegenden Teil aus nicht kooperativ durchgeführten Projekten entstanden sind, allerdings nicht zwingend erforderlich.

Peer Review: Erwartungskonträr ist keine Zunahme von Beiträgen in begutachteten Zeitschriften über die Zeit feststellbar; lediglich in der Psychologie, die hypothesenkonform in beiden Zeitblöcken den größten Anteil an Artikeln mit Peer-Review aufweist, zeigt sich ein leichter, aber nicht signifikanter Anstieg im Zeitverlauf. Die Psychologie scheint diesem Leistungsindikator generell ein stärkeres Gewicht beizumessen als die anderen Disziplinen, worauf auch der Anstieg erziehungswissenschaftlicher Zeitschriftenpublikationen bei einem konstanten Anteil an Peer Review-Zeitschriftenbeiträgen hinweist.

Die Ergebnisse legen nahe, dass der Indikator aufgrund der verschiedenartigen disziplinspezifischen Publikationstraditionen in der Bildungsforschung allein nicht geeignet ist, um das Publikationsverhalten dieses Feldes angemessen beschreiben oder auf dieser Grundlage Forschungsleistung bewerten zu können. Dass allein 90 der 270 Projekte keine Zeitschriftenveröffentlichungen aufweisen, legt die Schlussfolgerung nahe, dass diesem Publikationstypus in der Bildungsforschung insgesamt keine so große Relevanz zukommt wie in naturwissenschaftlichen Disziplinen.

5.2 Zur Methodik der Datenerhebung und -auswertung

Methodische Probleme bei der Projektanalyse ergeben sich aus der Repräsentativität des Datensatzes. Trotz des mehrstufigen und systematischen Erhebungsverfahrens der SOFIS-Redaktion ist von einem lückenhaften Projektbestand auszugehen, worauf schon Weishaupt und Merkens (2000) im ersten Datenreport Erziehungswissenschaft aufmerksam machen. Mit ihrem Erfassungsschwerpunkt auf sozialwissenschaftliche Disziplinen, denen die Psychologie nur zum Teil angehört, bleiben Projekte aus Entwicklungspsychologie, Kognitionsforschung oder Neurowissenschaften, für die eine vergleichbare Projektdokumentation nicht vorliegt, eher unberücksichtigt. Auch generelle Einschränkungen der Datenbasis wie über die Zeit veränderte Erfassungskriterien sowie unterschiedliche Feinheitsgrade und Qualitäten der Informationen durch differierende Indexierungskonsistenz spielen eine Rolle (Schui & Krampen, 2010; Dees & Rittberger, 2009).

Methodische Schwierigkeiten bei der Publikationsanalyse leiten sich aus der Stichprobengröße und Vollständigkeit der erhobenen Veröffentlichungen ab. Es ist zu vermuten, dass

trotz intensiver und umfänglicher Recherche relevante Publikationen nicht aufgefunden wurden, was vor allem Graue Literatur und Literatur des ersten Untersuchungszeitraumes betrifft. Auch eine Erfassung nicht relevanter Publikationen kann nicht ausgeschlossen werden. Die geringe Stichprobengröße von 270 Projekten führt zu schiefen Verteilungen der abhängigen Variablen durch Ausreißer-Projekte, weshalb die Effekte mit verteilungsfreien statistischen Verfahren gegengeprüft wurden. Eine Erweiterung der Stichprobengröße war aufgrund des Aufwandes einer Vollerhebung von Projektpublikationen jedoch nicht möglich.

6 Ausblick

Die Analyseergebnisse zeigen, dass auf Basis von infrastrukturellen Informationssystemen, die nicht primär zum Zweck der Forschungsbeobachtung aufgebaut wurden, Indikatoren zum Monitoring von Forschungsaktivitäten gebildet und wesentliche Aussagen zur Entwicklung eines Forschungsfeldes generiert werden können. Zugleich ist der Zusammenhang zwischen einerseits empirischer Grundlage sowie Spezifika des Untersuchungsgegenstandes und den andererseits daraus resultierenden Erkenntnismöglichkeiten deutlich geworden: je verschiedener die zugrundeliegenden Daten und kommunikativen Praxen innerhalb von Forschungsfeldern und Fachdisziplinen ausfallen, umso notwendiger ist die Adaption der verwendeten Indikatoren und umso eingeschränkter ist ihre Aussagekraft.

Auf Fachinformationssysteme, die einen wichtigen Bestandteil des wissenschaftlichen Forschungs- und Kommunikationsprozesses sichtbar machen, wird bereits als Datengrundlage in Evaluationsverfahren, wie etwa dem CHE Hochschulranking (FIS Bildung) oder dem Forschungsrating Soziologie (SOLIS), zurückgegriffen. Mit der veränderten Zielsetzung, in Zukunft verstärkt als Anbieter von Monitoringdienstleistungen zu fungieren, werden sich auch die Anforderungen an Datenbankproduzenten verändern. Um den Schritt von reinen Rechercheinstrumenten zur „sekundäranalytischen Nutzung“ (Herfurth, 1994, S.86) umsetzen zu können, rücken Fragen der Datenqualität in den Aspekten Datenbankinhalte, Datenstrukturen, Konsistenz und Genauigkeit der Datenerfassung in den Fokus, wie Dees und Rittberger (2009) am Beispiel der Literaturdatenbank FIS Bildung zeigen. Besondere Bedeutung kommt diesen Aspekten zu, wenn verschiedene Kollektionen zu Auswertungszwecken zusammengeführt werden, was insbesondere in Zeiten von Open Data mehr und mehr an Relevanz gewinnt.

Eine Entwicklung von Standards, wie sie vom Wissenschaftsrat angestoßen wurde (WR 2013) und die disziplin- bzw. feldbezogene Spezifika berücksichtigt, verringert die Schwierigkeiten bei der Zusammenführung von Daten zu Kollektionen und hat Auswirkungen auf die Bildung von Indikatoren und deren Aussagekraft. Für Datenbankproduzenten eröffnete eine Qualitätssteigerung, wie sie sich durch eine in dieser Form verstärkte Inputkontrolle erzielen ließe, neben einem verbesserten Retrieval zusätzliche Verwendungspotentiale für Evaluation und Wissenschaftsforschung.

Literatur

BMBF (2007). *Rahmenprogramm zur Förderung der empirischen Bildungsforschung*. Bonn: BMBF. URL: http://www.bmbf.de/pubRD/foerderung_der_empirischen_bildungsforschung.pdf [15.03.2014]

Block, H.-J., Hornbostel, S. & Neidhardt, F. (1992). *Drittmittel als Indikator der Hochschulforschung. Ergebnisse aus dem Forschungsprojekt „Vergleichende Bewertung von Leistungen der Hochschulen“*. Bonn: BMBF.

Buchhaas-Birkholz, D. (2009). Die „empirische Wende“ in der Bildungspolitik und in der Bildungsforschung: Zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft*, 20 (39), 27-33.

Dees, W. (2008). Transparenz und Evaluierbarkeit des erziehungswissenschaftlichen Publikationsaufkommens. Eine anwendungsorientierte Studie. *Erziehungswissenschaft* 19 (37), 27-32.

Dees, W. & Rittberger, M. (2009). Anforderungen an bibliographische Datenbanken im Hinblick auf szientometrische Analysen am Beispiel der FIS Bildung Literaturdatenbank. In: R. Kuhlen (Hg.), *Information: Droge, Ware oder Commons*? (S. 243-257). Boizenburg: Hülsbusch.

Faul, F., Erdfelder, E., Buchner, A. & Lang, A. -G. (2009). Statistical Power Analysis Using GPower 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41* (4), 1149-1160.

Herfurth, M. (1994). Voraussetzungen und Entwicklungsperspektiven scientometrischer Analysen auf der Grundlage von Datenbanken. In W. Neubauer (Hrsg.), *Deutscher Dokumentartag 1993: Qualität und Information* (S. 83-92). Frankfurt: DGD.

Hicks, D. (2005). The Four Literatures of Social Science. In F. Moed, W. Glänzel, & U. Schmoch (Hrsg.), *Handbook of Quantitative Science and Technology Research* (S. 473-496). Dordrecht: Kluwer.

Hinze, S. (2010). Forschungsförderung in Deutschland. In D. Simon, A. Knie, S. Hornbostel (Hrsg.), *Handbuch Wissenschaftspolitik* (S. 244-261). Wiesbaden: VS Verlag.

Hornbostel, S. (1997). *Wissenschaftsindikatoren. Bewertungen in der Wissenschaft*. Opladen: Westdeutscher Verlag.

Hornbostel, S. (2001). Wissenschaftsindikatoren: Mittel zur Selbstbeobachtung oder Schiedsrichter im Verteilungskampf? In E. Keiner (Hrsg.), *Evaluation (in) der Erziehungswissenschaft* (S. 23-42). Weinheim: Beltz.

Jüttemann, G. (2006). *Wilhelm Wundts anderes Erbe: Ein Missverständnis löst sich auf*. Göttingen: Vandenhoeck & Ruprecht.

Keiner, E. (2006). Erziehungswissenschaft, Forschungskulturen und die "europäische Forschungslandschaft". In L.A. Pongratz (Hg.). *Bildungsphilosophie und Bildungsforschung* (S. 180-199). Bielefeld: Janus.

Krause, J., Riege, U., Stahl, M. & Zens, M. (2009). Stand und Perspektiven der Erfassung sozialwissenschaftlicher Publikationen: Erfahrungen aus der Pilotstudie Forschungsrating Soziologie. *Soziologie,* 38 (3), S. 316-330.

Krull, W. (1999). *Forschungsförderung in Deutschland: Bericht der Internationalen Kommission zur Systemevaluation der Deutschen Forschungsgemeinschaft und der Max-Planck-Gesellschaft* (Internationale Kommission zur Systemevaluation, Hrsg.). Hannover: Scherrer.

Meyer, W. (2004). *Indikatorenentwicklung. Eine praxisorientierte. Einführung*. Saarbrücken: Centrum für Evaluation.

Mutz, R., Bornmann, L. & Daniel, H. (2013). Types of Research Output Profiles: A Multilevel Latent Class Analysis of the Austrian Science Fund's Final Project Repost Data. *Research Evaluation*, 22 (2), S. 118-133.

Parthey, H. (2006). Strukturwandel der bibliometrischen Profile wissenschaftlicher Institutionen im 20. Jahrhundert. In P. Hauke, K. Umlauf (Hrsg.), *Vom Wandel der Wissensorganisation im Informationszeitalter. Festschrift für Walther Umstätter zum 65. Geburtstag* (S. 91-105). Bad Honnef: Bock + Herchen.

Piro, F. N., Asknes, D. W. & Rørstad, K. (2013). A Macro Analysis of Productivity Differences across Fields: Challenges in the Measurement of Scientific Publishing. *Journal of the American Society for Information Science and Technology, 64 (2)* , 307-320.

Schmoch, U. (2008). *The Role of Universities in Economic Growth – The German Situation*. (UniDev Discussion Paper Series Paper Nr. 1). Lund : Lund University, Research Policy Institute

Schui, G. & Krampen, G. (2010). Möglichkeiten und Grenzen der Anwendung szientometrischer Indikatoren in Evaluationen sowie ihre Integration in ein allgemeines Modell der Wissenschaftsevaluierung. *Forschung über Qualität in der Wissenschaft, 4,* 86-90.

Schui, G., Hoffmann, J.M. & Krampen, G. (2013). *ZPID-Monitor 2011 zur Internationalität der Psychologie aus dem deutschsprachigen Bereich: Der ausführliche Bericht.* ZPID Science Information Online, 12 (1).

Weingart, P. & Winterhager, M. (1984). *Die Vermessung der Forschung. Theorie und Praxis der Wissenschaftsindikatoren.* Frankfurt: Campus-Verlag.

Weishaupt, H. & Merkens, H. (2000). Forschung und wissenschaftlicher Nachwuchs, *Datenreport Erziehungswissenschaft* (S. 117–134). Opladen: Leske u. Budrich.

Wissenschaftsrat (1980). *Empfehlung zur Forschungsberichterstattung der Hochschulen.* Köln : WR.

Wissenschaftsrat (2000). *Thesen zur künftigen Entwicklung des Wissenschaftssystems in Deutschland*.

Wissenschaftsrat (2006). Bewertungsmatrix Soziologie – Fassung vom 14.10.2006 –. URL: http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/Pilotstude_Forschungsrating_Soziologie/Bewertungsmatrix_Soz.pdf [15.03.2014]

Wissenschaftsrat (2008). *Pilotstudie Forschungsrating Soziologie. Abschlussbericht der Bewertungsgruppe.* Köln, 15.02.2008. URL: http://www.wissenschaftsrat.de/download/Forschungsrating/Dokumente/Grundlegende%20Dokumente%20zum%20Forschungsrating/8422-08.pdf <abgerufen am 15.03.2014>

Wissenschaftsrat (2013). *Empfehlungen zu einem Kerndatensatz Forschung*. Berlin, 25.01.2013. URL: http://www.wissenschaftsrat.de/download/archiv/2855-13.pdf [15.03.2014]

Wuchty, S., Jones, B. F. & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316* , 1036-1039.

Entwicklung eines Informationssystems für Lehrpläne - die *Curricula Workstation*

Jessica Drechsler, Robert Strötgen & Esther Chen

Für Universitätsbibliotheken stellt der Einsatz von Discovery Services eine sinnvolle Alternative zu herkömmlichen Suchwerkzeugen dar. Für Spezialbibliotheken reicht dies häufig nicht aus, um eine effektive Recherche anbieten zu können. Hier kann es sinnvoll sein, ein eigenes Informationssystem aufzubauen, das sich bestmöglich an den Anforderungen der Zielgruppe, an die Gegebenheiten der Bestände sowie an der vorhandenen Infrastruktur orientiert. Das Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung hat sich für diesen Weg entschieden und bietet mit der im Rahmen eines DFG-Projekts geförderten *Curricula Workstation* ein selbstentwickeltes Informationssystem für Lehrpläne an. Die Fragestellungen des vorliegenden Artikels konzentrieren sich insbesondere auf Methoden zur Einbindung der Zielgruppe in den Entwicklungsprozess sowie auf den Aufbau einer nachhaltigen (technischen) Infrastruktur für ein selbstentwickeltes Informationsangebot. Dazu werden die während der Konzeption und Umsetzung des Projekts gewonnenen Erfahrungen herangezogen.

Die Nutzung von Discovery Services als Ergänzung oder Ersatz des OPACs ist in Bibliotheken ein aktuelles Thema (siehe u.a. Summan, Pietsch & Pieper, 2013; Reh & Steenweg, 2013; Kohl-Frey, 2012)[1]. Insbesondere für Universitätsbibliotheken liefert die Möglichkeit unter einer Oberfläche diverse Suchindexe einzubinden, die über die angebotenen Facetten vom Nutzer möglichst intuitiv eingegrenzt und durchsucht werden können, einen Mehrwert für ihre Kunden (siehe u.a. Lennard & Surkau, 2011; Luca 2011). Für Spezialbibliotheken reicht der Einsatz einer Discovery Software aber häufig nicht aus um ihren Nutzern eine komfortable und effektive Recherche anbieten zu können, die die besonderen Eigenschaften der Bestände der Bibliothek sowie die Anforderungen der Zielgruppe berücksichtigen. So kann es bspw. sein, dass der Bestand keine sinnvolle Facettierung zulässt, da gängige Merkmale wie Verfasser, Sprache oder Dokumenttyp zur Eingrenzung des Bestandes nicht ausreichen oder für den betreffenden Bestand für die Recherche schlicht keinen hohen Stellenwert einnehmen. Häufig ist auch, und dies geht mit dem vorherigen Punkt einher, der gängige Sucheinstieg über eine freie Suche nicht geeignet, um in den Beständen zu recherchieren. Dies sind nur einige wenige Beispiele, die die Herausforderungen verdeutlichen sollen, die spezielle Bestände mit sich bringen können. Für Einrichtungen, die über solche Bestände verfügen, kann es daher sinnvoll sein, ein eigenes Informationsinfrastrukturangebot zu erstellen, das sich bestmöglich an die Gegebenheiten und Eigenschaften des Bestandes sowie an die Bedürfnisse der Zielgruppe anpasst. Das Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung (GEI) hat sich für diesen Weg entschieden und stellt mit der *Curricula Workstation* ein selbst-

[1] So bietet z.B. die Verbundzentrale des Gemeinsamen Bibliotheksversbunds (VZG) unter der Bezeichnung „GBV Zentral“ einen zentralen Solr-Index als Datenbasis für Discovery Systeme an, bei der Discovery Software setzt die VZG auf die Open Source Software Vufind und bietet für Bibliotheken einen kostenpflichtigen VuFind-Service für lokale Discovery-Systeme an. Auch das Hessische BibliotheksInformationsSystem (HeBIS) setzt auf den Aufbau und die Entwicklung einer Portallösung inklusive Resource Discovery Service (RDS), Oberfläche ist auch hier VuFind.

entwickeltes Informationssystem für aktuell ca. 5.000 digitale und gedruckte Lehrpläne bereit.

Die Bibliothek des GEI verfügt über die weltweit umfangreichste internationale Sammlung von Schulbüchern der Fächer Geschichte, Geographie, Sozialkunde/Politik und Religion/Philosophie/Ethik sowie von Fibeln und Lesebüchern aus 160 Staaten und hat sich in den letzten Jahren zu einer anerkannten modernen Forschungsbibliothek entwickelt, die vor allem aufgrund ihrer hohen Erschließungsqualität Wissenschaftler/innen aus dem In- und Ausland anzieht. Mit einem Bestand von rund 250.000 Medieneinheiten, der sich zu 70% aus Schulbüchern und 30 Prozent aus Fachliteratur sowie ca. 150 laufenden Abonnements allgemeiner und spezifischer internationaler Fachzeitschriften zusammensetzt, verfügt die Bibliothek über einen in Größe und Zusammensetzung europa- und weltweit einzigartigen Bestand zur Erforschung von Schulbüchern und weiteren schulischen Bildungsmedien. Zur Schulbuchsammlung gehören im Sinne der erweiterten Schulbuchforschung auch Lehrpläne, die, beschränkt auf die oben genannten Unterrichtsfächer, ebenfalls von der Bibliothek des GEI gesammelt und erschlossen werden. Allerdings handelt es sich bei Lehrplänen um besonders schwer zugängliche Forschungsquellen, die zwar teilweise online recherchierbar sind, es fehlte bisher jedoch weitgehend an einer wissenschaftsgerechten Erschließung, forschungsadäquaten Recherchierbarkeit und Archivierung dieser Materialien. Der Lehrplanbestand zeichnet sich außerdem dadurch aus, dass Metadaten, die normalerweise herangezogen werden, um nach Medien zu suchen vergleichsweise unwichtig sind. Lehrpläne verfügen bspw. sehr häufig über ähnliche oder gleich lautende Titel, die Verfasser sind meist nicht bekannt oder es handelt sich für ein Land immer um dieselbe Körperschaft. Eindeutige (und dem Suchenden bekannte) Nummern wie z.B. ISBN werden nur in den seltensten Fällen vergeben. Aus diesem Grund wurde im Rahmen der DFG-Förderlinie „Förderung herausragender Forschungsbibliotheken“ das Teilprojekt „*Curricula Workstation*“ geplant und realisiert. Neben der Konzeption einer nachhaltigen Vorgehensweise zur Sammlung und Archivierung von Curricula, der Klärung ihrer Rechtesituation als amtliche Werke und der Erstellung eines Konzepts zur einheitlichen und wissenschaftsgerechten Erschließung der Lehrpläne wurde im Rahmen dieses Projekts ein Informationsinfrastrukturangebot entwickelt, welches den besonderen Anforderungen, die sich aus dem Lehrplanbestand ergeben, gerecht werden sollte.

