BASE – Eine Suchmaschine für OAI-Quellen und wissenschaftliche Webseiten

Dirk Pieper und Sebastian Wolf, Bielefeld

Dieser Aufsatz beschreibt die Entwicklung der Suchmaschine BASE (Bielefeld Academic Search Engine) seit 2005. In dieser Zeit wurde der Index um ein Vielfaches ausgebaut und auch die Nutzungszahlen stiegen deutlich. Der Schwerpunkt liegt auf der Indexierung von Dokumentenservern, die ihre Daten über das „Protocol for Metadata Harvesting" (OAI-PMH) bereitstellen. Im Gegensatz zu speziellen OAI-Suchmaschine wie OAIster verfügt BASE jedoch über weitergehende Suchmöglichkeiten und indexiert auch wissenschaftliche Webseiten mit Hilfe eines integrierten Web-Crawlers und andere Quellen, wie zum Beispiel den Bibliothekskatalog der Universitätsbibliothek Bielefeld. BASE kommt darüber hinaus als Suchsystem auch in anderen Bereichen, zum Beispiel im Bielefeld eScholarship Repository, als Suchmaschine der Universität Bielefeld und im EU-Projekt DRIVER via Schnittstellen zum BASE-Index zum Einsatz.

BASE - A search engine for OAI sources and scientific web pages

This article describes the development of BASE (Bielefeld Academic Search Engine) since 2005. During this time, the BASE index grew considerably and so did the usage frequency. The vast majority of the indexed resources originate from OAI conforming repositories. In contrast to special OAI search engines such as OAIster, BASE provides advanced search options. Scientific web sites and similar resources are indexed, e.g. the library catalogue of Bielefeld University Library. But the area of application is wider than that; it is also used for the Bielefeld eScholarship Repository, it serves as the university's search engine and is in operation via interfaces to the BASE-index in the DRIVER project of the EU.

1 Einleitung

Als das erste Themenheft „Suchmaschinen" dieser Zeitschrift Anfang 2005 erschien, war BASE[1] ein gutes halbes Jahr unter Produktionsbedingungen im Netz (Summann/Wolf 2005). Das zweite Themenheft bietet nun einen willkommenen Anlass, eine Zwischenbilanz zu ziehen und es zeigt auch, dass Suchmaschinen als Plattform bibliothekarischer Informationsangebote ein zunehmend stärkeres Gewicht bekommen.

Kommerziell betriebene Suchmaschinen wie zum Beispiel Google, Google Scholar, Scirus oder auch Windows Live Search Academic werden von Studierenden gegenüber Bibliotheksangeboten, wie zum Beispiel Bibliothekskatalogen (Lewandowski 2006) und Fachdatenbanken (Urquhart et al. 2005) bevorzugt genutzt. Produkte wie Google Book Search oder Amazon Search Inside führen zu höheren Ansprüchen an Bibliothekskataloge. Die ersten Bibliotheken und Bibliotheksverbünde haben darauf reagiert: Bibliographische Angaben werden mit Daten aus Inhaltsverzeichnissen angereichert (Großgarten 2005), bei der Katalogaufbereitung wird zunehmend Suchmaschinentechnologie eingesetzt. Sichtbar wird dies zum Beispiel an (wie BASE) ebenfalls auf FAST[2] basierenden Produkten wie dem Dreiländerkatalog des HBZ[3], der Entwicklung eines neuen Webkatalogs für Lokalsysteme durch OCLC/PICA[4] oder in ersten Ansätzen zur Integration von Suchmaschinentechnologie in Vascoda[5].

Darüber hinaus gibt es inzwischen eine Reihe von Projekten im Bibliotheksumfeld, in denen die Open-Source-Software Lucene zum Einsatz kommt[6].Ziel beim Aufbau der Bielefeld Academic Search Engine (BASE) ist die Realisierung eines suchmaschinenbasierten Informationsangebots, das Zugriff auf verschiedensten über das Internet zugänglichen wissenschaftlichen Content ermöglicht. Die Vorteile beim Einsatz von Suchmaschinentechnologie – einfache, „google-ähnliche" Nutzung, hohe Performanz, Volltextsuche und hohe Relevanz der Suchergebnisse – sollen verbunden werden mit den aus der Datenbankwelt bekannten Vorteilen der Berücksichtigung bibliographischer Suchaspekte sowie hoher Datenqualität. Damit wird der an der Universitätsbibliothek Bielefeld seit den neunziger Jahren verfolgte Ansatz der Integration von externem wissenschaftlichen Content und lokaler Datenbankproduktion weiterentwickelt.

