Fachinformationssystem Informatik (FIS-I) und Semantische Technologien für Informationsportale (SemIPort)

Peter Fankhauser[7], Norbert Fuhr[8], Jens Hartmann[2], Anthony Jameson[5], Peter Klas[8], Stefan Klink[6], Agnes Koschmider[2], Sascha Kriewel[8], Patrick Lehti[7], Peter Luksch[1], Ernst W. Mayr[3], Andreas Oberweis[2], Paul Ortyl[4], Stefan Pfingstl[3], Patrick Reuther[6], Ute Rusnak[1], Guido Sautter[9], Klemens Böhm[9], André Schaefer[8], Lars Schmidt-Thieme[10], Eric Schwarzkopf[5], Nenad Stojanovic[2], Rudi Studer[2], Roland Vollmar[4], Bernd Walter[6], Alexander Weber[6]

[1]Informationsdienste, FIZ Karlsruhe, 76344 Eggenstein-Leopoldshafen
[2]Institut AIFB, Universität Karlsruhe (TH), 76128 Karlsruhe
[3]LS für Effiziente Algorithmen, TU München, 85748 Garching
[4]LS Informatik für Ingenieure und Naturwissenschaftler, Universität Karlsruhe (TH), 76128 Karlsruhe
[5]DFKI, Saarbrücken, 66123 Saarbrücken
[6]DBIS, Universität Trier, FB4 Informatik, 54286 Trier
[7]FhG IPSI, Darmstadt, 64293 Darmstadt
[8]FG Informationssysteme, Universität Duisburg
[9]IPD, Universität Karlsruhe (TH)
[10]Institut für Informatik, Universität Freiburg, 79110 Freiburg

Abstract: Der Workshop soll Forscher und Praktiker aus den Bereichen Informationsportale, Semantische Technologien und Maschinelles Lernen zusammenbringen, um neue Methoden und Technologien für Informationsportale vorzustellen. Der Workshop wird von Partnern der beiden bmb+f-geförderten Projekte Fachinformationssystem Informatik (FIS-I) und Semantische Methoden und Werkzeuge für Informationsportale (SemIPort) durchgeführt. Im Projekt FIS-I erstellen die Gesellschaft für Informatik zusammen mit dem Fachinformationszentrum Karlsruhe ein Informationsportal für die Informatik. Das Projekt SemIPort begleitet diesen Aufbau mit der Entwicklung semantischer Methoden und Werkzeuge. Der Workshop soll über die Projektgrenzen hinaus eine offene Diskussion projektrelevanter Themen ermöglichen. Aufgrund der direkten Einbindung der GI in die Projekte fanden projektbegleitende Workshops bereits auf der INFORMATIK 2003 und INFORMATIK 2004 statt.

Einleitung

Informationsportale stellen eines der anspruchsvollsten Einsatzfelder für semantische Technologien dar. Die Menge der zu verwaltenden Informationen ist typischerweise groß, neue Informationen müssen schnell integriert werden und die Darstellung muss sich flexibel an neue Entwicklungen sowie heterogene Benutzergruppen anpassen. Informationsportalen stellt sich das Problem der Integration von Daten aus heterogenen Quellen in besonderem Maße, insbesondere wenn die Nutzer aktiv an der Informationserfassung mitwirken. Semantische Technologien bieten hierfür die Basis, Methoden des Maschinellen Lernens müssen für eine automatische oder computergestützte Strukturierung der Informationsbestände herangezogen werden. Zu den adressierten Themen gehören im Einzelnen:

- Visualisierung und Browsing komplexer Datenbestände
- Personalisierung, User Modelling und Recommender-Systeme
- Web-basierte Informations Integration
- Ontologie- und Metadaten-Modellierung
- Knowledge Warehousing für große Datenbestände
- Semantic Web Mining
- Wissensintegration

Der Workshop greift die Themen auf und zeigt Methoden und Lösungen, die in den Projekten Fachinformationssystem Informatik (FIS-I) und Semantische Methoden und Werkzeuge (SemIPort) entwickelt und realisiert wurden. Weitere Projekte aus dem thematischen Umfeld zeigen zusätzliche Lösungsansätze.

Der erste Themenblock **Fachinformationssystem Informatik (FIS-I)** behandelt Aspekte aus Sicht des Nutzers sowie des Informationsanbieters. Es werden Informationssysteme vorgestellt, die unterschiedliche Ansätze und Funktionalitäten für die Informationssuche bieten. Im ersten Beitrag stellen Peter Luksch und Ute Rusnak *io-port.net*, das neue Informationsportal für die Informatik, vor. Im folgenden Beitrag beleuchten Agnes Koschmider und Andreas Oberweis den Einfluss von Google scholar auf *io-port.net*. Für *io-port.net* wurden Werkzeuge zur Datensammlung und Informationsgewinnung entwickelt, welche die Grundlage für leistungsfähige Informationssysteme bilden. Mit der Frage, wie Daten schnell erfasst und bearbeitet werden können, beschäftigen sich Ernst W. Mayr und Stefan Pfingstl in ihrem Beitrag. Paul Ortyl und Roland Vollmar beschreiben in ihrem Beitrag heuristische Verfahren für die semantische Anreicherung unstrukturierter bibliographischer Daten.

Der zweite Themenblock **Semantische Methoden und Werkzeuge für Informationsportale (SemIPort)** stellt die Methoden und Werkzeuge für die Nutzung und den Betrieb von Informationsportalen vor, die im Projekt entwickelt wurden. Hierbei handelt es sich um (i) Werkzeuge für das Backend, mithilfe derer ein Portal aufgebaut werden kann: der SWQL-Prozessor, der das Laden von Daten in die Ontologie sowie das Abfragen von Wissen aus der Ontologie erlaubt, der Metis-Crawler, der das Internet gezielt nach Informationen durchsucht und in die Ontologie einspeist, sowie Werkzeuge zur Datenintegration und Datenanreicherung wie der Duplicate Detector und die Artemis Data Mining Workbench, sowie um (ii) Werkzeuge für Nutzer-Dienste, mithilfe derer

Nutzer auf die Informationen des Portals zugreifen können. Diese lassen sich nochmals in serverseitige Dienste gliedern, die auf der Website laufen, wie etwa der Librarian Agent, der Nutzer beim sukzessiven Suchen berät, und die Semantic Recommendation Services, die z.B. interessante neue Publikationen vorschlagen, sowie in clientseitige Dienste, die beim Benutzer selbst laufen, wie etwa der DBL-Browser zum Durchstöbern der Wissensbasis, dem Document Manager zum Anlegen einer reichhaltigen persönlichen Bibliographie. Der Einsatz semantischer Technologien gewährleistet dabei die Interoperabilität und einfache, gemeinsame Erweiterbarkeit all dieser Werkzeuge. Auch die Anpassung an andere Wissensdomänen, etwa für andere Arten von Informationsportalen, lässt sich für die meisten Werkzeuge durch den Austausch der Metadaten-Ontologie bewerkstelligen.