Dieses Angebot ist seit Juni 2013 unter http://curricula-workstation.edumeres.net/ erreichbar und bietet eine strukturierte Suche nach den Eigenschaften eines Lehrplans wie Geltungsland, Unterrichtsfach, Bildungsstufe und Erscheinungsjahr. Ergänzt wird dies um eine freie Suche in den Metadaten sowie für die digitalen Ressourcen zusätzlich in den Volltexten. Die Weiterentwicklung der Workstation sieht in der zweiten Projektphase zudem interaktive Zusatzfunktionen, wie z.B. die Möglichkeit Datensätze zu annotieren oder eine Verknüpfung zwischen Lehrplänen herzustellen, vor.

Beim Aufbau eines solchen digitalen Forschungsinfrastrukturangebots ist die Orientierung am Nutzer sowie die Nutzung bereits vorhandener (technischer) Infrastrukturen essentiell, um ein Angebot zu schaffen, das sich bestmöglich in die bereits existierenden Dienste und Workflows eingliedert und von der Zielgruppe in optimaler Weise genutzt werden kann. Daher konzentriert sich der vorliegende Praxisbericht im weiteren Verlauf auf die folgenden zentralen Fragestellungen, die sich das Projektteam bereits während der Konzeptionsphase der *Curricula Workstation* stellte:

1. Mit welchen Methoden und Werkzeugen kann die Zielgruppe des Angebots in den Prozess der Erstellung des neuen Infrastrukturangebots einbezogen werden, um einen

Dienst zu schaffen, der unter Berücksichtigung der vorhandenen Ressourcen möglichst optimal auf die Bedürfnisse der Nutzer abgestimmt ist?

2. Wie kann die vorhandene (technische) Infrastruktur genutzt werden, um einerseits den Aufwand für die Umsetzung so gering wie möglich zu halten und andererseits keine technische „Insellösung“ zu schaffen?

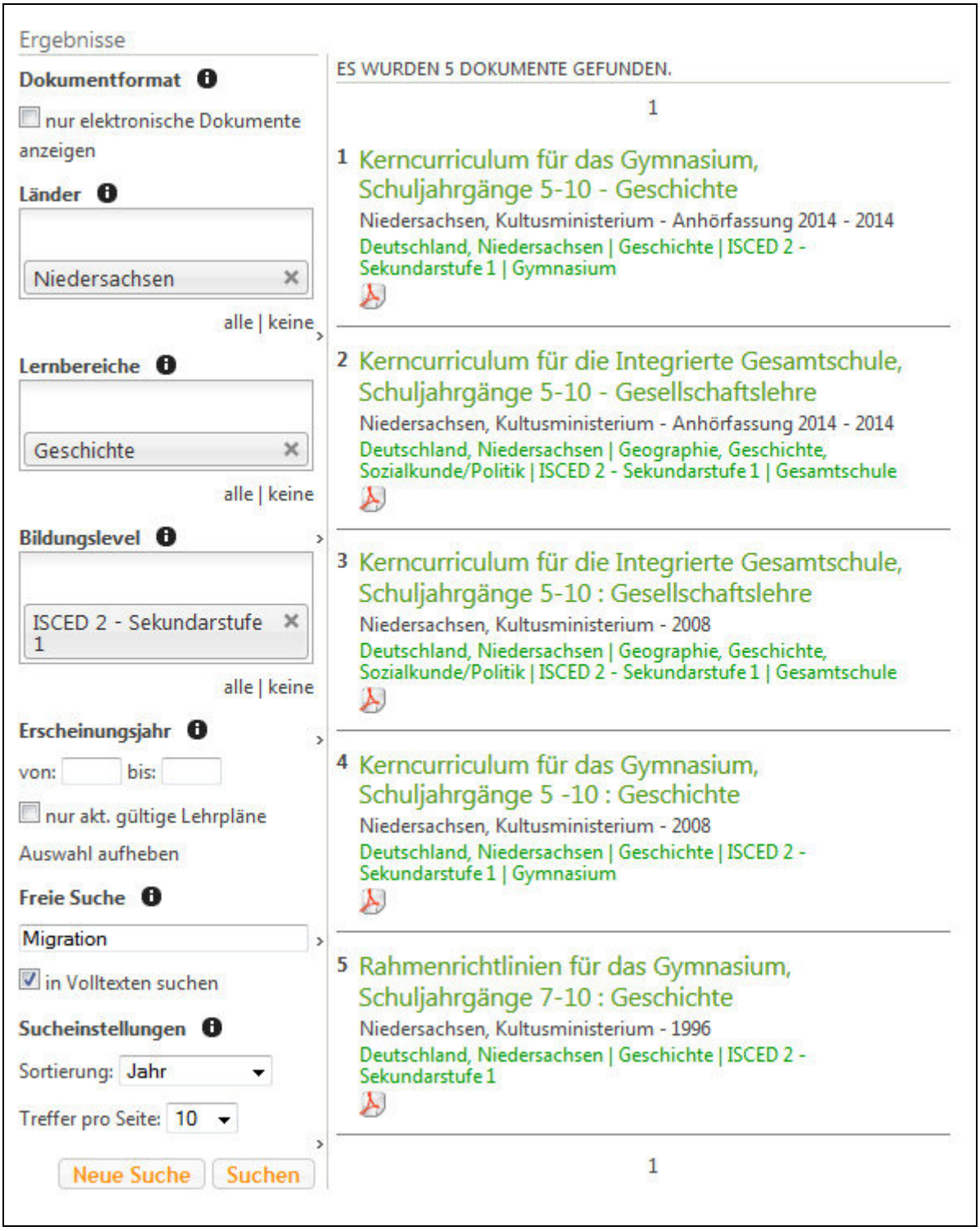


Abbildung 3: Ausschnitt aus der Trefferliste der *Curricula Workstation* mit Filteroptionen zur weiteren Eingrenzung der Suche.

Einbeziehung der Zielgruppe bei Konzeption und Umsetzung

Dem Projektteam war es besonders wichtig die Zielgruppe in allen Phasen der Konzeption durch geeignete Methoden in die Gestaltung und Umsetzung der *Curricula Workstation* zu involvieren.

Das Angebot richtet sich in erster Linie an Wissenschaftler aus dem Bildungsmedienbereich, kann und soll aber auch Informationsbedürfnisse aus der Bildungspraxis bedienen. Mit der erstgenannten Zielgruppe wurde daher bereits in der Konzeptionsphase ein Workshop durchgeführt, in dem intensiv die Anforderungen an ein Informationssystem für Lehrpläne aus unterschiedlichen Sichtweisen diskutiert wurden. Dieses Zusammentreffen wurde auch dazu genutzt anhand konkreter Beispiele aus der Praxis und eigenen Visualisierungen die aus Sicht der Zielgruppe wünschenswerten von den weniger wichtigen bzw. überflüssigen Eigenschaften und Funktionen des geplanten Informationsangebots abzugrenzen. Nach Abwägung der technischen Machbarkeit und der vorhandenen Ressourcen, flossen die Ergebnisse des Workshops in das detaillierte Anforderungsprofil der *Curricula Workstation* ein. Auf diese Weise erhielt z.B. die Umsetzung einer Volltextsuche in den digital verfügbaren Lehrplänen einen besonders hohen Stellenwert, da die Zielgruppe dies als essentiellen Mehrwert des Angebots einordnete. Ebenfalls ergab der Workshop, dass, statt des Erscheinungsjahrs eines Lehrplans, dessen Zulassungsjahre von großer Bedeutung bei der Arbeit mit diesen Materialien ist. Dies führte zu einer Anpassung des Konzepts zur Erschließung der Lehrpläne, welches nun neben der Angabe des Erscheinungsjahres auch Angaben zur Gültigkeitsdauer enthält. Als weniger elementar wurden dagegen z.B. das Abspeichern eines Suchprofils oder einer Suchhistorie erachtet, so dass diese Funktionen erst in einer zweiten Phase des Projekts umgesetzt werden sollen. Mit Hilfe des detaillierten Anforderungsprofils konnte dann anschließend unter zu Hilfenahme eines externen Dienstleisters die *Curricula Workstation* realisiert werden.

Auch nach der Entwicklung einer ersten Version des Informationsangebots und dem Übergang in den öffentlichen Testbetrieb, wurde die Fokusgruppe erneut hinzugezogen und zur Umsetzung befragt.

Durch diese Rückkopplung konnten wiederum Probleme in der Bedienung und Benutzerfreundlichkeit der Workstation aufgedeckt und behoben werden. Auch zukünftig sind solche Evaluierungsschleifen mit der Fokusgruppe geplant, um das Angebot fortlaufend und nachhaltig auf die Bedürfnisse der Zielgruppe auszurichten.

Die projektfremde Sicht der Zielgruppe erweitert damit während der gesamten Laufzeit den Blick des Projektteams auf das Produkt und ermöglicht ein Angebot, welches bestmöglich an die Anforderungen der Forschungs-Community angepasst ist. Dazu sind bereits im Dezember 2013 weitere Tests mit Benutzern sowie die Durchführung einer heuristischen Evaluation geplant. Ebenso wird die geplante Fortentwicklung der *Curricula Workstation* bereits im Rahmen der aufgeführten Evaluierungswerkzeuge vorbereitet. Dazu werden insbesondere mögliche interaktive Optionen, wie z.B. die Annotation oder die Herstellung einer Relation zwischen zwei Lehrplänen durch die Nutzer auf ihre Zustimmung hin geprüft und neue Funktionen erhoben.

Umsetzung unter Berücksichtigung der vorhandenen technischen Infrastruktur

Neben der starken Fokussierung auf die Zielgruppe, stellte sich für die Konzeption der *Curricula Workstation* die Frage, wie die vorhandene Infrastruktur möglichst optimal genutzt werden kann, um einerseits den Aufwand so gering wie möglich zu halten und

andererseits keine technische „Insellösung“ zu schaffen. Letzterer Punkt begünstigt die Gestaltung einer anschluss- und damit ausbaufähigen Informationsinfrastruktur.

Nach einer umfassenden Sichtung des Ist-Zustandes unter Berücksichtigung vorhandener Schnittstellen, Metadatenschemata und Software sowie der Einbeziehung und Abwägung möglicher neuer Lösungen, stellte sich die folgende Vorgehensweise für das Projektteam als passende Option zur technischen Realisation der *Curricula Workstation* dar: Die Metadatenerfassung der Lehrpläne findet auch weiterhin im Gemeinsamen Bibliotheksverbund (GBV) statt, lokale Daten wie z.B. zur Kategorisierung der Lehrpläne nach Geltungsland, Bildungslevel und Lernbereich oder die Erfassung des Gültigkeitszeitraums können dort auf der Lokal- und Exemplarebene abgelegt werden. Zusätzlich werden die Metadaten der Lehrpläne, die im elektronischen Volltext vorliegen, durch teilweise automatisierte Abläufe im DSpace Dokumentenserver eingepflegt und die Volltexte manuell hochgeladen. Die Lehrpläne werden dadurch mit einem persistenten Identifier ausgestattet, der die Langzeitarchivierung sowie die dauerhafte Zitierbarkeit der Materialien sicherstellt. Die Metadatenhaltung der Lehrpläne wird folglich komplett über die Gemeinsame Verbunddatenbank realisiert. Dies hat den Vorteil, dass die Sichtbarkeit der Lehrpläne über bekannte Werkzeuge wie z.B. den OPAC und den GVK erhalten bleibt und vertraute und damit eingespielte Arbeitsabläufe und Metadatenschemata von den Mitarbeitern weiterhin genutzt werden können. Der Mehraufwand durch das zusätzliche Einpflegen der elektronisch vorliegenden Lehrpläne in das DSpace Repository ist notwendig, um die Langzeitverfügbarkeit sicherzustellen, wird aber gleichzeitig durch ein Mapping der aus dem Verbundkatalog extrahierten Metadaten über ein Excel Makro mit dem in DSpace verwendete Dublin Core gering gehalten. Zur Auswahl der Software für den Dokumentenserver wurden Kriterien aus gängigen Leitlinien, wie z.B. dem DINI Zertifikat (DINI 2011) herangezogen. Mit Hilfe dieser Kriterien konnte eine Eingrenzung auf einige wenige Softwareangebote erfolgen. Zusätzlich hospitierten Mitglieder des Projektteams an verschiedenen Einrichtungen, wie z.B. der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW) mit dem Open-Access-Server EconStor[2] oder dem Deutschen Institut für Internationale Pädagogische Forschung (DIPF) mit dem Dokumentenserver peDOCS[3], um praktische Erfahrungen zu Konfiguration und Arbeitsabläufen mit den dort aufgebauten Dokumentenservern zu erhalten. Die auf diese Weise gewonnen Eindrücke aus der Praxis verfestigten die Entscheidung für DSpace als Repository Software für die *Curricula Workstation*. Mit DSpace[4] fiel die Wahl außerdem auf eine Open-Source-Software mit einer großen Anwendercommunity, die bei Herausforderungen, die durch eigene Ressourcen nicht gelöst werden können, bereitwillig und schnell Hilfestellung geben kann. Die Erfahrungen, die in diesem Projekt mit DSpace gesammelt werden können, nützen außerdem weiteren Projekten des GEI, wie z.B. bei dem Aufbau eines Institutionellen Repositoriums.

Dadurch, dass die Datenhaltung der Lehrpläne komplett über den Verbundkatalog des GBV gelöst wurde, konnte für den Aufbau eines Suchindex die SRU-Schnittstelle des OPAC genutzt werden. Hierüber werden die Metadaten und, falls vorhanden, die Volltexte der Lehrpläne in regelmäßigen Abständen abgefragt und in einen Solr-Index geschrieben. Über ein Frontend, welches als eine Typo3 Extension realisiert wurde, kann dieser Index über ein Suchformular benutzerfreundlich durchsucht werden. Mit dem Aufbau des Solr-Index und der Umsetzung des Frontends als Typo3 Extension wurde ein externer Dienstleister beauftragt.

[2] http://www.econstor.eu/dspace/about
[3] http://www.pedocs.de/
[4] http://dspace.org

Im Zusammenspiel zwischen Vorhandenem und Neuem bei Workflows und technischer Infrastruktur konnte im Projekt eine praktikable Lösung zur technischen Realisation der *Curricula Workstation* gefunden werden, die die Schaffung technischer Inseln umgeht und damit anschluss- und ausbaufähig bleibt. Für die Erstellung dieser Lösung war eine ausführliche Anforderungsanalyse, die Betrachtung des Ist-Zustands mit den verfügbaren Optionen sowie das Hinzuziehen externer Beratung durch Hospitation und Dienstleister wesentlich. Ebenso wichtig sind ausführliche Dokumentationen der Arbeitsabläufe, die die Mitarbeiter bei ihren neuen Aufgaben unterstützen. Zukünftig soll weiter daran gearbeitet werden, die Workflows zu vereinfachen.

Die während der Konzeption und Umsetzung der *Curricula Workstation* gesammelten Erfahrungen haben deutlich gezeigt, dass der spezielle Bestand von Lehrplänen auch besondere Anforderungen an Erschließung, Archivierung und Recherche dieser Materialien mit sich bringt. Dabei stellten sich die oben aufgelisteten Fragen bereits in der Anfangsphase des Projekts als zentral heraus. Die durchgängige Orientierung an diesen Fragen half dabei ein Informationsinfrastrukturangebot zu konzipieren, welches durch die fortwährende Einbindung einer Fokusgruppe und das Aufsetzen auf bestehende Infrastrukturen, sowohl Anklang bei der angestrebten Zielgruppe findet, als auch problemlos in die vorhandenen Arbeitsabläufe integriert werden kann.

Hinsichtlich der Einbeziehung der Zielgruppe haben sich insbesondere die folgenden Methoden und Vorgehensweisen als besonders nutzbringend erwiesen und können daher für ähnliche Projekte empfohlen werden:

- Möglichst frühe und regelmäßige Einbindung der Zielgruppe in den Konzeptions- und Umsetzungsprozess.
- Workshops als adäquate und wenig zeit- und kostenintensive Methode, um insbesondere in der Anfangsphase des Projekts Anforderungen und Bedürfnisse der Zielgruppe zu erheben.
- Evaluierung mit experten- und benutzerzentrierten Methoden (z.B. Heuristische Evaluation, Benutzertest) nach der Umsetzung des Informationsinfrastrukturangebots.
- Auch nach dem Online-Gang des Angebots regelmäßige Evaluierungsschleifen einplanen und das Feedback der Nutzer bei der Weiterentwicklung berücksichtigen.

Für den Aufbau des Angebots unter Berücksichtigung der vorhandenen technischen Infrastruktur zeigten sich die folgenden Herangehensweisen und Aspekte als besonders effizient und zielführend, weswegen sie als Hinweise für ähnliche Infrastrukturprojekte festgehalten werden sollen:

- Eine ausführliche Ist-Analyse der vorhandenen Workflows, technischen Infrastrukturen und der verwendeten Metadaten(formate) schafft Klarheit und erleichtert im weiteren Prozess die Abwägung der (technischen) Optionen.
- Anfertigung eines detaillierten Anforderungsprofils für das aufzubauende System und Gegenüberstellung von Anforderungsprofil und Ist-Zustand, um Potenziale zur Nachnutzung von Systemen und/oder Schnittstellen zu entdecken.
- Bei der Abwägung der unterschiedlichen Lösungen die Anschluss- und Ausbaufähigkeit der Systeme als wesentliche Kriterien berücksichtigen.
- Externe Beratung durch z.B. Hospitationen und/oder Dienstleister kann den Aufbauprozess wesentliche Impulse geben, ihn beschleunigen und gibt die Möglichkeit auf vorhandenen Erfahrungen aufzubauen bzw. fehlende Kompetenzen auszugleichen.

