Datenbank-Spektrum manuscript No.
(will be inserted by the editor)

BASE (Bielefeld Academic Search Engine)

Eine Suchmaschinenlösung zur Indexierung wissenschaftlicher Metadaten

Amelie Bäcker · Christian Pietsch · Friedrich Summann · Sebastian Wolf

Eingegangen: 19. Oktober 2016 Revidierte Fassung vom 2. Januar 2017

Zusammenfassung Wissenschaftliche Publikationen und ihre beschreibenden Metadaten stehen in stetig zunehmender Anzahl über Plattformen für elektronische Zeitschriften oder digitale Repositorien frei über das Internet zur Verfügung und lassen sich nachnutzen. Die Metadaten können über OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) abgerufen werden (Harvesting). Durch Weiterverarbeitung und Indexierung der Metadaten lassen sich Services wie die „Bielefeld Academic Search Engine“ (BASE) entwickeln.

Dieser Artikel beschreibt die praktischen Erfahrungen, die seit über 10 Jahren im Rahmen des Betriebs von BASE an der Universitätsbibliothek Bielefeld gewonnen wurden. BASE sammelt Millionen von Metadatensätzen aus Tausenden von Quellen weltweit. Die Metadaten werden während des Indexierungsprozesses teilweise korrigiert, normalisiert und um weitere Informationen angereichert. BASE sammelt zudem Metadaten zu den indexierten Quellen, die ebenfalls für Retrieval und Anzeige verwendet werden.

Eine Übersicht der Services vergleichbarer Suchdienste zeigt auf, welche Unterschiede und Gemeinsamkeiten es zwischen BASE und anderen wissenschaftlichen Suchdiensten gibt.

BASE wird auf vielfältige Weise verwendet, und die enthaltenen Daten werden über Schnittstellen nachgenutzt. Neue Herausforderungen ergeben sich insbesondere aus der Erweiterung des Umfangs der Metadaten und ihrer Bereitstellung mithilfe detaillierterer Datenformate jenseits von Dublin Core. Durch die Anreicherung der Metadaten um Informationen wie Personenattribute (IDs, Affiliationen), Förderorganisationen und verknüpfte Forschungsdaten entsteht die Notwendigkeit, das Datenschema für die Indexierung in geeigneter Weise zu erweitern.

Keywords academic search engine · vertical search · scholarly search service

A. Bäcker, C. Pietsch, F. Summann und S. Wolf
Universitätsbibliothek Bielefeld
Universität Bielefeld
Postfach 10 02 91
33502 Bielefeld
Deutschland
E-Mail: base.ub@uni-bielefeld.de

1 Einleitung

Die Möglichkeit, auf elektronischem Wege wissenschaftliche Fachliteratur aufzufinden, beschränkte sich bis Anfang der 2000er Jahre auf die Recherche in Fachdatenbanken oder Bibliothekskatalogen. Diese waren jedoch oft nicht frei zugänglich, verlangten von den Nutzern Vorkenntnisse hinsichtlich der Recherchestrategien und waren insgesamt aufgrund komplexer Trefferlisten, hoher Suchzeiten und komplizierter Zugangswege zum eigentlichen Dokument deutlich weniger komfortabel zu nutzen als Internetsuchmaschinen. Diese hatten sich schon Mitte der 90er Jahre mit AltaVista und später mit dem Start von Google im Jahr 1998 zum Recherchetool Nr. 1 entwickelt [5]. In aller Regel genügte die Eingabe weniger Suchbegriffe in ein Suchfeld, um ein gutes Suchergebnis zu erzielen. Zudem waren fast alle gefundenen Treffer sofort zugänglich.

Da wissenschaftliche Fachliteratur zunehmend auch über das WWW zur Verfügung stand und somit von Internetsuchmaschinen indexiert und auffindbar gemacht werden konnte, wurden Suchmaschinen wie Google frühzeitig auch für die Suche nach wissenschaftlicher Literatur genutzt – spätestens seit Google damit begann, auch PDF-Dateien zu indexieren. Als Nachteil erwies sich, dass wissenschaftliche Literatur häufig in den großen Treffermengen allgemeiner Suchmaschinen unterging und diese nur selten eine (funktionierende) Suche z.B. nach Titeln, Autoren oder Erscheinungsjahren ermöglichten. Entsprechend

gab es einen Bedarf an Suchdiensten, die sowohl die aus Fachdatenbanken bekannte gezielte Suche in bestimmten Suchfeldern boten als auch den Recherchekomfort einer Internetsuchmaschine. Der erste Dienst, der diese beiden Welten miteinander kombinierte, war die wissenschaftliche Suchmaschine Scirus, betrieben vom Elsevier-Verlag. Diese ging bereits im Frühjahr 2001, also nur gut $2\frac{1}{2}$ Jahre nach Google, an den Start. In Scirus konnten neben Zeitschriften des Elsevier-Verlags auch Websites von Hochschulen oder Preprint-Server durchsucht werden. Mit OAIster (2002), Google Scholar (2004) und BASE (2004) folgten weitere Dienste, die eine Recherche nach wissenschaftlicher Literatur ermöglichten. Im Laufe der Zeit kamen Dutzende weitere Dienste hinzu. Scirus selbst wurde Anfang 2014 vom Elsevier-Verlag wieder eingestellt; statt dessen verweist der Elsevier-Verlag auf die eigenen kommerziellen Angebote ScienceDirect und Scopus[1].

