

# Tiefenerschließung

## Katalog in der 3. Dimension

von FALK NIEDERLEIN

Das digitale Zeitalter hat die Bibliotheken fest im Griff. Traditionelle Geschäftsprozesse und Dienstleistungen müssen reformiert und den neuen Gegebenheiten angepasst werden. Insbesondere der Bibliothekskatalog steht dabei im Blickpunkt und verdient besondere Aufmerksamkeit. Ein Baustein auf dem Weg zum modernen Katalog ist das EU-Projekt „Maschinelle Tiefenerschließung“, das in diesem Beitrag vorgestellt werden soll.

Obwohl der Name Tiefenerschließung das Projektziel bereits treffend bezeichnet, ist es nicht in zwei Sätzen erklärt. Die Idee entstand während des Aufbaues des Datenbankdienstes DBoD (<http://www.dbod.de> und [http://prezi.com/k-lhbipz7nva/dbod\\_20120321/](http://prezi.com/k-lhbipz7nva/dbod_20120321/)), der den orts- und zeitunabhängigen Zugriff auf CD/DVD- und Online-Datenbanken im Internet sicherstellt. Bis heute wurden über 1.200 Datenbanken auf DBoD veröffentlicht, ein riesiger Schatz an wissenschaftlichen Kenntnissen. Das Projekt „Maschinelle Tiefenerschließung“ baut auf dem viel beachteten Dienst DBoD auf und verfolgt im Wesentlichen drei Ziele:

### Voraussetzungslose Benutzung fördern

Für eine Datenbankrecherche wird dem Benutzer üblicherweise viel Vorwissen abverlangt. Ihm muss im Vorfeld bekannt sein, dass eine Datenbank mit der gewünschten Information überhaupt existiert, wie sie genau heißt bzw. auffindbar ist und wie man in der speziellen Anwendung recherchiert.

Würde zum Beispiel bislang eine DIN-Norm gesucht, musste der Benutzer wissen, dass es eine Datenbank namens Perinorm gibt. In einem nächsten Schritt durfte er ahnen, dass diese Datenbank im Datenbankinformationssystem (DBIS) verzeichnet und dort ein Verweis auf das DBoD-Angebot hinterlegt ist. Nach dem Start der Datenbank konnte unter einer spezifischen Suchoberfläche mit der Recherche begonnen werden: Umständlich, zeitintensiv und nicht gerade intuitiv.

Mit der Integration der Tiefenerschließung in den Katalog können gleichzeitig alle erschlossenen Datenbanken durchsucht werden, ohne dass eine

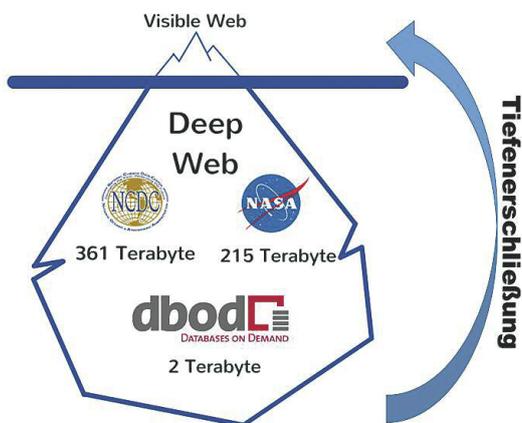
The screenshot shows the SLUB Dresden search results page for the query "Norm Zitierregeln". The interface includes a search bar at the top with the query entered. Below the search bar, there are navigation tabs for "Bücher, Bilder", "Aufsätze", and "Webseite". The search results are displayed in a list format, showing the title, author, and publication year for each result. On the right side, there is a sidebar with filters for "Ergebnis einschränken", "Verfügbarkeit", "Medientyp", "Thema", "Fachgebiet", "Urheber", "Erscheinungsdatum", "Sprache", and "Standort". The results list includes items like "Titelangaben von Dokumenten; Zitierregeln", "Abkürzungsverzeichnis der Rechtssprache", "Zitierfibel für Juristen", "Propylen-Weltgeschichte - Reg. : Alphabetisches Gesamt-Register / Gesamt-Inhaltsverzeichnis aller zehn Bände ; LT...", "Deutsches Grenzland Obersachsen / ein Literaturnachweis", "The Chicago manual of style / [the essential guide for writers, editors, and publishers]", "Spielen in der Schule / Sachstandsbericht und systematischer Literaturnachweis", and "Lehrerbücher der Deutscher und Tschechoslowaken".

