Wirtschaftlich, zuverlässig, weitreichend

Moderne Datenbankinformation mit DBoD und Deep linking

von **ACHIM BONTE, FALK NIEDERLEIN** und **SVEN SZEPANSKI**

Lizenzpflichtige elektronische Nachschlagewerke und Volltextdatenbanken wie die Normendatenbank „Perinorm“, „Chemical Abstracts“ oder „Journal Citation Reports“ bilden auch im Google-Zeitalter den unverzichtbaren Kern der wissenschaftlichen Fachinformation. Angesichts des teilweise sehr hohen Aufwands für Bereitstellung, Updates, Zugriffs- und Rechteverwaltung solcher Datenbanken wurden aus Mitteln des Europäischen Fonds für Regionale Entwicklung (EFRE) seit 2008 die einzelnen Datenbankserver der sächsischen Bibliotheken durch das zentrale Informationssystem Databases on demand (DBoD) abgelöst. Als Entwicklungsziele galten verringerte Betriebskosten, nachhaltiger Service auch für kleinere, bis dahin gar nicht erreichte Informationseinrichtungen, höhere Benutzerfreundlichkeit sowie deutliche Erweiterung der online angebotenen Inhalte. Wenn man so will, entsprach das schon damals dem heute allgegenwärtigen Gedanken des „Cloud Computing“, wonach IT-Infrastruktur nicht mehr individuell betrieben, sondern über einen Dienst vielen Anwendern zur Verfügung gestellt wird.

Bis zum Projektschluss hatte das Team von DBoD alle genannten Ziele glänzend verwirklicht. Der Dienst integriert heute über 1.000 Datenbanken für 26 Hochschulen aus Sachsen und Thüringen und verzeichnet mehr als 300.000 Zugriffe im Jahr. Durch eine DBoD-Kommunikationsschnittstelle ist die Recherche auch von außerhalb des Campus gewährleistet. Neben spezielleren Werken für einzelne Bibliotheken oder Betriebsgemeinschaften werden in DBoD die aus zentralen Mitteln des Freistaats Sachsen finanzierten Titel (Landeslizenzen) sowie die Zugriffsrechte aus landesübergreifenden Einkaufsgemeinschaften zusammengeführt. Mit diesem umfänglichen Angebot sowie der modernen Benutzeroberfläche hat DBoD sein ursprüngliches Vorbild, die „Regionale Datenbankinformation Baden-Württemberg“, bereits erreicht, wenn nicht übertroffen. Neben den Bibliotheken in Thüringen sollen sukzessive weitere Informationseinrichtungen aus anderen Bundesländern von DBoD profitieren. Während der Projektlaufzeit entstand darüber hinaus die Idee, die versammelte Datenbankinformation durch eine integrierte maschinelle Tiefenerschließung noch produktiver werden zu lassen. Seit 2011 arbeitet das DBoD-Team in einem EFRE-Folgeprojekt daran, den Servicenutzen von DBoD auf diese Weise nochmals signifikant zu steigern.

Tiefenerschließung?

„Tiefenerschließung“ weckt Assoziationen zum sogenannten „Deep Web“ bzw. „Versteckten Web“, das heißt zu jenem Teil des World Wide Web, der bei einer Recherche über normale Suchmaschinen nicht auffindbar ist. Der Umfang dieses Deep Web übersteigt die Größe des uns bekannten Internets um ein Vielfaches. So bedeuten etwa die 4,6 Milliarden Datensätze allein des Informationsanbieters LexisNexis (internationale Nachrichten sowie Branchen- und Firmeninformationen) mehr als die Hälfte sämtlicher Datensätze des weltweiten Suchmaschinenprimus Google. Auch DBoD ist mit einer

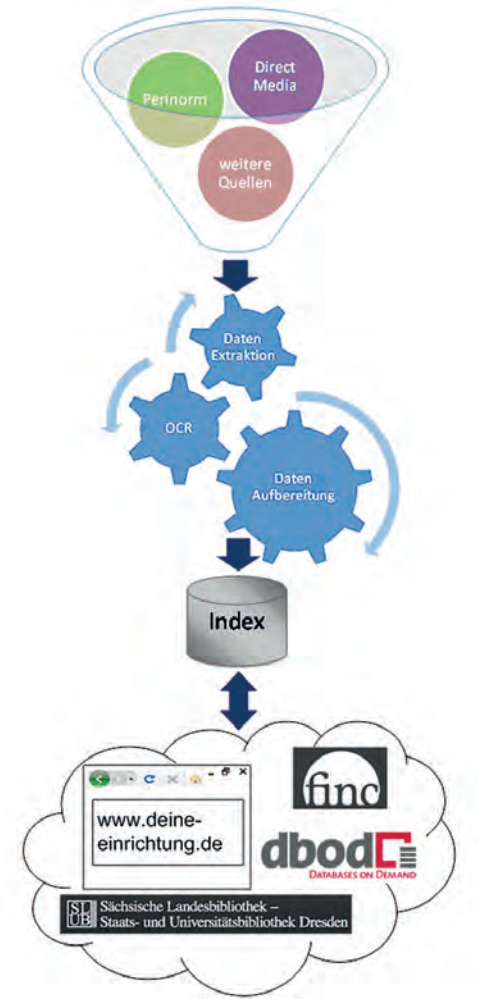
Datengröße von circa zwei TeraByte bislang weitgehend Bestandteil des Deep Web. Unter sorgfältiger Beachtung der geltenden urheberrechtlichen Grenzen soll mit dem Projekt „Tiefenerschließung“ das in DBoD verborgene Wissen für eine normale Katalogrecherche verfügbar gemacht werden.
Bislang benötigen DBoD-Benutzer ein relativ hohes Maß an Vorkenntnissen, um an die gewünschten Informationen zu gelangen. Sucht jemand zum Beispiel eine DIN-Norm, muss er zunächst wissen, dass solche in einer Datenbank namens Perinorm enthalten sind. Daneben sind Praxiserfahrungen in gegebenenfalls mehreren verschiedenen Suchoberflächen notwendig. Nach dem Konzept „Tiefenerschließung“ stellt der Benutzer seine spezifische Suchanfrage einfach im SLUB-Katalog, ohne die passenden Fachdatenbanken oder deren spezifischen Rechercheoberflächen notwendigerweise zu kennen. Die Abbildung veranschaulicht den komplexen Weg, den eine Datenbank durchlaufen muss, um im Katalog recherchierbar zu werden:

1. Datenextraktion
2. Maschinenlesbarkeit der gewonnenen Daten
3. Datenaufbereitung
4. Indexierung
5. Integration ins Katalogsystem

Es werde Licht!

Unter Beachtung der jeweils geltenden Nutzungsrechte werden in der ersten Station die inhaltlichen Daten einer Datenbank von den programmtechnischen getrennt. Soweit die gewonnenen Daten nicht in einem maschinenlesbaren Format vorliegen, erfolgt anschließend eine Texterkennung mittels OCR. Bei eingescannten Dokumenten oder Werken im Bildformat ist dieser zweite Schritt erforderlich. Zum Einsatz kommt dieselbe Software, die etwa auch das Google books-Programm zur Texterkennung verwendet. Die Datenaufbereitung schafft die Grundlage für die Indexierung. Ein wesentlicher Bestandteil dieser Bearbeitungsstufe ist die automatische Generierung von Metadaten zum Volltext. Diese Informationen sind wichtig für die Trefferanzeige und beeinflussen das Relevanz-Ranking. Unter Verwendung der mächtigen Datenintegrations-Software Talend Open Studio können sogenannte Jobs entwickelt werden, die die immensen Datenmengen effizient und in hoher Qualität verarbeiten. Vor der Integration in den Katalog werden die in den vorangegangenen Stationen strukturierten Daten mit Hilfe der Open Source Software Solr in einem Index zusammengefasst. Der Index bildet sozusagen eine Zwischenschicht und ist niemals direkt öffentlich erreichbar. Eingebunden in umfangreichere Suchwerkzeuge nimmt er vielmehr Anfragen entgegen und unterstützt die Auslieferung einer geeigneten Treffermenge. Das Format der Auslieferung kann beeinflusst werden und dient folglich als Schnittstelle für die Integration in den SLUB-Katalog und andere suchmaschinenbasierte Systeme.

Die bisher beschriebenen Methoden garantieren die parallele Suche über alle erschlossenen Datenbanken. Bei Vorliegen der Nutzungslizenz führt ein entsprechender Link in der Trefferanzeige sofort zum Aufruf der Datenbank. Darüber hinaus sollen als Ergebnis des Projekts nicht nur die einschlägigen Datenbanken aufgelistet und auf Wunsch gestartet, sondern auch relevante Treffer in den Datenbanken selbst unmittelbar angesteuert werden können. Sucht ein Benutzer zum Beispiel nach „Schwingbeanspruchung“ erhält er auf der Basis der neuen Technik den Hinweis, dass der Sachverhalt in den Datenbanken Perinorm und Aluminium Taschenbuch thematisiert wird und wird zugleich in den entsprechenden Kontext gelenkt.



Aktueller Projektstand

Das auf zwei Jahre veranschlagte Vorhaben hat inzwischen den Prototypen einer Volltextsuche auf der Webseite von DBoD hervorgebracht. Dabei wurden bereits 263 Datenbanken tiefgehend erschlossen, darunter auch die vielgefragte Perinorm. Allein für diese Datenbank liegen so circa zwei Millionen Dokumentenseiten volltextindexiert vor. Ausgedruckt ergäbe das einen Papierstapel von über fünfzig Metern Höhe. Daneben wurde die Mandantenfähigkeit entwickelt, die sicherstellt, dass die Benutzer der teilnehmenden Einrichtungen jeweils tatsächlich nur den für sie zugänglichen Datenbankbestand durchsuchen. Die Recherche bietet zudem schon viele aus der Internetnutzung geschätzten Funktionen wie Facetten, Snippets und gängige Suchstrategien. Eine Autovervollständigung befindet sich im Aufbau.

ACHIM BONTE

Wie geht es weiter? Um es einmal bildlich auszudrücken: Das Haus der Tiefenerschließung ist gebaut, mit der Einrichtung wurde auch schon begonnen. Nun kann Leben einziehen, wozu möglichst viele der auf DBoD angeboten Datenbanken integriert werden sollen. Die Übertragung des Konzepts auf Quellen jenseits der CD-/ DVD-Datenbanken, zum Beispiel Retrodigitalisate, ist denkbar und erwünscht.

FALK NIEDERLEIN

SVEN SZEPANSKI