Brillante Erweiterung des Horizonts

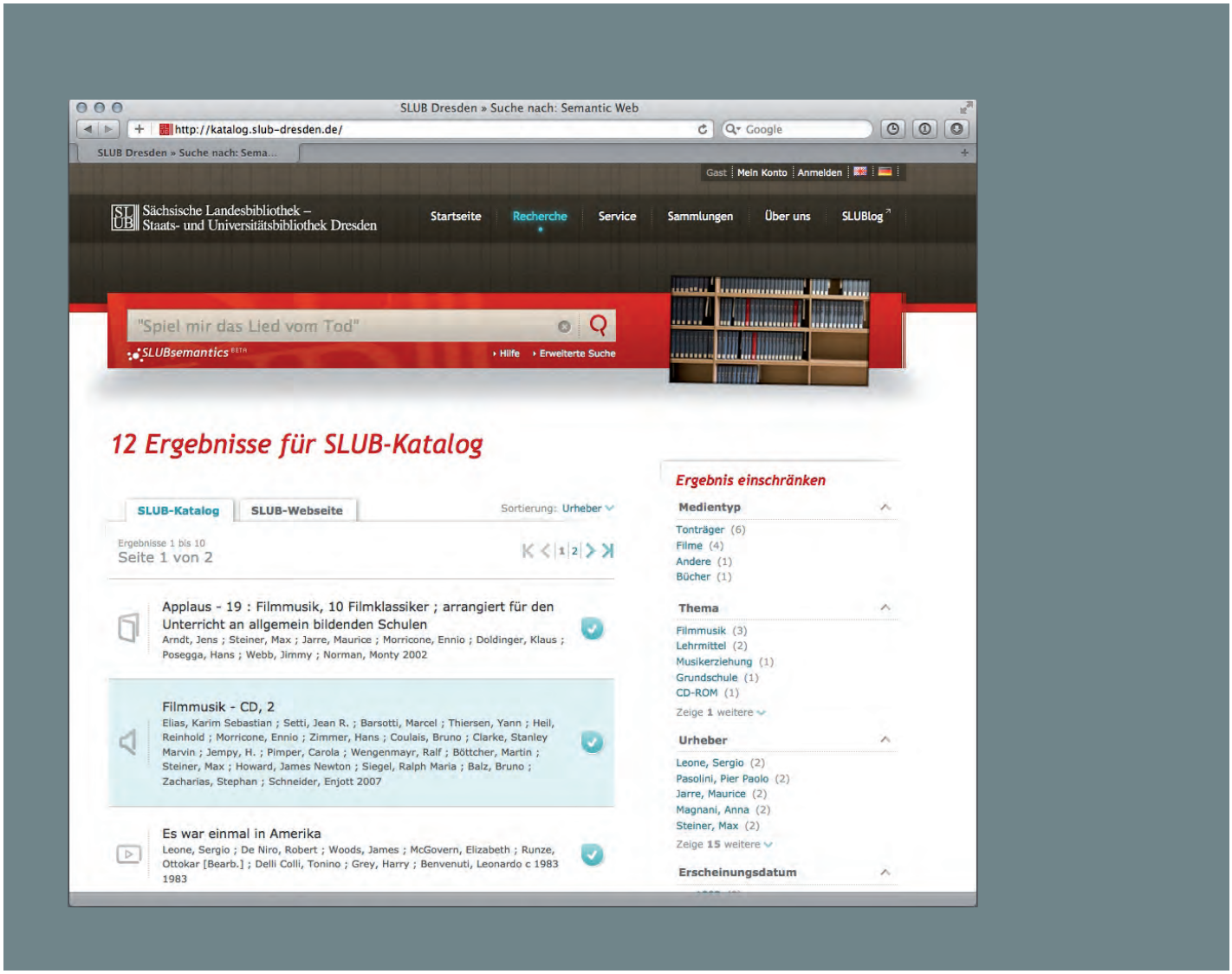
Eine multilinguale semantische Suche für den SLUB-Katalog