2 Von der Idee zum Dienst

2.1 Ausgangssituation und politische Ziele

Um die Jahrtausendwende gab es im Bereich der Hochschulbibliotheken einige Ansätze, integrierte Zugangssysteme zu digitalen Inhalten zu etablieren. Zum Einsatz kam dabei oftmals eine Mischung aus eigenen Nachweisdatenbanken mit Metasuchsystemen, die verteilte Suchen auf externen und internen Datenbanken abwickelten. Der gleichzeitige Siegeszug der Internetsuchmaschinen in diesem Zeitraum und deren zunehmende Nutzung auch im Hochschulbereich führte schnell zu Überlegungen, die damit neu gesetzten Standards bei Suchkomfort und Indexierungstechniken aufzunehmen, in die Bibliothekssuchumgebungen zu integrieren und damit konkurrenzfähig zu bleiben. Mit dieser Zielsetzung wurde an der Universitätsbibliothek Bielefeld seit 2001 eine praxiserprobte kommerzielle Suchmaschinensoftware der norwegischen Firma FAST getestet, um damit ein innovatives Suchsystem der Zukunft für Hochschulen und Hochschulbibliotheken zu entwickeln. Begonnen wurde entsprechend mit dem üblichen Crawlen, eingeschränkt auf wissenschaftliche Websites. Durch die in etwa zeitgleich ab 2001 aufkommende Open-Access-Bewegung und der damit einhergehenden OAI-PMH-Verbreitung ergaben sich alternative Optionen, um Metadaten zu wissenschaftlichen Publikationen in vergleichsweise guter Qualität zu erhalten. Im Wettbewerb zwischen zahlreichen via Crawling eingesammelten Webseiten sehr unterschiedlichen Inhaltstyps mit den vergleichsweise klar strukturierten und dokumentbezogenen bibliographischen Daten hat sich der zweite Ansatz als geeigneter herausgestellt. Weitergehende Strategien, Inhalte von vorhandenen bibliographischen Datenbanken zu indexieren (z.B. den lokalen Bibliothekskatalog und einzelne Spezialdatenbanken) und auf diese Weise zu integrieren, haben sich damals als zu ineffizient erwiesen. Diese Quellen wiesen jeweils proprietäre und ungleich umfangreichere Formate auf, so dass sich der jeweilige Transformationsaufwand als zu hoch herausstellte. Die Konzeptions- und Umsetzungsphase von der Grundidee bis zur Aufnahme des Regelbetriebs hat immerhin drei Jahre gedauert, bis am 24. Juni 2004 die Bielefeld Academic Search Engine (BASE) mit 19 indexierten Quellen gestartet werden konnte. Während die damals implementierten Systemstrukturen im Kern noch heute erkennbar sind, hat es zahlreiche Ergänzungen und Verfeinerungen bei Datenstruktur und Services gegeben. Gleichzeitig hat sich durch die kontinuierlich zunehmende Zahl von Datenquellen ein Index mit einer Größe von mehr als 100 Mio. Datensätzen entwickelt [8,9].

2.2 Besonderheiten der zu indexierenden Metadaten

Universelles Pflichtformat in OAI-PMH ist Dublin Core (DC), ein flaches Datenformat, das nach bibliographischen Kriterien zwar unscharf, dennoch aber viel konkreter und klarer im Hinblick auf die beschriebenen Objekte (überwiegend textbasierte Formen von Publikationen) ist als die per Web extrahierbaren Daten. Die fünfzehn Standardelemente des Dublin-Core-Formats erlauben eine umfassende Beschreibung verschiedener Objekttypen, zumal jedes Feld beliebig oft wiederholbar ist.

Die Unschärfe des Dublin-Core-Formats äußert sich vor allem in der leicht unterschiedlichen Verwendung der Elemente – je nach zum Einsatz kommender Repository-Software. Als Beispiel sei der Link zum Volltext genannt, der sich sowohl im Element „identifier“ als auch in „relation“ oder „source“ befinden kann. Im Element „subject“ kommen sowohl Schlagwörter und Schlagwortketten als auch Notationen unterschiedlicher Klassifikationen vor. Die Wiederholbarkeit der Felder und ihre fehlende Kontextualisierung führen auch dazu, dass bei Datumsangaben (Feld „date“), die häufig mehrfach pro Datensatz ausgeliefert werden, nicht eindeutig feststellbar ist, bei welcher Angabe es sich um das Publikationsdatum handelt.

Ein weiteres Problem des Datenformats besteht in seiner mangelnden Fähigkeit, hierarchische Relationen wiederzugeben. Insbesondere kommt dies bei ergänzenden Informationen zu Personen und beteiligten Körperschaften zum Tragen. Angaben wie Identifier, Affiliationen, Lebensdaten und Kontaktadressen müssen entweder im selben Feld wie der Autorenname angegeben werden (was dessen korrekte Indexierung erschwert) oder in einem Wiederholungsfeld, was eine nachträgliche Zuordnung bisher unmöglich macht. Dasselbe Problem betrifft auch die Angabe referenzierter Literatur (in „source“), weshalb dies meist in unstrukturierter Form erfolgt.

[1] http://www.sciencedirect.com/scirus/

3 Konzeption der Suchlösung: Backend und Frontend

3.1 Zum Einsatz kommende Software und Hardware

Von 2001 bis 2011 wurde in Bielefeld die kommerzielle Suchmaschinenlösung FAST (Fast Search And Transfer) eingesetzt, die einen hohen Grad an Professionalität aufwies. Im Kern agierte eine Serverfarm auf Linux-Basis mit einem verteilten Index, wie sie von größeren Suchmaschinen bekannt ist. Dazu wies die Software zahlreiche Merkmale auf, die die Datenverarbeitung sinnvoll unterstützten und damit viele Erweiterungen ermöglichten. Nach dem Verkauf von FAST zeichnete sich aufgrund der Preisentwicklung und schlechterer Supportbedingungen folgerichtig die Notwendigkeit einer Migration in Richtung Open-Source-Software ab und wurde nach relativ kurzer Vorbereitung im Jahr 2011 vorgenommen. Seither kommt eine Lucene/Solr-Lösung für das Backend und eine auf VuFind basierende Eigenentwicklung für das Frontend zum Einsatz. Trotz der immer noch exponentiell steigenden Dokumentenzahlen und des Ausbaus der Schnittstellen sowie deren zunehmender Nutzung arbeitet das Backend-System noch als Single-Node-Index, wobei die Weiterentwicklung der Lucene/Solr-Basis die Grundlagen geschaffen hat, um die wachsenden Datenmengen bewältigen zu können.