Im dritten Themenblock werden zwei **Projekte aus dem thematischen Umfeld** vorgestellt. Norbert Fuhr und Peter Klas stellen Daffodil vor, ein Zugangssystem für heterogene digitale Bibliotheken. Guido Sautter und Klemens Böhm erläutern in ihrem Beitrag ein Semantik-basiertes Retrieval Biosystematischer Dokumente.

1 Projekt Fachinformationssystem Informatik (FIS-I)

1.1 Das Informatikportal *io-port.net*

Das Informationsportal für die Informatik *io-port.net* wurde im Rahmen des vom Bundesministerium für Bildung und Forschung (bmb+f) geförderten Projekts „Fachinformationssystem Informatik“ entwickelt. Daran sind neben der Gesellschaft für Informatik e.V. (GI) und dem FIZ Karlsruhe weitere Wissenschaftler der Universitäten Karlsruhe und Trier sowie der Technischen Universität München beteiligt. Mit dem Fachportal *io-port.net* wird ein effizienter und nachhaltiger Zugang zu weltweit verteilter wissenschaftlich-technischer Fachinformation in der Informatik und verwandten Teilgebieten angeboten. Alle Schritte zur Informationsbeschaffung werden unter einer einfach zu bedienenden Oberfläche zentralisiert und durch leistungsfähige Werkzeuge unterstützt: fachbezogene Informationsrecherche und Navigation, Auswahl relevanter Informationen sowie anbieterübergreifender Zugang zum Volltext. Authentifizierungs- und Autorisierungsmechanismen ermöglichen das Angebot von personalisierten Diensten sowie die Integration von kostenpflichtigen Mehrwert-Diensten [Ko04].

io-port.net integriert Informatik-Wissen aus bislang separat verfügbaren Informationsquellen unter einer einheitlichen Oberfläche. Die Datenbasis mit derzeit ca. 2 Mio. Publikationen bilden die Datensammlungen CompuScience, DBLP, LEABiB und Collection of Computer Science Bibliographies (CCSB).

Eine strukturierte Aufbereitung mit standardisierten Metadaten, eine gezielte Auswertung und Verknüpfung der Daten verbessern und beschleunigen die fachliche Informationsgewinnung. Die Ergebnisse einer Trefferliste sind vielfältig miteinander verknüpft

und ermöglichen eine weitere Navigation im Datenbestand. Vom Autor einer Publikation gelangt man per Mausklick auf dessen individuelle Publikationsliste, von einem Zeitschriftentitel zum Inhaltsverzeichnis oder zum Jahresregister der Zeitschrift. Mehrfachtreffer im Suchergebnis werden durch ein Werkzeug weitestgehend ermittelt und in einem einheitlichen Datensatz zusammengeführt. Die Verweise auf die Originalquellen, die unterschiedlichen Umfang haben, werden angezeigt. Ausgehend von den Suchergebnissen wird eine Volltextvermittlung oder der direkte Zugriff auf elektronische Volltexte angeboten, z.B. auf Tagungsberichte, Dissertationen und Zeitschriften verschiedener GI-Fachgruppen sowie weiterer Anbieter (u.a. Verlage, Fachgesellschaften, Hochschulen). Der direkte Zugriff auf die elektronischen Volltexte der LNI-Reihe (Lecture Notes in Informatics) der GI wird exklusiv über *io-port.net* angeboten.

Für ausgewählte Fachbereiche der Informatik wurden exemplarisch Themenseiten eingerichtet, die einen kompakten Überblick über aktuelle Forschungs- und Entwicklungsergebnisse, laufende Projekte, Publikationen, sowie relevante Links und aktuelle Veranstaltungen geben. Eine redaktionelle Betreuung dieser Themenseiten ist Voraussetzung für deren Erfolg. Als weitere Komponente von *io-port.net* wird ein Werkzeug angeboten, das die ortsunabhängige Erstellung und Verwaltung von persönlichen bzw. institutsbezogenen Publikationslisten ermöglicht. Weitere Fachinformationen in *io-port.net* sind: Informatik-Lexikon, allgemeine Linklisten sowie ein umfassender Konferenzkalender.

Semantische Werkzeuge und personalisierte Dienste unterstützen eine effiziente Recherche. *io-port.net* dient als Anwendungsumgebung für die in dem forschungsorientierten Projekt „Semantische Methoden und Tools für Informationsportale (SemIPort)“ neu entwickelten Werkzeuge und integriert diese in die wissenschaftliche Informationsversorgung.

1.2 Suche nach Fachinformation im Zeitalter von Google Scholar

Seit November 2004 bietet Google unter scholar.google.com einen speziellen Suchdienst für wissenschaftliche Recherchen an. Google scholar durchsucht nach Angaben der Betreiber Bücher, wissenschaftliche Abhandlungen, technische Dokumente, Fachzeitschriften und sonstige Literatur aus dem Umfeld von Forschung und Lehre. Einen bedeutenden Vorteil gegenüber Google [Cu05] erlangt dieser Suchdienst durch die Verwendung einer Zitationsanalyse für Publikationen (Cited by). Allerdings ist der Suchalgorithmus noch nicht effizient, weil Suchergebnisse nicht immer zu Volltexten führen und damit wenig nutzbringend sind. Außerdem sind viele Suchergebnisse nicht anzeigbar, weil es sich um Zitate handelt. Das Informationsportal *io-port.net* hat sich mit seinen Informationsangeboten und integrierten Suchfunktionalitäten im Gegensatz zu Google scholar auf ein Fachgebiet spezialisiert - die Informatik - und deckt dieses Gebiet umfassend ab [Ko04]. In *io-port.net* sind auch semantische Werkzeuge integriert, die den Benutzer bei seiner Suche nach relevanten Inhalten unterstützen. Die Literaturnachweise sind aufgrund vollständiger Angaben und Abstracts qualitativ hochwertig und nach der Anzeige von Suchergebnisse gelangen Nutzer zur Volltextvermittlung. Gegenüber Google scholar finden Nutzer in *io-port.net* Dokumentationen zur Nutzung der Suchfunktionalitäten und zu den semantischen Werkzeugen. Ein weiterer Vorteil von *io-*

port.net besteht in der Klärung rechtlicher Fragestellungen zum Volltextdownload oder zur Referenz auf Volltexte.