VON ACHIM BONTE, ROBERT GLASS UND JENS MITTELBACH

Mit der Einführung ihres neuen SLUB-Katalogs auf der Basis der Discovery-Software Primo der Firma Ex Libris hat die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB) im Dezember 2010 die zunehmend unzulängliche Welt der traditionellen elektronischen Bibliothekskataloge hinter sich gelassen. Innerhalb von neun Monaten entstand ein übergreifendes Katalogfrontend, das auf älteren Systemen aufsetzt (zum Zweck des Data Harvesting oder auch zur Inanspruchnahme der lokalen Benutzerverwaltung), zugleich aber davon weitgehend unabhängig ist. Eine besondere Herausforderung bedeutete der Anspruch, Primo nicht „out of the box", das heißt als gesichtsloses Fertigprodukt einzusetzen, sondern als Herzstück des gesamten Informationsangebots individuell zu gestalten und weitgehend in die allgemeinen Webseiten zu integrieren. Auch die Ausleihbenutzerverwaltung sollte möglichst bruchlos in das Gesamtkonzept finden. Der SLUB-Katalog bietet heute unter einer attraktiven Benutzeroberfläche ein sehr gutes Trefferranking, Rechtschreibkorrektur, vielfältiges Drilldown, flexible Sortieralgorithmen und weitere, von Suchmaschinen gewohnte Funktionen. Über den Datenbestand des alten LIBERO-OPAC erheblich hinausreichend, integriert er rund 65 Millionen Bestandsnachweise zu gedruckten wie elektronischen Medien, zu Büchern, Zeitschriften und Aufsätzen, Tonträgern und Filmen, Noten und grafischen Dokumenten (Fotos, Karten, Zeichnungen). Zugleich organisiert er mit einem bequemen Authentifizierungs- und Autorisierungsverfahren (Shibboleth) den Zugriff auf alle online verfügbaren Ressourcen.

Seit seiner Einführung wird der neue SLUB-Katalog funktional wie inhaltlich kontinuierlich weiterentwickelt. Letzte technische Verbesserungen betrafen zum Beispiel die Literaturverwaltungs- und Exportfunktionen oder die individuelle Einschränkung von Suchergebnissen. Im Datenbestand wuchs der Katalog unter anderem durch die Berücksichtigung der bislang separaten Aufsatztiteldaten aus der „Sächsischen Bibliografie" oder die Übernahme von über 200.000 E-Book-Titeldaten nach dem Konzept der Patron-driven acquisition (kundengesteuerter Erwerb von Nutzungslizenzen).

Die wachsenden Möglichkeiten des sogenannten Semantic Web, der Bereicherung des World Wide Web um die Dimension der Bedeutung von Informationen, inspirierten im Sommer 2010 dazu, mit Hilfe eines speziellen Werkzeugs drei besonders



anspruchsvolle Aufgaben im SLUB-Katalog anzupacken: Erstens die unterschiedliche Erschließungstiefe der integrierten Dokumente. Während zum Beispiel lizenzierte E-Journals im Volltext durchsucht werden können, stehen für den gedruckten Kernbestand der Bibliothek nur die klassischen Metadaten (bibliographische Beschreibung, einige Sacherschließungsdaten) zur Verfügung. Zweitens die in Bibliotheken fast immer unzureichende thematische Suche, die von Benutzern eigentlich bevorzugt, gegenüber titelgenauen Abfragen aber in der Regel sehr viel schlechter bedient wird. Im Ergebnis recherchieren nicht wenige Bibliotheksbenutzer zunächst in Internetsuchmaschinen oder bei Online-Buchhändlern, bevor sie mit dem ermittelten Titelmaterial „ihren" Bibliothekskatalog befragen. Beinahe als Nebeneffekt sollte drittens der Tatsache Rechnung getragen werden, dass viele Benutzer einer wissenschaftlichen Bibliothek zwar fremdsprachige Dokumente verstehen, aber mit der fremden Sprache nicht aktiv arbeiten können. Ziel war mithin, bei Eingabe von „Automatikgetriebe" zum Beispiel auch Texte zu „automatic transmission" zu finden.