Der Einsatz von Suchmaschinentechnologie ist dabei allerdings kein Selbstzweck. Sie soll in erster Linie dem Nutzer helfen, in den immer größer werdenden Datenmengen das zu finden, was er sucht, zum Beispiel durch den Einsatz von linguistischen Methoden (Finden von ähnlichen Wortformen, multilinguale Suche, Ontologien, usw.), durch verschiedene Verfahren der Relevanzbewertung oder durch die Möglichkeit, nach einer Suche große Treffermengen einzuschränken („Search Refinement"). BASE beinhaltet im Vergleich zu 2004 inzwischen eine Vielzahl unterschiedlicher wissenschaftlich relevanter Datenquellen. Es ist jedoch kein Zufall, dass sich im Projektverlauf der Schwerpunkt auf die Indexierung von Repository-Servern, die ihre Daten nach dem Standard des „Protocol for Metadata Harvesting" (OAI-PMH) bereitstellen, herausgebildet hat. Zum einen besitzen diese Daten – trotz aller Probleme, die beim Sammeln und Aggregieren von OAI-Daten in der Praxis auftreten können (Pieper/Summann 2006; Summann/Wolf 2006) – gegenüber unstrukturiertem Webcontent eine vergleichsweise hohe Qualität, die es ermöglicht, Suchen nach verschiedenen Aspekten wie Autor, Titel, Schlagwort usw. anzubieten, zum anderen bieten sie im Vergleich zu reinen Nachweisdaten häufig einen Link zum Volltext. Ein weiterer Grund dieser Schwerpunktbildung liegt in der notwendigen Positionierung auf dem Suchma-

1 http://base.ub.uni-bielefeld.de [7.2.2007]
2 Fast Search & Transfer: www.fastsearch.com [7.2.2007]
3 http://suchen.hbz-nrw.de/dreilaender [7.2.2007]
4 www.oclcpica.org/dasat/images/4/100764-fast.pdf [7.2.2007]
5 http://lists.vascoda.de/presse.pl?id=5 [7.2.2007]
6 Zum Beispiel die Entwicklung eines neuen Online-Katalogs an der UB Heidelberg, siehe www.ub.uni-bayreuth.de/SISIS/SSV/SAT/sat_20061024_files/Lucene.pdf [7.2.2007]

schinenmarkt, wobei es auf der Hand liegt, dass schon aufgrund des ungleich geringeren möglichen Einsatzes ökonomischer und technischer Ressourcen bibliothekarische Suchmaschinenangebote mit den oben genannten Produkten nur schwer konkurrieren können. BASE versucht daher mit einem inhaltlichen Profil und mit über den Standard hinausgehenden Suchfunktionen eine Nische zu besetzen. Dazu gehören im Vergleich zu Google unter anderem folgende Merkmale:

- intellektuelle Auswahl wissenschaftlich relevanter (OAI-)Quellen
- hohe Datenqualität, Transparenz der indexierten Quellen über ein Quellenverzeichnis
- Volltextindexierung und Verbindung mit den zugehörigen Metadaten
- differenzierte Anzeige von bibliographischen Daten
- mehr Suchoptionen, Suchhistorie, Sortierung

Zum Konzept des Einsatzes von Suchmaschinentechnologie an der Universitätsbibliothek Bielefeld gehört jedoch nicht nur die Positionierung von BASE als suchmaschinenbasierter OAI-Service-Provider. Die FAST-Software ermöglicht es auch, mit Hilfe eines integrierten Web-Crawlers und mit Datenbankkonnektoren Dokumente zu indexieren und den Index flexibel in unterschiedlichen Kontexten zu nutzen. Die Nutzung der Daten eines (externen) Suchmaschinenindex in unterschiedlichen Portalen ist in der kommerziellen Suchmaschinenwelt schon seit längerem Praxis (Lewandowski 2005, S. 21 ff.) und sollte auch im bibliothekarischen Umfeld realisiert werden können, was nicht zuletzt auch die weiter unten kurz beschriebene Teilnahme der Universitätsbibliothek Bielefeld am EU-Projekt DRIVER[7] zeigt.

