Konzeption und Betrieb eines Kompetenz- und Dienstleistungsnetzes für die Informatik

Lutz Horn[7], Michael Ley[6], Peter Luksch[7], Jörg Maas[1], Ernst W. Mayr[5], Andreas Oberweis[3], Paul Ortyl[4], Stefan Pfingstl[5], Enzo Rossi[7], Felix Rüssel[7], Ute Rusnak[7], Daniel Sommer[2], Wolffried Stucky[2], Roland Vollmar[4], Marco von Mevius[3]

[1] Gesellschaft für Informatik e.V., D-53175 Bonn
[2] Institut AIFB, Universität Karlsruhe (TH), D-76128 Karlsruhe
[3] LS für Entwicklung betrieblicher Informationssysteme, Universität Frankfurt, D-60054 Frankfurt am Main
[4] LS Informatik für Ingenieure und Naturwissenschaftler, Universität Karlsruhe (TH), D-76128 Karlsruhe
[5] LS für Effiziente Algorithmen, TU München, D-85748 Garchingen
[6] Gruppe Datenbanken- und Informationssysteme, Universität Trier, D-54286 Trier
[7] FIZ Karlsruhe, D-76344 Eggenstein-Leopoldshafen

Abstract: Ziel des Projekts Fachinformationssystem Informatik (FIS-I) ist die Konzeption und der Betrieb eines Kompetenz- und Dienstleistungsnetzes für die Informatik. Das Portal soll den Zugriff auf weltweit publiziertes Informatikwissen zentralisieren, Publikationen strukturiert und standardisiert mit Metadaten erfassen und langfristig die Verfügbarkeit der archivierten Informationen absichern. An dem Projekt arbeiten neben der Gesellschaft für Informatik (GI) und dem FIZ Karlsruhe, Wissenschaftler der Universitäten Frankfurt, Trier und Karlsruhe sowie der Technischen Universität München mit.

1. Einleitung

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt Fachinformationssystem Informatik (FIS-I) hat die Aufgabe, Wissenschaftlern, Studierenden, industriellen Nutzern, aber auch interessierten Laien alle für sie relevanten Informationen zur Informatik auf einer Plattform (IO-Port) verfügbar zu machen. Das IO-Port soll den Zugriff auf weltweit publiziertes Informatikwissen zentralisieren, Publikationen strukturiert und standardisiert mit Metadaten erfassen und langfristig die Verfügbarkeit der archivierten Informationen absichern. Ein besonderes Merkmal des Projekts ist, dass die verschiedenen Interessenlagen der an dem Publikationsprozess beteiligten Personengruppen und Institutionen (Autoren, Bibliotheken, Fachinformationszentren und Nutzer) bei der Gestaltung des neuen Informatik-Portals IO-Port berücksichtigt werden sollen. An dem Projekt arbeiten neben der Gesellschaft für Informatik (GI) und dem FIZ Karlsruhe, Wissenschaftler der Universitäten Frankfurt, Trier und Karlsruhe sowie der Technischen Universität München mit. Zusätzlich besteht

eine enge Kooperation zu dem ebenfalls vom BMBF geförderten Forschungsprojekt "Semantische Methoden und Tools für Informationsportale" (SemIPort) (http://km.aifb.uni-karlsruhe.de/semiport).

Die vorliegende Arbeit ist folgendermaßen strukturiert. Zunächst wird eine kurze Einführung in die identifizierten Anforderungen an das IO-Port gegeben. Im nächsten Abschnitt wird die Architektur des erstellten Prototyps und des zu implementierenden Systems dargestellt. Darauf aufbauend erfolgt die Beschreibung des integrierten Nachweissystems. Ein besonderer Fokus liegt hier auf den Aspekten der Datenhaltung und des Datenaustauschs. Zusätzlich werden die Datenbanken der einzelnen Projektpartner präsentiert. Anschließend werden die Grundlagen zur Konzeption eines Geschäftsmodells skizziert. Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick.

2. Anforderungen an das IO-Port

Anforderungen an ein Portal für die Informatik ergeben sich aus verschiedenen Perspektiven. Die typische Nutzergruppe ist die der klassischen Nutzer, die über das System Literatur suchen und lesen. Weitere Anforderungen ergeben sich aus der Sicht der Informationsanbieter, die Informationen in das System einpflegen, und aus der Sicht der Betreiber des Systems, die die technische Verfügbarkeit (Betrieb) sicherstellen. Im Folgenden werden die Basisanforderungen an das IO-Port unter Berücksichtigung dieser verschiedenen Sichten beschrieben.

2.1 Anforderungen aus der Sicht der Nutzer

Die klassischen Nutzer verwenden das System zur Literaturrecherche oder beispielsweise zur Suche im Veranstaltungskalender. Wissenschaftler benötigen etwa Informationen zu aktuellen Trends, neuen oder laufenden Forschungsprojekten oder suchen aktuelle Veröffentlichungen innerhalb ihres Forschungsbereichs. Sie suchen dabei entweder gezielt nach einer bestimmten Veröffentlichung oder aber ganz allgemein nach Literatur zu einem gewissen Thema. Ein benutzerfreundliches Portal muss also die Bereitstellung von bibliografischen Informationen ermöglichen, die Bezugsquellen für die gewünschten Volltexte liefern und zusätzlich den Zugriff auf Volltexte unterstützen. Dabei sollte die Anfragesprache intuitiv verständlich sein (auch für Laien). Andererseits muss die Anfragesprache mächtig genug sein, um bei Bedarf auch komplexe Anfragen formulieren zu können. Für Anwendungssituationen, in denen Nutzer des Systems nicht gezielt nach Veröffentlichungen oder anderen Informationsquellen suchen, sind Navigationsmöglichkeiten ("Browsing") zum Auffinden der Informationen wünschenswert. Durch eine solche Funktionalität kann sich der Anwender den Dokumentenbestand – nach Teilbereichen der Informatik geordnet – auflisten lassen. Neben diesen Basisfunktionalitäten sollte das IO-Port weitere Anforderungen abdecken, z.B.:

- Anpassung der Benutzerschnittstelle an bestimmte Benutzergruppen. Die Anwender des Portals haben verschiedene Rollen und benötigen daher rollenbasierte Sichten.

- Zusammenführen interner und externer Informationsquellen über eine Hyperlinkoberfläche.

- Push-Mechanismen zur automatisierten Bereitstellung von aktuellen Informationen für bestimmte Personen und Personengruppen.