SLUBsemantics

Ein dezidiert kooperativer Ansatz mit Öffnung auch gegenüber innovativen High-Tech-Unternehmen zählt seit einigen Jahren zum Markenzeichen der Dresdner Bibliotheksstrategie. Während die meist jungen Firmen ihre Ideen testen und sich erste Referenzen erarbeiten, profitiert die Bibliothek von deren besonderer Innovationskraft und Einsatzbereitschaft. Im intensiven Austausch der auf beiden Seiten beteiligten Experten entstehen auf diese Weise besonders passgenaue Produkte. Entsprechend wäre auch die hier vorgestellte semantische Suche ohne die enge Zusammenarbeit mit dem Dresdner Unternehmen „Avantgarde Labs" keines-

DAS SEMANTISCHE WEB

Die Idee des "semantischen Web" geht auf Tim Berners-Lee zurück, den Erfinder des World Wide Web. Sie wirbt für die Entwicklung von Technologien, mit denen Computer die Inhalte von Musik, Bildern und Videos besser verarbeiten können sollen. Ziel ist, nicht nur die Bedeutung einzelner Wörter zu erkennen, sondern auch die Beziehungen zu anderen Bedeutungen. Daraus ergeben sich hierarchische Klassen von Bedeutungen, die miteinander in Beziehung stehen oder sich gegenseitig ausschließen. Beispiel: Eine Bibliothek ist ein Gebäude, aber weder Kirche noch Hotel.

Solche semantischen Klassifizierungen werden den Inhalten als Metadaten hinzugefügt. Dafür sind die Web Ontology Language (OWL) sowie das Resource Description Framework (RDF) entwickelt worden, zwei maschinenlesbare Sprachen zur formalen Beschreibung von Multimedia-Inhalten. Das heutige Web besitzt bislang nur zu einem sehr geringen Teil solche semantischen Beschreibungen.

Das Problem ist daher, die etwa 98 Prozent Webinhalte anzureichern, die noch keine semantischen Beschreibungen aufweisen. SLUBsemantics bietet für den Bestand der SLUB eine softwaretechnische Lösung, die nicht zuletzt unter Beteiligung der Anwender (social tagging) sukzessive verbessert werden soll.

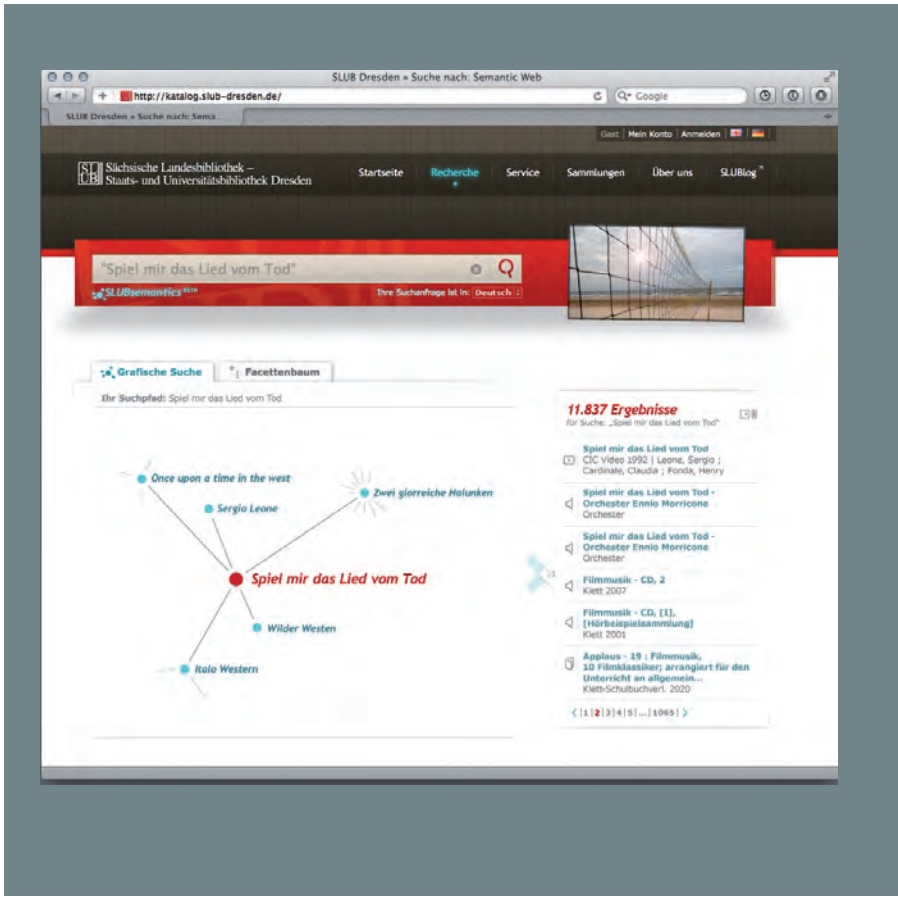
wegs möglich gewesen. Avantgarde Labs ist aus forschungsstarken Lehrstühlen der Technischen Universität Dresden hervorgegangen und beschäftigt sich seit 2008 mit Datenintegration, Data Mining und Softwareentwicklung. Die Gestaltung der Benutzeroberfläche lag in den bewährten Händen des Webdesigners und SLUB-Mitarbeiters Thomas Jung.