Literatur

DINI, Deutsche Initiative für Netzwerkinformation e.V., Arbeitsgruppe "Elektronisches Publizieren" (2011). *DINI-Zertifikat Dokumenten- und Publikationsservice 2010*. Göttingen: DINI (DINI-Schriften ; 3). URN: urn:nbn:de:kobv:11-100182794

Kohl-Frey, O. (2012). *KonSearch: Eine Literatursuchmaschine für alles? Erfahrungen mit der Einführung von Summon als Discovery-Tool an der Universität Konstanz*. URN: urn:nbn:de:bvb:29-opus-37830

Lennard, H. & Surkau, M. (2011). *Benutzerevaluation und Usability-Test zur neuen Suchoberfläche Primo (ExLibris) Berlin.* Berlin: Institut für Bibliotheks- und Informationswissenschaft der Humboldt - Universität zu Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 309)

Luca, H. (2011). *Usability-Studie zu KonSearch : Evaluation der neuen Literatursuchmaschine der Universität Konstanz*. URN: urn:nbn:de:bsz:352-168427

Reh, U. & Steenweg, H. (2012). *Der lange Weg von HeBIS zum Discovery Service*. Hamburg: Presented at the 101. Deutscher Bibliothekartag. URN: urn:nbn:de:0290-opus-13144

Summann, F., Pietsch, C. & Pieper, D. (2013). *Die Cloud im lokalen Bibliothekskatalog : eine integrative lokale Portallösung mit nahtlos eingebundenem Discovery Service der Universitätsbibliothek Bielefeld*. Leipzig: Presented at the 102. Deutscher Bibliothekartag. Retrieved from http://pub.uni-bielefeld.de/publication/2563307

Notizen

Wenn Informationswelten und Weltwissen zusammenwachsen: Das Internet der Dinge erreicht die Welt der Gegenstände

Reinhard Karger

Business Intelligence und Predictive Analytics haben die Recherche verändert, Social Media, Sentiment Analysis, Opinion Mining, Informationsextraktion und Big Data haben die Wissensquellen quantitativ und qualitativ erweitert. Group Ware ermöglicht gratis die gemeinschaftliche Informations- und Dokumentarbeit. Die DGI und die Information Professionals können und müssen wichtige und notwendige Beiträge zur Orientierung in den entgrenzten digitalen Ökosystemen leisten. Damit können sie die Gegenwart erträglicher machen, die Vorteile der Digitalisierung erhalten und die Zukunft gewinnen.

Überwundene Medienbrüche schaffen neue Perspektiven

Es gibt heute kaum noch Telefonzellen, keine Schallplattenläden, aber an jeder Ecke einen Geldautomaten. Alle telefonieren mobil, alle fotografieren digital, bezahlen elektronisch, Musik ist auf dem Smartphone, Bücher sind auf dem Tablet. Der Brockhaus wird nicht mehr gedruckt, dafür ist die Wikipedia umfangreicher als es jemals eine gedruckte Enzyklopädie war - und aktueller. Das Internet verbindet Menschen, Wissen und Dienste. Das ist bekannt. Was wird in nächster Zeit passieren? Weitere Medienbrüche werden überwunden, die Welt in den Grundzügen verändert.

Die Gesellschaft entwickelt sich substantiell zu einer dialogischen Wissensgesellschaft. Die Wirtschaft setzt auf durchgehend digitalisierte Wertschöpfungsketten. Die Warenwirtschaft wurde durch eine Datenwirtschaft ergänzt und manchmal durch diese ersetzt. Business Intelligence, Predictive Analytics haben die Recherche verändert, Social Media, Sentiment Analysis, Opinion Mining, Informationsextraktion und Big Data haben die Wissensquellen quantitativ und qualitativ erweitert. Cloud Computing hat Datenzugriff und -pflege standortunabhängig gemacht. Group Ware ermöglicht gratis die gemeinschaftliche Informations- und Dokumentarbeit, Echtzeit-Interaktionskanäle stehen zur Verfügung, die Nutzung ist komfortabel und selbst mobil überraschend preiswert. Industrie 4.0 bringt diese Entwicklungslinien zusammen und revolutioniert die Produktion.

Internet und Web sind in den letzten Jahren zur industriellen Produktivitätsplattform und zum gesellschaftlichen Betriebssystem geworden. Im nächsten Schritt werden wir mit dem Internet der Dinge eine tiefgreifende und hochaufgelöste Vernetzung der Dinge in unserer Umgebung erleben: Das Internet der Dinge verändert das Planen und Produzieren, das Säen und Ernten, das Einkaufen und Abrechnen, das Konsumieren und Recyceln, das Archivieren und Finden, ändert Beruf und Alltagsleben. Wenn Dinge sich mitteilen, wird die Welt interaktiv.

Technisch entsteht das Internet der Dinge durch Gegenstände mit angehafteten Etiketten oder Smart Label, über die die Eigenschaften dieser Dinge digital auswertbar sind. Die Label können außerordentlich klein sein und – wenn tatsächlich jeder Alltagsgegenstand involviert sein soll - werden außerordentlich preiswert sein müssen. Im Ergebnis werden die Dinge durch Smart Label eineindeutig identifizierbar – nicht nur als Produkttyp, sondern als individuelle Produktinstanz, nicht nur ein Fahrrad, sondern mein Fahrrad, nicht nur ein Joghurt, sondern genau dieser Joghurtbecher. Eine Waschmaschine wüsste, wie oft und was

sie gewaschen hat und wieviel Strom und Wasser sie in ihrem bisherigen Produktleben verbraucht hat; eine Jeans wüsste, wie oft sie mit welcher Temperatur gewaschen wurde, ein Ersatzteil kann seine Echtheit beweisen. Ein Haus könnte seine historische Energieverbrauchskurve mitteilen, archivierte Rechnungen über Reparaturarbeiten und Versicherungsbelege, einen Bauplan oder ein 3D-Modell anzeigen. Die dingliche Beschreibung, die die Dinge zur Verfügung stellen, ist möglichst umfassend und notwendigerweise standardisiert, so dass die Information über den einzelnen Gegenstand in beliebigen Anwendungen eingesetzt oder angereichert, weiter verarbeitet und digital genutzt werden kann.

Noch ist das Internet der Dinge weder präsent oder alltäglich. Noch gibt es kein Gerät, kein Gadget, das ikonenhaft für das Konzept steht und es dadurch konkretisiert. Viel ist über das mobile Internet vorab geschrieben worden – erst als das iPhone erhältlich war, war auch das mobile Internet da, und plötzlich waren Mailen und Browsen mobil, aber auch die „Tagesschau“ und der „Tatort“ in der Mediathek. Der zukünftige Alltagseinfluss des Internet der Dinge ist heute schwer vorstellbar. Natürlich gibt es bereits „runtastische“ Laufschuhe, die Daten übermitteln, wo jemand und wie lange er gelaufen ist oder Personenwaagen, die persönliche Gewichtsdaten sammeln, über WLAN in Verläufen darstellen und den Body Mass Index berechnen.

Die Fusion der Welten zum Internet der Dinge und Dienste

Das Internet der Dinge wird nicht punktuell sein, sondern flächendeckend. Es wird den Medienbruch zwischen physischen Objekten und virtuellen Welten überwinden. Der Informationsreichtum und die Detailauflösung wachsen, der Dialograum wird um eine neue Dimension erweitert. Die Dinge des analogen Alltags werden Teil des digitalen Dialogs. Das Internet der Dinge wird mit einer erstaunlich umfassenden Durchdringung der Lebenswirklichkeit die nächste große Informationslücke schließen und neue Informationsozeane eröffnen. Informationskompetenz wird dadurch zu einer Schlüsselkompetenz in jeder Lebenslage.

Das Internet der Dinge bezieht die meisten Dinge des täglichen Gebrauchs ein. Daraus ergeben sich fundamental neue Anwendungschancen und Innovationsperspektiven. Wir haben in anderen Bereichen ähnlich gravierende Veränderungen kennen gelernt. Das Navigationssystem hat die Landkarte abgelöst, wird nun durch das moderne Smartphones ersetzt und bietet mit ortsbasierten Dienste vollkommen neue Geschäftschancen.

Die ursprüngliche Motivation für das Internet der Dinge kommt aus unterschiedlichsten Richtungen und zahlreichen Quellen: aus der Produktions-, Verpackungs- und Transportindustrie, der Logistik- und Tourismusbranche, aus der Welt der Unternehmenssoftware und der Warenwirtschaftssysteme, von den Hotels, den Speditionen und Handelsunternehmen, aber auch aus Bereichen wie Telematik, Sport oder Gesundheit. Der eigentliche Wunsch: Wenn die Dinge digital adressierter sind, sind sie erreichbar für Dienste, Prozesse und Abläufe, z.B. für Produktion und Logistik, Lagerhaltung und Bestellung, Garantie oder Recycling, Tausch oder Verkauf.

Jedes Vorprodukt, jeder Werkstückträger, jedes Produkt wird mit einem Chip bzw. einer eineindeutigen ID gelabelt und kann bei Herstellung und Transport, bis zum Konsumenten fortwährend mit Daten angereichert werden. Dadurch erzählt jedes Produkt seine individuelle Entstehungs-, Weiterverarbeitungs- Transport- und möglicherweise dann auch Nutzungsgeschichte (vgl. SemProM, Produkte führen Tagebuch, http://www.semprom.org).

Nachvollziehbar wird z.B. der Temperaturverlauf, von Beginn der Produktion über die Lagerung bis zur Lieferung nach Hause. Berechenbar wird auch die CO2-Emission durch Herstellung, Ressourceneinsatz oder z.B. Transport, Kühlung oder Lagerung. Transparent wird, woher das Produkt kommt, welchen Zweck es erfüllen, wie es bedient werden soll, wie lange es unterwegs war und wie viel Schadstoffe freigesetzt wurden. Natürlich könnte das Produktgedächtnis auch die weitere Nutzung berücksichtigen, so dass Daten über den gesamten Produkt-Lebenszyklus gesammelt abrufbar sein können. Smart Services werden das Internet der Dinge ergänzen und die digital erschlossenen dinglichen Informationen zu Mehrwerten weiter verarbeiten. Dienste sind Anwendungen, mit denen man Ziele erreichen kann, z.B. etwas bestellen oder Tickets buchen, in Erfahrung bringen, ob ein Flugzeug sich verspätet hat oder ob die gewünschten Knöpfe lieferbar sind.

Das Internet der Dinge und Dienste wird einen ganzheitlichen Einfluss auf unseren Umgang mit Dingen haben – wie das Internet der Daten unseren Zugang zu Wissen und den Umgang mit Wissensfragen grundsätzlich verändert hat. Der Realitätscheck ist notwendig. Sind das Technikfantastereien oder sehr wahrscheinliche Veränderungen?

Megatrends und Eigendynamik

Der aktuell erreichte Zustand der digital vernetzten Gesellschaft – Alltagsweltdurchdringung, Vernetzungsintensität, Netzabdeckung oder Breitbandinfrastruktur – lässt sich auf das Zusammenwirken von vier Megatrends zurückführen. Miniaturisierung (die Größe schrumpft, die Leistung steigt), Skalierung (Unikate sind teuer, Serie preiswert), Digitalisierung (die Objektinformation löst sich vom Objekt und kann in beliebigen Anwendungen Mehrwerte entfalten), Vernetzung (Netzeffekte füttern den Netzwerkerfolg).

Die Megatrends haben sich in den vergangenen Jahrzehnten wechselseitig und zunehmend verstärkt. Sie werden auch die nächste große Veränderung der Informationsinfrastruktur bewirken. Dabei sind Netzeffekte die treibende Kraft und Miniaturisierung die technologische Grundlage. Skalierung macht das technisch Mögliche erschwinglich und wirtschaftlich erfolgreich. Digitalisierung liefert die Informationsbasis für inhaltlich sinnstiftende und wirtschaftlich erfolgreiche Anwendungen.

Die Innovationsintensität der letzten Jahre vorausgesetzt wird folgendes geschehen: Hersteller werden sich entscheiden können, ob sie – zu den gleichen Kosten – grüne Farbe oder grüne Elektronik auf beliebige Objekte drucken wollen. Grüne Schrift kann man lesen. Grüne Elektronik könnte man zusätzlich auch auslesen. Das Internet der Dinge verändert nicht die gegenständliche Welt, aber es eröffnet den Zugriff auf die Gegenstände der Welt. Aus stummen Gegenständen werden mitteilsame Dinge, aus sprachlosen Sachen smarte Objekte und das Verhältnis von Mensch und Gegenstand kehrt sich um. Nicht der Mensch erklärt den Gegenstand, sondern das Ding gibt Auskunft über seine Eigenschaften. Der Nutzer kann auf diese Eigenschaften zugreifen und sie mit seinen Plänen und Wünschen verbinden. Zusammen werden die gegenständliche Welt der Dinge und die digitale Welt der Dienste zu einem gigantischen und nahtlos zugänglichen Informationsraum.

Die Karten werden neu gemischt. Neue Suchräume bewirken auch neue Services. Neue Mehrwerte, neue Angebote, neue Anbieter und neue Kollateralerfolge werden entstehen. Neue Produktionsweisen werden ermöglicht, Logistik und Konsum werden verändert, das Web expandiert - erneut. Eine noch unbekannte Zahl von unbeabsichtigten Zufallserfolgen mit aktuell noch nicht vorhandenen Geschäftsmodellen wird entstehen.

Eine neue Dimension für den digitalen Dialog

Dennoch: Gesellschaftliche Eruptionen sind nicht plan-, nicht vorhersehbar. Manchmal fordern sie, dass man als Person oder auch als Fachgesellschaft eine neue Rolle einnimmt und Herausforderungen annimmt. Seit der Aufdeckung der aggressiven NSA-Datensammlung haben sich ab Juni 2013 Kommunikationsparadigmen erneut und grundsätzlich geändert, kulturelle Gewissheiten schon wieder verschoben. Persönliche, gesellschaftliche und wirtschaftliche Handlungsspielräume werden neu reflektiert. Die internationalisierte Diskussion um Vorratsdatenspeicherung und Datenanalyse, zu Privatheit und Persönlichkeitsrechten führt zu interkulturellen Grundsatzdebatten, sie polarisiert und Kompromisse sind noch nicht gefunden. Es gibt eine neue Konfliktzone. Welche Position wird die DGI einnehmen?

Die DGI hat Internet und WWW, Web 2.0 und Social Media weniger euphorisch umarmt, als viele es erwartet hätten, aber Information und Wissen sind ihre Themen seit 66 Jahren. Das Internet der Dinge wird die vorhandenen Infrastrukturen ergänzen und zusätzliche Informationsräume erschließen. Informationskompetenz wird zu einer persönlichen, gesellschaftlichen und erwerbswirtschaftlichen Schlüssel- und Zukunftsqualifikation.

Erschien manchen die DGI als zu wenig web-affin, ist sie nun weniger infiziert von einem Alles-Immer-Überall Datenversprechen. Während manche Geschäftsmodelle, die auf einem unbegrenzten Austausch von Daten aufbauen, in ihrer Basis bedroht sind, hat die DGI zurückhaltend agiert - manchmal vielleicht etwas reserviert. Das bedeutet für den Moment, dass sie sich und ihre Balance nicht neu finden muss. Wenn Gewissheiten schmelzen, strahlen beständige Handlungskompetenzen und Urteilsfähigkeit in einem neuen Licht.

Die DGI greift die aktuellen Debatten auf und wird ihre gesellschaftliche Relevanz bei der Erkundung dieser neuen Informations- und Lebenswelten verstehbar, wissenschaftlich nachvollziehbar und aber auch hörbar unter Beweis stellen. Die DGI und die Information Professionals können und müssen wichtige und notwendige Beiträge zur Orientierung in den entgrenzten digitalen Ökosystemen leisten. Damit können sie die Gegenwart erträglicher machen, die Vorteile der Digitalisierung erhalten und die Zukunft gewinnen.

Business Information Quality und organisationale Wissensgenerierung: Informationsanalyse, Informationsqualität und Business Intelligence

Ronald Billen

Die Begriffe Datenqualität und Datenanalyse sind mit dem Business Intelligence-Konzept auf das engste verbunden. Intention des Verfassers ist es, die Beziehungen und Wechselwirkungen der Begriffe aufzuzeigen, wobei sich die Datenqualität als das integrierende Element herauskristallisiert. Spielt diese bereits bei der täglichen Routine des Online Transaction Processing (OLTP) eine wesentliche Rolle, so ist sie bei Business Intelligence geradezu erfolgskritisch, da an dieser Stelle die Weichen für zukünftige Unternehmensstrategien gestellt werden. Nachdem einleitend die strategische Bedeutung der Ressource „Information" im Unternehmen beleuchtet wird, stellen das betriebliche Informationssystem und dessen Unterteilung in operative und analyseorientierte Subsysteme den Einstiegspunkt der Betrachtung dar. Dem folgt ein erneuter Versuch, die Begriffe „Daten" und „Information" voneinander abzugrenzen, nur um festzustellen, dass deren korrekte Verwendung mehr auf der richtigen Intuition als auf detaillierter Definition beruht. Im dritten Kapitel wird detailliert die Rolle der Qualität von Daten im Unternehmen erläutert und die im Umgang mit ihnen zu beachtenden gesetzlichen Regelungen. Besonderes Augenmerk richtet sich auf die Datenqualitätsanforderungen und auf die Wahrung der Datenintegrität. Abschließend wird die historische Entwicklung der betrieblichen Informationssysteme besprochen und die Rolle der Datenqualität im BI-Konzept. Als Beispiel für Datenanalyse hat der Verfasser den OLAP-Ansatz gewählt. Für die Wahl war entscheidend, dass nach seiner Auffassung diese Analysetechnik der alltäglichen Realität in Unternehmen entspricht und ohne tiefer gehende Statistikkenntnisse für Fragestellungen aus Vertrieb, Marketing, Controlling etc. vielfach einsetzbar ist.

1 Einleitung

Die Bedeutung von Informationen und Wissen, oder präziser Expertise, für den langfristigen wirtschaftlichen Erfolg ist im Unternehmensumfeld unbestritten: Wissen gilt als der wichtigste Produktionsfaktor in Unternehmen. Dies ist keine Erkenntnis unserer Tage, obwohl es oft genug so verkauft wird. Das Informationspotenzial eines Unternehmens wird nicht umsonst als dessen „Schmierstoff" angesehen, und dem Zusammenhang zwischen Wissen und dem Erwerb von Wettbewerbsvorteilen wird eine strategische und somit erfolgskritische Bedeutung beigemessen. Obwohl aufgrund eines häufig unausgewogenen Verhältnisses der beteiligten Dimensionen Mensch, Organisation und Technik oft als Totgeburt bezeichnet, wurde somit dem Strategiekonzept Wissensmanagement zu spätem Ruhm verholfen und hat ihm entweder bereits zu einem festen Platz in der Unternehmensorganisation verholfen, oder doch zumindest an entscheidender Stelle das Bewusstsein dafür geschärft.