2 Entwicklung der Quellen- und Nutzerzahlen

BASE startete im Juni 2004 mit einer Einzelplatz-Installation unter FAST Data Search 3.0. Nachdem in der Zwischenzeit die FAST-Software auf die Version 4.1 aktualisiert wurde, läuft BASE seit Februar 2006 auf einer Server-Farm und weist damit – im Kleinen – die von großen Suchmaschinen bekannte skalierbare Systemarchitektur auf.

Zum Zeitpunkt der Abfassung des Artikels für das erste Themenheft enthielt BASE ca. 800.000 indexierte Dokumente aus 35 Quellen. Mittlerweile (Stand: Dezember 2006) umfasst der Index 4,3 Millionen Dokumente aus 307 Quellen. Die Quellen werden laufend aktualisiert und die Zahl indexierter Dokumente wächst täglich im Durchschnitt um 5.000 bis 10.000.

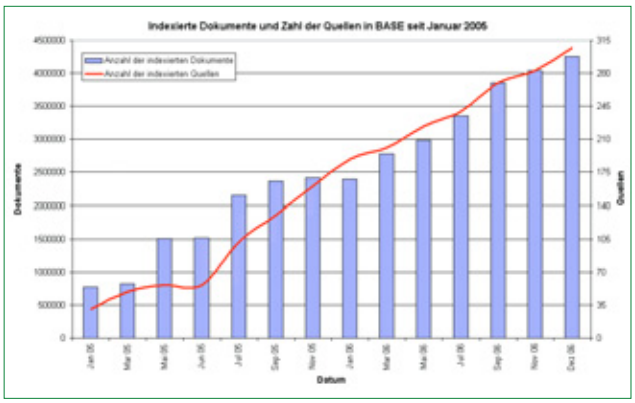

Abbildung 1: Entwicklung der Zahl indexierter Dokumente und Quellen in BASE seit Januar 2005

Im Gegensatz zu wissenschaftlichen Suchmaschinen wie Google Scholar wird bei BASE der Umfang des Index und der Abdeckungsgrad der Indexierung durch eine Liste der indexierten Quellen[8] deutlich, die inzwischen täglich automatisch aktualisiert wird und nach verschiedenen Kriterien sortierbar ist.

Auch die Nutzungszahlen von BASE steigen stetig. Derzeit werden über die BASE-Suchmaske im Durchschnitt 3.000 bis 4.000 Suchanfragen täglich gestellt. Die Zahl aller Suchabfragen liegt noch deutlich höher, da BASE inzwischen auch von externen Quellen abgefragt wird, so zum Beispiel von der Metasuchmaschine Definero[9] als auch über bereitgestellte und sich zum Teil noch in der Entwicklung befindlichen HTTP- und SOAP-Schnittstellen.

3 Neue Suchmöglichkeiten und Integration von Google Scholar

Die Suchoberfläche von BASE folgt konsequent dem erfolgreichen Google-Konzept und ist so einfach wie möglich gehalten, ohne dabei auf ein ansprechendes Design und weitere Funktionalitäten zu verzichten. So ist es möglich, einen Suchbegriff um beliebig viele Zeichen zu erweitern (Rechtstrunkierung). Durch die automatische Lemmatisierung von Suchbegriffen werden bei einer Suchanfrage standardmäßig ähnliche Wortformen (Plural, Genitiv) mit abgesucht. Diese Funktion kann der Nutzer über eine Checkbox („zusätzliche Wortformen finden“) auch abwählen. Weiterhin ist es möglich, gezielt in Metadatenfeldern (Titel, Autor, Schlagwort, usw.) zu suchen.

Die Anzeige der Treffer orientiert sich an der aus Internetsuchmaschinen gewohnten Darstellung und ist neben der von FAST bereitgestellten Möglichkeiten der Suchverfeinerung („Search Refinement“) um die Funktionalitäten Sortierung und Suchhistorie erweitert worden. Bei jedem Treffer ist unterhalb des Teasers (einem automatisch generierten Textauszug) bzw. der Anzeige der Metadaten (sofern vorhanden) der Link „Diesen Titel in Google Scholar suchen“ zu sehen.