3.2 Workflow und Indexdefinition im Überblick (Backend)

Der grundlegende Ablauf der Datenbehandlung wurde auch bei der Systemmigration 2011 beibehalten, insbesondere um den Migrationsaufwand möglichst gering zu halten. Die Harvesting-Ergebnisse werden nach Möglichkeit um Dewey-Dezimalklassifikationscodes (DDC) angereichert. Ausgehend davon wird u.a. ein Browsing nach DDC realisiert. Anschließend werden auch die Metadaten zu den Quellen erfasst und liefern dabei relevante Hinweise für die nachfolgenden Index-Aktivitäten (insbesondere die verwendete Repository-Software hat Auswirkungen auf die gelieferten Metadatenstrukturen). Unter Berücksichtigung dieser Informationen schließt sich der Pre-Processing-Schritt an, der aus den Metadaten mithilfe von Normalisierung und Anreicherung ein selbst definiertes XML-Format erzeugt, das der DC-orientierten Indexstruktur angepasst ist. Abschließend werden diese Daten mit der konfigurierten Lucene/Solr-Technologie indexiert.

3.3 Technische Realisierung im Detail (Backend)

Die Ergebnisse der (technisch unabhängigen) Harvesting-Aktivitäten liegen in einem fest definierten Festplattenbereich, wo die abgerufenen Metadaten zu jeder Quelle separat abgelegt werden. Vor der Speicherung haben bereits kleinere Korrekturen wie die Bereinigung von Zeichensatz- und XML-Fehlern stattgefunden. Bei ausreichenden Metadaten (benötigt wird ein Abstract ausreichender Länge) schließt sich die automatische Klassifikation an, bei der die einzelnen Sätze mit DDC-Klassifikationscodes angereichert werden [6, 10]. Mit diesen Daten findet der aufwändigste Schritt, das Pre-Processing, statt, wobei für jede Quelle eine individuelle Behandlung mit einem dezidierten Skript vorgenommen wird. Notwendig ist dies, weil erfahrungsgemäß individuelle Eigenheiten jeder Quelle Anpassungen erfordern. Natürlich lassen sich bei Quellen mit identischen oder ähnlichen Metadatenstrukturen Prozeduren ableiten und effizient nachnutzen. Da die Daten OAI-PMH-konform als XML-Dateien vorliegen, lassen sich zur effizienten Weiterverarbeitung die zahlreich verfügbaren XML-Tools verwenden. Hierbei werden insbesondere XSLT- und Perl-Skripte eingesetzt.

Aufgrund der weiten Verbreitung des Dublin-Core-Formats orientiert sich das interne Zielformat und damit auch die interne Datenstruktur von BASE an diesem Modell. Beim Pre-Processing werden neben den fünfzehn Originalfeldern, die in ein einheitliches Format gebracht werden, auch eine Reihe von normalisierten und ergänzten Informationen indexiert und gespeichert. Zusätzlich in normalisierter Form indexiert werden – soweit möglich – die Informationen zu Sprache, Dokumentart und Publikationsjahr einer Veröffentlichung sowie die Angaben zu Lizenzinformationen und dem damit verbundenen Zugriffsstatus (open/restricted access). Zur Realisierung kommen hierbei intellektuell erstellte Tabellen zum Einsatz, über die die heterogenen Angaben aus den Originaldaten dem jeweiligen Oberbegriff zuordnet werden. So werden z.B. die Begriffe *buchkapitel, book chapter, book section* oder *parte de libro*, die sich in den originären Metadaten im Feld *type* befinden, der Dokumentart *Teil eines Buches* zugeordnet.

Darüber hinaus werden die Metadaten der einzelnen Datensätze um Herkunftsinformationen (Quellenname, Kontinent, Land) ergänzt, mit deren Hilfe auch der BASE-interne Identifier gebildet wird. Eine gesonderte Behandlung erfahren darüber hinaus die Personen- und Körperschaftsnamen aus „creator“ und „contributor“, die in einem gemeinsamen Personenfeld suchbar gemacht werden.

Abschließend kann der Indexierungsvorgang durch den Lucene-Indexer basierend auf der Indexdefinition durchgeführt werden. Zur Zeit wird der Index trotz exponentiell gewachsener Dokumentzahlen noch immer einmal wöchentlich am Wochenende komplett neu aufbereitet und dann zu Beginn der Woche durch einen Kopiervorgang aktiviert.

3.4 Frontend

Die Benutzeroberfläche wird mit VuFind realisiert und nutzt die Query-Schnittstelle, um die vom Index gelieferten Informationen aufzubereiten und in der Trefferanzeige auszugeben. Die Lucene-Query-Behandlung wird dabei eingesetzt, um das sogenannte OA-Boosting und die multilinguale Suche zu unterstützen. Durch das (abschaltbare) OA-Boosting werden Dokumente, die als Open Access gekennzeichnet wurden, vor vergleichbaren Treffern gerankt, die keinen Open-Access-Status besitzen. Die multilinguale Suche wird durch die Integration des Eurovoc-Thesaurus[2] realisiert, über den ein Begriff in bis zu 27 (europäischen) Sprachen gefunden werden kann. Hierbei wird im Hintergrund eine Suche nach Synonymen durchgeführt.