Mit zunehmendem Einblick in die Materie, insbesondere, wenn man sich mit den Aspekten der Wissensbausteine und der Identifikation von Wissensfeldern vertraut macht, stellt man fest, dass bezüglich der Gewichtung der Wissensmanagementthemen eine gewisse Diskrepanz besteht: Ungeachtet aller theoretischen Ansätze von Willke, Nonaka und Takeuchi etc. zur organisationalen Wissensgenerierung entsteht der Eindruck, dass Wissensmanage-

ment sich vornehmlich mit der Formulierung von Wissenszielen, dem Identifizieren von vorhandenem Wissen, dessen Bewertung, Speicherung, Verteilung etc. befasst. Der Frage nach Möglichkeiten der Informationsanalyse und damit der Wissensgenerierung kommt offensichtlich eine untergeordnete Bedeutung zu. Doch gerade diese spielt bei Business Intelligence (BI) eine wesentliche Rolle: Gesucht werden bislang unbekannte Muster und Zusammenhänge von Daten, um so zu neuen Erkenntnissen zu gelangen, die für die strategische Ausrichtung der Geschäftsprozesse, vor allem Controlling, Marketing und Verkauf, maßgeblich sind. Immerhin ist es die wichtigste Aufgabe des betrieblichen Informationsmanagements, und hierzu zählt BI, Daten für Nachfrager verfügbar zu machen. Deshalb muss es die tägliche Suche nach diesen unterstützen. Und auch wenn im Titel die Begriffe Informationsqualität und Informationsanalyse[1] auftauchen, so spricht man im BI-Kontext eher von Datenqualität und Datenanalyse. Daher werden die Begriffe Informationsqualität und Datenqualität bzw. Informationsanalyse und Datenanalyse im Folgenden synonym verwendet.

Insbesondere bei der Wissensgenerierung bilden Informationsqualität und Informationsanalyse Größen von strategischer und daher erfolgskritischer Bedeutung für das Unternehmen, unabhängig von dessen Größenmerkmalen. Die inhaltliche Deutung dieser Begriffe, ihre wechselweise Beziehung sowie ihre Bedeutung im BI-Kontext als einem analyseorientierten Informationssystem darzustellen, ist im Folgenden zu klären.

2 Das Informationssystem - Ausgangspunkt der Betrachtung

Informationssysteme (IS) im engeren Sinne sind computergestützte Anwendungssysteme bzw. Softwaresysteme zur Ausführung betrieblicher Aufgaben. Im weiteren Sinne werden sie als Informations- und Kommunikationssysteme (IuK-Systeme) bezeichnet und integrieren als solche Hard- und Software, Menschen und Anwendungen.[2]

Als technisches System umfasst ein Informationssystem bzw. informationstechnisches System (IT-System) alle geregelten betriebsinternen und -externen Informationsverbindungen sowie deren technische und organisatorische Einrichtung und deren technische und organisatorische Einrichtung zur Informationsgewinnung und -verarbeitung entsprechend dem Informationsprozess. Dabei bildet das IT-System den formalen Teil des gesamten betrieblichen Kommunikationssystems.[3] Gemäß Maxime der Informationslogistik ist es seine Aufgabe, dem Informationsnachfrager bzw. Handlungsträger im Unternehmen die zur Erledigung seiner Aufgaben erforderlichen, relevanten Information zum richtigen Zeitpunkt in der geeigneten Art und Weise, bezogen auf Qualität und Menge, am richtigen Ort zur Verfügung zu stellen.[4] IuK-Systeme unterteilen sich in zwei Typen:

Den einen bilden die ***Administrations- und Dispositionssysteme***, d. h. die ***operativen Informationssysteme*** zur Unterstützung der täglichen Aufgabenbearbeitung, die vom Sachbearbeiter auf operativer Ebene verwendet werden.

---

[1] Zwahr, A. (Hrsg. 2006). Meyers großes Taschenlexikon, Band 1 (10. Auflage). Mannheim: Bibliographisches Institut. Analyse (grch. „analysis“ = „Auflösung“): Systematische Untersuchung eines Sachverhalts unter Berücksichtigung seiner Teilaspekte.

[2] Gabriel, R. (o. J.). Informationssystem. ENZYKLOPÄDIE DER WIRTSCHAFTSINFORMATIK ONLINE-LEXIKON. Verfügbar unter: http://enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/uebergreifendes/ Kontext-und-Grundlagen/Informationssystem.

[3] Lackes, R., Schewe, G., Siepermann, M. (o. J.). Informationssystem. Gabler Wirtschaftslexikon. Verfügbar unter: http://wirtschaftslexikon.gabler.de/Definition/informationssystem.html.

[4] o. V. (o. J.) Leitthema Informationslogistik. Verfügbar unter: http://www.informationslogistik.org.

Zum zweiten Typ zählen die ***analyseorientierten Systeme*** **(AIS)** für die ***entscheidungsunterstützende Informationsversorgung*** bei strategischen Planungs- und Entscheidungsprozessen (Analyseaufgaben) der Fach- und Führungskräfte.[5] Diese wurden früher oft **Management Support Systems (MSS)** genannt; heute werden sie mit dem Synonym **Business Intelligence Systems** bezeichnet. Die Entwicklung betrieblicher Informationssysteme wird im vierten Kapitel detailliert dargestellt.

3. Daten im Unternehmen

The best way to reduce the amount of data - delete it."
Sheila Childs, Research Vice President, Gartner Group

3.1 Vom Wesen eines diffusen Begriffs

"Data is the new oil. Unfortunately, the technology has evolved faster than the workforce skills to make sense of it, and organizations across sectors must adapt to this new reality or perish."
Andreas Weigend, Head Social Data Lab, Stanford University

Daten, Informationen Wissen - unsere Sprache weist oft Begriffe auf, die selten eindeutig definiert werden (können). Abstrakt wie sie sind, ist ihre Bedeutung mehr zu erahnen als zu artikulieren, und dennoch benutzten wir sie intuitiv oft richtig und sogar im richtigen Zusammenhang. Daten (Sing. „Datum") werden aus den Zeichen eines Zeichenvorrats nach definierten Ordnungs- bzw. Syntaxregeln gebildet, die Sachverhalte symbolisieren. Diese Daten sind reine Fakten, und selbst wenn man sie mit weiteren Angaben wie Maßeinheiten oder anderen Metadaten versehen würde, sind sie ohne Zusammenhang und Hintergründe nicht interpretierbar. So wird aus den Zeichen „1, 0, 0" und den Angaben „cm" das Datum „100 cm", doch ohne Aussage, worauf sich dieses bezieht.[6] Zu Informationen werden sie, wenn sie in Form von bestimmten Strukturen miteinander in Beziehung gesetzt werden, wodurch sie einen Zweckbezug erhalten und ihnen eine eindeutige, unverwechselbare Bedeutung oder Semantik zugeordnet wird, beispielsweise in Form von Metadaten.[7] Informationen wiederum sind die Vorstufe des Wissens und beziehen sich auf Sachverhalte, Ereignisse, Funktionen, Arbeitsweisen, Verfahren und Vorgänge. Sie bilden insbesondere Inhalt und Bedeutungsgehalt einer Nachricht in textlicher, grafischer oder audiovisueller Form ab, können durch Daten dargestellt, auf Datenträgern gespeichert, in Computern verarbeitet und über Ausgabegeräte ausgegeben werden. Umgekehrt lassen sich Informationen aus Daten reproduzieren, wenn sie entsprechend interpretiert werden können. Als bedeutsame Ressource für den Unternehmenserfolg werden sie für die Realisierung strategischer Unternehmensziele verwendet und durch das Informationsmanagement in vorhandene und neue Anwendungen eingebracht. Sie enthalten keine irrelevanten oder redundanten Teile.[8] In diesem Sinne ist das IT-System ein Medium zur Entscheidungsfindung und -durchsetzung des Managements und damit Grundlage für den gesamten

[5] Gabriel, R. Gluchwoski, P., Pastwa, A. (2011). Data Warehouse & Data Mining. Herdecke Witten: W3L GmbH. S. 3.
[6] Billen, R. (2013). Wissensmanagement und Business Intelligence. Methoden zur Wissensgenerierung im Vergleich. S. 32-40 Verfügbar unter: http://www.gfwm.de/files/gfwmTHEMEN6-Sep2013.pdf
[7] Bodendorf, F. (2003). Daten- und Wissensmanagement. S 1. Berlin Heidelberg New York Springer-Verlag.
[8] o. V. (o. J.). Information. In: ITWissen. Das große Online-Lexikon der Informationstechnologie. Verfügbar unter: http://www.itwissen.info/definition/lexikon/Information-information.html.

Managementprozess[9], und die von ihm produzierten, verteilten und verarbeiteten Daten bzw. Informationen sind aus verschiedenen Gründen zu schützen.

3.2 Daten und Datenqualität

Für Unternehmen ist Datennutzung bezüglich Effizienzsteigerung, Umsatzwachstum und Wettbewerbsfähigkeit ein erfolgskritischer Faktor. Für diese Nutzung ist Datenqualität die grundlegende Voraussetzung. Sie beschreibt Korrektheit, Relevanz und Verlässlichkeit von Daten. Datenqualität dient dem Zweck, unternehmensweit optimierte Geschäftsprozesse zu realisieren, die dadurch ökonomische Vorteile schaffen.[10] Dabei müssen Daten jederzeit den betrieblichen und gesetzlichen Anforderungen hinsichtlich vergleichbarer und reproduzierbarer Ergebnisse, Aktualität und Vollständigkeit standhalten.[11] Datenqualität (Data Quality - DQ), Stammdatenmanagement (Master Data Management MDM) und Steuerung der Datenprozesse (Data Governance – DG) sollen dies gewährleisten. Verschärfend wirkt sich dabei ein wachsendes ordnungspolitisches Umfeld mit seinen regulatorischen Anforderungen aus. So hat der Gesetzgeber das Thema Datenqualität mittels Regelungen wie GoBS[12], GDPdU[13], KonTraG[14], Basel II[15], Solvency II[16], SOX[17] usw. gesetzlich verankert.[18] Insbesondere Banken sind hier gefordert. Um den Forderungen nach Transparenz durch Aufsichtsbehörden und anderen Stakeholdern Folge zu leisten, müssen sie ein integriertes System zur Risikosteuerung aufbauen. So sind beispielsweise über das Basel III-Lastenheft hinaus parallele Datensilos aufzulösen, um eine ganzheitliche Sicht auf den gesamten Geschäftsbetrieb zu gewährleisten. Ähnlichen Herausforderungen ist auch die Versicherungsbranche mit Solvency II ausgesetzt, zumal diese schon 2013 umgesetzt sein muss.

Vor diesem Hintergrund wird die Bedeutung von Informations- oder Datenqualität begreifbar: Datenqualität wird als Produktionsfaktor und Teil des Unternehmenswerts angesehen, der aktiv bewirtschaftet wird mit dem Zweck, unternehmensweit optimierte Geschäftsprozesse zu realisieren und dadurch wirtschaftliche Vorteile erzielen. Sie bezeichnet ein Qualitätsmaß für Relevanz und Korrektheit von Daten und Datensätzen und wie gut sie die Realität abbilden. Dadurch wird ihre Verlässlichkeit als Planungsgrundlage für Unternehmensaktionen ersichtlich. Datenqualität erfordert, dass alle Personen, die mit Daten

---

[9] Lackes, R., Schewe, G., Siepermann, M. (o. J.). Informationssystem. Gabler Wirtschaftslexikon. Verfügbar unter: http://wirtschaftslexikon.gabler.de/Definition/informationssystem.html.

[10] o. V. (o. J.). Datenqualität: Warum ist das wichtig? Verfügbar unter: http://www.omikron.net/Datenqualitaet.html.

[11] o. V. (2008). Datenqualität: Erfolgsfaktor für Compliance. Verfügbar unter: http://www.compliancemagazin.de/ complianceservices/beratung/deducto240708.html.

[12] GoBS = Grundsätze ordnungsmäßiger DV-gestützter Buchführungssysteme.

[13] GDPdU = Grundsätze zum Datenzugriff und zur Prüfbarkeit digitaler Unterlagen.

[14] KonTraG = Gesetz zur Kontrolle und Transparenz im Unternehmensbereich.

[15] Basel II = Rahmenvereinbarung über Eigenkapitalempfehlung für Kreditinstitute. Ersetzt seit 1.1.2007 die seit 1988 geltende Eigenkapitalvereinbarung (Basel I). Empfehlungen des Basler Ausschusses stützen sich auf drei Pfeiler: Mindesteigenmittelanforderungen, Aufsichtsrechtlicher Überprüfungsprozess und Kontrolle durch den Markt.

[16] Solvency II = Reformprojekt der Europäischen Kommission zur Festlegung neuer Solvabilitätsvorschriften für Versicherungsunternehmen.

[17] SOX = Sarbanes-Oxley Act; 2002 vom US-Kongress erlassen, soll öffentliches Vertrauen in Wertpapiere wieder herstellen, Führungspraktiken in Unternehmen verbessern, ethische Geschäftspraktiken fördern, Transparenz und Vollständigkeit von Bilanzen erhöhen. Hintergrund sind die Unternehmenszusammenbrüche von Enron, Worldcom und weitere Fälle von Bilanzfälschungen. Wichtigste Neuordnung der Finanzberichterstattung für börsennotierte Unternehmen in den USA seit den 1930er Jahren.

[18] o. V. (o. J.). Datenqualität: Erfolgsfaktor für Compliance. Verfügbar unter: http://www.compliancemagazin.de/ complianceservices/beratung/deducto240708.html.

arbeiten oder dafür verantwortlich sind, die gleiche Auffassung des Qualitätsbegriffs haben. Deshalb sind Kriterien festzulegen, an denen sich Informationen messen lassen. Speziell bei der Daten- bzw. Informationsanalyse spielt Datenqualität eine bedeutende Rolle, und Daten müssen demgemäß eine Reihe von Kriterien erfüllen. Dies betrifft in besonderem Maße Stammdaten. So dürfen z. B. Kundennummern nur in bestimmte Wertebereiche fallen, oder Produktnummern müssen ein bestimmtes Format einhalten.

Dennoch ist es in vielen Unternehmen Realität, dass Datenqualität allein unter operativen Gesichtspunkten betrachtet wird, statt als erfolgskritische Größe. So konstatiert der Präsident von Navesink Consulting im Harvard Business Review, dass die meisten Manager wüssten, dass die Datenqualität in ihren Unternehmen mangelhaft sei, das Thema jedoch keine ausreichende Priorität erhielte, solange keine wirtschaftlichen Argumente dagegen sprächen.[19] Das sich ergebende Dilemma: Ist für Datenqualitätsprobleme die IT-Abteilung zuständig, wird ihnen oft eine niedrige Priorität zugeordnet. Somit müssen sich Geschäftsbereiche selbst um Data Governance und Datenqualität kümmern, da sie tag-täglich damit konfrontiert werden. Einer Gartner-Studie zufolge ist mangelnde Datenqualität dafür verantwortlich, dass 40 Prozent aller Geschäftsinitiativen ergebnislos enden und sich die Mitarbeiterproduktivität um bis zu 20 Prozent verringert. Eine TDWI[20]-Umfrage ergab, dass mangelhafte Datenqualität zu Umsatzeinbußen von 54 Prozent führt und eine Zusatzkostensteigerung von 72 Prozent verursacht bei einer Kundenzufriedenheit von nur 67 Prozent.[21] Um dem entgegenzuwirken, müssen Unternehmen eine Kultur entwickeln, in der Datenqualität hohe Priorität hat.[22] Als nächsten Schritt zu einem unternehmensweiten Datenqualitätsmanagement müssen alle, die für Daten verantwortlich sind, den gleichen Qualitätsbegriff teilen. Um die Datenqualitätsanforderungen zu erfüllen, müssen Daten u. a. folgenden Kriterien entsprechen:[23]

1. Darstellung: Verständliche, übersichtliche, einheitliche, interpretierbare, widerspruchsfreie Darstellung der Daten.

2. Korrektheit: Die Daten müssen mit der Realität übereinstimmen.

3. Aktualität: Alle Datensätze müssen jeweils dem aktuellen Zustand der abgebildeten Realität entsprechen.

4. Eindeutigkeit: Jeder Datensatz muss eindeutig interpretierbar sein.

5. Einheitlichkeit: Einheitliche Struktur der Datensätze.

6. Genauigkeit: Daten müssen in der jeweils geforderten Exaktheit vorliegen, z. B. Anzahl der Nachkommastellen).

7. Konsistenz: Ein Datensatz muss in sich und zu anderen Datensätzen konsistent sein und darf keine Widersprüche aufweisen

8. Redundanzfreiheit: Innerhalb der Datensätze dürfen keine Dubletten vorkommen.

---

[19] o. V. (2013). *Die wahren Kosten fehlerhafter Daten. Sechs einfache Schritte zur Lösung von Datenqualitätsproblemen.* S. 5. Verfügbar unter: http://whitepaper.cio.de/uploads/files/ 69b061f5f397259ab4c81b67553112a1299d184.pdf.

[20] TDWI = **T**he **D**ata **W**arehouse **I**nstitute; URL: www.tdwi.eu

[21] o. V. (2013). Die wahren Kosten fehlerhafter Daten. Sechs einfache Schritte zur Lösung von Datenqualitätsproblemen. S. 1. Verfügbar unter: http://whitepaper.cio.de/uploads/files/ 669b061f5f397259ab4c81b67553112a1299d184. pdf.

[22] S. ebenda, S. 12.

[23] S. ebenda, S. 5.

9. Verständlichkeit: Die Datensätze müssen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen der Fachbereiche übereinstimmen

10. Relevanz: Der Informationsgehalt von Datensätzen muss den jeweiligen Informationsbedarf erfüllen.

11. Struktur: Die Informationen eines Datensatzes müssen einheitlich strukturiert sein.

12. Vollständigkeit: Ein Datensatz muss alle notwendigen Attribute enthalten

13. Zuverlässigkeit: Nachvollziehbare Entstehung der Daten.[24]

Mittels Qualitätskriterien werden Datenprofile für kritische Systeme und Datenquellen erstellt, anhand derer der Ursprung fehlerhaften Daten erkennbar wird und wie gravierend die Probleme sein können. In der Praxis wird Datenqualität als Maß für korrekte Datenspeicherung verstanden und dafür, wie ihre Darstellung mit den erwarteten Ergebnissen übereinstimmt. Dafür muss die Datenqualität sowohl im Erfassungssystem als auch bei der Übertragung in Zielsysteme, z. B. Data Warehouse-Systeme, und alternative Darstellungen, z. B. Auswertungen, überwacht werden. Dazu müssen in jeder Stufe des Datenflussprozesses geeignete Maßnahmen zur Qualitätssicherung umgesetzt werden.[25] Wie und in welchen Bereichen sich mangelnde Datenqualität auswirkt, veranschaulicht Abbildung 1.