Datenbanksitzung gestartet werden müsste. Im SLUB-Katalog ist die Tiefenerschließung seit April produktiv.

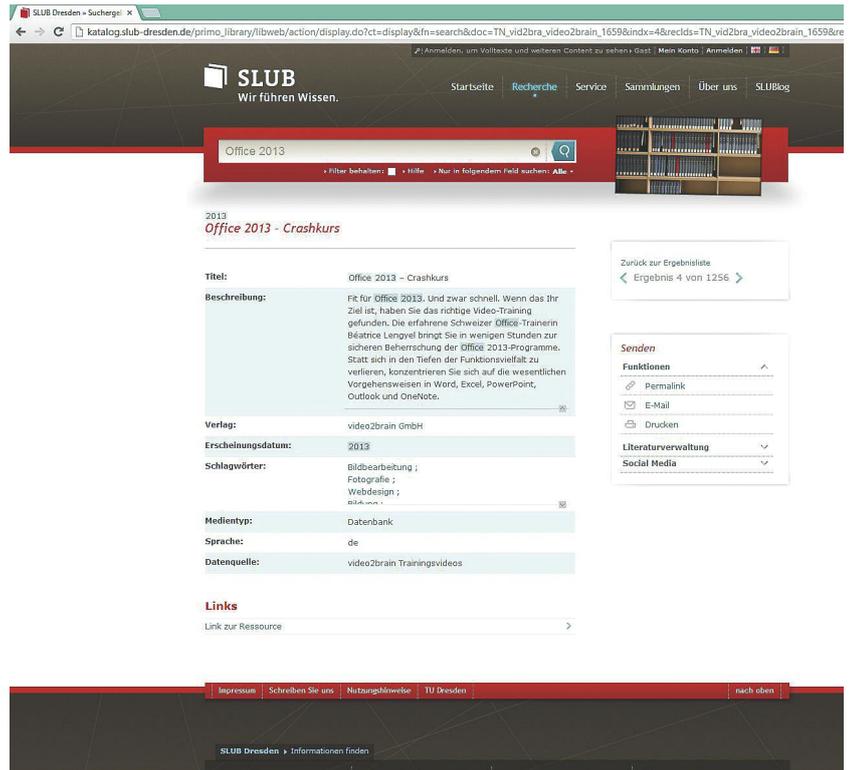
Um noch einmal das Beispiel der Normensuche aufzugreifen: Dass DIN-Normen nicht nur für die Ingenieur-Ausbildung relevant sind, zeigt die hohe Nachfrage nach Informationen zur Gestaltung von wissenschaftlichen Arbeiten. Geisteswissenschaftliche Studienfächer werden in ihrer Ausbildung kaum etwas von der Existenz der Perinorm erfahren. Da aber die Tiefenerschließung bereits im SLUB-Katalog verfügbar ist, verweisen Suchanfragen wie Norm Zitierregel oder Gestaltung von Forschungsarbeiten auf die entsprechende Norm. Der Kreis schließt sich, wenn der Benutzer für seine Arbeit das neu installierte Microsoft Office 2013 verwenden möchte und dafür eine Anleitung benötigt. Durch die Integration der Online-Datenbank video2brain findet der Benutzer durch Eingabe von Office 2013 schnell das passende Trainingsvideo.