In einem Kurztest ergab die Suche in Google scholar nach dem Begriff „evolutionäre Algorithmen“ 538 Treffer, wobei auf den ersten beiden Seiten schon 15 Treffer nicht angesehen werden konnten – die Suchmaschine hatte die Texte nur als Zitate in anderen wissenschaftlichen Werken gefunden. In *io-port.net* ergab die Suche nach diesem Begriff 65 Treffer. In Google scholar werden zwar mehr Suchtreffer angezeigt, allerdings handelt es sich um viele Zitate oder Buchtreffer und erfahrungsgemäss beachten Suchmaschinennutzer auch nur die ersten Treffer einer Anzeige. Die englischsprachige Suche nach „evolutionary algorithms“ ergab eine Trefferzahl von 183.000, hier sind unter den ersten 100 Treffern 21 Zitate und 10 Büchertreffer. In *io-port.net* werden 4617 Treffer an wissenschaftlichen Publikationen und Zeitschriftenartikeln angezeigt. Ähnliche Unterschiede in der Treffergröße zwischen Deutsch und Englisch gibt es, wenn nach dem Begriff „Entwurfsmuster“ beziehungsweise „design pattern“ gesucht wird. Der englische Begriff liefert 831.000 Treffer und der deutsche 941. In *io-port.net* ergeben sich ähnliche Unterschiede bei der Trefferanzahl zwischen Englisch und Deutsch: für den englischen Begriff können 3555 und den deutschen 39 Treffer angesehen werden. Damit Nutzer in *io-port.net* bei hohen Trefferzahlen relevante Inhalte finden, können durch die Integration von semantischen Werkzeugen verfeinerte Begriffe der Suche angezeigt werden: bei der Suche nach dem Begriff „design patterns“ werden noch Begriffe wie „design patterns learning“, „design patterns problem“, „design patterns applications“, „system design patterns“ oder „software design patterns“ vorgeschlagen.

Sowohl Google scholar als auch *io-port.net* befinden sich momentan in der Test- bzw. Anfangsphase und müssen sich gegenüber verschiedenen Herausforderungen (u.a. Dublettenerkennung, Extraktion der Autorennamen, Integration von Volltexten) positionieren. Als Fazit des Kurztests lässt sich feststellen, dass der Betastatus Google scholar anzumerken ist und die Ergebnisse mancher Suchbegriffe lückenhaft und nicht aktuell sind. Die Recherche in einem Informationsportal wie *io-port.net* kann Google scholar auf keinen Fall ersetzen.

1.3 Erfassung und Bearbeitung bibliographischer Daten

Die klassische Erfassung bibliographischer Daten erfordert viel Aufwand und Zeit. Mit geeigneten Methoden können diese Daten aus Inhaltsverzeichnissen im HTML-Format halbautomatisch extrahiert und aufbereitet werden. Werkzeuge unterstützen die Erfassung und Korrektur der gewonnenen Daten und ermöglichen so eine schnelle und korrekte Erfassung bibliographischer Daten.

Die Anwendung **DataGen** extrahiert aus Inhaltsverzeichnissen, die im HTML-Format vorliegen, bibliographische Daten mit Hilfe von Wrappern. Vordefinierte Masken ergänzen die gewonnen Daten um weitere Inhalte. DataGen ist in der Lage, weitere Daten aus dem Internet nachzuladen, zu verarbeiten und mit den bisher gewonnenen Daten zu vereinigen. Mit nachgeschalteten Skripten erfolgt eine automatische Aufbereitung und

Umwandlung der Daten (Zeichen-Kodierung in BibTeX, Normalisierung der Autorennamen) ins LEABib-eigene SRC-Format.

Im Rahmen des Projektes FIS-I wurde der ***io-port.net*-Editor** zur Erfassung und Bearbeitung bibliographischer Daten entwickelt. Durch die Unterstützung verschiedener Formate kann der *io-port.net*-Editor flexibel eingesetzt werden. Er vereinfacht die einheitliche Erfassung und Korrektur der Daten durch vordefinierte Wertelisten, einer Autorendatenbank und weiterer Hilfsmittel. Korrekte LaTeX-Formeln können auf einfache Weise mit dem integrierten LaTeX-Formel-Editor eingefügt werden.

Datenkorrektur - Ein Java-Programm prüft die erfassten Daten mit Hilfe verschiedener Verfahren auf Richtigkeit durch Abgleichen mit vordefinierten Wertelisten, Test auf richtigen Zeichensatz (ASCII) und gültige LaTeX-Formatierungen und Klammerungen. Mittels des GUI können die gefundenen Fehler schnell korrigiert werden.

1.4 Heuristische Verfahren für die semantische Anreicherung unstrukturierter bibliographischer Daten

Im Projekt FIS-I wird der Zugriff auf Informatik-Literatur zentralisiert. Der Projektpartner Universität Karlsruhe – Collection of Computer Science Bibliographies (CCSB) ist einer der Datenlieferanten für das Projekt. Die bibliographischen Daten der CCSB sind sehr unterschiedlicher Qualität. Sie sind oft mehrmals konvertiert, aus verschiedenen Quellen maschinell extrahiert und nicht selten in einem Dateiformat, das die Daten nicht vollständig semantisch beschreibt. In dem Vortrag zeigen wir Beispiele von realen, bei uns aufgetretenen Problemen und deren Lösung, welche die semantische Qualität und damit die Nutzbarkeit der bei uns gesammelten bibliographischen Daten wesentlich verbessert.