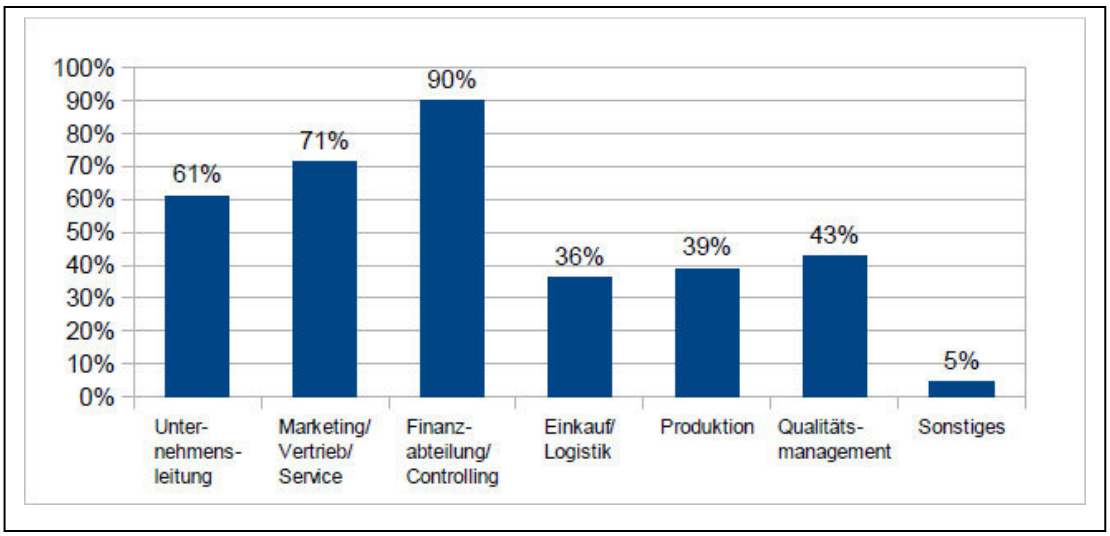


Abbildung 1: Von Datenqualität besonders betroffene Unternehmensbereiche.[26]

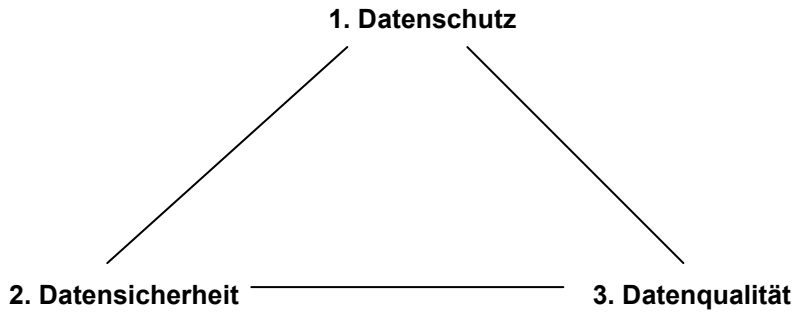
[24] o. V. (2010). Der K(r)ampf mit der Datenqualität. Verfügbar unter: http://www.cfoworld.de/der-krampf-mit-der-datenqualitaet.

[25] o. V. (o. J.). Datenqualität. In: ITWissen. Das große Online-Lexikon der Informationstechnologie. Verfügbar unter: http://www.itwissen.info/definition/lexikon/Datenqualitaet-data-quality.html.

[26] o. V. (2013). Die Bedeutung von Datenqualität. Unabhängige Anwenderbefragung von TDWI und emagixx, März 2013. S. 8/13. Verfügbar unter: http://www.tdwi.eu/fileadmin/user_upload/wissen/whitepaper/Emagixx_Auswertung_Studie_Datenqualit%C3%A4t_2013.pdf.

3.3 Datenschutz im Unternehmen

Der Schutz von Unternehmensdaten bezieht sich auf drei Bereiche:



1. Datenschutz (Privacy Protection, Data Protection): Nach DIN 44300 eine Sachlage, bei der die schutzwürdigen Belange Betroffener vor Beeinträchtigung, die von der Verarbeitung der Daten ausgeht, bewahrt sind. Es geht demnach nicht um Daten per se, wie man annehmen könnte, sondern um ***Personen***, deren vertrauliche, weil personenbezogene Daten verarbeitet werden. Rechtlicher Ausgangspunkt ist das Grundrecht auf informationelle Selbstbestimmung: Jede Person soll selbst bestimmen, wer wann welche Informationen über sie erhält.[27]

Desweiteren umfasst Datenschutz den Schutz

- dieser Daten vor Missbrauch, Schutz der Privatsphäre und Schutz der Persönlichkeit
- personenbezogener Daten vor Missbrauch beim Speichern, Übermitteln, Verändern und Löschen
- vor Ausspähung
- des Einzelnen vor Eingriffen in seine Grundrechte.

2. Datensicherheit (Data Security): Sie beschreibt nach DIN 44300 einen genormten Zustand und beinhaltet

- den Schutz vor Datenverlusten und unberechtigter Einsicht und Manipulation
- die Sicherung vor Verlust durch technisches Versagen oder physische Ereignisse.
- Datensicherheit muss den Schutz der Daten vor fahrlässigen bzw. technischen Fehlern, physikalischen Schäden, externen Schadensquellen und Spionage bzw. Sabotage garantieren. In der betrieblichen Datenverarbeitung betrifft dies alle technischen und organisatorischen Maßnahmen zum Schutz der Daten vor Verfälschung, Zerstörung und unzulässiger Weitergabe. Insbesondere soll Datensicherheit die ***Verlässlichkeit*** (dependability) der Daten gewährleisten. Zu dieser zählen die Kenngrößen (attribtes)
- Vertraulichkeit (confidentiality)
- Integrität (integrity)
- Verfügbarkeit (availiability)
- Verlässlichkeit (reliability)

[27] o. V. (o. J.). Was ist Datenschutz? Verfügbar unter: http://www.datenschutz.de/recht/fundament/datenschutz.

- Sicherheit (safety)
- Wartbarkeit (maintainability).

Aus Sicht des Deutschen Bundesamtes für Sicherheit in der Informationstechnik (BSI) und der ITSEC[28] sind dabei vor allem die Bereiche Vertraulichkeit, Integrität und Verfügbarkeit von Bedeutung.

3. Datenqualität (Data Quality): Voraussetzung ist, dass alle Personen, die mit Daten arbeiten oder dafür verantwortlich sind, die gleiche Auffassung des Qualitätsbegriffs haben. Deshalb sind Kriterien festzulegen, an denen sich Informationen messen lassen. Insbesondere im Kontext der Daten- bzw. Informationsanalyse spielt Datenqualität eine bedeutende Rolle, und Daten müssen demgemäß eine Reihe von Kriterien erfüllen. So dürfen z. B. Kundennummern nur in bestimmte Wertebereiche fallen oder Produktnummern müssen ein bestimmtes Format einhalten.

3.3.1 Datenqualität - antizipatives Handeln erforderlich

Fatalerweise reagiert man in vielen Unternehmen nur auf schlechte Datenqualität. Maßnahmen zur Lokalisierung und Bereinigung ungültiger oder beschädigter Daten werden erst dann ergriffen, wenn diese bereits aufgetreten sind, anstatt sie zu vermeiden. Doch dann haben fehlerhafte Daten in Geschäftsprozessen oft schon gravierenden, irreparablen Schaden angerichtet und möglicherweise bereits andere Systeme infiziert. Daher sollten kontaminierte Daten bereits am Entstehungsort eliminiert werden, um eine Kontamination jener Datenbestände zu vermeiden, in die die fehlerhaften Daten einfließen.[29] Nachdem ein Datenqualitätsbeauftragter (Data Steward) ernannt wurde, die problematischen Daten sind und deren korrekter Zustand feststeht, wird mit IT-Unterstützung Datenbereinigung (engl. Data Cleansing oder Data Scrubbing) durchgeführt. Dazu gehören Verfahren zum Entfernen oder Korrigieren von Datenfehlern in Form von inkorrekten, redundanten, inkonsistenten, veralteten oder falsch formatierten Daten. Vor allem die Redundanzbeseitigung durch Zusammenlegen gleicher Datensätze sowie die Datenfusion durch Vervollständigen lückenhafter Datensätze sind von besonderer Bedeutung.

3.3.2 Datenintegrität - wesentlich für Datenqualität

Umfassende Integrität[30] oder Korrektheit der Daten ist Teil der Datenqualität, denn nur korrekte und zuverlässige Daten sind eine verlässliche Grundlage für richtige Entscheidungen und Handlungen. Aber: Was bedeutet in diesem Zusammenhang „richtig“? Das Prinzip der Richtigkeit und Vollständigkeit von Informationen besagt: Informationen sind richtig, wenn sie Umstände und Tatsachen sachgerecht (bei Personendaten auf die Person bezogen) wiedergeben. Ob eine bestimmte Information oder ein bestimmtes Personendatum

[28] ITSEC = Information Technology Secutity Evaluation Criteria; europäischer Standard für Bewertung und Zertifizierung von Software und Computersystemen hinsichtlich ihrer Funktionalität und Vertrauenswürdigkeit bez. Daten- und Computersicherheit. Zertifizierung nach ITSEC erfolgt in Deutschland durch das Bundesamt für Sicherheit in der Informationstechnik.

[29] o. V. (2013). Die wahren Kosten fehlerhafter Daten. Sechs einfache Schritte zur Lösung von Datenqualitätsproblemen. S. 12. Verfügbar unter: http://whitepaper.cio.de/uploads/files/ 69b061f5f397259ab4c81b67553112a1299d184.pdf.

[30] o. V. (o. J.). Integrität die, allgemein: Makellosigkeit, Unbescholtenheit, Unbestechlichkeit. Verfügbar unter: http://www.enzyklo.de/Begriff/INTEGRIT%c3%84T.

richtig ist, kann deshalb nicht allgemein, sondern nur im konkreten Anwendungsfall beantwortet werden. Die Richtigkeit kann sich nur auf Tatsachen beziehen, die auch objektiv festgestellt werden können. In der Regel erkennbare, subjektive Werturteile sind relativ und lassen auch eine andere Meinung zu.[31]

Dagegen beschreibt nach DIN 44300 Datenintegrität einen genormten Zustand, der die ***Widerspruchsfreiheit*** von Daten bezüglich Plausibilität, Datenmanipulation und inhaltlicher Richtigkeit bezeichnet. Diese ist dann gegeben, wenn sich jedes gespeicherte Datenelement in einem konsistenten, widerspruchsfreien Zustand befindet, so wie es vom letzten autorisierten Benutzer verlassen wurde. Datenintegrität (Syn. Datenkonsistenz) umfasst alle Aspekte, die korrektes und zuverlässiges Arbeiten mit Datenbanken sicherstellen und unterstützen, d. h. Schutz vor Datenverlust, unberechtigter Einsicht und Manipulation ist dann gewährleistet, wenn Informationen über einen bestimmten Zeitraum nicht verändert werden, z. B. beim Erfassen und Speichern in einem Datenbanksystem. Rechtsgrundlage der Datenintegrität sind die Grundsätze ordnungsmäßiger IT-gestützter Buchführungssysteme (GoBS), Kap. 6 Dokumentation und Prüfbarkeit, Abschnitt 6.2.4. Als Maßnahmen zur Wahrung der Datenintegrität sind alle Vorkehrungen zu beschreiben, durch die gewährleistet wird, dass Daten und Programme nicht von Unbefugten geändert werden können. Hierzu gehört neben der Beschreibung des Zugriffsberechtigungsverfahrens der Nachweis der sachgerechten Vergabe von Zugriffsberechtigungen.

3.3.3 Datenintegrität und Datensicherheit - die Wechselwirkungen

Datenintegrität und Datensicherheit beeinflussen sich gegenseitig, weil Datensicherheit die Voraussetzung für Datenintegrität ist: Sie muss den Schutz der Daten vor fahrlässigen bzw. technischen Fehlern, physikalischen Schäden, externen Schadensquellen und vor Spionage bzw. Sabotage garantieren. Zur Datensicherheit zählt der Schutz der Daten vor unberechtigtem Zugriff und absichtlichem oder unabsichtlichem Verändern oder Zerstören. Ziel der Datensicherheit ist es, Verfügbarkeit, Vertraulichkeit, Integrität und Verbindlichkeit der Daten zu gewährleisten. Datenintegrität wiederum ist ein Teilaspekt der Datensicherheit, denn nur integre Daten können Datensicherheit gewährleisten.

3.3.4 Datenintegrität und Datenmanagement

Datenmanagement umfasst alle Prozesse der Planung, Steuerung und Kontrolle im Zusammenhang mit:

- Datendefinition und -modellierung
- Datenbankentwurf und -betrieb
- Bedarfsgerechter Datenaufbereitung und -präsentation
- Gewährleistung von Datensicherheit und Datenschutz
- Externer Datenbeschaffung.

Anforderungen an das Datenmanagement sind:

- Gewährleistung der Datenintegrität
- Minimale Zugriffs- und Übertragungszeiten

[31] Definition lt. International Data Group (IDG), Boston/USA.

- Minimale Redundanz
- Anpassbarkeit der Datenstruktur und der Daten
- Möglichkeiten der Auswertung und Verknüpfung
- Wirtschaftlichkeit.

4 Business Intelligence - Wissensgenerierung aus Unternehmensprozessen

4.1 Von MIS zu ESS - Entwicklung der Management Support Systems

Computergestützte Informationssysteme lassen sich in operative und analyseorientierte Systeme[32] unterteilen, wobei letztere hier der Betrachtungsgegenstand sind.

Die Anfänge von IT-Systemen zur Aufbereitung betriebswirtschaftlicher Daten für die Managementunterstützung liegen in den späten 1960ger Jahren. Diese als Management Information Systems (MIS) bezeichneten Dialog- und Transaktionssysteme zur Speicherung großer Datenmengen waren die Antwort auf das Verlangen von Führungskräften nach Informationssystemen zur automatischen Informationsgenerierung für Planungs- und Kontrollaufgaben mit folgenden Funktionen:

- Erstellung periodischer, standardisierter Berichte
- Verfügbarkeit für alle Managementebenen
- Aggregierte und zentralisierte Informationen über alle Geschäftsaktivitäten
- Größtmögliche Aktualität und Korrektheit.

Das letzte Kriterium verdeutlicht, dass Datenqualität und -integrität von Anfang an eine grundlegende Anforderung an das Datenmanagement sind.

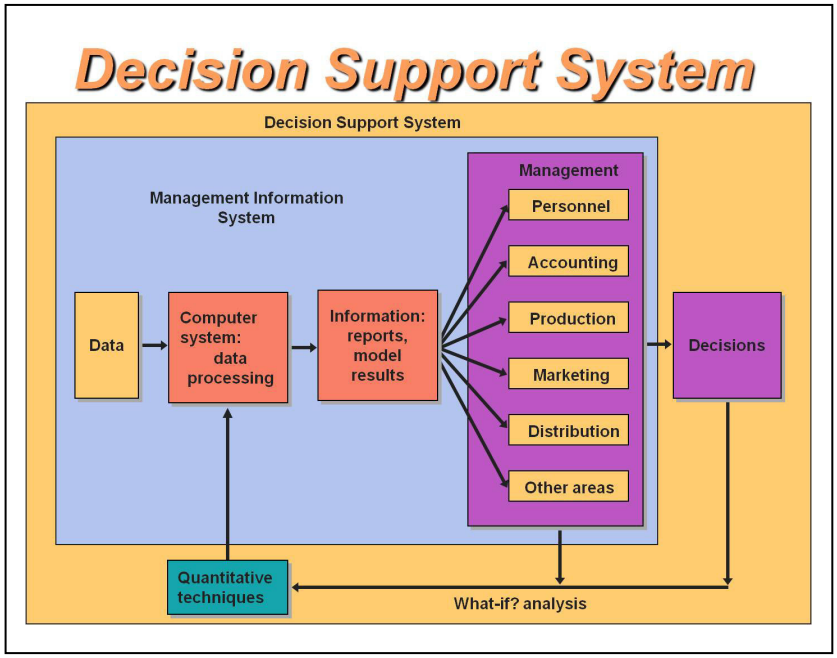


Abbildung 2: Schematische Darstellung eines Decision Support System.[33]

[32] S. Gabriel, R., Gluchwoski, P., Pastwa, A. (2011). Data Warehouse & Data Mining. Herdecke Witten: W3L GmbH. S. 16.
[33] Verfügbar unter: http://sarjeevansainbhi.webs.com/apps/photos/photo?photoid=136220064.

Diese Administrations- und Dispositionssysteme ermöglichten es Führungskräften, über Hierarchieebenen hinweg verdichtete Daten unterschiedlicher Granularität aus der operativen Datenbasis ohne aufwändige Datenmodellierung und Analyseverfahren zu extrahieren.[34] Da diese Systeme aber nicht alle Funktionen abdeckten und eher Insellösungen mit allen Nachteilen darstellten, wurde bereits Ende der 1960er Jahre über andere Lösungen nachgedacht. Der Hauptgrund für das frühe Ende der MIS lag in der fehlenden Datenfilterung, -säuberung und -verdichtung auf operativer Ebene. Dadurch kam es trotz einer Flut von Daten zu einem Informationsmangel aufgrund der Menge irrelevanter Informationen.[35]

Diesen Mangel sollten ab Mitte der 1970er Jahre interaktive Decision Support Systems (DSS) bzw. Entscheidungsunterstützungssysteme (EUS) beheben. Diese computergestützten Planungs- und Informationssysteme (Syn. computergestütztes Planungssystem, Führungsinformationssystem (FIS)), unterstützen die Entscheidungsvorbereitung auf den Führungsebenen, indem entscheidungsrelevante Informationen verdichtet und in Tabellen oder Grafiken dargestellt werden. Im Gegensatz zu MIS sollten DSS dass Management nicht allein mit Informationen versorgen, sondern den Planungs- und Entscheidungsprozess aktiv unterstützen, mittels Modellen, Methoden und problembezogenen Daten das Urteilsvermögen der Entscheidungsträger steigern und damit die Entscheidungsqualität verbessern. DSS haben eine eigene Datenbank – ein bedeutsamer Unterschied zu MSS –, in der Daten der operativen Datenbasis aufbereitet werden, typischer weise:

- Wöchentliche Verkaufs- bzw. Umsatzzahlen
- Geplante Umsatzzahlen neuer Produkte
- Mögliche Auswirkungen verschiedener Entscheidungsalternativen auf Basis von Erfahrungswerten
- Alle Daten, auf die unternehmensweit zugegriffen werden kann, d. h. auch solche aus Alt-(Legacy-) Beständen, relationalen Datenbanken, Cubes und Datenlagern.