### Das Deep Web sichtbar machen

Das Problem der erschwerten Erreichbarkeit betrifft nicht nur die Datenbanken auf DBoD. In der Fachwelt hat sich für dieses Phänomen der Begriff Deep Web etabliert. Das Deep Web beschreibt Inhalte, die



entweder nicht frei zugänglich sind oder Inhalte, die von Suchmaschinen nicht indiziert werden. Das Visible (sichtbare) Web enthält vergleichsweise nur einen geringen Teil der Informationen des Deep Web. Gerne wird in diesen Zusammenhang das Bild eines Eisberges zur Visualisierung verwendet. Das Gros an Informationen ist hierbei in themenspezifischen Datenbanken gespeichert. Beispiele sind die Datensammlung des National Climatic Data Center (361 Terabyte), die Daten der NASA (215 Terabyte), aber eben auch DBoD mit circa 2 Terabyte an Daten. Allerdings handelt es sich bei den Erstgenannten eher um Bilddaten, die Datenbanken auf DBoD sind überwiegend Volltextdatenbanken. Die Tiefenerschließung tritt nun an, um im Rahmen der jeweils geltenden lizenzrechtlichen Rahmenbedingungen möglichst viele Informationen aus den DBoD-Datenbanken dem Visible Web zuzuführen.

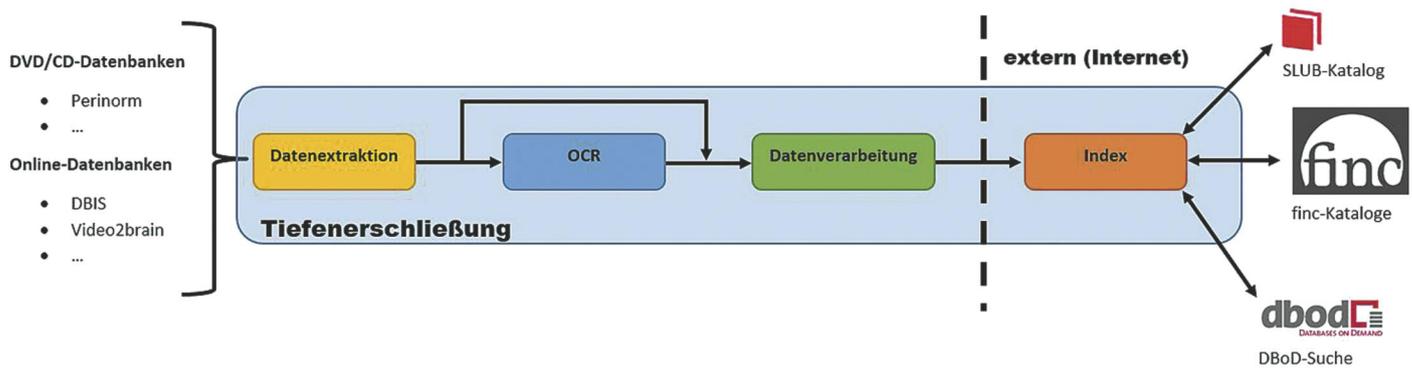


### Volltextsuche ermöglichen

Der Volltext spielt im Kontext der Tiefenerschließung eine bedeutende Rolle und ist eine weitere innovative Komponente im Projekt. Abbildung 4 zeigt, welche Dimensionen die Einbeziehung des Volltextes in den Katalog hat. Konnten für die Suche bislang weitgehend nur Metadaten wie Titel, Autor oder Verlag herangezogen werden, ist nun der vollständige Text im Katalogindex hinterlegt. Aus einer zweidimensionalen ( $x \cdot y = \text{Metadaten}$ ) wird eine dreidimensionale ( $x \cdot y \cdot z = \text{Volltext}$ ) Suche. Mit der Tiefenerschließung steht folglich eine Technik zur Verfügung, mit welcher die z-Achse, also die Tiefe, erschlossen werden kann.