- Abwicklung finanzieller Transaktionen durch das Angebot elektronischer Zahlungsverfahren (z.B. Kreditkarten) oder durch Abbuchung. Aus der Sicht der Nutzer sollte die Anonymisierung solcher Transaktionen ermöglicht werden.

Zielgruppe des Systems sind neben Einzelnutzern insbesondere auch institutionelle Nutzer, wie z. B. Fachbereiche einer Universität, Schulen oder Entwicklungs- und Forschungsabteilungen von Unternehmen. Diese Institutionen müssen Lizenzen für relevante Teile des Systems erwerben können.

2.2 Anforderungen aus der Sicht der Informationsanbieter und Systembetreiber

Aus der Sicht der Informationsanbieter ist es von besonderer Bedeutung, dass der Aufwand beim Verwalten des Dokumentenbestandes möglichst gering ist. Da in das IO-Port Quellen aus verschiedenen Standorten einpflegt werden, sind effiziente und zuverlässige Einfüge-, Lösch- und Änderungsfunktionalitäten eine unmittelbare Voraussetzung für die Aktualität und Qualität und damit den nachhaltigen Erfolg des Portals. Ein zusätzlicher kritischer Aspekt für die Informationsanbieter ist der Schutz des Systems und seiner Inhalte (Datensicherheit), wozu insbesondere der Schutz vor illegalem Zugriff durch sichere Authentifizierungsverfahren für unterschiedliche Nutzergruppen zählt. Die langfristige Archivierung der Informationen (Datensicherung) sollte ebenso gewährleistet sein.

Aus der Sicht der Betreiber des Portals ist ein stabiles System mit geringem Wartungsaufwand für eine zeit- und kostensparende Administration der kritische Erfolgsfaktor. Von zentraler Bedeutung ist weiter, dass das System im laufenden Betrieb gewartet werden kann. Die Installation von möglichen nutzerspezifischen Funktionalitäten sollte möglichst einfach sein, und die Basisanwendungen des Systems sollten – soweit möglich – auf Standardkomponenten beruhen.

3. Architektur

Zahlreiche Beiträge aus der Fachliteratur behandeln das Thema Software-Architektur mehr oder weniger umfassend, dennoch existiert derzeit keine allgemein akzeptierte,

standardisierte Definition des Begriffs selbst (http://www.sei.cmu.edu/architecture/definitions.html). Eine oft zitierte Definition ist folgende: Die Architektur eines Softwaresystems stellt die Strukturen des Systems dar, bestehend aus Software-Komponenten, deren extern sichtbare Eigenschaften (u. a. Performance, Verfügbarkeit, Funktionalität, Fehlerbehandlung) und deren Beziehungen untereinander [BCK98]. Software-Architektur ist damit gleichzeitig Bauplan und Ablaufplan für Software. Sie soll ein stabiles Grundgerüst bereitstellen, um Entwicklung, Betrieb und Wartung des gesamten Software-Systems sicherzustellen. Dabei muss sie auch flexibel und erweiterbar sein, um die möglichst einfache Umsetzung neuer und geänderter Anforderungen zuzulassen (nach [St02]). Software-Architektur bildet somit den schwierigen Übergang von der Analysephase zur konkreten technischen Realisierung.

Die Projektteilnehmer haben einen Anforderungskatalog erstellt, der aufgrund der Vielzahl der einzelnen Komponenten und deren Beziehungen untereinander hohe Anforderungen an die Software-Architektur und die Software-Entwicklung stellt. Die Praxis hat bestätigt, dass sich Spezifikationen, Randbedingungen und Einflussfaktoren gerade zu Beginn eines Projekts häufig ändern, wodurch die Vorgehensweise "in einem Schritt von der Spezifikation zum Produkt" nicht möglich ist. Die Entwicklung von Software-Systemen wird deshalb auch mit der Verfolgung von beweglichen Zielen *(moving targets)* [St02] verglichen, bei der sich Zwischenlösungen und das bewegliche Ziel im Projektverlauf durch Iterationen immer weiter nähern. Für die Realisierung des IO-Ports wurde der Ansatz gewählt, grundlegende Komponenten in zunächst einfachen Strukturen zu realisieren und in einem fortlaufenden iterativen Entwicklungsprozess Zwischenlösungen in Form von entscheidungsreifen Zwischenprodukten zu erstellen. Die einzelnen Zwischenprodukte haben das Ziel, die Machbarkeit zu belegen und eine Rückkopplung von Testnutzern für die weitere Entwicklung zu berücksichtigen.

Ein Prototyp soll im Rahmen der GI-Jahrestagung INFORMATIK 2003 vorgestellt werden ("Prototyp2003"). Die Architektur des Prototyp2003 sieht eine Zweiteilung des Gesamtsystems in Datenmanagementsystem und Portalsystem vor. Im Datenmanagementsystem werden die Daten gesammelt, aufbereitet und archiviert, im Portalsystem werden die Daten den Nutzern mit verschiedenen Abfragemechanismen angeboten.

Das *Portalsystem* soll dem Nutzer die Möglichkeit bieten, über eine Weboberfläche auf statische und dynamische Inhalte mit Such- oder Navigationsfunktionalität ("Browsing") zuzugreifen. Es soll Personalisierung ermöglichen und Aspekte der Internationalisierung (Mehrsprachigkeit) berücksichtigen. Die von Nutzern formulierten Anfragen werden über eine generische Anfragekomponente je nach Anwendungsfall an spezialisierte Anfragekomponenten weitergeleitet, die auf lokale Datenbestände zugreifen. Diese Datenbestände beruhen auf unterschiedlichen Technologien (z.B. DBMS Oracle (http://www.oracle.com), Suchmaschine Lucene (http://jakarta.apache.org/lucene)) und sollen so kurze Antwortzeiten für eine Vielzahl unterschiedlicher Anwendungsfälle gewährleisten (s. auch Abschnitt 4.1). Weitere externe (nicht lokale) Datenquellen sollen mittels einer speziellen Anfragekomponente für verteilte Datenquellen in das Portal eingebunden werden. Das Portalsystem soll zusätzlich ein Content Management System (CMS) erhalten, um unstrukturierte Informationen einpflegen, verwalten und dem Nutzer in geeigneter Form anbieten zu können.