Auch DSS konnten sich nicht dauerhaft durchsetzen, zum einen, weil auch sie lokal ausgerichtet waren und keine unternehmensweiten Modelle zur Simultanplanung erstellten, zum anderen stießen hier die IT-Kenntnisse vieler Manager an ihre Grenzen.[36]

Die neue MIS-Generation der Executive Information Systems (EIS) gewährte besseren Datenzugang und Kommunikationsunterstützung[37] durch Zugriff auf E-Mail, Fax und Intranet. Als konzeptionelle und funktionale Weiterentwicklung der MIS sind sie nicht nur für das Top Management nutzbar, sondern von allen Mitarbeitern mit Analyseaufgaben. Entwickelt wurden sie in den 1980er Jahre, als im Zuge einer stärkeren DV-Vernetzung und leistungsstarker PC's mit anwenderfreundlicher Benutzeroberfläche eine bessere Basis für das MIS-Konzept boten. Darauf basierend, wurden mit EIS (Syn. Führungsinformationssystem FIS) leistungsstarke, dialog- und datengestützte Informationssysteme mit ausgeprägten Kommunikationselementen zur Entscheidungsunterstützung[38], die aktuelle entschei-

[34] S. ebenda, S. 20 - 21
[35] S. ebenda, S. 24
[36] S. ebenda, S. 27
[37] Kommunikationsunterstützung = Unterstützung jeglicher Art von Informationsaustausch zwischen menschlichen und/oder elektronischen Kommunikationspartnern innerhalb eines Unternehmens und mit externen Partnern durch Zugriff auf E-Mail, Fax, Intranet etc.
[38] Entscheidungsunterstützung = Unterstützung menschlicher Entscheidungen durch intelligente wissensbasierte Anwendungssystemfunktionen.

dungsrelevante interne und externe ohne Entscheidungsmodell zur Auswahl und Analyse über individuell anpassbare Benutzeroberflächen anbieten.[39]

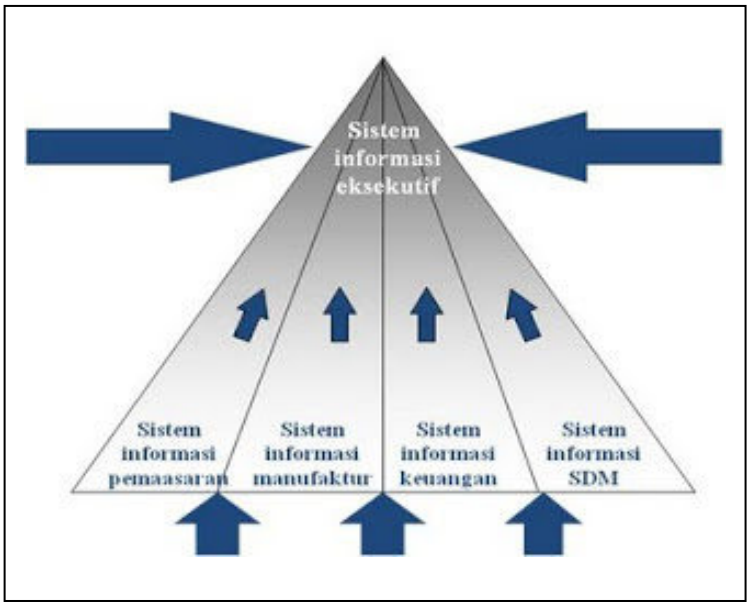


Abbildung 3: Executive Information Model (EIS).[40]

Durch Integration von präsentations- und kommunikationsorientierten EIS und problemlösungsorientierten DSS entstanden die Executive Information Systems ESS und damit die Verbindung von vergangenheitsorientierter Informationsversorgung und zukunftsorientierter Analyse.[41] Als Kombination aus Data Support und Decision Support sind sie merklich leistungsstärker, da durch Benutzerfreundlichkeit, grafische Informationsaufbereitung und Reduktion komplexer Informationszusammenhänge eines EIS einerseits und entscheidungsunterstützende Analyse- und Modellfunktion eines DSS Synergiepotentiale voll ausgeschöpft werden und es kann sowohl vergangenheitsbezogen dokumentiert als auch zukunftsorientiert analysiert werden.[42]

4.2 BI - Versuch einer Begriffserklärung

Der Begriff „Business Intelligence“ (BI) ist nicht neu und wird ebenso heterogen verwendet wie die Begriffe Daten oder Informationen. BI hat seit langem einen festen Platz in Unternehmen wie auch in der Informatik, der Informationswissenschaft oder der Betriebswirtschaftslehre und ist in die jeweiligen Fachterminologien eingegangen. Dabei existieren ebenso unterschiedliche Definitionen wie Begriffsabgrenzungen. Erstmals verwendet wurde der Begriff 1958 vom IBM-Infomatiker Hans Peter Luhn in seiner im IBM Journal erschienenen Abhandlung *“A Business Intelligence System“*. Dort beschreibt er eine Methode zur Aufbereitung unstrukturierter Informationen in Form von internen und externen Dokumenten zur Verteilung an Mitarbeiter als Grundlage für schnellere und genauere Entscheidungen.[43] 1989 wurde der Begriff von Howard Dresner, einem ehemaligen

[39] Gabriel, R. Gluchwoski, P., Pastwa, A. (2011). Data Warehouse & Data Mining. Herdecke Witten: W3L GmbH. S. 28 - 30.
[40] Verfügbar unter: http://aboutinformationsystem.blogspot.de/2011/07/characteristics-of-executive.html
[41] S. ebenda, S. 33.
[42] S. ebenda, S. 34 - 35
[43] Raaz (o. J.). S. 1/6.

Analysten der Gartner Group, aufgegriffen. Er deiniert BI als *"...process of transforming data into information and, through discovery into knowledge."* (Sic!).[44] In diesem Sinne sieht Dresner BI als Oberbegriff für die Verdichtung und Analyse von Informationen zu nutzbarem Wissen.

Heute wird BI im Allgemeinen mit „Geschäftsanalytik" übersetzt. Es bezeichnet einen (idealerweise) unternehmensweiten, integrierten, unternehmensspezifischen, IT-basierten Ansatz zur betrieblichen Managementunterstützung. Er basiert auf der systematischen Erfassung, Speicherung, Analyse, Präsentation und Kommunikation von Geschäftsdaten. Durch Einsatz von Ad-hoc-Abfragen (Queries), Online-Analysen (Online Analytical Processing – OLAP), Data Mining, Reporting etc. können Unternehmen fachbereichs- und managementspezifische Auswertungen und Analysen der internen wie externen Markt- und Wettbewerbsdaten erstellen und sogar Prognosen abgeben, um zukünftige Chancen und Potenziale wie auch mögliche Bedrohungen zu erkennen. Diese können beispielsweise Absatzzahlen eines neuen Produktes innerhalb eines bestimmten Zeitraumes und Verkaufsgebietes betreffen, die Wirksamkeit von Marketingaktionen, die Liquidität oder die mögliche Abwanderung von Kunden, also strategische Entscheidungen über den Kurs des Unternehmens. So gesehen ist BI nichts Neues: Schon immer wurden aus Geschäftsprozessen Rückschlüsse auf die Geschäftsentwicklung gezogen. Daher hat BI zunächst nichts mit aufwändigen IT-Lösungen zu tun. Der Mehrwert von BI besteht darin, die richtigen Informationen zum richtigen Zeitpunkt den Entscheidern als Entscheidungsgrundlage zur Verfügung zu stellen. Demgemäß kann diese Entscheidungsgrundlage ein handgeschriebener Zettel aus der Fertigung mit der Tagesproduktion sein. Auch hier werden aus Geschäftsprozessen aggregierte Informationen gewonnen, in diesem Falle über die Produktivität des Unternehmens.

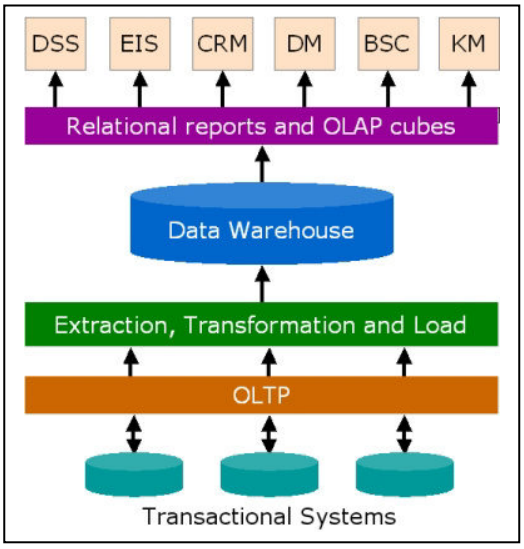


Abbildung 4: Beispiel einer BI-Architektur.[45]

4.3 Business Intelligence und Datenqualität - der Input als kritische Größe

BI ist ein Prozess zur systematischen, IT-gestützten Datenanalyse mit dem Ziel, qualitätvollere operative oder strategische Entscheidungen im Unternehmen zu treffen durch Gewinnung von Erkenntnissen. Mit diesen können Unternehmensprozesse rationaler gestal-

[44] S. ebenda.
[45] Verfügbar unter: http://www.b2.adm.br/componentsandarchitectureofabusinessintelligencebisystem.htm.

tet, Kunden- und Lieferantenbeziehungen optimiert, Kosten gesenkt, Risiken minimiert und die Wertschöpfung gesteigert werden. So betrachtet ist es selbst erklärend, dass der Output nur so gut sein kann wie der Input: Die Qualität der zu analysierenden Daten ist ebenso maßgebliche wie auch erfolgskritische Größe. Sie entscheidet über Effektivität, Erfolg und Akzeptanz von BI-Projekten. Die Realität zeigt jedoch, dass oft genug fehlerhafte Daten in das zentrale Data Warehouse fließen und somit die Qualität der Ergebnisse verringern, da niemand geschäftliche Entscheidungen auf mangelhafte Daten stützen will. Dies belegt auch die Befragung von mehr als 2.600 BI-Nutzern, Anbietern und Beratern der BARC[46]-Studie 2010 "The BI Survey 9": Galt zuvor die Abfragegeschwindigkeit als größtes Problem von BI-Software, so ist es jetzt die Datenqualität.[47] Folgende Fakten belegen dies nachdrücklich:

- Unabhängig von der Datenqualität veralten jeden Monat zwei Prozent des Datenbestands.
- Aufgrund mangelnder Datenqualität bleiben bis zu 70 Prozent aller Data Warehouse-Prozesse[48] (engl. Data Warehousing) wirkungslos.[49]
- Schlechte Datenqualität lässt Unternehmen bis zu 25 Prozent des operativen Gewinns einbüßen.[50]

Diese Zahlen sind umso alarmierender, da Entscheider bereits seit vielen Jahren das Datenqualitätsmanagement (DQM) im BI-Umfeld als wichtigstes Thema identifiziert haben und das Problem als gelöst gelten sollte. Und wenngleich unternehmensweite Datenqualität als Produktionsfaktor und Teil des Unternehmenswerts betrachtet wird, ist eine fehlerhafte Datenbasis immer noch ein kritischer Faktor bei BI-Projekten.[51] Die Ursache für schlechte Datenqualität ist oft nicht bei den eingebenden Mitarbeitern zu suchen. Schlechte Datenqualität ist oft das Ergebnis mangelnder Einweisung, unzureichender Qualifizierung und eine ungenügende Bildung.[52] Häufig sind es nicht gepflegte und daher fehlerhafte Kundendaten, die direkte Kosten verursachen, wenn Artikel nicht rechtzeitig geliefert werden, oder es wegen falscher Adressen zu Rechnungsausfällen kommt. Indirekte Kosten entstehen dagegen z. B. durch Imageverlust.[53]

Abhilfe könnte nur die Einführung eines Datenqualitätsmanagement (DQM) schaffen. Erster Schritt ist die unternehmensweite Integration eines dreistufigen DQ-Prozesses:

1. Fehler erkennen: In ersten Schritt bleibt die Fehlerquote hoch. Es werden nur die Schwachstellen aufgezeigt, ohne dass ein messbarer Nutzen nicht erzielt wird.

---

[46] BARC = Business Application Research Center; URL: www.barc.de

[47] o. V. (2010). Datenqualität ist größtes Problem. Verfügbar unter: http://www.monitor.at/index.cfm/storyid/13122.

[48] Data Warehouse (DWH) = zentrales oder verteiltes Datenlager, in dem Daten aus verschiedenen Quellen in einer separaten Datenbank zusammengefasst werden. Im Data Warehouse-Prozess wird dieses Datenlager mittels folgender Schritte bewirtschaftet und ausgewertet: 1. ETL-Prozess, d.h. Extraktion, Transformation, Laden von Daten aus den Quellsystemen ins Data Warehouse, 2. Langfristige Datenspeicherung, 3. Bereitstellung der für die aktuelle Abfrage benötigten Daten, 4. Datenanalyse im jeweiligen Bereichs-Data Mart bzw. Versorgung nachgelagerter Anwendungssysteme, z. B. Dashboards.

[49] o. V. (2010). Datenqualität & Business Intelligence: Analysen benötigen hochwertige Daten. BI-SPEKTRUM 02.2010. Verfügbar unter: http://www.trilliumsoftware.com/uploadedFiles/Deutsch_Site/Dokumente_and_Downloads/ BI_Spektrum.pdf.

[50] Buck, D. (2012). Datenqualität K.o.-Kriterium für Business Intelligence. Verfügbar unter: http://www.computerwoche.de/a/k-o-kriterium-fuer-business-intelligence,1938325.

[51] S. ebenda.

[52] o. V. (o. J.). Business Intelligence und Datenqualität. Verfügbar unter: http://www.izug.de/?p=11.

[53] Buck, D. (2012). Datenqualität K.o.-Kriterium für Business Intelligence. Verfügbar unter: http://www.computerwoche.de/a/k-o-kriterium-fuer-business-intelligence,1938325.

2. Fehler beheben: Im zweiten Schritt sinkt die Fehlerquote anfangs. Danach ist die Quote gewissen Schwankungen ausgesetzt. Die permanente Bindung von Ressourcen schränkt das Tagesgeschäft. Aufwand und Nutzen sind gleich hoch.
3. Fehler vermeiden: Im letzten Schritt sinkt die Fehlerquote nachhaltig auf ein niedriges Niveau. Der Nutzen steigt bei gleichzeitiger Verminderung des Aufwands.[54]

Es ist selbstverständlich, dass bei der Untersuchung der Datenbasis ausnahmslos alle Daten berücksichtigt werden, und dass nicht z. B. nur die Anschrift, sondern auch die Anrede gepflegt sein muss.

4.4 Datenanalyse mit OLAP: Abfrage multidimensionaler Datenmodelle

4.4.1 OLAP - Begriff und Bedeutung

Geprägt wurde der Begriff **On**line **A**nalytical **P**rocessing OLAP (Online-Analyse dispositiver Daten) 1993 von dem englischen „Datenbankpapst“ und Entwickler des relationalen Datenmodells Dr. Edgar F. Codd in seinem Artikel *“Beyond Decision Support“*. Gedacht war OLAP als Gegenstück zu OLTP (Online Transaction Processing). Bezieht sich OLTP auf die Verarbeitung operativer Daten und damit auf die Automatisierung von Geschäftsprozessen wie Bestellabwicklung oder Verkauf, so werden bei OLAP meist viele verschiedene, auch aggregierte und historische Daten angesprochen und analysiert. Die daraus hervorgehenden Informationen dienen der Entscheidungsunterstützung.[55] Damit war OLAP die Reaktion auf den Ruf nach benutzerfreundlichen, flexiblen Ad-hoc-Analysen. Die Beziehung zwischen OLAP und OLTP ist dennoch eng, denn die Daten, die in ein OLAP-System geladen werden, haben ihren Ursprung im OLTP.

OLAP steht für eine Vielzahl von Methoden, mit denen unterschiedliche Anwendungsbedürfnisse bei unternehmensweiten Geschäftsprozessanalysen erfüllt werden sollen.[56]

Um Informationen für eine effiziente Unternehmensplanung umgehend bereitzustellen, ermöglicht OLAP den permanenten Zugriff auf unterschiedliche Datenquellen. Dadurch lassen sich Daten aus verschiedenen Sichtweisen multidimensional darstellen, Geschäftsprozesse analysieren und mögliche Gefahrenpotentiale eruieren.[57] Hierzu Codd: *„OLAP sind zahlreiche „Was-wäre-wenn“ und/oder „Warum“-Szenarien, welche innerhalb einer spezifischen historische Basis und Perspektive ausgeführt werden. Dynamische Unternehmensanalysen sind notwendig, um Informationen aus Unternehmensdatenmodellen zu kreieren, zu manipulieren, anzuregen und herzustellen. Das beinhaltet die Fähigkeit, neue oder unvorhergesehene Verbindungen zwischen Variablen zu erkennen, die nötigen Parameter zu identifizieren, um große Mengen an Daten zu handhaben und eine unendliche Anzahl an Dimensionen zu kreieren und kreuzdimensionale Bedingungen bzw. Ausdrücke zu spezifizieren.“*[58] Die Anforderungen an ein OLAP-System hat Codd in seinen bekannten zwölf Regeln beschrieben, zu denen später aufgrund gestiegener Anforderungen die „6 erweiterten Coddschen Regeln“ hinzukamen.

---

[54] o. V. (o. J.). Business Intelligence und Datenqualität. Verfügbar unter: http://www.izug.de/?p=11.
[55] o. V. (o. J.). OLAP in: HMD Handbuch der Modernen Datenverarbeitung. Verfügbar unter: http://hmd.dpunkt.de/glossar/gesamt/a.php#agile_business_intelligence.
[56] o. V. (1999). OLAP-Leitfaden. S. 2. Verfügbar unter: http://www.klesse.de/files/Applix_OLAP_leitfaden.pdf.
[57] S. ebenda, S. 3.
[58] Richter, W. (2001). Virtual Communities und Customer Relationship Management. Zitiert nach Kurz, A. (1999). Data Warehousing Enabling Technologies. Verlag MITP.
Verfügbar unter: http://www.iicm.tugraz.at/Teaching/theses/2001/_idb8c_/wrichter/thesis-final/node58.html.

4.4.2 Datenmodellierung in drei Dimensionen

Grundlegend OLAP-Funktionalität ist die multidimensionale Sichtweise auf die Daten.