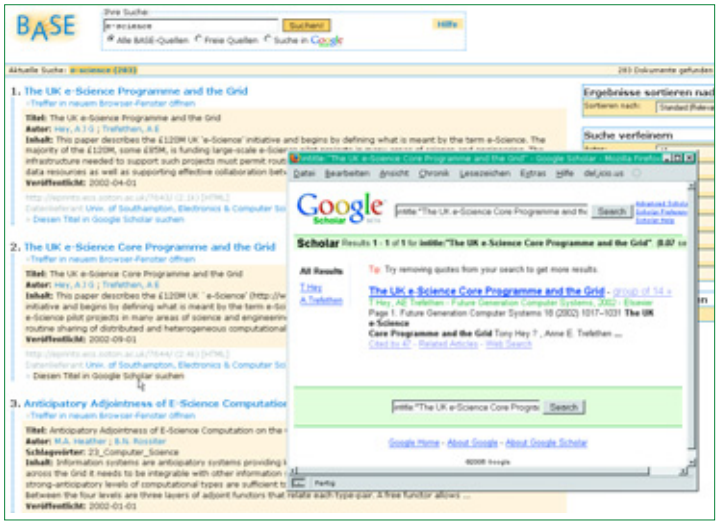

Abbildung 2: Die Treffer aus Google Scholar erscheinen in einem neuen Fenster. Der Nutzer hat nun unter anderem die Möglichkeit, sich die zitierenden Artikel („Cited by“) anzusehen

Wird der oben erwähnte Link ausgewählt, wird eine Suche nach dem Titel des Treffers in Google Scholar gestartet. Sinn dieser Funktion ist es, den Nutzern die zitierenden Artikel („Cited by“), die ebenfalls von Google Scholar indexiert wurden, anzuzeigen, was insbesondere für die wissenschaftlichen Nutzer von BASE von Interesse ist. Zudem werden verschiedene Fassungen des Artikels (der zum Beispiel als Preprint, Postprint und als Konferenzbericht vorliegen kann) und/oder die verschiedenen Server, auf denen ein Artikel abgerufen werden

7 Digital Repository Infrastructure Vision for European Research, www.driver-repository.eu [7.2.2007]
8 http://base.ub.uni-bielefeld.de/about_sources.html [7.2.2007]
9 www.definero.de [7.2.2007]

kann, zu einer Gruppe zusammengefasst, die über den Link „Group of X“ angezeigt werden.
Die Einbindung dieser Funktion konnte dank E-Mail-Kommunikation mit Anurag Acharya, dem Chefentwickler von Google Scholar, aktiviert werden. Für das Jahr 2007 wurde von ihm die Möglichkeit in Aussicht gestellt, trefferbezogene Links direkt auf die zitierenden Artikel („Cited by“) setzen zu können.

Neben der trefferbezogenen Integration kann eine Suche in Google Scholar auch direkt von der Suchmaske in der Trefferanzeige gestartet werden („Suche in Google Scholar“), um weitere nicht durch BASE indexierte wissenschaftliche Quellen finden zu können und die beschriebenen Google-Scholar-Funktionen, die von BASE selbst derzeit nicht bereit gestellt werden können, als Add-On zu nutzen. Google Scholar wurde deshalb gewählt, da es sich um das populärste Angebot im wissenschaftlichen Bereich handelt und – beispielsweise im Gegensatz zur allgemeinen wissenschaftlichen Suchmaschine Scirus – eine hohe inhaltlich Übereinstimmung bei den indexierten Quellen besteht. Zudem erscheinen bei Google Scholar – sofern die Bibliothek einen Linkresolver einsetzt und dieser entsprechend konfiguriert ist – trefferbezogene Links auf die eigenen Bibliotheksbestände („Library links“). Wissenschaftlern und Studierenden, die Google Scholar als primäre Quelle der Informationsbeschaffung nutzen, können auch, ohne dass sie im Bibliothekskatalog recherchieren, auf die eigenen Bibliotheksbestände gelenkt werden. Suchmaschinen sollten daher nicht nur als Konkurrenz zu den eigenen Angeboten betrachtet werden, sondern vielmehr als Möglichkeit, die eigenen Bestände für ein breiteres Publikum sichtbar zu machen.
Als letzter Punkt soll hier noch erwähnt werden, dass auch die barrierefreie Gestaltung der BASE-Webseiten ausgebaut wurde. Mittlerweile erfüllen die Informationsseiten von BASE die Kriterien der Priorität 2, Suchmaske und Trefferlisten erfüllen die Kriterien der Priorität 1 der entsprechenden Richtlinien, die im Bundesgesetzblatt festgehalten sind (Bundesministerium der Justiz 2002).