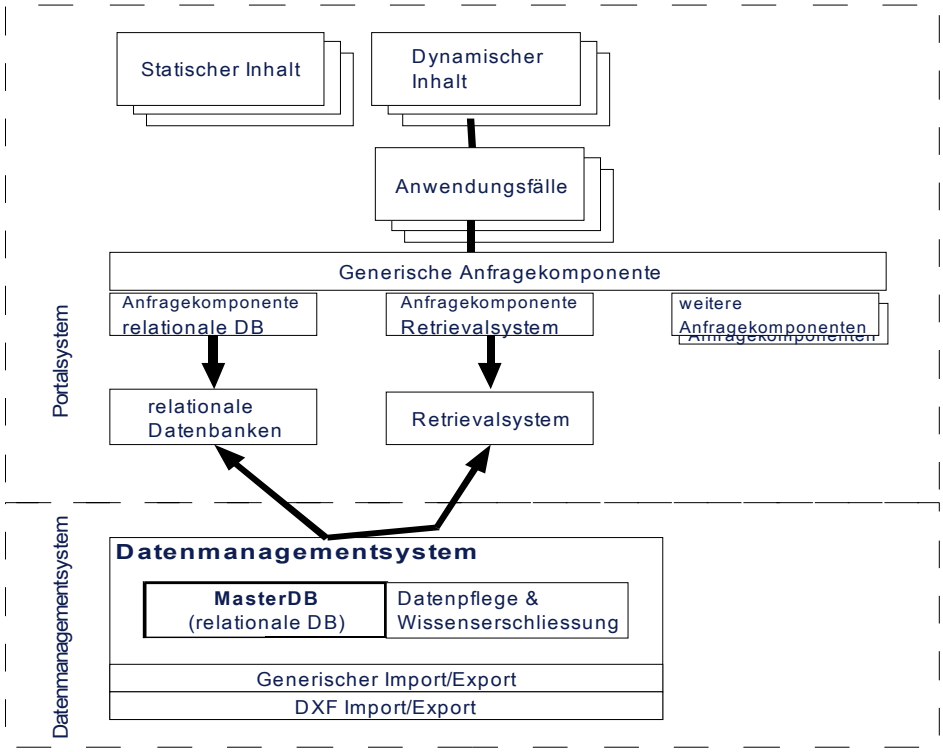


Abbildung 1: Architektur des Prototyp2003 (Herbst2003)

Das *Datenmanagementsystem* übernimmt die Sammlung, Aufbereitung und Pflege der Daten für die zentrale Nachweisdatenbank ("MasterDB") (s. auch Abschnitt 4.1). Hier befinden sich die verfügbaren Informationen in einer strukturierten Form. Eine Komponente für Datenpflege und Wissenserschließung übernimmt die Kontrolle und Verbesserung der Datenqualität (z.B. Vereinheitlichung bei Namensgebungen, Dublettenerkennung und deren Verarbeitung). Die MasterDB importiert die kompletten Datenbestände aller Informationsanbieter in einer einheitlich aufbereiteten und strukturierten Form. Die Import-/Export-Komponente ist für den Import und Export von Daten in und aus der MasterDB verantwortlich. Das wichtigste Import-Format ist das Data Exchange Format ("DXF"; s. auch Abschnitt 4.1), in dem die Projektpartner ihre Daten zur Verfügung stellen. Weitere relevante Daten aus dem Web liefert ein ontologiebasierter Crawler, der am Institut AIFB (Universität Karlsruhe) im Rahmen des SemIPort-Projekts entwickelt wird. Die Verwaltung aller strukturierten Daten in einer zentralen Nachweisdatenbank (MasterDB) ermöglicht es, die für Abfragen im Portal benötigten Informationen regelmäßig in speziell aufbereiteter Form an das Portalsystem zu exportieren.

Zur Entwicklung des Portals wurde eine *Software-Entwicklungsumgebung (SEU)* aufgebaut, um die Projektanforderungen nachvollziehbar und flexibel umzusetzen. Der Prototyp2003 wurde mit einer klassischen SEU erstellt, deren wichtigste Komponente

die Java-basierte Entwicklungsumgebung "Eclipse" (http://www.eclipse.org) ist. Zur Automatisierung der Builds und anderer Aufgaben wird das Werkzeug "ANT" (http://ant.apache.org) eingesetzt. Als Zielplattform wurde der Web-Applikationsserver "Tomcat" (http://jakarta.apache.org/tomcat/index.html) und das Webframework "Struts" (http://jakarta.apache.org/struts/index.html) gewählt. Auf den Einsatz eines J2EE Applikationsservers und der dazugehörigen Managementkomponente ("Jboss" - http://www.jboss.org/index.html) wurde zunächst verzichtet.

Im nächsten Schritt soll der Entwicklungsprozess um generative Aspekte erweitert und der Fokus der Entwicklungsarbeit auf Modellierung gelegt, d.h. modellzentrisch gearbeitet, werden (vgl. Model Driven Architecture – http://www.omg.org/mda/). Die SEU soll um geeignete Werkzeuge für Modellierung und Generierung erweitert werden, wozu folgende Alternativen näher evaluiert werden:

- Einsatz von kostenloser bzw. Open Source Software: AndroMDA (http://www.andromda.org/), XDoclet (http://xdoclet.sourceforge.net/) oder Poseidon in der Community Edition (http://www.gentleware.de/) mit geringen Lizenzkosten.

- Entwicklung nach der Model Driven Architecture (MDA) mit einem typischen MDA Werkzeug ArcStyler (http://www.arcstyler.de/) mit vergleichsweise hohen Lizenzkosten.

- Durchgängige Entwicklung ausgehend vom Modell bis zum Quellcode mit den IBM-Produkten Rational Rose/Rational XDE (http://www-3.ibm.com/software/rational/) oder dem Borland-Produkt Together (http://www.borland.com/together/index.html) mit vergleichsweise hohen Lizenzkosten.

Zur *Teamkommunikation* und *Projektdokumentation* wurde eine Infrastruktur aufgebaut, das so genannte "ProjectWeb". Dieses basiert auf "WikiWiki" Technologie (http://zwiki.org) und dient u.a. als Dateiserver.

4. Integriertes Nachweissystem

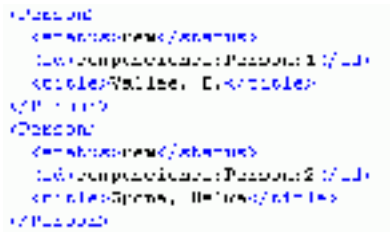
Das vom BMBF geförderte Projekt FIS-I soll ein möglichst umfassendes Informationsportal für Informatik-Fachliteratur (IO-Port) aufbauen. Ausgangspunkt ist dabei die Integration der vier bestehenden Datenbanken CompuScience, DBLP, LEABIB und CSB. Ziel ist die Abdeckung aller in den großen digitalen Bibliotheken verfügbaren Veröffentlichungen sowie traditioneller nur auf Papier veröffentlichten Arbeiten.