Die Verfahren kann man sich auf drei Ebenen vorstellen. Die Daten müssen konvertiert, extrahiert und bereinigt werden. Der größte Teil der Konvertierung geht zur Zeit von XML (mit DublinCore Namensraum [WK98]) und (X)HTML Dateiformaten aus. Der Datenbestand, den wir aus XML/DublinCore in CCSB integrieren, beträgt ca. 780 000 Einträge – 35% der Gesamtanzahl (Stand: Juni 2005). Das DublinCore Schema ist aus Prinzip sehr allgemein und ungenau. Die existierenden Vorschriften für Qualified DublinCore [DCMI] und Richtlinien für die Kodierung der bibliographischen Einträge in DublinCore [Ap05] werden leider in der Praxis noch nicht verwendet und fast alle Einträge kommen in Unqualified DublinCore [DCES]. Das bedeutet, dass nicht festgestellt werden kann, was aus semantischer Sicht die benannten Knoten des XML-Baums eigentlich beinhalten. Die Abbildung von Knotennamen in einen anderen Bibliographiespezifischen Namensraum (z.B. BibTeX) macht Datenextrahierung notwendig. Die einzigen Felder, die potenziell direkt umgeschrieben werden könnten, sind DC.TITLE und DC.CREATOR. Fast alle anderen Informationen sind sehr oft zu DC.DESCRIPTION umgeleitet. Es gibt aber noch Felder DC.IDENTIFIER, DC.SOURCE, DC.LINK und DC.RELATION, die, falls vorhanden, normalerweise die notwendigen Informationen bieten, um den Eintrag brauchbar zu machen. Die obengenannten Felder werden inkonsistent benutzt. Das erfordert die Erkennung von Inhalten und Extrahierung aller relevan-

ter Informationen wie z.B. Zeitschriftenname, Band (Vol.), Ausgabe (No.), Seitenzahl, ISSN, ISBN, Verlag, Herausgeber. Auch das Datum muss aus vielen verschiedenen Schreibarten in „Jahr-Monat-Tag“ umgewandelt werden. Der Typ der Publikation kann manchmal nur aus der URL erraten werden. Man kann annehmen, dass die extrahierten Daten vollständig semantisch korrekt und in dem Extrahierungsprozess bereits bereinigt worden sind. Für die übrigen Felder, wie Autoren, Titel und Zusammenfassung muss dies erst durchgeführt werden. Das bedeutet Entfernung aller überflüssigen Textauszeichnungen (HTML und LaTeX Strukturen), Korrektur und Umschreibung von Sonderzeichen (UTF-8, HTML Entities, TeX Darstellungsart, falsch genutzter Mathematikmodus von TeX). Die Autorennamen, die als zusammengesetztes Textelement vorliegen, müssen getrennt werden und entschieden, welcher Teil der Teilkette der Vorname ist, welcher der Nachname und was nicht zum Namen gehört und entfernt werden soll.

Die obengenannten Probleme sind nur ein Teil von in der CCSB aufgetretenen Aufgaben, die im Projekt größtenteils gelöst wurden. Mehrere zusammengesetzte heuristische Regeln, mit regulären Ausdrücken als Werkzeug, haben die Datenextrahierung und -bereinigung möglich gemacht.

2 Projekt Semantische Methoden und Werkzeuge für Informationsportale (SemIPort)

2.1 Ontologie-basiertes Web Mining zum Aufbau großer Informationsportale

Die Erkennung und Extraktion relevanter Daten im Internet wird zunehmend durch den rapiden Zuwachs an Dokumenten erschwert. Bestehende Ansätze, denen aktuelle Suchmaschinen in der Regel folgen, entgegnen den anfallenden Datenmengen mit immer neuer Rechenleistung. Diese Vorgehensweise wird sich jedoch nicht beliebig fortsetzen lassen. In dem SemIPort Projekt wurde der fokussierte Web-Crawler METIS (http://ontoware.org/projects/metis) zur Identifikation und Extraktion kontextrelevanter Informationen aus dem Internet entwickelt, welcher Hintergrundwissen in Form von Ontologien verwendet.

Grundsätzlich wird zwischen mehreren Arten von Ontologien unterschieden. Zum einen wird eine **Web-Ontologie** modelliert. Diese beschreibt die Struktur und Eigenschaften von Dokumenten im Internet, sowie deren Verknüpfungen mittels sog. *Hyperlinks.* Sie repräsentiert außerdem *Hosts,* auf denen Internet-Dokumente gespeichert werden. In der **Domänen-Ontologie** wird die eigentliche Domäne beschrieben. Das dort gespeicherte Wissen stellt letztendlich das Ziel der fokussierten Suche dar. Zum Aufbau eines Informationsportals für wissenschaftliche Publikationen aus der Informatik beschreibt die Domänen-Ontologie z. B. Fachrichtungen, Eigenschaften von Publikationen und beschreibt u.a. Personen und Forscher.

Im Gegensatz zu Informationsextraktionsmechanismen, die eine Bewertung von Res-

sourcen erst nach der eigentlichen Extraktion zulassen[1], zielt der von uns entwickelte Ansatz auf eine Bewertung von Ressourcen während der eigentlichen Suche ab. Dies ermöglicht eine effektive und effiziente Nutzung vorhandener Kapazitäten[2]. Die fokussierte Suche nach Ressourcen basiert auf der Bestimmung der Relevanz einer Ressource zu einer bestimmten Domäne oder Teildomäne. Generell wird als Suchstrategie die Verfolgung von Relationen zu Ressourcen mit möglichst hoher Relevanz verwendet (fokussiertes Crawlen). Die Berechnung der Relevanz setzt sich zusammen aus der **inhaltlichen Analyse** des Dokumenteninhalts und einer **Bewertung der Verlinkungsstruktur** der Ressource. Das Ergebnis ist ein numerischer Wert, welcher zur Erzeugung einer sortierten Menge an Ressourcen verwendet wird. Diese Menge dient zur weiteren Suche, wobei Ressourcen mit höherer Relevanz vorrangig verarbeitet werden.

Zusammenfassend werden Ontologien zur Modellierung der Umgebungswelt und zur Modellierung der eigentlichen Anwendung (Domäne) eingesetzt. Die Domänen-Ontologie beschreibt dabei Konzepte zu denen weitere Instanzen identifiziert werden sollen. Die Suche nach diesen Instanzen wird mittels einer semantischen Bewertung durchgeführt.

2.2 Personalisierte Benutzerinteraktion mit wissenschaftlichen Informationsportalen

Die Interaktion mit Information besteht nicht nur aus Information Retrieval, sondern auch aus dem Organisieren und Verstehen der gesammelten Daten. Gerade bei wissenschaftlichen Informationssystemen ist eine entsprechende Werkzeugunterstützung wertvoll. Wird solch ein Werkzeug in ein Informationsportal integriert, bietet es zudem die Möglichkeit, dem Benutzer bei der Informationssuche gezielter zu unterstützen, da mehr Indizien über dessen Informationsbedarf im System vorhanden sind, beispielsweise der Teil der persönlichen Dokumentsammlung, mit der sich der Benutzer gerade beschäftigt. Diese zusätzlichen Daten können auch zur Unterstützung anderer Benutzer verwendet werden; hier bieten sich kollaborative Empfehlungssysteme oder sogar einfach nur das Verfügbarmachen der von anderen Benutzern definierten Annotationen, wie eine Themenzuordnung zu Dokumenten, an.