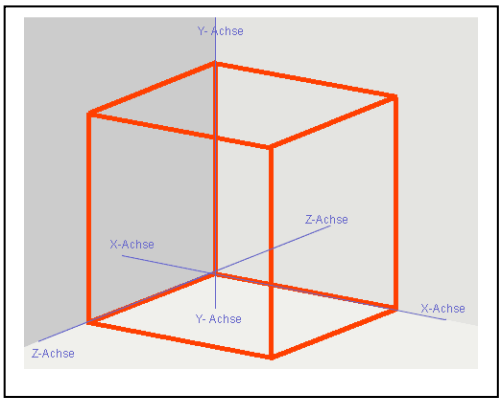
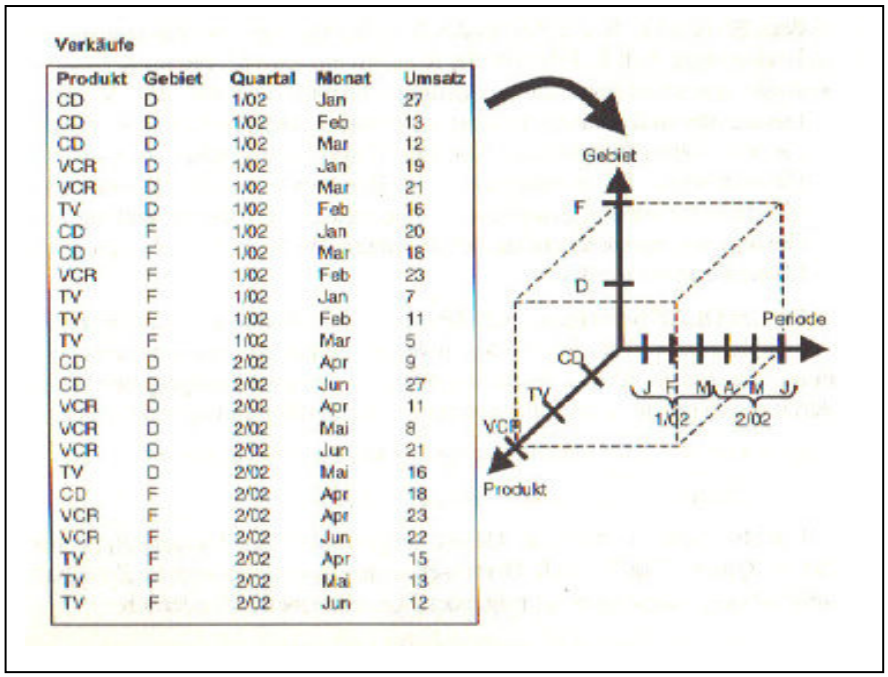


Abbildung 5: Würfeldarstellung mit x-, y- und z-Achse.[59]

Dabei geht OLAP von zwei Prinzipien aus: Zum einen zeigt es den Übergang von einer relationalen bzw. zweidimensionalen Darstellung hin zu einer dreidimensionalen, wie in Abbildung 6 dargestellt. Eine Tabelle ist die einfachste Form der Mehrdimensionalität, wobei die einzelnen Zellen bzw. Datenfelder durch ihre Koordinaten bestimmt werden. Fügt man eine weitere Dimension hinzu, wird die Tabelle dreidimensional. Dies wird durch einen imaginäre Tiefenachse, die z-Achse erreicht, die man sich nur durch räumliches Denken vorstellen kann, da auf dem Papier oder am PC eine solche Darstellung bestenfalls als kaum handhabbares Hilfskonstrukt möglich ist.



Verkäufe

Produkt	Gebiet	Quartal	Monat	Umsatz
CD	D	1/02	Jan	27
CD	D	1/02	Feb	13
CD	D	1/02	Mar	12
VCR	D	1/02	Jan	19
VCR	D	1/02	Mar	21
TV	D	1/02	Feb	16
CD	F	1/02	Jan	20
CD	F	1/02	Mar	18
VCR	F	1/02	Feb	23
TV	F	1/02	Jan	7
TV	F	1/02	Feb	11
TV	F	1/02	Mar	5
CD	D	2/02	Apr	9
CD	D	2/02	Jun	27
VCR	D	2/02	Apr	11
VCR	D	2/02	Mai	8
VCR	D	2/02	Jun	21
TV	D	2/02	Mai	16
CD	F	2/02	Apr	18
VCR	F	2/02	Apr	23
VCR	F	2/02	Jun	22
TV	F	2/02	Apr	15
TV	F	2/02	Mai	13
TV	F	2/02	Jun	12

Abbildung 6: Von der zweidimensionalen Tabelle (Relation) zum dreidimensionalen Würfel.[60]

[59] Verfügbar unter: http://www.at-mix.de/y_achse.htm
[60] Bodendorf, F. (2003). Daten- und Wissensmanagement. S. 33. Berlin Heidelberg New York Springer-Verlag.

Synonym für OLAP-Datenbanken ist der Würfel (engl. cube), der auch aus drei Dimensionen besteht, wenngleich OLAP-Datenbanken oft bedeutend mehr als drei Dimensionen umfassen. Werden mehr als drei Dimensionen kombiniert, spricht man von Multi- oder n-Dimensionalität[61] bzw. MOLAP.

4.4.3 Dimensionen und Kennzahlen

Der zweite Gedanke ist die logische Trennung zwischen unabhängigen Attributen (Dimensionen), z. B. Produkt, Verkaufsgebiet, Zeit, und abhängigen betriebswirtschaftlichen Kennzahlen (Fakten) wie Umsatz, Erlöse, Gewinne und Verluste, die zu analysieren sind. So kann etwa der Umsatz eines Produktes innerhalb eines bestimmten Gebiets und Zeitraums analysiert werden. Jedem unabhängigen Attribut wird eine eigene Dimension zugewiesen, die dadurch einen Vektorraum aufspannen. Zumeist sind es die drei „klassischen“ Dimensionen Produkte, Regionen und Zeit, die in die Auswertungen einfließen, aber auch Auswertungen nach Kunden, Filialen, Abteilungen, Vertretern usw. sind möglich. Die Werte der abhängigen Kennzahlen können in diesem Vektorraum angeordnet werden, weil deren Position durch die Kombination der Werte entlang der Dimensionen eindeutig bestimmt ist. Bei mehr als drei unabhängigen Dimensionen erhält man einen hyper- oder multidimensionalen Datenwürfel.

Gemäß der 12. Regel nach Codd sollte die Anzahl der Dimensionen theoretisch unbegrenzt sein. Diese Forderung ist allerdings utopisch; wie sollte man solch einen Datenwürfel darstellen bzw. modellieren? Daher findet man normalerweise fünf bis acht Dimensionen.[62] Die Anordnung der Dimensionen in Hierarchien und der abhängigen Kennzahl Umsatz im Würfel zeigt Abbildung 7. OLAP analysiert diese multidimensionalen Datenwürfel und erlaubt dem Anwender eine dynamische Datenauswertung nach frei definierbaren Auswahlkriterien oder Dimensionen.[63]

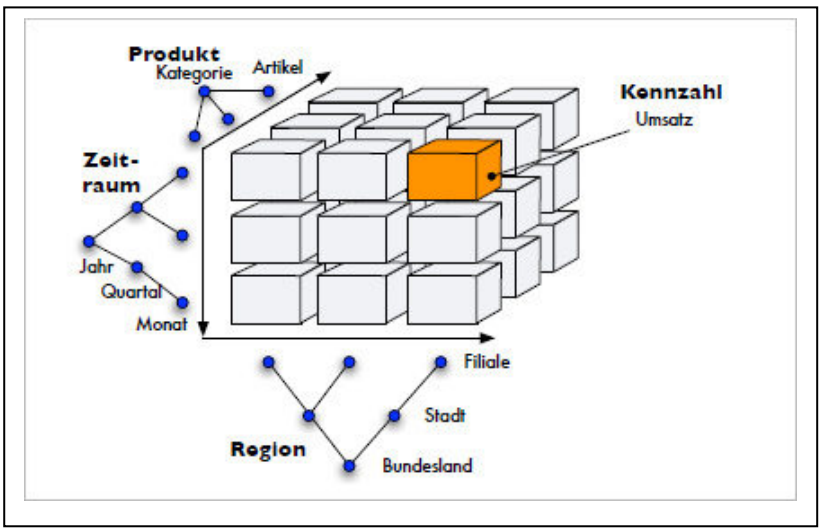


Abbildung 7: Verteilung der Dimensionen und der zu analysierenden Kennzahl im Datenwürfel.[64]

[61] o. V. (1999). OLAP-Leitfaden. S. 5. Verfügbar unter: www.klesse.de/files/Applix_OLAP_leitfaden.pdf.
[62] Bilke, P. und St. (2007). Unternehmensdaten gekonnt auswerten. S. 17. (2. durchgesehene Auflage:2007-09) KnowWare Vertrieb GmbH: Osnabrück.
[63] Bodendorf, F. (2003). Daten- und Wissensmanagement. S. 33. Berlin Heidelberg New York Springer.
[64] o. V. (o. J.). Multidimensionales Datenmodell. S. 1, Folie 3. Verfügbar unter: http://www.inf.uni-konstanz.de/ fileadmin/dbis/DWOLAP/4-Mddm-4.pdf.

Dynamische Abfragen bzw. Analysen ermöglichen die individuelle Anpassung der Fragestellung des Benutzers und die Navigation innerhalb des Datenbestandes anstelle statischer Berichte.

5 Schlussbetrachtung

An dieser Stelle schließt sich der Kreis. Es ist selbsterklärend, dass bei all diesen Prozessen die Qualität der zu modellierenden und zu analysierenden Daten nie aus dem Blickfeld geraten darf: Sie ist das integrierende Glied zwischen operativer und analysierender Datenhaltung und letztlich verantwortlich für ein qualitätvolles Ergebnis. Dennoch betreiben mehr als 90 Prozent der Unternehmen das Thema Datenqualität nicht systematisch: Stichwort „Data Governance"? Fehlanzeige! Dabei bezeichnet gerade dieser Ansatz die Datenverwaltung mit dem Ziel, valide und konsistente Informationen bereitzustellen und dies nach dem Grundsatz der Informationslogistik *„Die benötigte Information zur richtigen Zeit, am richtigen Ort, im richtigen Format und in der erforderlichen Qualität"*. Darunter subsumieren sich Disziplinen wie Datenintegration, Datenqualität oder Stammdatenmanagement, die zyklisch für die Verarbeitung, Analyse und Überwachung der Daten verantwortlich sind. Speziell die Datenqualität kann erst im Rahmen eines Zyklus Datenqualität gewährleisteten. So nutzen laut Umfrage des BARC nur neun Prozent (sic!) der Befragten Data Governance-Programme zur langfristigen Sicherung und Kontrolle der Datenqualität ihrer SAP-Systeme oder Applikationen. 91 Prozent dagegen verzichten darauf.

Literatur

Bodendorf, F. (2003). *Daten- und Wissensmanagement*. Berlin Heidelberg New York Springer-Verlag

Bilke, P. und St. (2007). *Unternehmensdaten gekonnt auswerten.* (2. durchgesehene Auflage: 2007-09) KnowWare Verlag GmbH: Osnabrück

Engels, Ch. (2010). *Basiswissen Business Intelligence.* Herdecke Witten: W3L GmbH

Gabriel, R. Gluchwoski, P., Pastwa, A. (2011). *Data Warehouse & Data Mining.* Herdecke Witten: W3L GmbH

Internetquellen

Buck, Daniel (2012). *Datenqualität K.o.-Kriterium für Business Intelligence.* Verfügbar unter: http://www.computerwoche.de/a/k-o-kriterium-fuer-business-intelligence,1938325.

Grosser, T., Bange, C. (2009*). Datenqualität in SAP-Systemen. Die unabhängige Anwenderbefragung über die Wahrnehmung der Datenqualität in SAP-Systemen.* BARC-Institut Würzburg, August 2009. Verfügbar unter: http://www.barc.de/datenqualit%C3%A4t-in-sap-systemen.

Lackes, R., Schewe, G., Siepermann, M. (o. J.). *Informationssystem.* Gabler Wirtschaftslexikon. Verfügbar unter: http://wirtschaftslexikon.gabler.de/Definition/informationssystem.html.

o. V. (o. J.). *Business Intelligence und Datenqualität.* Verfügbar unter: http://datenanalyse.org/ ndex.php/blog/informations-und-datenqualitaet/21-business-intelligence-und-datenqualitaet.

o. V. (o. J.). *Business Intelligence und Datenqualität.* Verfügbar unter: http://www.izug.de/?p=11.

o. V. (o. J.). *Business Software. Definitionen und Abkürzungen aus der IT-Branche.* Verfügbar unter: http://www.softselect.de/business-software-glossar.

o. V. (2008). *Datenqualität: Erfolgsfaktor für Compliance.* Verfügbar unter: www.compliancemagazin.de/complianceservices/beratung/deducto240708.html.

o. V. (o. J.). *Datenqualität: Warum ist das wichtig?* Verfügbar unter: http://www.omikron.net/ Datenqualitaet.html.

o. V. (o. J.). *Datenqualität ist größtes Problem.* Verfügbar unter: http://www.monitor.at/index.cfm/ storyid/13122_Business-Intelligence-Datenqualitaet_ist_groesstes_Problem.

o. V. (2010). *Datenqualität & Business Intelligence: Analysen benötigen hochwertige Daten.* BI-SPEKTRUM 02-2010. Verfügbar unter: http://www.trilliumsoftware.com/uploadedFiles/Deutsch_Site/Dokumente_and_Downloads/ BI_Spektrum.pdf.

o. V. (2010). *Der K(r)ampf mit der Datenqualität*. Verfügbar unter: http://www.cfoworld.de/der-krampf-mit-der-datenqualitaet.

o. V. (o. J.). *Die wahren Kosten fehlerhafter Daten. Sechs einfache Schritte zur Lösung von Datenqualitätsproblemen.* Verfügbar unter: http://whitepaper.cio.de/uploads/files/ 669b061f5f397259ab4c81b67553112a1299d184.pdf.

oV. (o. J.). *Erfolgsfaktor DATENQUALITÄT*. Verfügbar unter: http://www.omikron.net/Kunden-nach-Branche.html.

o. V. (o. J.). *Information.* In: *ITWissen. Das große Lexikon der Informationstechnologie*. Verfügbar unter: http://www.itwissen.info/definition/lexikon/Information-information.html.

o.V. (o. J.) *Leitthema Informationslogistik*. Verfügbar unter: www.informationslogistik.org.

o. V. (o. J.). *Multidimensionales Datenmodell.* S. 1, Folie 3. Verfügbar unter: http://www.inf.uni-konstanz.de/fileadmin/dbis/DWOLAP/4-Mddm-4.pdf.

o. V. (o. J.). *OLAP.* In: *HMD **H**andbuch der **M**odernen **D**atenverarbeitung.* Verfügbar unter: http://hmd.dpunkt.de/glossar/gesamt/a.php#agile_business_intelligence.

o. V. (o. J.). *Rahmen von Solvency II.* Verfügbar unter: http://www.solvency-ii-kompakt.de/content/ rahmen-von-solvency-ii.

o. V. (o. J.) *Was ist Datenschutz?* Verfügbar unter: http://www.datenschutz.de/recht/fundament/ datenschutz/.

Datenqualität in der Deutschen Digitalen Bibliothek

Francesca Schulze

Die „Deutsche Digitale Bibliothek“ (DDB) ist ein Portal und eine Plattform für Kultur und Wissenschaft in Deutschland. Sie bietet einen kostenfreien und zentralen Zugang zu digitalen Angeboten aus Bibliotheken, Archiven, Museen, Denkmalpflegeämtern, Mediatheken sowie Forschungs- und Wissenschaftseinrichtungen. Dafür bringt sie die Kultur- und Wissenschaftseinrichtungen in einem Netzwerk zusammen und ermöglicht so die spartenübergreifende Präsentation und Vernetzung der digitalen Bestände. Seit Abschluss der öffentlichen Testphase im März 2014 ist das Portal mit neuen Funktionalitäten und Datenbeständen regulär online (www.deutsche-digitale-bibliothek.de). Die DDB ist jedoch mehr als ein Portal. In ihrer Rolle als digitale Plattform sorgt sie für die Zusammenführung, Aufbereitung und nachhaltige Vernetzung der heterogenen Datenbestände. Von diesem Mehrwert profitieren nicht nur die Besucher des DDB-Portals, sondern auch weitere Nutzergruppen. Als nationaler Aggregator ist die DDB beispielsweise für die Weitergabe der Daten an das europäische Kulturportal „Europeana“ (www.europeana.eu) zuständig, wodurch sich die Sichtbarkeit der Einrichtungen mit ihren Beständen in anderen Internetangeboten erhöht. Zudem hat sie mit der Veröffentlichung der Anwendungsprogrammierschnittstelle (API) Ende 2013 ein Angebot geschaffen, mit dem externe Anwender die Metadaten weiterverwenden und in unterschiedliche Kontexte einbinden können. Besondere Herausforderungen bestehen darin, die heterogenen Metadaten der Einrichtungen zu vereinheitlichen, mit Normdaten aus externen Datenquellen anzureichern und in standardisierte Datenformate, wie dem DDB-Anwendungsprofil des Europeana Data Models zu überführen. Dabei spielt die Anwendung von Qualitätskriterien und Prinzipien des Linked Open Data eine wichtige Rolle. Der Vortrag gibt einen kurzen Überblick darüber, wie die DDB die Daten qualitativ verbessert und illustriert anhand von Beispielen im DDB-Portal sowohl Herausforderungen als auch bereits umgesetzte Anwendungen von vernetzen und angereicherten Daten. Davon ausgehend wird ein Ausblick in die Vorhaben der DDB im Bereich der Datenanreicherung und -bereinigung geboten.

Semantische Online-Suchsysteme
Prozesse, Methoden und Komponenten

Andrea Horch

Semantische Suchsysteme liefern i.d.R. weitaus präzisere Trefferlisten als herkömmliche Suchmaschinen. Diese Verbesserung der Suchergebnisse wird durch Funktionen wie z.B. die Verarbeitung von Synonymen oder das logische Schlussfolgern aus der Benutzeranfrage erzielt.

Zur Bewältigung der zusätzlichen Bearbeitungsschritte benötigen diese Suchmaschinen neue bzw. veränderte Prozessschritte sowie eine verbesserte Datenbasis. Diese Verarbeitungsschritte umfassen beispielsweise die Analyse der Sucheranfrage mithilfe semantischer Daten wie Linked Open Data (LOD) oder domänen-spezifischen Ontologien. Hierbei kann die Benutzeranfrage entweder durch weitere Suchterme aus der semantischen Datenbasis angereichert werden oder es werden logische Schlussfolgerungen anhand der Durchführung von Ableitungsregeln auf der Wissensbasis gezogen. Auch die Datenbasis kann mithilfe von semantischen Daten aufbereitet und mit weiteren Metadaten angereichert werden.