Im Mittelpunkt der neuen Anwendung mit dem aufschlussreichen Namen „SLUBsemantics“ steht die Idee, den Benutzer seine Anfrage völlig zwanglos formulieren zu lassen – in seiner Muttersprache, mit seinem persönlichen Wortschatz – und ihm alle semantisch relevanten Katalogeinträge in geordneter Form zurück zu liefern. Das bedeutet konkret, dass die Terme aus der Suchanfrage nicht zwingend in den Metadaten der relevanten Katalogisate vorhanden sein müssen. Die Software erkennt automatisch inhaltliche Zusammenhänge und gibt strukturierte, auf Wunsch auch grafisch aufbereitete Trefferlisten aus. Gibt der Benutzer zum Beispiel das Wort „Bank“ ein, werden ihm sowohl relevante Katalogeinträge zu (einzelnen) Kreditinstituten, der Bankenkrise oder dem Eurosystem als auch zum Sitzmöbel, einer Sandbank oder einzelnen Personen, wie dem Historienmaler Heinrich Bank, vorgeschlagen. Sucht er „Hauptstadt Belgien“, werden auch alle Treffer mit „Brüssel“ berücksichtigt. Lautet die Eingabe „Once upon a time in the west“, erhält er zugleich Treffer zum deutschen Filmtitel „Spiel mir das Lied vom Tod“, zum Regisseur Sergio Leone oder zu „Zwei glorreiche Halunken“, einem anderen Italo-Western Sergio Leones. Die einzelnen möglichen Bedeutungen und Kontexte eines Wortes werden als Konzepte bezeichnet. Alle Ressourcen des SLUB-Katalogs, die einen Bezug zu den gefundenen Konzepten haben, werden als Ergebnismenge vorgeschlagen werden, unabhängig von der Originalsprache. Der Benutzer spart damit Übersetzungsaufwand in externen Quellen und eine mehrfache Eingabe von gleichen Suchanfragen. Darüber hinaus wird er auf bislang verdeckte Inhaltsbeziehungen zu anderen Bibliotheksbeständen hingewiesen und erhält so während der Katalogsuche reichlich Gelegenheit, kontinuierlich hinzu zu lernen bzw. sich von den Funden neu anregen zu lassen. Seit dem amerikanischen Soziologen Robert Merton (1910 – 2003) kennen wir dieses Phänomen in der internationalen Wissenschaft als Serendipity. Serendipity steht für das ursprünglich gar nicht Gesuchte, das sich als unerwartete, bereichernde Entdeckung erweist. Der SLUB-Katalog wird in diesem Sinne nicht nur zu einem noch mächtigeren Auskunftsmittel, sondern buchstäblich zu einem Ort des Entdeckens und Lernens.

Die magisch wirkende Funktionsweise von SLUBsemantics ist Resultat eines so einfachen wie einleuchtenden Prinzips – des Rückgriffs auf große, sozial gepflegte und netzwerkartig angelegte Informationsstrukturen zum Zweck der automatischen Anreicherung und Verknüpfung von Katalogdaten. In unseren Bibliothekskatalogen weisen bibliografische Repräsentationen von Objekten – abgesehen von wenigen, mit viel Aufwand gepflegten Normdaten wie SWD-Schlagworten oder Personennormdaten – gewöhnlich kaum Verknüpfungen und fast gar keine semantische Relationen auf. Obgleich Linked Open Data (d.h. frei verfügbare Daten im WWW) heute ein gängiges Schlagwort ist, lässt sich dieser grundsätzliche Mangel nicht etwa durch die bloße Bereitstellung der bibliografischen Daten als sogenannte offene RDF-Tripel beheben, da die notwendigen Verknüpfungen fehlen, um wirklich semantische Qualität zu erzielen. Der hier vorgestellte