Im Rahmen von SemIPort wurde ein persönliches Dokumentmanagementwerkzeug und ein zugehöriges Empfehlungssystem entwickelt [SJ04,Sc04], welche in Informationsportale integriert werden können. Die Kernidee dabei ist, dass die Benutzer persönliche Wissensbasen auf Grundlage einer gemeinsamen Ontologie aufbauen und diese Wissensbasen vom Empfehlungssystem gesammelt und zur Generierung von Empfehlungen genutzt werden. Die gemeinsame Ontologie ermöglicht hier eine einfache Integration der Daten aus den verschiedenen Quellen, sowie die Herstellung des Bezugs zwischen den Benutzerdaten und dem Informationsraum des Portals auf einer abstrakteren als der reinen Inhaltsebene. Für Empfehlungen kann so z.B. das Wissen genutzt werden, dass

---

[1] An dieser Stelle sei auf Methoden wie HITS-Algorithmus und PageRank-Algorithmus verwiesen.
[2] Vorwiegend Reduktion des Bedarfs an Speicherplatz, Bandbreite und Rechenzeit.

sich der Benutzer gerade mit einem bestimmten Themengebiet beschäftigt, und das eine bestimmte Menge von Dokumenten im Portal unter dieses Themengebiet fallen.

Bevor Benutzerdaten genutzt werden können, müssen diese vorhanden sein. Dazu muss der Dokumentmanager von den Benutzern akzeptiert und dessen Annotationsfunktionalität genutzt werden. Bei der Entwicklung haben wir uns deshalb zum einen an Benutzeranforderungen orientiert, die wir anhand von Interviews und Literaturrecherchen identifiziert haben, und zum anderen gerade in Bezug auf die Annotationen die Richtlinie verfolgt, dass jegliche vom Benutzer investierte Arbeit einen direkten, unmittelbaren Nutzen für diesen haben soll. Einige Merkmale des resultierenden Systems sind die Möglichkeit zur räumlichen Anordnung von Dokumenten (siehe Abbildung1), die Suche anhand von Annotationen, die kontextspezifische Relevanzbewertung , und das Durchstöbern der Dokumentsammlung basierend auf zwischen Dokumenten definierten Beziehungen.

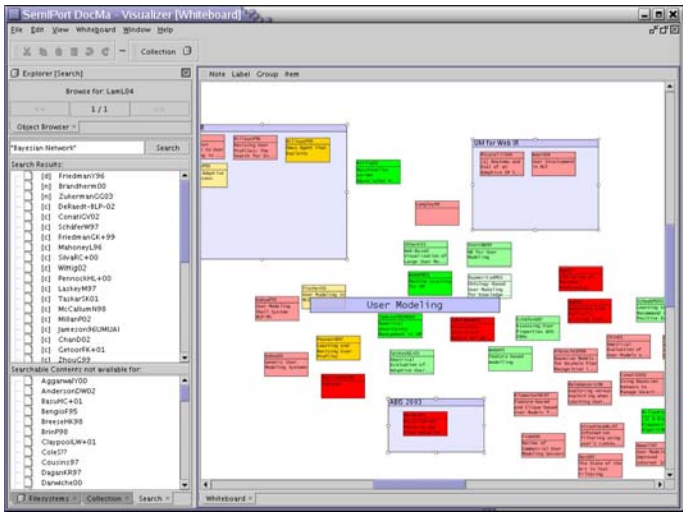


Abbildung 1: SemIPort Document Manager mit geöffnetem Whiteboard (rechts), auf dem Dokumente durch Karteikarten repräsentiert räumlich angeordnet werden können.

2.3 Finden und Browsen von bibliographischen Daten mit SWQL

Das Aufspüren von relevanten Informationen ist eine komplexe Aufgabe. Hierzu ist das Finden relevanter Publikationen wichtiger Konferenzen oder (Fach-)Journalen zur Befriedigung eines Informationsbedürfnisses wichtiger denn je. Der DBLP-Bibliographie-Server mit seinen mehr als 650.000 Publikationen von über 400.000 Autoren ist ein Service, der von tausenden Wissenschaftlern auf der ganzen Welt genutzt wird und eine fundamentale Hilfestellung für das Finden von Publikationen, Konferenzen/Journalen oder anderer Wissenschaftlern in ähnlichen Gebieten bietet [Ley02].

Im Rahmen des SemIPort-Projektes wurde für die Bereiche Browsing und Visualisierung ein Mensch-Maschine-Interface entwickelt, welches dem Benutzer bei seiner Suche nach den gewünschten Informationen unterstützt und leitet (siehe Abbildung 2). Der DBL-Browser ermöglicht durch die vollkommene Verlinkung der bibliographischen Daten ein Navigieren durch den gesamten Datenbestand und ist ähnlich einfach zu bedienen wie bekannte Internet-Browser. Durch die IR-Schnittstelle sind auch komplexere Suchen möglich. Der DBL-Editor bietet die Möglichkeit weitere Datensätze auf hohem Qualitätsniveau manuell zu erfassen und dabei „on-the-fly“ nach bereits bekannten Daten zu suchen, um Duplikate schon bei der Eingabe zu vermeiden [KLRRWW04].

Neben den Daten des DBLP-Servers, welche in den Hauptspeicher des Computers eines Benutzers geladen werden können, besteht die Möglichkeit über das Internet die Daten des ***io-port.net*** (siehe Abschnitt 1) mit Hilfe von SWQL (siehe Abschnitt 10) zu browsen und nach Autoren sowie nach Wörtern im Titel einer Publikation zu suchen.