4.1 Zentrale Nachweisdatenbank

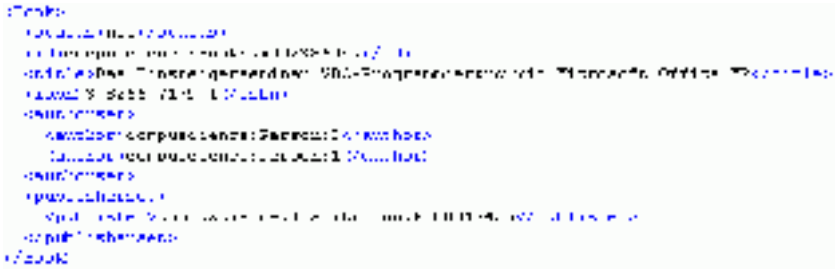
Die Projektpartner haben festgelegt, dass die derzeit verteilt vorliegenden bibliografischen Daten vereinheitlicht und in einer zentralen Nachweisdatenbank zusammengeführt und angeboten werden. Die Identität der Quelldatenbanken bleibt dabei erhalten. Es gibt zwei wichtige Gründe für diese Entscheidung: erstens die Möglichkeit, reich modellierte Daten in verschiedenen Formen und Aggregationsstufen präsentieren zu können, zweitens die Kooperation mit dem Partnerprojekt SemIPort, das diese Daten für Tests neu zu entwickelnder semantischer Werkzeuge verwendet.

Zusammen mit dem Partnerprojekt SemIPort wurde ein Modell der Fachdomäne "The World of Scientific Publication" entwickelt. Das Ziel dieser Modellierung ist es, alle für die Domäne relevanten Konzepte zu identifizieren. Zu diesen Konzepten gehören Zeitschriften, Bücher, Konferenzen, Verlage, Institutionen und Personen. Für jedes dieser Konzepte werden Eigenschaften seiner Instanzen identifiziert und die Beziehungen bestimmt, in denen es mit anderen Konzepten steht. Dabei wird z.B. festgestellt, dass eine Instanz des Konzepts "Buch" mit null oder mehr Instanzen des Konzepts "Person" in einer Beziehung steht, die sich als "Autorenschaft" bezeichnet lässt. Ausgangspunkt dieser Modellierung sind die bei den Projektpartnern existierenden Daten, deren explizite und implizite Struktur analysiert wird. Es wird angestrebt, im Modell auch die Beziehungen explizit zu machen, die in den existierenden Daten nur implizit vorhanden sind.

Die bibliografischen Daten der vier Projektpartner sind unterschiedlich strukturiert, verwenden syntaktisch und semantisch unterschiedliche Dateiformate und sind unterschiedlich normalisiert. Für die Integration dieser Daten ist eine Vereinheitlichung notwendig. Für den Austausch der Daten zwischen den Datenlieferanten und der zentralen Nachweisdatenbank wird ein gemeinsames Austauschformat, das Data-Exchange-Format (DXF), verwendet. Dieses Format soll den einfachen Import von Daten in die zentrale Nachweisdatenbank ermöglichen. Das DXF enthält die Instanzen der Konzepte in einer einfachen XML-Syntax. Alle Beziehungen, die zwischen Instanzen bestehen, werden durch die Verwendung von Schlüsseln dargestellt. So enthält etwa das folgende Beispiel die Daten zu zwei Instanzen des Konzepts "Person", die durch Schlüssel im Feld "id" identifiziert sind.



Die Instanz des Konzepts "Book" im folgenden Beispiel referenziert die obigen Personen als Autoren. Da ein Buch mehrere Autoren haben kann, werden diese in einem Container-Element zusammengefasst.



Diese Form des DXF erfordert, dass die Instanzen der Konzepte von den Datenlieferanten identifiziert werden. Dies ist für die existierenden Daten teilweise schwierig zu leisten, da z.B. unterschiedliche Schreibweisen von Namen oder Titeln nur mit erheblichem manuellen Aufwand vereinheitlicht werden können. Die im DXF gelieferten Daten werden in die zentrale Nachweisdatenbank importiert und dort in einer relationalen und normalisierten Form, dem Data-Storage-Model (DSM), gespeichert. So existiert für jedes Konzept eine eigene Tabelle, deren Einträge mit den Einträgen anderer Tabellen verknüpft sind. Die Daten aus dem obigen Beispiel werden in drei Tabellen gespeichert: "Institution", "Person" und "Book". Die Beziehungen von einem Buch zu seinen Autoren und seinem Verlag werden über N-zu-M-Beziehungen hergestellt. Ähnliche Relationen existieren auch für viele weitere Phänomene, die in der Domäne "World of Scientific Publications" beobachtet werden können.

Die Normalisierung der Daten im DSM hat zu Folge, dass bei einer Suche in den Daten und bei der Anzeige der Details, z.B. einer Publikation, Joins zur Zusammenführung der Tabellen verwendet werden müssen. Je nach Anzahl der beteiligten Tabellen kann diese Operation in einem RDBMS zu unakzeptablen Antwortzeiten des Systems führen. Da die meisten im DSM gespeicherten Daten nur selten – und dabei kontrolliert – geändert werden, erscheint es möglich, die Normalisierung im DSM teilweise wieder aufzugeben und sowohl Redundanz als auch uneindeutig bestimmte Feldinhalte zu akzeptieren. Ziel ist es, Daten, die häufig zusammen abgefragt werden, z.B. der Titel einer Publikation zusammen mit den Autorennamen und dem Zeitschriftentitel, möglichst schnell und ohne Joins zugreifbar zu machen.

4.2 Datenbanken der Partner

DBLP (Uni Trier)

Die an der Universität Trier betriebene Bibliografie DBLP (http://dblp.uni-trier.de) ist ein Prototyp für ein erfolgreiches Web-Portal für Informatik-Fachpublikationen. DBLP entstand bereits Ende 1993. Der Web-Server enthielt zunächst einfache bibliografische Informationen über Veröffentlichungen zu den Spezialgebieten "Datenbanksysteme" und "Logikprogrammierung" ("DBLP"). Später wurde die enge fachliche Fokussierung zugunsten einer breiten Abdeckung fast der gesamten Informatik aufgehoben.