Sowohl die gängigen Suchoperatoren (*AND*, *OR*, *NOT*, Phrasensuche) als auch die Möglichkeit, Platzhalter zu verwenden (das Sternchen ersetzt beliebig viele Buchstaben), werden unterstützt (siehe auch Tabelle 1). Die Suchoberfläche bietet die aus Bibliothekskatalogen oder Fachdatenbanken bekannte Möglichkeit, gezielt z.B. nach Titeln, Autoren oder Zeiträumen zu suchen. In der erweiterten Suchoberfläche kann der Suchbereich zudem auf Dokumentarten, Nachnutzungsmöglichkeiten (Creative-Commons-Lizenzen), Zugangsart (Open Access) und auf Quellen aus bestimmten Kontinenten bzw. Ländern (z.B. nur europäische Quellen, nur Quellen aus Deutschland) eingegrenzt werden.

Die Trefferliste bietet eine differenzierte Anzeige der Metadaten des Treffers (Titel, Autor, Inhaltsbeschreibung, Erscheinungsdatum). Sind keine Metadaten vorhanden, wird stattdessen ein Auszug (Teaser) aus dem Inhalt des Treffers ausgegeben. Darunter wird die URL des Dokuments und der Datenlieferant angezeigt. Von einem Treffer ausgehend ist z.B. eine direkte Suche nach weiteren in BASE indexierten Werken des Verfassers oder nach weiteren Dokumenten aus der Quelle (Datenlieferant) möglich. Über den Link „In Google Scholar suchen“ kann eine Suche nach dem entsprechenden Titel in Google Scholar ausgeführt werden, über den dann weitere Informationen, Zitationen und Versionen ermittelt werden können. Die Reihenfolge der Trefferausgabe erfolgt nach Relevanz – hierbei spielen u.a. Position und Häufigkeit der verwendeten Suchbegriffe und der Open-Access-Status eine Rolle –, kann aber auch nach verschiedenen Kriterien (Titel, Autor, Erscheinungsjahr) sortiert werden.

Es gibt verschiedene Möglichkeiten, einzelne Treffer zu exportieren und zu speichern, um die Metadaten z.B. in ein Literaturverwaltungsprogramm zu übernehmen. Auch die Einrichtung eines RSS-Feeds zu einer Suche ist möglich, über den der Nutzer regelmäßig automatisch über neue Titel zu dieser Suche informiert wird. Der Nutzer kann zudem einen persönlichen Account erstellen, um z.B. einzelne Treffer als „Favorit“ abzuspeichern oder Suchanfragen zur Suchhistorie hinzuzufügen.

Die Trefferliste kann über Drilldowns eingeschränkt werden. Hierbei handelt es sich um Auswahlmenüs, mit denen sich die Trefferliste auf Autoren, Schlagwörter, DDC, Erscheinungsjahr, Quelle, Sprache, Dokumentart, Zugang und Nachnutzung einschränken lässt. Welche Möglichkeiten der Suchverfeinerung angeboten werden, ist vom Suchergebnis abhängig. Sind zum Beispiel alle gefundenen Treffer frei zugänglich, wird die Einschränkungsmöglichkeit „Zugang“ nicht angeboten.

Alternativ zur Suche steht ein Browsing über verschiedene Bereiche (DDC, Dokumentart, Nachnutzung und Zugangsart) zur Verfügung.

4 BASE im Vergleich zu anderen wissenschaftlichen Suchdiensten

4.1 Überblick: Wissenschaftliche Suchmaschinen

Neben BASE gibt es zahlreiche weitere Suchdienste, die eine gezielte Suche nach und in wissenschaftlichen Dokumenten anbieten. Die englischsprachige Wikipedia führt in einer Liste[3] über 130 Dienste auf, wobei hier u.a. auch Verlagsangebote, Fachdatenbanken und E-Journal-Portale zu finden sind. 30 der aufgeführten Dienste sind interdisziplinär, wobei von diesen wiederum ein gutes Dutzend frei zugänglich ist. Beispielhaft sollen hier Unterschiede und Gemeinsamkeiten zwischen BASE und den wissenschaftlichen Suchlösungen Google Scholar, Microsoft Academic, OAIster sowie OpenAIRE aufgezeigt werden, die wie BASE den wissenschaftlichen Suchraum weltweit abdecken.

4.2 Harvesting von XML-Metadaten via OAI-PMH

Wie BASE bezieht OAIster die Metadaten wissenschaftlicher Repositorien über OAI-PMH. OAIster ging bereits im Jahr 2002 an der University of Michigan online und wurde 2009 von der Non-Profit-Organisation OCLC (Online Computer Library Center) übernommen. Der Index hat nach eigenen Angaben[4] derzeit einen Umfang von etwa 30 Mio. Dokumenten aus 1500 Quellen. Seit der Übernahme durch OCLC ist OAIster ein Bestandteil des WorldCat, dem weltweit größten Bibliothekskatalog mit über 2 Milliarden Bestandsnachweisen. Das OAIster-Segment kann separat durchsucht werden[5], die Trefferanzeige ist jedoch

[2] http://eurovoc.europa.eu/

[3] https://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines

[4] http://www.oclc.org/oaister.en.html

[5] http://oaister.worldcat.org/

vollständig in den WorldCat integriert und somit auch in erster Linie auf bibliographische Angaben an physikalischen Standorten optimiert. Eine Trefferanzeige kombiniert daher den Online-Zugang mit dem Hinweis auf vorhandene, gedruckte Exemplare in Bibliotheken weltweit. Nach der Recherche ist eine Einschränkung auf Formate möglich (dies umfasst sowohl Zugangsarten als auch eine grobe Unterteilung nach Dokumenttypen, z.B. „Downloadbares Archivmaterial“). Eine Sortierung der gesamten Treffermenge, z.B. nach Erscheinungsjahren, ist dagegen nicht möglich.