Ansatz stellt bibliografische Daten mittels Data-Mining-Technologien in einen thematischen Kontext. Der Kontext stammt aus dem größten sozial erzeugten Informationsnetzwerk, das die Menschheit bislang erschaffen hat, der freien Enzyklopädie „Wikipedia“. Die Vorteile dieses Netzwerks liegen auf der Hand: Es wird pausenlos von Millionen von Menschen mit Spezialwissen aktualisiert und erweitert, es ist in allen Sprachen verfügbar und bildet das Menschheitswissen so umfassend und so detailliert ab wie keine andere Ressource auf diesem Planeten. Allein die englischsprachige Version enthält über drei Millionen Artikel, gefolgt von der deutschen mit über einer Million. Jedes Ereignis, jede Erfindung, jeder Sachverhalt erscheint innerhalb von Minuten nach Bekanntwerden in Form eines Artikels oder Textabschnittes einschließlich Links zu verwandten Themen und Einordnungen in Kategorienbäume in der Wikipedia. Von hier aus finden die Informationen dank der neuen Technologie den Weg in den SLUB-Katalog.

Der neue Ansatz der multilingualen Suchtechnologie vergrößert die Menge relevanter Suchergebnisse dramatisch. Bei Orientierung und Auswahl hilft eine grafische Visualisierung, die ein mehrfaches hierarchisches Einschränken der Suchergebnisse unterstützt. Mit dem intuitiv zu bedienenden Entscheidungsbaum kann sich der Benutzer zielgerichtet seinen Interessen nähern. Bei der Anfrage „Bank“ würde die Auswahl in der höchsten Hierarchieebene etwa „Wirtschaft“, „Wissenschaft“ und „Kultur“ anbieten. In einer tieferen Hierarchieebene entscheidet sich der Anwender zwischen „Sozialwissenschaft“, „Politikwissenschaft“ und „Naturwissenschaft“ für eine konkrete Interpretation. Dabei werden in Echtzeit die relevanten Katalogeinträge eingeschränkt und entsprechend ihrer aktuellen Relevanz neu sortiert.



Ausblick

SLUBsemantics ist seit Mitte Dezember in einer Beta-Version auf der Webseite der SLUB verfügbar und eröffnet in der Welt der Katalogrecherche einen neuen Horizont. Die bereits in der ersten Ausgabe erstaunlich präzise Suchfunktion wird in den nächsten Monaten konsequent weiterentwickelt. Neben der Verbesserung der Anreicherungs- und Suchalgorithmen sollen die Benutzer systematisch in die Datenoptimierung einbezogen werden, womit die automatisch erzeugten Anreicherungskonzepte durch menschliche Intelligenz auf Plausibilität geprüft und inhaltlich ergänzt würden. Dieses Konzept des sogenannten User Tagging wird funktionieren, sofern die Benutzer hinreichende Beteiligungsanreize, das heißt eine spielerisch zu bedienende Oberfläche sowie eine sichtbare Qualitätssteigerung, erhalten.

Ihre volle Macht kann die Lösung zudem entfalten, wenn bei der Anreicherung nicht allein die knapp vier Millionen Katalogisate der SLUB Berücksichtigung finden, sondern auch die im SLUB-Katalog verfügbaren lizenzierten Volltexte. Der Recall, also die Anzahl der gefundenen Dokumente, würde sprunghaft steigen; ebenso die Precision, die Genauigkeit, mit der ein Dokument zur formulierten Suchanfrage passt, da bei Volltextdokumenten das Ergebnis der Termextraktion und -anreicherung per se um ein Vielfaches besser ausfällt als bei rein bibliografischen Daten.

Schließlich soll das bislang noch gezielt anzuwählende SLUBsemantics mittelfristig mit dem einfachen Suchschlitz und der Standardfacettierung des SLUB-Katalogs vereint und damit für jede Katalogsuche automatisch produktiv werden. Auf der Basis von SLUBsemantics arbeiten die SLUB und Avantgarde Labs gegenwärtig mit internationalen Partnern an der Vorbereitung eines EU-Drittmittelprojekts. Über dessen Inhalt und technische Details von SLUBsemantics wird es im nächsten Jahr einen ausführlichen Beitrag in der Fachpresse geben. Für die Leserinnen und Leser von BIS lautet unsere Empfehlung schon heute: SLUBsemantics aufrufen und testen. Sie werden beeindruckt sein.


ACHIM BONTE


ROBERT GLASS


JENS MITTELBACH