In den ersten Jahren war der Umfang der Bibliografie relativ bescheiden [Le97, Le02]: Erst Ende 1996 wurde eine Größe von 50000 bibliografischen Sätzen erreicht. In den letzten Jahren konnte jedoch der Umfang der Sammlung stark vergrößert werden. Die Schwelle von 100000 bibliografischen Sätzen wurde Ende 1998 erreicht, 200000 im März 2001, 300000 im Juli 2002 und 400000 im Juli 2003.

Zentrale Publikationen der Informatik wurden in DBLP komplett erfasst. Zeitschriften wie Theoretical Computer Science, Acta Informatica, Journal of Computer and Systems Science, JACM, AI, ACM Transactions on ..., Informatik Spektrum und viele mehr sind ab Band 1 ohne Lücken indiziert. Ebenso werden für die Serien "Lecture Notes in Computer Science" und "Informatik Fachberichte" vollständige Bibliographien als Teil von DBLP angeboten. Trotzdem ist DBLP eine sehr aktuelle Sammlung: z.Zt. (Juli 2003) beschreiben 27.8% der bibliografischen Sätze Publikationen, die seit der Jahrtausendwende erschienen sind. 69.7% der Sätze beschreiben Publikationen der Jahre 1993 bis 2003.

Die Benutzerschnittstelle von DBLP bietet neben der üblichen Suchmaske zahlreiche Möglichkeiten des "Browsing" an (http://dblp.uni-trier.de): Von jedem Zitat kann zum Inhaltsverzeichnis des betreffenden Zeitschriften- oder Tagungsbandes navigiert werden. Von dort kann man zu Seiten für die gesamte Zeitschrift bzw. Tagungsserie gelangen. Natürlich kann die Navigation auch umgekehrt - 'top-down' - erfolgen. Besonders populär ist das Browsing innerhalb des "Personen-Publikationen-Netzes": Sämtliche Vorkommen von Autoren- oder Herausgebernamen sind in DBLP mit Hyperlinks auf "Personen-Seiten" unterlegt. Eine Personen-Seite führt alle für die jeweilige Person bekannten Publikationen auf und enthält Verweise auf die Koautoren.

Schwierigstes Problem bei der Wartung von DBLP ist die Gewährleistung einer befriedigenden Datenqualität. Neben möglichst korrekten Angaben zum Titel, Publikationsorgan, usw. sind Personennamen besonders kritisch. Die Generierung von Personen-Seiten ist nur sinnvoll, wenn Variationen in der Schreibweise von Personennamen vermieden werden und andererseits Personen mit gleichen Namen erkannt und unterschieden werden. In DBLP sind z.Zt. Publikationen von mehr als 260000 verschiedenen Personen aufgeführt, eine vollständige Korrektheit ist bei dieser Größenordnung unrealistisch. Um eine hohe Datenqualität zu gewährleisten, wird bei der Eingabe neuer bibliografischer Sätze versucht, die betreffenden Personen in der Datenbank zu finden und die Schreibweisen der Namen zu vereinheitlichen. Ein großer Teil der korrigierten Fehler wird während der Erfassungsarbeit gefunden. Zusätzlich werden die Betreiber von DBLP täglich von Benutzern, oft den Autoren selbst, per E-Mail auf Fehler aufmerksam gemacht.

Es ist verständlich, dass Wissenschaftler, Organisationen, Zeitschriften oder Tagungsreihen bewertet und verglichen werden. Ein einfaches Zählen von Publikationen ist natürlich immer unsinnig, ein Ranking auf Grundlage der DBLP-Daten (siehe http://database.cs.ualberta.ca/coauthorship, http://citeseer.nj.nec.com/impact.html, [Ne01]) ist zusätzlich problematisch [Ma02, Le03], weil DBLP und jede andere existierende Bibliografie von einer fairen, gleichmäßigen Abdeckung der gesamten Informatik (noch) weit entfernt sind. Primäres Ziel für DBLP (und IO-Port) ist die möglichst umfassende

Information über Fachpublikationen. Die DBLP-Daten stehen Dritten für bibliometrische Untersuchungen zur Verfügung.

LEABIB (TU München)

Die bibliografische Datenbank LEABIB (http://wwwmayr.informatik.tu-muenchen.de/leabib/) des Lehrstuhls für Effiziente Algorithmen der Technischen Universität München besteht seit 1983 und enthält mehr als 68000[1] Literatureinträge aus dem Bereich der Theoretischen Informatik. Die Datenbank deckt dabei einen Zeitraum von 1926 bis heute ab.

Die Daten werden in einer Textdatei mit festem Format (SRC-Format) gespeichert. Dabei werden folgende Daten erfasst: citkey, author, title, booktitle, editor, institution, journal, number, organization, pages, publisher, type, volume, year, keywords, note, series und abstract. Für die weitere Verwendung und die Webschnittstelle wird das obige Format in BiBTEX konvertiert. Für die Suche in der BiBTEX-Datei wurde das Tool leagrep entwickelt. Dieses bietet eine gezielte, fehlertolerante Suche nach bestimmten Feldinhalten.

Eine bibliothekarische Kraft erfasst zur Zeit noch per Hand die Literaturdaten. Es werden ca. 500 Einträge pro Monat erfasst. Vor kurzem wurde ein Tool fertiggestellt, das es erlaubt, Literaturdaten aus dem Internet halbautomatisch zu erfassen. Dieses Wrapper-Tool extrahiert die relevanten Daten aus einer Webseite und speichert die erfassten Daten im SRC-Format (weitere Ausgabeformate, z.B. SQL, können auf einfache Weise realisiert werden). Mit Hilfe dieses Tools soll in Zukunft ein großer Teil der Literaturdaten erfasst werden. Dadurch sollte es möglich sein, ca. 1000 - 1500 Einträge im Monat zu erfassen.

Die Daten werden nach der Erfassung zuerst automatisch überprüft. Dabei werden bestimmte Feldinhalte auf Gültigkeit und Richtigkeit durch Vergleich mit vordefinierten Werten überprüft. Aktuell werden die Felder institution, organization, publisher, journal und series auf diese Weise geprüft. Anschließend erfolgt die Prüfung der Autoren und Editoren. Nach Abschluss der Korrektur generiert ein Programm die Schüssel (citkey) der Einträge. Weitere Tools dienen zum Erkennen und Löschen von doppelten Einträgen. Nach abgeschlossener Korrektur werden die neu erfassten Daten an die bestehende SRC-Datei angehängt, nach dem Feld citkey sortiert und ins BiBTEX-Format konvertiert.