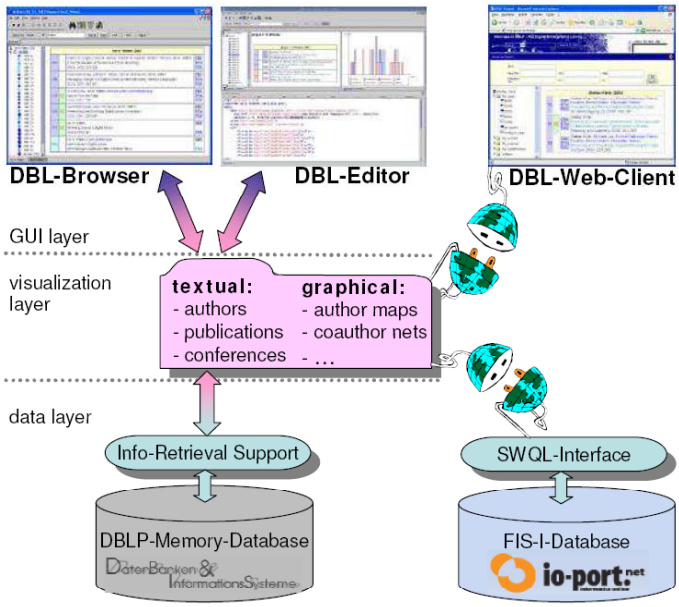


Abbildung 2: Schematischer Aufbau der DBL-Komponenten

2.4 SWQL – Eine OWL-basierte Abfragesprache zur Datenintegration

Zahlreiche Anwendungen besonders aus dem „Semantic Web“ Umfeld basieren auf einer Ontologie, welche die Domäne der Anwendung modelliert. Die Mehrzahl der existierenden Datenquellen beziehen sich jedoch nicht direkt auf eine Ontologie, sondern sind z.B. als Relationen oder in XML modelliert. Diese Datenmodelle haben ihre eigenen Schema- und Abfragesprachen, wie ein relationales Schema und SQL bzw. XML Schema und XQuery. Um solche Datenquellen über eine Ontologie abzufragen, wird

eine Abfragesprache benötigt, die in der Lage ist, Abfragen an Ontologie-basierte Daten zu stellen, d.h., Instanzen eines Konzepts auszuwählen und aufgrund von Attributwerten zu filtern, zwischen Instanzen über Beziehungen zu navigieren sowie Attributwerte auszuwählen. Hierzu wurde die Abfragesprache Semantic Web Query Language (SWQL) entwickelt [LSF04]. Sie basiert auf der Syntax der derzeit vom W3C standardisierten XML-Abfragesprache XQuery. Allerdings wurde die Semantik, das Typsystem und das Datenmodell von XQuery vollständig ersetzt.

SWQL nutzt die Web Ontologie Sprache OWL als ihr Typsystem, d.h. eine Abfrage wird mit Hilfe des Vokabulars einer Ontologie formuliert und kann gegen diese auch geprüft werden. Das zugehörige Datenmodell für SWQL ist ein Graph-basiertes Datenmodell, das ähnlich dem RDF-Datenmodell ist. Außerdem wurde XPath, die Sprache in XQuery zur Selektion einzelner Knoten im XQuery-Datenmodell, durch SWQLPath, einer Sprache zur Selektionen einzelner Knoten in einem SWQL-Graphen, ersetzt. Im Gegensatz zu existierenden RDF-Abfragesprachen, die von einem ähnlichen Datenmodell und Typsystem ausgehen, ist SWQL eine vollständige Abfragesprache, die neben Selektion und Navigation auch die Konstruktion neuer Datenstrukturen erlaubt. Dies erlaubt es, Datenstrukturen zu ändern, anzupassen und zu übersetzen, was Voraussetzung zum Einsatz als Datenintegrationssprache ist.

SWQL ist außerdem eine streng getypte Sprache, was eine statische Prüfung von SWQL-Abfragen auf Fehler erlaubt. So kann beispielsweise die Kompatibilität der Parameter von Funktionsaufrufen oder der Typen bei der Konstruktion von Konzepten und Beziehungen bereits vor der Ausführung der Abfrage geprüft werden. Dies ist besonders dann nützlich, wenn Abfragen nicht direkt ausprobiert werden können, z.B. im Fall von Funktionsbibliotheken oder sehr lang dauernden Abfragen.

Um SWQL-Abfragen gegen verschiedene reale Datenquellen wie XML-, RDF- oder Relationale Datenbanken stellen zu können erwies sich eine Schichtenarchitekur als sehr flexibel und performant, dabei wird für jeden Datenquellentyp ein Übersetzungsmodul für das Datenmodell sowie eine Funktionsbibliothek zum Datenbankzugriff geschrieben. Diese Architektur erlaubte es, innerhalb kürzester Zeit SWQL-Abfragen an eine XML- und eine RDF-Datenbank sowie gegen die von FIS-I eingesetzte Lucene-Datenstruktur auszuführen.

Für die Integration konkreter Datenquellen in ein Ontologie-basiertes Portal müssen u.a. die Schemata der einzelnen Datenquellen integriert werden, dazu werden SWQL-Funktionsbibliotheken genutzt. Diese Funktionsbibliotheken bestehen aus feingranularen Funktionen zur Übersetzung einzelner Konzepte aus dem lokalen Schema in die globale Ontologie. Die vollständigen Funktionsbibliotheken können mit Hilfen der statischen Typüberprüfung garantieren, nur korrekte Instanzen der Zielontologie zu erzeugen. Als lokale Schemata können neben OWL-Ontologien auch XML-Schemata dienen [LF04]. Diese Schemaintegrationsmethode lässt sich sowohl für das Füllen eines Datawarehouses, als auch für eine Mediator-Wrapper-**A**rchitektur einsetzen.

2.5 Konzeptionelle Anfrage-Verfeinerung

Das Hauptproblem klassischer Methoden für die Anfrage-Verfeinerung besteht darin, dass Anfragen als „bag of words“ betracht werden, so dass die erzeugten Verfeinerungen keine semantische Verbindung mit der Anfrage haben (d.h. der Verfeinerungs-Prozess wird auf syntaktischer Ebene durchgeführt).

Im Rahmen des Projekts wurde eine neue Methode, die sogenannte Konzeptionelle Anfrage-Verfeinerung, entwickelt, die mit Hilfe von Ontologien versucht, die semantische Bedeutung einer Anfrage nachzuvollziehen. Die Methode basiert auf dem Librarian Agent Query Refinement Process. Das ist ein drei-stufiger Prozess bestehend aus: 1. der Messung der Mehrdeutigkeiten einer Anfrage, 2. der Empfehlung der Verfeinerungen der Anfrage und 3. dem Ranking von Verfeinerungen.