Auch OpenAIRE bezieht die Metadaten aus wissenschaftlichen Repositorien (sowohl Literatur als auch Forschungsdaten) vorrangig über OAI-PMH. OpenAIRE ist ein von der EU gefördertes Projekt mit dem Ziel, den Zugang zu allen Open Access publizierten wissenschaftlichen Ergebnissen im europäischen Forschungsraum zu ermöglichen[6]. Es werden jedoch auch Repositorien außerhalb Europas indexiert, sofern sie den von OpenAIRE definierten Metadaten-Guidelines genügen, d.h. „OpenAIRE-kompatibel“ sind[7]. Material, welches nicht per Open Access zur Verfügung steht, wird nur akzeptiert, wenn es in Zusammenhang mit EU-geförderter Finanzierung steht[8]. OpenAIRE konzentriert sich also in wesentlich stärkerem Maße als die anderen hier genannten Suchmaschinen auf Open-Access-Dokumente. Über 90% der derzeit 17 Mio. Dokumente aus knapp 800 Quellen sind per Open Access zugänglich[9]. OpenAIRE bietet zudem als Besonderheit die Möglichkeit an, Publikationen auf Forschungsförderer einzuschränken. Diese und weitere Facetten stehen unter der Suchmaske und nach einer Suche in der Trefferliste zur Verfügung. Die Facettierung nach Dokumenttypen ist recht feingliedrig und umfasst derzeit 28 Dokumenttypen. Eine Sortierung der gesamten Treffermenge ist nicht möglich. In stärkerem Maße als bei BASE kommen Textmining-Verfahren (auf Basis von Volltexten) zur Metadatenanreicherung zum Einsatz. Mit ihrer Hilfe werden Projektinformationen extrahiert, Dokumente klassifiziert, inhaltlich ähnliche Dokumente, referenzierte Literatur und in Beziehung stehende Forschungsdaten ermittelt. Metadatensätze von Dubletten werden in OpenAIRE zusammengeführt und so bereinigt.

Die Universitätsbibliothek Bielefeld ist technischer Partner im OpenAIRE-Projekt und bringt dort ihre im Rahmen von BASE erworbenen Expertise im Bereich Metadaten-Harvesting und -Aggregation ein. Sie koordiniert zudem Aufgaben im Bereich der OpenAIRE-Guidelines und Services für Nutzungsstatistiken.

[6] https://www.openaire.eu/project-factsheets

[7] OpenAIRE Guidelines https://guidelines.openaire.eu/

[8] OpenAIRE's Content Acquisition Policy https://www.openaire.eu/openaire-s-content-acquisition-policy

[9] https://www.openaire.eu/search/find?keyword=

4.3 Webseiten-Indexierung

Google Scholar indexiert neben den Metadaten z.T. auch die Volltexte (PDFs) wissenschaftlicher Dokumente. Da Google Scholar keine bibliographisch strukturierten Datenformate indexiert, sondern Webseiten und PDF-Dokumente, gibt Google Scholar selbst Empfehlungen zur Art der Bereitstellung der Metadaten und zur Gestaltung von Titelblättern und Literaturverzeichnissen in PDF-Dokumenten, um eine korrekte Indexierung von Titeln, Autorennamen und Zitationen zu gewährleisten[10]. Diese Art der Indexierung führte insbesondere in den ersten Jahren zu Problemen [2, 3], da viele Webseiten oder PDFs nicht gemäß den Vorgaben von Google Scholar gestaltet waren bzw. die entsprechenden Vorgaben erst im Laufe der Jahre von Google Scholar entwickelt wurden. Durch verfeinerte Indexierungsmethoden und breiter gewordene Unterstützung der „Google Scholar Metatags“ in Repository- und E-Journal-Software wie z.B. dem Open Journal System (OJS) sind diese Probleme in den letzten Jahren weitestgehend ausgeräumt worden.

In Google Scholar werden unterschiedliche Versionen eines Artikels an einer Stelle zusammengefasst. Dabei findet eine Zusammenführung von Dubletten und verschiedenen Fassungen statt (Preprint, vollständiger Artikel, Artikel als Beitrag in einem Sammelband etc.). Als „Haupteintrag“ fungiert dabei i.d.R. das Dokument mit den meisten Zitationen. Die zitierenden Dokumente können über einen Link angezeigt werden (dies umfasst nur die Dokumente, die ebenfalls in Google Scholar indexiert wurden). Von indexierten Artikeln wertet Google Scholar die Literaturverzeichnisse aus und zeigt die Einträge ggf. als eigene Treffer in der Trefferliste an, gekennzeichnet durch den Hinweis „Zitation“. Bei diesen Einträgen fehlt ein Link auf eine Eingangsseite zum Dokument (Frontdoor) oder ein PDF. Diese Einträge können von der Trefferanzeige ausgeschlossen werden, indem die Option „Zitate einschließen“ abgewählt wird. Google Scholar indexiert auch lizenzpflichtige Dokumente aus Fachdatenbanken und zeigt an, ob das Dokument von der eigenen Bibliothek lizenziert wurde und somit für einen Nutzer dieser Bibliothek ggf. online erreichbar ist. Der Nutzer kann hierzu in den Einstellungen unter „Bibliothekslinks“ die eigene Bibliothek auswählen. Google Scholar bietet nur wenige Möglichkeiten, das Suchergebnis einzugrenzen oder zu sortieren. Lediglich eine Sortierung oder Eingrenzung auf Erscheinungsjahre bzw. einen Erscheinungszeitraum ist möglich. Bei der Sortierung nach Erscheinungsjahren werden nur die im letzten Jahr hinzugefügten Artikel (sortiert nach Datum) angezeigt.