1/2

4 BASE als Suchsystem für verschiedene Anwendungen

Der konsequente Einsatz von Cascading Style Sheets (CSS) zur Trennung von Inhalt und Layout erwies sich als großer Vorteil, nicht nur bezüglich der Integration neuer Funktionen, sondern auch bei der Entwicklung eigener Sichten, wie zum Beispiel für das Bielefeld eScholarship Repository, in dem BASE als Suchsystem zum Einsatz kommt. Dabei basieren Dokumentenablage, OAI-Schnittstelle und Browsing-Funktionen auf OPUS, die Indexierung der Daten erfolgt jedoch mit FAST und die Suchoberfläche ist ein eigenständiger View der BASE-Oberfläche. Das Bielefeld eScholarship Repository stellt die Forschungsergebnisse der Universität Bielefeld der wissenschaftlichen Community weltweit zur Verfügung. Von der Suchmaske aus ist es möglich, unter anderem nur im Repository oder auch in allen BASE-Quellen zu suchen. Das Suchergebnis wurde an das Design des Bielefeld eScholarship Repositories angepasst. Das Design der verschiedenen Oberflächen wird dabei ausschließlich durch den Einsatz unterschiedlicher CSS-Dateien gesteuert.

Seit 2005 werden mit FAST die Webseiten der Universität Bielefeld und der Universitätsbibliothek indexiert, wobei dieser Teil nicht in BASE einfließt. Hierbei wurden die Stärken des Systems bei der Indexierung von Webquellen deutlich. Durch den Einsatz eines Tools zur Relevanzbewertung, die vom FAST-System bereitgestellt wird, konnte eine wesentliche Verbesserung der Relevanz bei den gefundenen Treffern erreicht werden.

10 http://europa.eu/eurovoc [7.2.2007]

5 EU-Projekt DRIVER

Das DRIVER-Projekt verfolgt das Ziel, in einer ersten Phase eine Testumgebung zur Vernetzung von mehr als 50 in Europa verteilten wissenschaftlichen Dokumentenserver (sog. Repositorien) zu entwickeln. Das Projekt, an dem zehn Partner aus acht Ländern beteiligt sind, wird durch die Abteilung „Forschungsinfrastruktur" der Europäischen Kommission gefördert (Lossau 2006). Die Universitätsbibliothek Bielefeld ist dabei verantwortlich für die Aggregation und Speicherung von OAI-Metadaten, die Indexierung mit FAST und die Bereitstellung einer HTTP- und SOAP-Schnittstelle. Über diese Schnittstellen wird es für Betreiber des zukünftigen DRIVER-Portals möglich sein, sich flexibel aus dem BASE-Index Quellen für das eigene Informationsangebot zusammenzustellen und in das eigene Portal zu integrieren. DRIVER wird entsprechend der Ausrichtung dann den Teil von BASE nutzen, der die Daten der europäischen Repository-Server repräsentiert.

6 Fazit, Ausblick und Perspektiven

Sowohl die steigenden Nutzerzahlen, die zunehmenden Anfragen von Repository-Betreibern zwecks Aufnahme in den BASE-Index als auch die Beteiligung am DRIVER-Projekt lassen das Fazit nach rund drei Jahren BASE positiv ausfallen. Mehr Aufwand als gedacht musste in den Bereich Harvesting und Normalisierung von OAI-Metadaten investiert werden. Beim kontinuierlichen Ausbau des Index durch Aufnahme weiterer OAI-Quellen steht insbesondere auch die weitere Verbesserung und Automatisierung des OAI-Datenflusses vom Harvesting bis hin zur Indexierung mit FAST auf der Agenda.

In der Entwicklung befinden sich derzeit ein Searchplugin, über das man BASE direkt über die Suchmaschinen-Toolbar im Browser durchsuchen kann, und das Browsing über die indexierten Quellen mithilfe der Dewey Decimal Classification (DDC). Alle Dokumente, die mit einer DDC-Nummer versehen sind, können über dieses Browsing ermittelt werden. Derzeit sind dies allerdings erst ca. 25.000 Dokumente und damit nur ein sehr kleiner Teil aller Dokumente im BASE-Index. Allerdings betrifft dies alle von BASE indexierten OPUS-Systeme, sowie andere von DINI zertifizierte Repository-Server, was die Sinnhaftigkeit von Standardisierung über Software oder Zertifizierung verdeutlicht.