Die Aufgabe des Verfeinerungs-Prozesses ist es, die Verfeinerungen, die zu der semantischen Bedeutung der Anfrage passen, zu generieren. D.h. eine Verfeinerung sollte eine konkrete semantische Rolle bezüglich der Anfrage spielen. Nach der Informationstheorie gibt es zwei generelle Relationen, welche die Bedeutung eines Begriffes erläutern können: Specialisation und Modification. Diese semantischen Relationen zwischen den Begriffen der Anfrage und den Begriffen des Textes zu produzieren, ist das Ziel der entwickelten Methode. Auf diese Weise sind die Verfeinerungen, die diese Methode generiert, nicht nur zusätzliche Begriffe in einer Anfrage, sondern semantische Erweiterungen der Anfrage. So ist z.B. für die Anfrage „code“ der Spezialisierungs-Begriff „binary“ (siehe oben) eine semantische Erweiterung: „code + binary“ bedeutet „binary code“ (binärer Code) und nicht eine beliebige Relation zwischen „code“ and „binary“. Infolgedessen produziert unsere Methode Verfeinerungen, die relevant für die Benutzer-Präferenzen sind.

Mit Hilfe von Natural Language Processing (NLP) kann man die Relationen Specialisation und Modification in einem Text entdecken. Weil NLP-Methoden oft sehr zeitintensiv sind, wurde eine neue Methode entwickelt, die flaches Tagging mit logischem Inferenzing kombiniert. Auf diese Weise wird von der linguistischen Verarbeitung (d.h. wir benutzen linguistische Strukturen, wie z.B. noun phrases) und vom logischen Inferenzing (d.h. mit Hilfe von Regeln produzieren wir neue Informationen, die für die Verfeinerungen wichtig sind) profitiert.

Die Hauptvorteile für einen Benutzer sind:

a) bessere Qualität des Suchprozesses: kürzere Suchzeit, bessere Exploration des Suchraums und bessere Berücksichtigung der Präferenzen des Benutzers sowie
b) bessere Qualität der Ergebnisse: die Ergebnisse sind geclustert und relevanter.

3 Weitere Projekte aus dem Umfeld

3.1 DAFFODIL - Nutzerorientiertes Zugangssystem für heterogene Digitale Bibliotheken

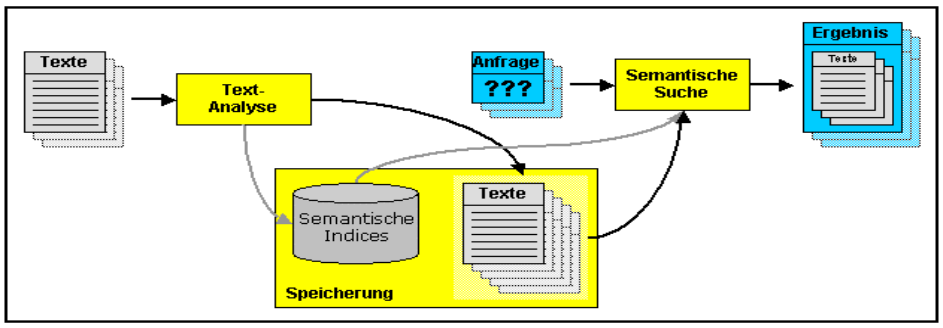
Daffodil ist ein virtuelles digitales Bibliotheks–System, das die strategische Unterstützung des Suchenden während des gesamten Information–Retrieval–Prozesses zum Ziel hat. Für einen Benutzer äußert sich diese in Form von höheren Suchfunktionen, sogenannten Strategemen, die Funktionalitäten über das allgemein übliche Maß hinaus bereitstellen. Der aktuelle Prototyp kann über www.daffodil.de genutzt werden. Dazu wird via Java Webstart Technologie die grafische, intuitiv bedienbare Benutzeroberfläche gestartet. Zur Zeit existieren eine Vielzahl von Diensten zum Suchen und Browsen in Digitalen Bibliotheken, sowie Dienste zur Verwaltung von gefundenen Objekten. An zentraler Stelle steht das Suchwerkzeug, das über eine formularbasierte Schnittstelle eine einheitliche Anfrageformulierung über aktuell 15 verteilte Datenquellen bereitstellt. Dabei wird die Anfrageformulierung durch proaktive Dienste unterstützt, um schon vor der eigentlichen Anfrageprozessierung Fehler zu erkennen, bzw. dem Benutzer Vorschläge zur aktuellen Anfrage zu unterbreiten. Das Ergebnis einer Anfrage wird anschließend zusammengefasst und einheitlich dargestellt.

Weitere Dienste zum Suchen und Browsen sind der Referenz- und Zitationsbrowser, der Journal- und Konferenzbrowser, das Coautorenwerkzeug, neben einer Reihe von unterstützenden Werkzeugen wie einem Thesaurus und dem Klassifikationsbrowser. Zur nachhaltigen Speicherung und Verwaltung von gefundener Information von Benutzern und Gruppen dient die persönliche Handbibliothek. Dort können alle bereitgestellten Digitalen Bibliotheksobjekte, darunter Metadaten, Volltexte, Autoren, Begriffe, Webseiten, etc. in strukturierter Weise abgelegt werden. Das Konzept der Awareness ermöglicht Gruppenarbeit und Langzeitverfolgung von Anfragen. Aktuell wird Daffodil im Rahmen des EU–Projekts DELOS zur Evaluation von Digitalen Bibliotheken als Basisframework eingesetzt werden, um grafische Werkzeuge, Dienste oder ganze Digitale Bibliotheken mit- und untereinander zu vergleichen. Zudem werden zur Zeit weitere Werkzeuge zur Kollaboration von Suchenden untereinander oder mit Experten (z.B. Bibliothekaren) in das System integriert, um den Benutzern effektivere Möglichkeiten zu bieten, ein bestehendes Informationsbedürfnis zu bearbeiten. Das Projekt Daffodil wurde durch die (DFG) innerhalb des Schwerpunktprogramms “Distributed Processing and Delivery of Digital Documents”(V 3D2) gefördert.