Die Erfassung der Daten soll in Zukunft durch verstärken Einsatz des Wrapper-Tools beschleunigt werden. Ebenso wird das Tool um eine Autoren-Datenbank erweitert, so dass eine schnellere Autorenidentifizierung möglich wird. Eine neue Version des Wrapper-Tools wird die Prüfung der Feldinhalte auf Richtigkeit beinhalten.

---

[1] Stand: Juli 2003

Die Erfassung der Daten wird weiterhin im SRC-Format erfolgen. Ein Import des SRC-Formates in eine SQL-Datenbank wird zur Zeit geprüft. Leider bietet SQL keine fehlertolerante Suche in den Daten. PostgreSQL lässt sich z.B. durch Implementierung eines geeigneten Moduls um eine fehlertolerante Suche erweitern. Ein weiteres Ziel ist die Auswertung der Autoren-Mitautoren-Beziehung.

The Collection of Computer Science Bibliographies (Uni Karlsruhe)

Die "Collection of Computer Science Bibliographies" ist die größte und älteste freie bibliografische Nachweisdatenbank im Bereich der Informatik und Mathematik. Sie ist eine Sammlung von im Internet kostenlos verfügbaren Einträgen.

Die "Collection of Computer Science Bibliographies" wurde von Alf-Christian Achilles am Lehrstuhl Informatik für Ingenieure und Naturwissenschaftler (Prof. Dr.-Ing. Roland Vollmar) an der Universität Karlsruhe (TH) seit 1992 entwickelt. Nachdem Anfragen anfangs per E-Mail gestellt werden konnten, ist sie seit 1994 im WWW erreichbar (http://liinwww.ira.uka.de/bibliography/). Die "Collection of Computer Science Bibliographies" bietet für Forscher und Entwickler seit nun mehr als 10 Jahren einen kostenlosen Service. Sie wird auf Platz 1 der Liste http://directory.google.com/Top/Computers/Computer_Science/Publications/ geführt. Von knapp 1000 Anfragen pro Tag im Jahre 1995 ist die Zahl der Zugriffe auf die Literaturdatenbank auf inzwischen mehr als 10000 täglich angestiegen. Im März 2003 wurde ein neuer Rekord von fast 350000 Zugriffen registriert.

Die Literaturdatenbank in Zahlen (Stand Juli 2003):

- Insgesamt über 1300000 Literatureinträge, davon
- mehr als 450000 Einträge, die Klassifizierungsdaten (ACM, MSC) oder Schlüsselwörter enthalten,
- mehr als 230000 Einträge, die Verweise auf Volltextdokumente enthalten, und
- mehr als 170000 Einträge, die eine Zusammenfassung enthalten.

Die bibliografischen Daten stammen aus derzeit mehr als 1300 Quellen. Die meisten von ihnen werden regelmäßig aktualisiert und über diese Aktualisierungen wird die "Collection of Computer Science Bibliographies" wöchentlich auf den neuesten Stand gebracht. Viele Bibliographien sind von Experten des jeweiligen Gebietes als Teil ihrer Forschungsarbeit zusammengestellt. Dies macht sie zu einem wertvollen Instrument für die Forschung. Andere Quellen enthalten vollständige Bibliographien von vielen Zeitschriften und Konferenzen, Sammlungen aller technischen Berichte von Universitäten, Forschungsgruppen und Forschungsabteilungen einiger großer Firmen. Seit kurzem sind auch die Daten, die von E-Print Servern über die Open Archive Initiative Schnittstelle zur Verfügung stehen, in der "Collection of Computer Science Bibliographies" integriert. Zum Beispiel werden bei Anfragen nun auch die Daten von http://www.arXiv.org, soweit sie Informatik-Relevanz haben, mit durchsucht.

Die Struktur der Datenbank und die Suchschnittstelle erlauben für jeden gefundenen

Eintrag die Identifizierung der Quellbibliografie. Das führt die Benutzer auch zu Bibliografien, die besonders für den interessierenden Forschungsbereich relevant sind.

Von den Arbeiten im Rahmen des Projekts ist eine weitere Steigerung der Qualität sowohl der zur Verfügung gestellten Daten, als auch der Suchergebnisse und ihrer Präsentation zu erwarten. Weitere qualitativ hochwertige Datenquellen sollen integriert werden. Die Suchschnittstellen sollen verbessert werden, und insbesondere sollen noch bessere Methoden entwickelt und implementiert werden, um Einträge, die die gleiche Literaturstelle beschreiben, zu erkennen und in der Präsentation der Suchergebnisse zu vermeiden.

CompuScience (FIZ Karlsruhe)

CompuScience ist eine bibliografische Datenbank für Veröffentlichungen auf dem Gebiet der Informatik und der Computer-Technologie. Sie wurde ab 1987 im Rahmen eines BMBF-Projekts entwickelt und wird vom FIZ Karlsruhe betrieben. Die Datenbank beinhaltet Literaturnachweise seit ca. 1965 aus Zeitschriften und Büchern (u.a. Lecture Notes in Computer Science). Auch nicht-konventionelle Literatur wird erfasst, wie z.B. Dissertationen, Reports und Preprints.

CompuScience enthält Daten verschiedener Provider: von ACM (insbesondere die Referate der ACM Computing Reviews), Infodata (Potsdam), der TU Dresden (Dresden) sowie vom FIZ Karlsruhe selbst erstellte Daten. Insbesondere werden die für die Informatik relevanten Daten aus den Datenbanken des Zentralblatts für Mathematik und des Zentralblatts für Didaktik der Mathematik übernommen und darüber hinaus zusätzliche Daten speziell für die CompuScience-Datenbank erfasst.

Die Datenbank weist Informatikliteratur aus den folgenden Sachgebieten nach:

- Theoretische Informatik
- Software
- Computer und Computersysteme, Netzwerke
- Datenstrukturen
- Kryptologie
- Informationssysteme
- Algorithmen und Komplexität
- Künstliche Intelligenz
- Robotik
- Computer-Grafik, Bildverarbeitung
- Simulation, Modellbildung
- Anwendungen in Medizin, Management, Ingenieurs- und Naturwissenschaften
- Informatik und Ausbildung

Jeder Eintrag in der Datenbank enthält die bibliografischen Angaben einer Veröffentlichung. Die meisten enthalten zusätzlich Zusammenfassungen oder Referate

sowie Schlagworte und sind darüber hinaus entsprechend dem ACM Klassifikationsschema klassifiziert. Die meisten neueren Einträge enthalten Links zu den im Internet verfügbaren Volltexten. CompuScience wird derzeit unter der Internet-Retrieval-Software von "Cellule MathDoc" angeboten, die schon von den Datenbanken MATH, MATHDI und ERAM verwendet wird. Sie garantiert eine einfache und sichere Funktionalität für den Nutzer, die eine Auswahl unter mehreren Suchmasken und Ergebnisanzeigen sowie die Möglichkeit des Browsens über das ACM-Klassifikationsschema zur Verfügung haben.