Zum Umfang des Indexes und der Zahl der indexierten Quellen macht Google keine Angaben. Schätzungen belaufen sich auf mindestens 100 Mio. englischsprachige Artikel

[10] Inclusion Guidelines for Webmasters: https://scholar.google.de/intl/de/scholar/inclusion.html

[4] und 160 Mio. Artikel insgesamt [7]. Auch wenn die von Google Scholar selbst angegebenen Treffermengen zweifelhaft erscheinen (siehe Tabelle 2), ergibt sich aus diesen und weiteren Untersuchungen [1], dass Google Scholar die umfangreichste wissenschaftliche Suchmaschine ist.

Die zweite große Suchmaschine, die wissenschaftliche Webseiten indexiert, ist Microsoft Academic. Nach eigenen Angaben umfasst sie 120 Mio. Nachweise[11]. Eine Besonderheit von Microsoft Academic ist, dass bereits auf der Startseite eine Auswahl an neuen Artikeln aus verschiedenen Forschungsgebieten angezeigt wird. Die Startseite erweckt daher eher den Eindruck eines wissenschaftlichen Nachrichtenportals. Microsoft Academic bietet einige Möglichkeiten an, das Suchergebnis nachträglich einzugrenzen, die es in dieser Form nicht in anderen Suchdiensten gibt. So kann der Nutzer z.B. einen Erscheinungszeitraum („Date range“) auswählen. Auswählbar sind dabei – anders als z.B. bei der freien Eingabe in Google Scholar – nur die Erscheinungsjahre, die in der Trefferliste tatsächlich vorkommen. Zudem werden dem Nutzer u.a. die Top-5-Autoren, Affiliationen, Forschungsgebiete und Zeitschriftentitel zur jeweiligen Recherche angezeigt. Diese Facetten können auf jeweils 20 Einträge erweitert werden. Microsoft Academic verweist in der Trefferliste ggf. auch auf weitere Treffer aus der eigenen Suchmaschine „Bing“ und liefert bei der Eingabe eines Begriffes z.T. einen Eintrag aus der englischsprachigen Wikipedia. Auch in der Anzeige eines Treffers unterscheidet sich Microsoft Academic von anderen Suchdiensten. Nach Auswahl eines Treffer wird der Nutzer nicht sofort zur Frontdoor oder zum PDF geführt, sondern zu einer „Detailanzeige“, die weitere Informationen zum Dokument enthält. Hier sind z.B. das Abstract, „Linked References“ (Einträge aus dem Literaturverzeichnis des Artikels) und „Top Citations“ zu finden. Die bibliographischen Angaben zur Publikationen werden angezeigt, ggf. auch weitere Informationen zum Autor, und es werden die Quellen aufgelistet, über die das Dokument abgerufen werden kann. Selbst durchgeführte Stichproben zeigten jedoch, dass – anders als in Google Scholar – die Volltexte von PDF-Dokumenten nicht oder nicht immer indexiert werden.

[11] http://academic.research.microsoft.com/

[12] Das Sternchen kann als Trunkierungszeichen innerhalb oder am Ende eines Suchbegriffs gesetzt werden und ersetzt beliebig viele Zeichen

[13] Google Scholar zeigt als Standard die „am besten“ zum Suchergebnis passenden Ergebnisse an – dabei müssen nicht alle gesuchten Begriffe im Dokument selbst vorkommen

[14] Außer der Freitextsuche bietet Microsoft Academic keine Suchfelder an. Das Suchergebnis kann aber nach verschiedenen Kriterien eingeschränkt werden

[15] # ersetzt exakt ein Zeichen innerhalb eines Suchbegriffs, ? mehrere, * ersetzt beliebig viele Zeichen am Ende eines Suchbgegriffs

[16] Microsoft Academic erlaubt keine Phrasensuche und keine Einschränkung auf Titel. OpenAIRE erlaubt keine Phrasensuche.

Tabelle 1 Vergleich der Indexgröße und Suchmöglichkeiten (Stand: 30.11.2016)

Suchdienst	Indexgröße	Operatoren	Suchfelder
BASE	100 Mio.	AND, OR, NOT, Phrase, *[12]	Autor, Dokumentart, Lizenz, Quelle, Schlagwort, Titel, URL, Verlag, Zeitraum, Zugang
Google Scholar	160 Mio.	Fuzzy-AND[13], OR, NOT, Phrase	Titel, Autor, Zeitschrift, Zeitraum
Microsoft Academic	120 Mio.	AND	-[14]
OAIster	30 Mio.	AND, OR, NOT, Phrase, # ? * [15]	Autor, Format, ISBN, ISSN, Inhalt, Sprache, Verlag, Zeitraum
OpenAIRE	17 Mio.	AND, OR	Autor, Schlagwort, Titel, Verlag

Tabelle 2 Vergleich der Treffermengen (Stand: 30.11.2016)

	Suchanfrage / Trefferzahlen[16]		
Suchdienst	autonomous robots	“autonomous robots”	title: “autonomous robots”
BASE	35.057	8.763	851
Google Scholar	592.000	91.400	2130
Microsoft Academic	3.888	–	-
OAIster	4169	1.023	189
OpenAIRE	2377	–	*315*

5 Nutzung und Nachnutzung

5.1 Nutzung der Website

BASE verzichtet auf den Einsatz von Technologien, die die Privatsphäre der Nutzer beeinträchtigen könnten. Darum setzt BASE keine Tracker ein, sondern analysiert nur die anonymisierten Server-Logfiles[17].