3.2 Semantik-basiertes Retrieval Biosystematischer Dokumente

Im Projekt "Collaborative Research: Development of New Digital Library Applications in the Context of a basic Ontology for Biosystematics Information Using the Literature of Entomology (Ants)" werden derzeit biosystematische Dokumente in großem Umfang digitalisiert. Um die sukzessive entstehende Kollektion sinnvoll nutzen zu können, müssen die Texte in XML annotiert und für den Zugriff über ein Retrieval-System bereitgestellt werden. Ohne die Annotationen würde sich das Retrieval als extrem schwierig

gestalten, da sich die Dokumente in Form, Sprache und Wortwahl sehr ähnlich sind, bestehen sie doch zum größten Teil aus anatomischen Beschreibungen von Insekten, die den einzelnen Körpermerkmalen Beschreibungen zuordnen, aus ihren Taxonomischen Namen und ihren Fundorten. Auf dem unmarkierten Text wären klassische Techniken für Volltext-Retrieval (etwa Boolesches oder Vectorspace-Retrieval) gerade für die Suche nach Insekten mit bestimmten Körpermerkmalen (bzw. Dokumenten, die Insekten mit bestimmten Merkmalen beschreiben) nicht geeignet, da sich hier die Zusammengehörigkeit von Merkmal und Beschreibung nicht hinreichend abbilden lässt. Zudem würden die Taxonomischen Namen aufgrund ihrer sehr geringen Dokumentfrequenz (im Idealfall ist diese eins, da die Namen ein Insekt eindeutig identifizieren) aus jedem klassischen Volltextindex herausfallen, wodurch zum Auffinden der Beschreibung einer ganz bestimmten Spezies eine aufwändige Volltextsuche ohne die Unterstützung durch einen Index ausgeführt werden müsste.



Auf der anderen Seite ist die manuelle Annotation sehr zeitaufwändig und daher kostenintensiv. Daher wird an der Universität Karlsruhe ein System entwickelt, das sowohl ein automatisches Markup leistet, als auch die Dokumente speichert und über eine Retrieval-Engine zugreifbar macht, wobei die markierten Teile zur Indizierung genutzt werden. Die Annotierung stützt sich auf mehrere dynamisch aktivier- und entfernbare Analysekomponenten, die unter anderem Pattern-Matching und NLP-Techniken einsetzen, um semantische Einheiten in den Dokumenten zu identifizieren und zu markieren. Derzeit umfasst dies Körpermerkmale und die ihnen zugeordnete Beschreibung, Ortsangaben und die aus mehreren Wörtern bestehenden Taxonomischen Namen.

Ausblick

Mit *io-port.net* steht am Projektende ein Kompetenz- und Dienstleistungsnetz für die Informatik zur Verfügung, das weltweit publiziertes Informatikwissen mit ca. 2 Millionen Einträgen strukturiert und standardisiert nachweist, Zugang zu Volltexten direkt (u.a. LNI-Reihe der GI) oder per Dokumentlieferdienste anbietet sowie weitere thematisch fokussierte Angebote der Informatik (Fachgruppen, Kompetenznetze) und personalisierte Dienste (Erfassung persönlicher Publikationslisten) integriert. *io-port.net* verwendet dabei eine tiefe Datenintegration, die über eine für viele Portale typische lose Kopplung verschiedener Anbieter weit hinausgeht und ermöglicht den Nutzern einheitli-

che Anfragen über einen Datenbestand in definierter Qualität. Semantische Werkzeuge unterstützen dabei sowohl die Nutzer bei Suche und Navigation in den Inhalten als auch den Portalbetreiber bei Aufbau und Betrieb des Portals.

Das Angebot des Informationsportals *io-port.net* ist derzeit fokussiert auf Endnutzer in Hochschulen und Forschungseinrichtungen. Um diese Nutzer langfristig zu halten und in das Portal aktiv einzubeziehen sowie neue Nutzer bzw. Nutzergruppen ansprechen zu können, ist die Realisierung weiterer, innovativer Mehrwertdienste notwendig. Im Vordergrund sollten dabei Mehrwertdienste stehen, die den Nutzer bei allen Schritten der Informationsbeschaffung und vor allem auch der Informationsorganisation am Arbeitsplatz effizient unterstützen: v.a. bei Recherche, Analyse und Bewertung von Ergebnismengen, beim direkten Zugriff auf relevante Quellen, bei der individuellen Informationsorganisation sowie der Kollaboration mit Fachkollegen.

Literaturverzeichnis

[Ap05] Apps, A: Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata, Dublin Core Metadata Initiative, June 2005, http://dublincore.org/documents/dc-citation-guidelines/

[Cu05] Cusomano, M.A: Google: what it is and what it is not. Communication of the ACM 48(2): 15-17 (2005)

[DCES] Dublin Core Metadata Element Set, Version 1.1: Reference Description, Dublin Core Metadata Initiative, December 2004, http://dublincore.org/documents/dces/

[DCMI] Dublin Core Metadata Terms, Dublin Core Metadata Initiative, June 2005, http://www.dublincore.org/documents/dcmi-terms/

[KLRRWW04] Klink, S.; Ley, M.; Rabbidge, E.; Reuther, P.; Walter, B. Weber, A.: Browsing and Visualizing Digital Bibliographic Data. Symposium on Visualization, Konstanz, Germany, May 19-21, Eurographics Association, 2004; S. 237-242.

[Ko04] Koschmider, A. et al.: Entwicklung eines Informationsportals für die Informatik, In P. Dadam; M. Reichert, INFORMATIK 2004 - Informatik verbindet, GI-Jahrestagung, Ulm, volume 50 of LNI, pp. 208-213. September 2004.

[Ley02] Ley, M.: The DBLP-Computer Science Bibliography: Evolution Research Issues, Perspectives. In 9th International Symposium SPIRE, 2002; S. 1-10.

[LF04] Lehti, P.; Fankhauser, P.: XML Data Integration with OWL: Experiences and Challenges. In Proceedings of the 2004 Symposium on Applications and the Internet (SAINT 2004); 26-30 January 2004, Tokyo, Japan

[LSF04] Lehti, P.; von Stackelberg, S., Fankhauser, P.: The Semantic Web Query Language SWQL; 2004, http://www.ipsi.fraunhofer.de/oasys/projects/semiport/SWQL-WD0704.doc

[SJ04] Schwarzkopf, E.; Jameson, A.: Personalized Support for Interaction with Scientific Information Portals. AMR 2003, LNCS3094, pp. 58-71, 2004

[Sc04] Schwarzkopf, E.: Enhancing the Interaction with Information Portals. Proceedings of IUI 2004, pp. 322-324, 2004

[WK98] Weibel, S.; Kunze, J.; Lagoze C.; Wolf, M.: Dublin Core Metadata for Resource Discovery, RFC 2413, September 1998, http://www.ietf.org/rfc/rfc2413.txt