Im Rahmen des FIS-I-Projekts wird die CompuScience einer der Bausteine der zentralen FIS-I Datenbank sein. In diesem Zusammenhang werden an der CompuScience umfangreiche qualitätsverbessernde Maßnahmen durchgeführt und die Datenbank völlig neu aufgebaut. Die Verbesserungen betreffen die überwiegend folgenden Punkte:

- Um eine bessere Aktualität zu erreichen sowie das Auftreten von Lücken und Dubletten zu vermeiden, wurde die Inputprozedur völlig umgestellt.

- Um den oben beschriebenen Scope möglichst genau zu treffen, wurde das Profil überarbeitet.

Es ist geplant, die noch nicht in das jetzt benutzte Datenformat umgesetzten älteren Daten bis zum Ende des Jahres wieder in die Datenbank einzuspielen. Darüber hinaus wird bis dahin ein regelmäßiges Update möglich sein, das monatlich ca. 2500 neue, aktuelle Einträge liefern wird. Zu den derzeit (Stand Juni 2003) enthaltenen ca. 170.000 Einträgen werden bis zum Ende des Jahres auf diese Weise noch einmal ca. 40.000 Einträge dazukommen. In den folgenden Jahren wird die Datenbank dann jährlich um ca. 30.000 Einträge erweitert werden können. Dabei wird es sich vor allem um Nachweise neuerer und neuester Literatur handeln, aber auch um das Auffüllen vorhandener Lücken.

CompuScience soll ferner durch Fortsetzung und Ausbau von Kooperationen mit Verlagen, der GI und deren Fachgruppen sowie universitären und außeruniversitären Informatik-Instituten thematisch, konzeptionell und zeitlich komplettiert und aktualisiert werden. Durch eine Intensivierung der Kontakte zu den Nutzern sollen die Zugriffsmöglichkeiten optimiert und den Bedürfnissen der Nutzer angepasst werden.

5. Geschäftsmodell

Die Erarbeitung eines Konzeptes zur finanziellen Bewertung des Erfolges des IO-Port ist eine wichtige Grundlage für den langfristigen Betrieb von IO-Port. Wird das IO-Port als eine Investition betrachtet, die der Nutzerbindung dient, können die negativen beziehungsweise positiven Investitionskonsequenzen als Kosten oder Nutzen interpretiert werden. Die Ermittlung der Wirtschaftlichkeit dieser Investition kann hierbei mittels der klassischen Investitionsrechnung erfolgen, wobei die Ausgaben den laufenden Einnahmen gegenübergestellt werden. Ein Beispiel für solch eine Berechnung liefert der Return On Investment (ROI) des IO-Port. Bei dieser Berechnung wird das

investierte Kapital dem er- warteten oder laufenden Gewinn gegenübergestellt. Voraussetzung für derartige Berechnungen ist eine möglichst genaue Abschätzung der aus der Einführung des Portals resultierenden Kosten und Nutzen. Dies ist allerdings aufgrund der Komplexität der Problemstellung nur bedingt möglich. Dennoch bildet eine solche investitionstheoretische Kosten/Nutzen-Schätzung die Grundlage der Profitabilitätsrechnung des IO-Port. Eine Schwäche der Investitionsrechnung liegt in der einseitigen Betrachtungsweise – sie berücksichtigt lediglich finanzielle Ergebnisse – und lässt weitere kritische Erfolgsfaktoren außer Acht. Die unberücksichtigten, nicht-monetären Leistungsbeiträge sind dabei kein Selbstzweck, sondern nur Teilziele auf dem Weg zu einer verbesserten finanziellen Situation. Angesichts der hohen Komplexität und vor allem der nicht unmittelbar finanziell messbaren Wirkungsmöglichkeiten des IO-Port (z.B. Nutzerzufriedenheit, Nutzerbindung usw.) erscheint eine ausschließliche Steuerung und Bewertung über Finanzkennzahlen nicht ausreichend. Aufgrund einer mehrdimensionalen Perspektive bei der Planung und Steuerung des IO-Port wird empfohlen, eine Kontrolle auf einem spezifischen Kennzahlensystem, dem sog. Portal-Kontroll-Cockpit (PKC), aufzusetzen. Im Gegensatz zur klassischen Betrachtung rein monetärer Effekte integriert das PKC auch "weiche" Faktoren in das Kontrollsystem. Die Finanzperspektive des PKC beinhaltet ausschließlich finanzielle ("harte") Faktoren, die sich als Wirkung aus den vorgelagerten Perspektiven ergeben. Diese können als Ursachen oder "weiche" Faktoren interpretiert werden.

Die zielgerichtete Aufbereitung von Informationen, wie beim IO-Port, ist die Grundvoraussetzung zur effektiven und effizienten Wissensbeschaffung für die Nutzer. Die zentrale Dienstleistung des IO-Port ist also die Vermittlung der gewünschten Informationen an den Nutzer. Erlösquellen sind der integrale Bestandteil eines Geschäftsmodells, denn sie bestimmen die Einkünfte und bilden die Basis der Wirtschaftlichkeit des IO-Ports. Die direkte Erlösquelle über den direkten Verkauf der Informationsprodukte ist im IO-Port mit Problemen behaftet, da die integrierten Datenquellen bisher großenteils kostenlos zur Verfügung gestellt wurden. Hier ist über die Mehrwerte eine Erlösquelle zu erschließen.