[17] Zum Anonymisieren von Logfiles nach EU-Recht kommt selbst geschriebene Software zum Einsatz, die frei nachgenutzt werden darf: https://github.com/pietsch/loganalyse

Tabelle 3 Nutzung der BASE-Website (Stand: 30.11.2016, ohne Bots)

Jahr	Besuche	Suchen	Seiten	Zugriffe
2014	865.690	4.379.515	8.868.915	22.834.482
2015	1.012.204	3.953.382	7.875.816	25.403.458
2016/1–11	1.270.678	3.706.948	7.155.036	27.206.030

Tabelle 4 BASE-Nutzung nach Ländern (2015)

Land	Seiten	Zugriffe
Deutschland	2.126.451	8.592.824
Frankreich	1.264.203	2.785.242
USA	1.080.889	2.085.242
Kanada	741.076	964.707
China	625.353	2.253.891
UK	164.102	560.455
Österreich	108.436	529.491
Schweiz	97.757	483.004
Belgien	94.532	645.723
Polen	82.029	396.169

Tabelle 5 Länder mit den meisten Dokumenten in BASE (Stand: 30.11.2016)

Land	Dokumente	Repositorien
USA	36.666.421	918
Deutschland	13.123.564	335
Frankreich	6.786.048	171
Großbritannien	4.506.821	266
Spanien	4.146.643	195
Italien	2.962.115	140
Polen	2.474.879	101
Australien	2.326.982	83
Japan	2.017.990	436
Schweiz	1.893.516	31

Das Webinterface von BASE erfährt in den letzten Jahren eine ungefähr linear steigende Nutzung, wie Tabelle 3 belegt. Die Zahl der Suchvorgänge innerhalb von BASE blieb in etwa konstant. Wenn Nutzer über eine allgemeine Suchmaschine zu BASE fanden, war dies in 85 % der Fälle Google. Ähnlich viele Besucher wie von Google kommen vom KVK[18] zu BASE.

Die stärkste Nutzung erfährt BASE aus Deutschland, Frankreich und den USA. Diese drei Länder stellen auch die meisten Dokumente und andere erfasste Objekte bereit. Auf den weiteren Plätzen gibt es aber starke Unterschiede zwischen Nutzung und Bereitstellung von Dokumenten. Aus welchen zehn Ländern im Jahr 2015 die meisten Zugriffe auf die Web-Oberfläche von BASE kamen, zeigt Tabelle 4. Welche zehn Länder derzeit die meisten Dokumente zu BASE beisteuerten, zeigt Tabelle 5.

[18] Karlsruher Virtueller Katalog ⟨https://kvk.bibliothek.kit.edu/⟩, eine Metasuchmaschine für Bücher und digitale Medien.

5.2 Nachnutzung durch andere Diensteanbieter

Die Nachnutzung von BASE-Suchergebnissen und aggregierten Daten ist in den letzten Jahren erheblich gestiegen. Suchergebnisse stellt BASE über die sogenannte HTTP-API bereit. Für die Verteilung von aufbereiteten bibliographischen Metadaten in zwei Metadatenformaten (oai_dc und base_dc) steht eine OAI-PMH-Schnittstelle zur Verfügung. Aufgrund der hohen Datenmenge (z.Z. 30 GB gzip-komprimiertes XML) stellt BASE diese Daten außerdem als Komplettabzüge (Dumps) zur Verfügung, die anschließend über die OAI-PMH-Schnittstelle inkrementell aktualisiert werden können.

Die Suchschnittstelle von BASE wird auf Anfrage freigeschaltet. Für nichtkommerzielle Diensteanbieter ist die Nutzung dieser API kostenlos. Im Oktober 2016 waren 101 Nutzer registriert. Neben vielen europäischen und amerikanischen Hochschulbibliotheken und Forschungseinrichtungen gehören dazu beispielsweise die Chinesische Akademie der Wissenschaften, Ex Libris, ResearchGate sowie Open-Access-Projekte wie OpenAIRE und Contentmine.

Ausgewählte Kooperationspartner können außerdem die OAI-Schnittstelle von BASE nutzen. Im Oktober 2015 hatten 40 externe Nutzer die nötigen Zugriffsrechte, darunter AuthorClaim, dissem.in, die Deutsche Nationalbibliothek, Europeana Cloud, SHARE (share.osf.io) und mehrere Bibliotheksverbünde.

Mehrere deutsche Fachportale nutzen BASE für ihre Metasuche, z.B. Fachportal Pädagogik, Germanistik im Netz, ilissAfrica, Livivo (ZB MED), Virtuelle Fachbibliothek Biologie (vifabio) und ViFa medien buehne film. Manche treffen dabei eine Vorauswahl bestimmter Repositorien unter fachlichen Gesichtspunkten. Weil BASE nicht nur klassische Hochschulschriftenserver, sondern auch Plattformen mit Digitalisaten von Fotos, Karten und anderen Quellenmaterialien indexiert, bietet BASE auf diese Weise auch Zugang zu Forschungsdaten.

Seit Dezember 2015 bindet der EBSCO Discovery Service (EDS) die von BASE gesammelten und aufbereiteten Daten in seinen Dienst ein. Auch der von der Universitätsbibliothek Leipzig betriebene nichtkommerzielle Artikelindex finc nutzt BASE-Daten.

Die nichtkommerzielle deutsche Suchmaschine MetaGer[19] war eine der ersten Nutzerinnen der BASE-API. Im Jahr 2016 kamen die Metasuchmaschinen eTools.ch und Searx[20] hinzu.

Eine besondere Rolle spielt BASE als Datenlieferant neuartiger Open-Access-Dienste: Open Access Button, dissemin und ImpactStory bieten Dienste an, die ohne BASE nicht möglich wären. Dazu gehören die ersten beiden alternativen DOI-Resolver DOAI.io und oaDOI.org, die beim

[19] MetaGer: https://metager.de

[20] Searx: https://github.com/asciimoo/searx

Auflösen eines Digital Object Identifiers (DOI) bevorzugt auf frei zugängliche Versionen eines Dokuments verweisen.