2007 wird die lokale Datenbankproduktion (Bibliothekskatalog, Aufsatzdatenbank JADE und weitere) mit FAST eine stärkere Rolle spielen, wobei sich interessante Perspektiven durch die Entwicklung des ebenfalls auf FAST basierenden Webkatalogs durch OCLC/PICA hinsichtlich einer möglichen Integration in die BASE-Umgebung ergeben. BASE wird dann für externe Nutzer hauptsächlich die Funktion eines OAI-Service-Providers wahrnehmen, während für Nutzer auf dem Campus Bielefeld zusätzlich die Integration lokaler Quellen angeboten werden kann. Ein weiterer Aspekt, der aufgrund des hohen Aufwands im OAI-Bereich zurückstehen musste, ist die Auseinandersetzung mit den linguistischen Features der FAST-Software. In einem ersten Schritt werden Teile des eurovoc-Thesaurus[10] in die FAST-Indexierungs-Stufen eingebunden, um in Zukunft eine multilinguale Suche in BASE zu ermöglichen.

Literatur

Bundesministerium der Justiz (Hrsg.) (2002): Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie Informationstechnik-Verordnung – BITV) : vom 17. Juli 2002. In: Bundesgesetzblatt Teil I 2002 (49), 2654-2660. URL: www.bgblportal.de/BGBL/bgbl1f/bgbl102s2654.pdf [7.2.2007].

Großgarten, A. (2005): Das 180T-Projekt in Köln oder wie verarbeite ich 180.000 Bücher in vier Monaten. In: Information, Wissenschaft und Praxis 56 (8), 454-456.

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

Lewandowski, D. (2006): Suchmaschinen als Konkurrenten der Bibliothekskataloge: Wie Bibliotheken ihre Angebote durch Suchmaschinentechnologie attraktiver und durch Öffnung für die allgemeinen Suchmaschinen populärer machen können. In: Zeitschrift für Bibliothekswesen und Bibliographie 53 (2), 71-78.

Lossau, N. (2006): DRIVER: Networking European Scientific Repositories. In: BI.research – Forschungsmagazin der Universität Bielefeld 2006 (29), 61-65. URL: www.uni-bielefeld.de/Universitaet/Einrichtungen/Pressestelle/dokumente/BI_research/29_2006/Forschungsmagazin_1_06_Driver_61_65.pdf [7.2.2007].

Pieper, D.; *Summann*, F. (2006): Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. In: Library Hi Tech 24 (4), 614-619.

Summann, F.; *Wolf*, S. (2005): BASE – Suchmaschinentechnologie für digitale Bibliotheken. In: Information, Wissenschaft und Praxis 56 (1), 51-57.

Summann, F.; *Wolf* S. (2006): Suchmaschinentechnologie und wissenschaftliche Suchumgebung. In: VÖB Online-Mitteilungen 2006 (86), 3-8. URL: www.univie.ac.at/voeb/php/downloads/om86.pdf [7.2.2007].

Urquhart, C.; *Thomas*, R.; *Spink*, S, *Fenton*, R.; *Yeoman*, A.; *Lonsdale*, R.; *Armstrong*, C.; *Banwell*, L; *Ray*, K.; *Coulson*, G.; *Rowley*, J. (2005): Student use of electronic information services in further education. In: International Journal of Information Management 25 (4), 347-362.

Suchmaschine, Maschinelles Indexierungsverfahren, Quelle, Bibliothek, UB Bielefeld, BASE, FAST

Die Autoren

Dirk Pieper

wurde am 30.03.1967 in Heide geboren und studierte Politikwissenschaft und Volkswirtschaftslehre an den Universitäten Bonn und Hamburg. Danach absolvierte er die Ausbildung zum Höheren Dienst an wissenschaftlichen Bibliotheken an der FH Köln und ist seit 2004 Erwerbungsdezernent und Koordinator im Bereich Digitale Bibliothek an der UB Bielefeld.

Telefon: (05 21) 106 40 10
dirk.pieper@uni-bielefeld.de

Sebastian Wolf

wurde am 26.4.1975 in Göttingen geboren und studierte Wissenschaftliches Bibliothekswesen an der FH Hannover. Seit 1999 ist er in der Universitätsbibliothek Bielefeld tätig, unter anderem im Bereich Pflege und Weiterentwicklung des Web-Angebots. Er betreut den Suchdienste-Kompass (www.ub.uni-bielefeld.de/biblio/search), ein Tutorial zur Nutzung und Funktionsweise von Suchmaschinen. Im BASE-Projekt arbeitet er seit 2002 im Bereich Webdesign, Usability und Funktionalitätstests.

Telefon: (05 21) 106 40 44
sebastian.wolf@uni-bielefeld.de

Universitätsbibliothek Bielefeld
Universitätsstraße 25
33615 Bielefeld