Grundsätzlich kann das IO-Port von seiner Funktionalität her als ein Online-Shop betrachtet werden, über den digitale Produkte angeboten und nachgefragt werden. Die Informationsanbieter nutzen das Portal dabei als zentrale Schnittstelle. Zentral ist neben der Bereitstellung von bibliografischen Nachweisinformationen und inhaltsbeschreibenden Mehrwertinformationen (z.B. Abstract, Klassifikation etc.) auch der unmittelbare Zugriff auf die Volltexte (Dokumente). Nutzer können Zeitlizenzen für Informationen oder Dokumente erwerben, über die sie für eine gewisse Zeitspanne online zugreifen können. Es lassen sich dabei für Einzelnutzer zwei Arten von Lizenzen unterscheiden. Bei *Kurzzeitlizenzen* ("Schnupperangebote") wird einem Nutzer die Möglichkeit gegeben, nur die Grundfunktionalitäten des Portals kennen zu lernen. Es wird für einen kurzen Zeitraum Zugriff auf lizenzpflichtige Teile des Portals gewährt (etwa eine Stunde). *Einzellizenzen* sind langfristig gültige Lizenzen mit einer Mindestdauer von zum Beispiel einem Jahr. Für institutionelle Nutzergruppen stellen *Campuslizenzen* erfahrungsgemäß das praktikabelste Lizenzmodell dar. Campuslizenzen erlauben allen Gruppenmitgliedern jederzeitigen Zugriff auf das Portal. Die Preise ergeben sich dabei aus der Größe der Nutzergruppe. Lizenzen für den Zugriff auf die kostenpflichtigen

Dienste des IO-Port können online erworben werden. Für die Bezahlung stehen unterschiedliche Verfahren zur Verfügung. Es könnte beispielsweise für jeden registrierten Nutzer ein privates IO-Port-Konto, das per Überweisung aufgeladen werden kann, eingerichtet werden. Alternativ könnte die Abrechnung auch mittels elektronischen Geldes erfolgen. Zusätzliche Optionen sind das Bezahlen per Geldkarte oder auf Basis von mobilen Endgeräten (z.B. Handy oder Handheld). Für institutionelle Nutzergruppen bieten sich zur Abrechnung sogenannte Gruppenkonten an. Ein Gruppenadministrator kann dort einen bestimmten Geldbetrag einzahlen. Gruppenmitgliedern kann anschließend erlaubt werden, bis zu einem pro Person festgelegten Betrag das Geld für den Erwerb von Kurzzeit- oder Einzellizenzen zu nutzen. So könnten beispielsweise Universitäten flexibel auf konkrete Nutzerbedürfnisse reagieren.

6. Zusammenfassung und Ausblick

Die in diesem Beitrag dargestellten Anforderungen führten zur ersten prototypischen Implementierung des IO-Port. Dabei standen das Zusammenführen der heterogenen Datenbestände in eine zentrale Nachweisdatenbank und die Realisierung der grundlegenden Recherchefunktionalität im Vordergrund. Die Architektur soll in der beschriebenen Art weiterentwickelt werden. Ziel ist die vollständige Realisierung der geplanten Softwarearchitektur, die die komplexen Anforderungen des Projekts in eine Anwendung mit verlässlichem und nachhaltigem Betrieb umsetzt. Dabei soll u.a. die Funktionalität für Recherche und Navigation in den Datenbeständen erweitert werden. Neben der Integration von Werkzeugen aus dem SemIPort-Projekt ist die Einbindung von eLearning-Komponenten geplant. Zusätzlich ist die Ableitung eines stimmigen Geschäftsmodells von zentraler Bedeutung für den nachhaltigen Erfolg des IO-Port. Die bisher entwickelten Konzepte werden dabei als Grundlage dienen.

Informatik-Fachpublikationen sind heute oft in elektronischer Form im Internet verfügbar. Für viele Wissenschaftler und Studenten sind leistungsfähige Suchmaschinen wie Google zum Auffinden von Forschungsliteratur wichtiger als Bibliothekskataloge oder traditionelle Bibliografien. Suchmaschinen haben jedoch ihre Grenzen. Auch die auf Informatik-Fachpublikationen spezialisierte Suchmaschine "CiteSeer" sieht nur den offen zugänglichen Teil der Publikationen. Die von CiteSeer für jedes indizierte Dokument automatisch durchgeführte Zitatanalyse liefert zwar Hinweise auf viele dem System nicht vorliegende Publikationen, eine auch nur annähernd vollständige Bibliographie kann jedoch nicht erwartet werden. Die wichtigsten kostenpflichtigen digitalen Bibliotheken der Informatik von ACM, IEEE Computer Society, Springer und Elsevier sind voneinander isoliert. Mediator-Systeme wie BibFinder können ein wirklich integriertes System nicht ersetzen. Allein das aus dem "ACM Guide to Computing Literature" entstandene ACM Portal versucht, Literatur anderer Verlage zu einem nennenswerten Teil abzudecken. Das Gesamtvolumen der Datenbank bleibt jedoch unbefriedigend. Nicht-englischsprachige Informatik-Literatur bleibt unberücksichtigt.

Die systematische Erfassung "grauer Literatur" erscheint dagegen zunehmend obsolet.

Wissenschaftler stellen ihre Arbeiten in der Regel direkt auf ihren persönlichen Web-Seiten bereit. Die Zusammenstellung von Preprints in Reihen, die von Universitäten oder Forschungsinstituten herausgegeben werden, spielt eine immer geringere Rolle. Die unüberschaubar große Anzahl von Fundorten, die sehr heterogene Struktur und Qualität der "Selbstverlage", die mangelnde Archivierung und der oft problematische rechtliche Status (Copyright) lassen eine auch nur teilweise manuelle Erfassung unsinnig erscheinen.

Dem Benutzer von IO-Port sollen neben dem Zugriff auf eine qualitativ möglichst hochwertige Datenbank Hilfestellungen zur Suche im Bereich der grauen Literatur gegeben werden. Dies kann in Form vorformulierter Anfragen an Suchmaschinen wie Google, CiteSeer usw. oder in Form einer Metasuchmaschine erfolgen. Neben der gezielten Suche nach bekannten Arbeiten ist die Suche nach "Home Pages" von Informatiker(inne)n besonders wichtig.

Literaturverzeichnis:

[BCK98] Bass, P.; Clements, P,; Kazman, R.: Software Architecture in Practice. Addison-Wesley, 1998.

[Le97] Ley, M.: Die Trierer Informatik-Bibliographie DBLP. GI Jahrestagung 1997, Informatik Aktuell, 1997, S. 257-266.

[Le02] Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. SPIRE 2002, LNCS 2676, 2002, S. 1-10.

[Le03] Ley, M.: ACM SIGMOD Contribution Award 2003 Acceptance Speech. http://www.acm.org/sigmod/dblp/db/about/contributionaward03.html

[Ma02] Mattern, F.: Zur Evaluation der Informatik mittels bibliometrischer Analyse. Informatik Spektrum 25(1), 2002, S. 22-32.

[Ne01] Nebel, B.: Ranking? Publikationen, Zitate, Drittmittelprojekte und Promotionen an deutschen Informatikfakultäten im Spiegel des WWW. Informatik Spektrum 24(4), 2002, S. 234-249.

[St02] Starke, G.: Effektive Software-Architekturen. Ein praktischer Leitfaden, Hanser Fachbuchverlag, 2002.