6 Ausblick und Weiterentwicklungen

Nach über 12 Jahren Laufzeit umfasst der BASE-Index inzwischen einen Bestand von rund 100 Mio. Dokumenten aus knapp 5.000 Quellen. Ebenso stetig ist in den letzten Jahren die Anzahl an Kooperationen, Komponenten und Aktivitäten im BASE-Umfeld gewachsen[21]. Der unter Punkt 3 aufgezeigte Vergleich mit anderen Suchlösungen zeigt jedoch auch, dass Dienste wie Google Scholar, Microsoft Academic oder OpenAIRE interessante und nützliche Funktionen anbieten, die in BASE bisher erst teilweise umgesetzt werden konnten oder noch fehlen. BASE konzentriert sich bei eigenen Weiterentwicklungen auf pragmatische Lösungen, die im Projektumfeld realisiert werden können. Ein Beispiel dafür ist der Link „In Google Scholar suchen“. Solche Links sind im Prinzip auch für andere Suchdienste denkbar.

Darüber hinaus ist in BASE selbst der Ausbau eigener Dienste und Funktionen geplant, z.B. die Einbindung eines „Claiming“-Dienstes für ORCID-iDs sowie die Indexierung alternativer Metadatenformate.

Mittlerweile als Standard etabliert, bietet ORCID (Open Researcher and Contributor ID) die Möglichkeit, Autoren mithilfe einer numerischen Kennung eindeutig zu identifizieren. Im Rahmen des DFG-geförderten Projektvorhabens „ORCID DE – Förderung der Open Researcher and Contributor ID in Deutschland“[22] sollen Publikationen in Zukunft auch über BASE ORCID-iDs zugeordnet werden können („Claiming“). Während es im ersten Schritt vor allem darum gehen wird, das ORCID-Profil um in BASE geclaimte Publikations- und Forschungsdatennachweise zu ergänzen, sollen später auch die OAI-Metadaten in BASE um die ORCID-Kennung angereichert werden, um eine Nachnutzung zu ermöglichen und damit auch eine redundante Zuordnung zu vermeiden. Dieses Vorgehen erfordert eine Anpassung und Erweiterung des bisherigen Indexprofils.

Eine Erweiterung der Indexstruktur ist nicht nur durch die Zunahme der Autoreninformationen naheliegend, sondern auch durch den allgemein zunehmenden Umfang der geharvesteten Metadaten, die inzwischen auch in Beziehung stehende Informationen wie Affiliationen oder Verweise auf verwandte Dokumente enthalten können. Diese in Dublin Core häufig nicht, oder – formatbedingt – nur unstrukturiert ausgelieferten Informationen sind in Datenformaten wie z.B. MARC oder MODS feingranularer dargestellt, was eine Betrachtung dieser alternativen Formate in den Fokus rücken lässt. Problematisch, da arbeitsintensiv, erscheint an dieser Stelle die Zahl der relevanten Formate und ihre unterschiedliche Verbreitung. In geringerem Umfang als bei Dublin Core gibt es auch hier unterschiedliche Formatinterpretationen.

Vorbereitet sind zudem die Integration der Crossref-Metadaten, die auch Open-Access-Publikationen enthalten, und die Bereitstellung von Profilanalysen von Repository-Servern und ihren globalen Netzwerkstrukturen (als Registry-Service).

Geplant ist der Einsatz von Big-Data-Technologien, um in den Bereichen Dublettenerkennung und Linked-Data-Aggregation die Grundlage für weitere Service-Angebote zu schaffen.

[21] Einen aktuellen Überblick liefert „BASE – a Next Generation Multi-Level Repository-based Service Provider“ https://www.base-search.net/about/download/base_poster.pdf

[22] https://dini.de/projekte/orcid-de/

Literatur

1. Harzing, A.: Microsoft Academic (Search): a Phoenix arisen from the ashes? Scientometrics pp. 1637–1647 (2016). DOI 10.1007/s11192-016-2026-y
2. Jacsó, P.: Google scholar's ghost authors. Libr J **134**, 26–27 (2009). URL http://lj.libraryjournal.com/2009/11/industry-news/google-scholars-ghost-authors/
3. Jacsó, P.: Metadata mega mess in Google Scholar. Online Inform Rev pp. 175–191 (2010). DOI 10.1108/14684521011024191
4. Khabsa, M., Giles, C.: The number of scholarly documents on the public web. PLoS ONE **9**(5), e93949 (2014). DOI 10.1371/journal.pone.0093949
5. Lewandowski, D.: Alles nur noch Google? Entwicklungen im Bereich der WWW-Suchmaschinen. BuB – Forum für Bibliothek und Information **54**(9), 558–561 (2002)
6. Lösch, M.: Automatische Sacherschließung elektronischer Dokumente. In: Proc. 100. Deutscher Bibliothekartag in Berlin (2011). URL http://nbn-resolving.de/urn:nbn:de:0290-opus-10992
7. Orduna-Malea, E., Ayllón, J., Martín-Martín, A., López-Cózar, E.: About the size of Google Scholar: playing the numbers. Scientometrics pp. 931–949 (2015). DOI 10.1007/s11192-015-1614-6
8. Pieper, D., Summann, F.: Bielefeld Academic Search Engine (BASE). An end-user oriented institutional repository search service. Libr Hi Tech **24**, 614–619 (2006). URL http://nbn-resolving.de/urn:nbn:de:0070-pub-27663089
9. Pieper, D., Summann, F.: 10 years of “Bielefeld Academic Search Engine” (BASE): Looking at the past and future of the world wide repository landscape from a service providers perspective. In: 10th International Conference on Open Repositories (OR2015) (2015). URL http://nbn-resolving.de/urn:nbn:de:0070-pub-27663089
10. Waltinger, U., Mehler, A., Lösch, M., Horstmann, W.: Hierarchical Classification of OAI Metadata Using the DDC Taxonomy, vol. 6699, pp. 29–40. Springer (2011). DOI 10.1007/978-3-642-23160-5_3