Für die Implementierung der semantischen Verarbeitungsprozesse existiert eine Vielzahl an Methoden und Werkzeugen, die auch bereits in bestehenden semantischen Suchmaschinen wie Swoogle eingesetzt werden. Beispiele für die Methoden und korrespondierende Werkzeuge sind Reasoner wie HermiT oder Pellet zum logischen Schlussfolgern auf OWL-Ontologien, OpenCalais für die automatische Erstellung von Metadaten für unstrukturierte Inhalte sowie Frameworks wie Apache Jena zur Entwicklung semantischer Anwendungen.

Diese Methoden und Werkzeugen können den einzelnen Komponenten semantischer Suchmaschinen zugeordnet werden. Eine Implementierung kann durch die Kombination bestehender Lösungen bzw. den Einsatz von Frameworks und verfügbarer Wissensbasen wesentlich schneller umgesetzt werden als durch eine Neuentwicklung der entsprechenden Komponenten. Die Studie »Semantische Suchsysteme für das Internet« stellt die Prozesse, Methoden und Komponenten semantischer Suchsysteme vor und zeigt auf welche bestehenden Softwarewerkzeuge für die Implementierung der einzelnen Komponenten genutzt werden können.

Hocheffiziente Tiefenerschließung multilingualer Textkollektionen mittels intelligentCAPTURE

Manfred Hauer

Mittels intelligentCAPTURE werden Texte, vorwiegend Inhaltsverzeichnisse von Monografien und Abstracts von Aufsätzen, aus bisher 55 Sprachen im Minutentakt ab Papier erfasst und automatisch mit Deskriptoren in deutscher und/oder anderen Zielsprachen ausgezeichnet. Bisher werden so jährlich rund 400.000 derartige Informationseinheiten produziert und über Bibliothekskataloge, http://www.dandelon.com und weitere Suchmaschinen öffentlich zugänglich gemacht.

Deutsches Textarchiv Qualitätssicherung (DTAQ)

Frank Wiegand

1 Projekt

Das DFG-geförderte Projekt Deutsches Textarchiv (DTA, 2007–2014) stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Um den historischen Sprachstand adäquat abzubilden, werden für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt. Das elektronische Volltextkorpus des DTA ist über das Internet frei zugänglich und dank seiner Aufbereitung durch (computer-)linguistische Methoden schreibweisentolerant durchsuchbar. Gesucht werden kann nach konkreten Wort- oder Grundformen (Lemmata), zudem kann die Suchanfrage beispielsweise durch Wortarten-, Kontext- oder Metadatenfilter verfeinert werden. Sämtliche Texte stehen zum Download im DTA-Basisformat (DTABf, ein TEI-XML-Dialekt) sowie in weiteren Formaten (u. a. für E-Books) unter einer Creative-Commons-Lizenz zur Nachnutzung bereit. Mit Stand März 2014 umfasst das DTA-Korpus 1.730 Werke (ca. 500.000 Seiten, ca. 115 Millionen fortlaufende Wortformen).

2 Problem

Dienstleister, die Transkriptionen im Double-Keying-Verfahren anbieten, versprechen meist eine Transkriptionsgenauigkeit von ca. 99,95 Prozent (durchschnittlich fünf Fehler in 10.000 Zeichen). Neben solchen Transkriptionsfehlern können in annotierten Texten weitere Fehlerarten auftreten, etwa in der Annotation (XML), in den Metadaten (jedes Werk wird mit ausführlichen Metadaten versehen), im Workflow (z. B. fehlerhafte Bilddigitalisate), Druckfehler und Fehler hinsichtlich der verschiedenen Darstellungsformen im DTA. In einem fortlaufenden Prozess zur Qualitätssicherung muss sichergestellt sein, dass diese Fehlerquellen zuverlässig erkannt, verwaltet und sukzessive behoben werden können.

3 Lösung

DTAQ (DTA-Qualitätssicherung) ist ein webbasiertes Tool, um in XML/TEI-annotierten Textdigitalisaten verschiedene Arten von Fehlern zu finden, zu kategorisieren und zu korrigieren. DTAQ steht DTA-internen und -externen Nutzern nach Anmeldung mit einem persönlichen Account zur Verfügung. Das Benutzerinterface von DTAQ ist anpassbar und bietet verschiedene Sichten auf die Bilddigitalisate, die Texttranskriptionen und weitere Präsentationsformen. Verschiedene Editiermöglichkeiten, die auf die jeweiligen Vorkenntnisse und Anforderungen der Nutzer abgestimmt sind, erleichtern das schnelle Arbeiten in den Texten. Jede einzelne digitalisierte Seite kann von den Benutzern als „korrekturgelesen“ markiert werden.

Aufbau von Terminologien aus Dokumentenkorpora - wie ein wenig Mathematik und Sprachverarbeitung Tage und Wochen manueller Arbeit vermeiden helfen

Stefan Geißler

Die semantische Erschließung von großen Textmengen für verbesserte Suche und Analyse setzt in hohem Maße auf strukturierte Wortschätze wie Thesauri. Die Verfügbarkeit von geeigneten Thesauri von ausreichender Abdeckung und Aktualität ist dabei jedoch nicht immer ohne weiteres gegeben und oftmals stehen große Anfangsaufwände an, um derartige Ressourcen zunächst zu erstellen, zu erweitern und zu pflegen.

Ein geeigneter Einsatz von linguistischen und quantitativen Verfahren kann helfen, diese Aufwände signifikant zu reduzieren und den Thesaurus-Erstellern Vorschläge und Hilfestellung liefern. Wir stellen eine Web-basierte Anwendung vor, die ausgehend von einer zunächst unstrukturierten Menge von domänenspezifischen Textdokumenten durch die Erzeugung von Kandidaten den Bearbeiter in die Lage versetzt, zügig mit dem Aufbau von entsprechenden strukturierten Wortschätzen fortzufahren.

Mehrwert schöpfen - Metasuche und Analyse von Fachinformationen

Manfred Jaksch

Die Ressource „Wissen“ wird als Entscheidungsgrundlage für unternehmerisches Handeln und für das Erkennen und die Verbesserung der eigenen Markt- und Wettbewerbsposition immer bedeutsamer. Um das global vorhandene Wissen für die eigenen Entwicklungen einzusetzen braucht es Werkzeuge, die effizient und möglichst redundanzfrei Informationen aufbereiten und verfügbar machen. Datenbanken sind solche Werkzeuge, werden aber oft gegenüber den Informationsangeboten aus dem Internet als überholt angesehen. Bei genauerer Betrachtung bieten sie jedoch deutliche Mehrwerte, die mit „Surfen“ im Internet und anschließender aufwendiger Bewertung nicht erzielt werden können. Der Vorteil der Datenbanken ergibt sich durch eine gute Vorarbeit: einen interdisziplinären und internationalen Korpus an Quellen anlegen, die Publikationen ordnen und strukturieren, lange, komplexe Texte komprimieren und sie suchoptimiert in Datenbanken ablegen. Solche gut strukturierten Daten lassen sich mit den heutigen leistungsfähigen Suchmaschinen und Analysetools schnell recherchieren und unter verschiedenen Gesichtspunkten auswerten. Zeitreihen zu einem bestimmten Thema aufzeigen, um technologische Entwicklungen zu verfolgen oder Potenziale neuer Technologien frühzeitig erkennen und zu nutzen, funktioniert sehr einfach. Aber auch zu konkreten technischen Fragen wie: „In welchen Fachgebieten werden kohlenstofffaserverstärkte Verbundwerkstoffe eingesetzt?“ oder „Welche Institutionen veröffentlichen zu einem Thema am häufigsten?“ können schnell Ergebnisse gezeigt werden. Um Datenbank-Mehrwerte zu schöpfen (erzeugen), benötigt man einerseits Schöpfwerkzeuge wie Software und Semantik und andererseits Schöpf-werkzeuge, um sie ergreifen zu können, vor allem aber Wissensaustausch.

Referenten, Moderatoren und Autoren

Sarah Baldin, MA (MA LIS)
FrauenMediaTurm - Archiv und Dokumentationszentrum
Am Bayenturm 2
50678 Köln
sarah.baldin@googlemail.com

Dr. Daniela Becks
CONTENiT AG
Harderhook 19
46395 Bocholt

Hermann Bense, Dipl. Inform.
bense.com Gesellschaft für digitales Publizieren mbH
Schwarze-Brüder-Straße 1
44137 Dortmund
hb@bense.com
www.ontology4.us

Ronald Billen, Diplom-Betriebswirt
Goethestraße 40
69214 Eppelheim
ronald.billen@gmx.de

Andreas Blumauer, MSc IT
Semantic Web Company GmbH
Mariahilfer Straße 70 / Neubaugasse 1
1070 Wien
Österreich
a.blumauer@semantic-web.at
www.semantic-web.at

Alexander Botte, Wiss.-Dok.
DIPF - Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt am Main
botte@dipf.de
www.dipf.de

Esther Chen, M.A., M.A. LIS
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
Celler Straße 3
38114 Braunschweig
chen@gei.de
www.gei.de

Jessica Drechsler, M.A.
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
Celler Straße 3
38114 Braunschweig
drechsler@gei.de
www.gei.de

Evelyn Dröge, M.A.
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft
Unter den Linden 6
10099 Berlin
evelyn.droege@ibi.hu-berlin.de
www.ibi.hu-berlin.de

Dr. Bernd Fachbach
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
bernd.fachbach@v2c2.at

Thorsten Förster, B.A.
Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information
Abteilung für Informationswissenschaft
Universitätsstraße 1
40225 Düsseldorf
thorsten.foerster@uni-duesseldorf.de

Stefan Geißler, M.A.
TEMIS Deutschland GmbH
Blumenstraße 15
69115 Heidelberg
stefan.geissler@temis.com
www.temis.com

Dipl.-Ing. Johannes Göllner, MSc
Leiter Wissensmanagement an der ZentDok/Landesverteidigungsakademie
Roßauer Lände 1
1090 Wien
Österreich
johannes.goellner@bmlv.gv.at
www.bmlv.gv.at

DI Heimo Gursch
Institut für Wissenstechnologien
TU Graz
Inffeldgasse 13/5
8010 Graz
Österreich
gursch@tugraz.at

Manfred Hauer, Dipl.-Inf.wiss., M.A.
AGI - Information Management Consultants
Mandelring 238 b
67433 Neustadt/Weinstraße
manfred.hauer@agi-imc.de
www.dandelon.com

Steffen Hennicke, M.A.
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und
Informationswissenschaft
Unter den Linden 6
10099 Berlin
steffen.hennicke@ibi.hu-berlin.de
www.ibi.hu-berlin.de

Andrea Horch, M.Sc.
Fraunhofer-Institut für Arbeitswirtschaft
und Organisation IAO
Competence Team Electronic Business
Services
Nobelstraße 12
70569 Stuttgart
andrea.horch@iao.fraunhofer.de
www.e-business.iao.fraunhofer.de

Julia Iwanowa, M.A.
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und
Informationswissenschaft
Unter den Linden 6
10099 Berlin
julia.iwanowa@ibi.hu-berlin.de
www.ibi.hu-berlin.de

Dipl.-Ing. Manfred Jaksch
WTI-Frankfurt eG
Ferdinand-Happ-Straße 32
60314 Frankfurt am Main
m.jaksch@wti-frankfurt.de
www.wti-frankfurt.de

Christian Kaiser
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
christian.kaiser@v2c2.at
www.vif.tugraz.de

Reinhard Karger, M.A.
Deutsches Forschungszentrum für
Künstliche Intelligenz
DFKI GmbH
Stuhlsatzenhausweg 3
Campus D3.2
66123 Saarbrücken
reinhard.karger@dfki.de
www.dfki.de

Dr. Andreas Kempf
GESIS − Leibniz-Institut für
Sozialwissenschaften
Unter Sachsenhausen 6-8
50667 Köln
andreas.kempf@gesis.org
www.gesis.org

Dr. Roman Kern
Know-Center
Inffeldgasse 13/6
8010 Graz,
Österreich
kern@know-center.at

Silvana Keutz, B.A.
Hochschule Darmstadt
Haardtring 100
64295 Darmstadt
silvana.keutz@stud.h-da.de
www.h-da.de

Prof. Dr. Günter Krampen
Leibniz-Zentrum für Psychologische
Information und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
krampen@zpid.de
www.zpid.de

Lennart Lamerz
Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information
Abteilung für Informationswissenschaft
Universitätsstraße 1
40225 Düsseldorf
lennart.lamerz@uni-duesseldorf.de
www.phil-fak.uni-duesseldorf.de/infowiss

Nikolas Leichner, Dipl.-Psych.
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
leichner@zpid.de
www.zpid.de

Agnes Mainka. M.A.
Heinrich-Heine-Universität Düsseldorf
Institut für Sprache und Information
Abteilung für Informationswissenschaft
Universitätsstraße 1
40225 Düsseldorf
agnes.mainka@uni-duesseldorf.de
www.phil-fak.uni-duesseldorf.de/infowiss

Ing. Mag. Klaus Mak
Oberst des höheren militärfachdienstlichen Dienstes
Leiter der Zentraldokumentation (ZentDok) an der Landesverteidigungsakademie
Roßauer Lände 1
1090 Wien
Österreich
klaus.mak@bmlv.gv.at

Prof. Dr. Thomas Mandl
Stiftung Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22
31141 Hildesheim
mandl@uni-hildesheim.de
www.uni-hildesheim.de

DI Nadja Marko
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
nadja.marko@v2c2.at
www.vif.tugraz.at

Dr. Anne-Kathrin Mayer
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
mayer@zpid.de
www.zpid.de

Dipl.-Ing. Christian Meurers
Referent Multimedia Dokumentation und Situation Awareness Center an der Zentraldokumentation/Landesverteidigungs akademie
Roßauer Lände 1
1090 Wien
Österreich
christian.meurers@bmlv.gv.at
http://www.bmlvs.gv.at/

Frederike Ohrem, M.A.
Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Informationstechnik für Führungssysteme (ITF)
Fraunhofer Straße 20
53343 Wachtberg
frederike.ohrem@fkie.fraunhofer.de
www.fkie.fraunhofer.de

Daniela Olenik, B.A.
Hochschule Darmstadt
Haardtring 100
64295 Darmstadt
daniela.olenik@stud.h-da.de
www.ha-da.de

Gabriele Patzer, Dipl. Dokumentarin
GBI- Genios Deutsche Wirtschaftsdatenbank GmbH
Freischützstraße 96
81927 München
gabriele.patzer@genios.de
www.genios.de, www.wiso-net.de, www.geniosschule.de

Johannes Peter, Dipl.-Psych.
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Universitätsring 15
54296 Trier
peter@zpid.de
www.zpid.de

Prof. Dr. Isabella Peters
ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften
Leibniz-Informationszentrum Wirtschaft
Düsternbrooker Weg 120
24105 Kiel
i.peters@zbw.eu
www.zbw.eu

Aida Raimbekova, B.A.
Hochschule Darmstadt
Haardtring 100
64295 Darmstadt
aida.raimbekova@stud.h-da.de
www.h-da.de

Prof. Dr. Marc Rittberger
Direktor
DIPF - Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt am Main
rittberger@dipf.de
www.dipf.de

Manfred Rosenberger
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
manfred.rosenberger@v2c2.at

apl. Prof. Dr. Ulrich Schade
Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Informationstechnik für Führungssysteme (ITF)
Fraunhofer Straße 20
53343 Wachtberg
ulrich.schade@fkie.fraunhofer.de
www.fkie.fraunhofer.de

Prof. Dr. Angela Schreyer
Fachhochschule Potsdam
Fachbereich Informationswissenschaften
Friedrich-Ebert-Straße 4
14467 Potsdam
schreyer@fh-potsdam.de
informationswissenschaften.fh-potsdam.de

Francesca Schulze, Dipl.-Dokumentarin
Germany Projektkoordination Deutsche Digitale Bibliothek
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main
f.schulze@dnb.de
www.ddb.de

Dr. Michael Schwantner
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
michael.schwantner@fiz-karlsruhe.de
www.fiz-karlsruhe.de

Lukas Sikorski, Dipl.-Inf.
Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Informationstechnik für Führungssysteme (ITF)
Fraunhofer Straße 20
53343 Wachtberg
lukas.sikorski@fkie.fraunhofer.de
www.fkie.fraunhofer.de

DI Selver Softic
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
selver.softic@v2c2.at
www.vif.tugraz.at/

Beate Renate Somorowsky, MA LIS, Dipl.-Bibl. (FH)
Adlerstraße
53881 Euskirchen
beatesomorowsky@t-online.de

Ute Sondergeld
DIPF - Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt am Main
sondergeld@dipf.de
www.dipf.de

Dr. phil. Christian Stein
Humboldt-Universität zu Berlin
Exzellenzcluster Bild Wissen Gestaltung
Unter den Linden 6
10099 Berlin
christian.stein@hu-berlin.de
www.interdisciplinary-laboratory.hu-berlin.de/

Dr. Alexander Stocker
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
alexander.stocker@v2c2.at
www.vif.tugraz.at/

Robert Strötgen, Dipl.-Inf.-Wiss, M.A.
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
Celler Straße 3
38114 Braunschweig
stroetgen@gei.de
www.gei.de

Julia Maria Struß
Stiftung Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22
31141 Hildesheim
julia.struss@uni-hildesheim.de
www.uni-hildesheim.de

Dr. Violeta Trkulja
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft
Unter den Linden 6
10099 Berlin
violeta.trkulja@ibi.hu-berlin.de
www.ibi.hu-berlin.de

Angela Vorndran, M.A., MLIS
DIPF - Deutsches Institut für Internationale Pädagogische Forschung
Schloßstraße 29
60486 Frankfurt am Main
vorndran@dipf.de
www.dipf.de

Dr. Luzian Weisel
Vizepräsident der Deutschen Gesellschaft für
Informationswissenschaft und Informationspraxis e.V.
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
luzian.weisel@fiz-karlsruhe.de

Frank Wiegand
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstraße 22/23
10117 Berlin
wiegand@bbaw.de
www.deutschestextarchiv.de
@textarchiv

Dr. Robert Woitsch
BOC GmbH
BOC Asset Management
Operngasse 20b
1040 Wien
Österreich
robert.woitsch@boc-eu.com

Prof. Dr. Christian Wolff
Lehrstuhl für Medieninformatik
Institut für Information und Medien, Sprache und Kultur (I:IMSK)
Universität Regensburg
93040 Regensburg
christian.wolff@ur.de
http://mi.ur.de

Prof. Dr. Christa Womser-Hacker
Stiftung Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22
31141 Hildesheim
womser@uni-hildesheim.de
www.uni-hildesheim.de

Mario Zechner
Know-Center
Inffeldgasse 13/6
8010 Graz,
Österreich
mario.zechner@know-center.at

DI Markus Zoier
Virtual Vehicle Research Center
Inffeldgasse 21a
8010 Graz
Österreich
markus.zoier@v2c2.at
www.vif.tugraz.at/