Die Tiefenerschließung beschränkt sich allerdings nicht nur auf Volltexte. Alle Daten, die über die klassischen Metadaten hinausgehen, werden indiziert. Das ist wichtig, da es Datenbanken gibt, die keinen Volltext enthalten oder er technisch bedingt nicht indiziert werden kann. Beispiele sind Fachwörter-





bücher wie „DUBBEL interaktiv“ oder Formelsammlungen mit umfangreichen Wertetabellen. Die Vorteile, Volltexte zu indexieren, liegen klar auf der Hand. Fachbegriffe führen nun auch über den Katalog zu relevanter Literatur. Die Suche nach „Gussmassel“, einem Begriff aus der Aluminiumproduktion, führt zum Beispiel zu einem Treffer. Selbst das umfangreichste Nachschlagewerk der Welt, die Wikipedia, kennt dieses Wort nicht. Häufig ist es so, dass der Titel eines Werkes nicht unmittelbar auf den Inhalt schließen lässt. Schlagwörter waren bisher die Bindeglieder zwischen Titel und Inhalt. Volltexte können aber eine viel detaillierte Themenzuordnung herstellen.

Darüber hinaus hat die Indexierung von Volltexten noch einen ganz besonderen Effekt. Wenngleich die Tiefenerschließung noch keine Technik zur Semantisierung einsetzt, führen Anfragen, die der natürlichen Sprache nahe kommen, ebenso zu Treffern. Somit ist die Suche nach „Anforderungen zum Bauen eines Swimming Pools“ tatsächlich erfolgreich. Der geübte Suchmaschinenbenutzer hätte wohl eher „Schwimmbad Anforderungen Neubau“ eingetippt, weil er durch Erfahrung gelernt hat, dass eine starke Abstraktion bislang erfolgversprechender ist.

#### Tiefenerschließung:

##### Solitär oder ein Dienst unter Vielen?

Nun werden einige Leser zu Recht einwenden, dass es doch mittlerweile Produkte gibt, die genau die drei genannten Ziele umsetzen. Zu nennen sind hier etwa Primo Central von Ex Libris, Summon 2.0 von SerialSolution und EBSCO Discovery Service. Indes unterscheidet sich unser Projekt von den Big-Playern in vielerlei Hinsicht. Zunächst ist festzuhalten, dass die Tiefenerschließung ursprünglich mit dem Fokus auf die in Bibliotheken ungeliebten, aber vielfach noch unersetzten DVD/CD-Datenbanken gestartet ist. Die drei Discovery-Lösungen beschränken sich ausschließlich auf Online-Datenbanken, während die Tiefenerschließung neben den großen online verfügbaren Titeln eben auch DVD/CD-Datenbanken wie die Perinorm oder die Werkausgaben der „Digitalen Bibliothek“ von Directmedia Publishing indexiert. Deutlich im Nachteil sind die

Discovery-Lösungen daneben bei der Umsetzung der intuitiven Benutzerführung, da die Verschmelzung der Discovery-Frontends mit dem Bibliothekskatalog bislang nicht überzeugend gelingen will. Beispiele sind ein vom lokalen Katalog abgekoppeltes Relevanzranking oder eben die mangelnde Möglichkeiten zur Anpassung an das vorhandene Katalogsystem. Als Konsequenz entscheiden sich viele Bibliotheken für die getrennte Präsentation ihrer lizenzpflichtigen Inhalte.

Zusammenfassend ist festzuhalten, dass die Tiefenerschließung derzeit ein Alleinstellungsmerkmal genießt. Vor allem die direkte Integration der Volltextsuche in den Katalog ist wegweisend und öffnet ein neues Kapitel in der Welt der Kataloge.

#### Ein Blick hinter die Kulissen

Es bedarf mehrerer Stationen, bevor eine Information aus einer Datenbank in den Katalogindex gelangt. Die Abbildung oben veranschaulicht dies schematisch.

**Datenextraktion:** Im ersten Schritt müssen die Daten für eine Weiterverarbeitung aus den Datenbanken extrahiert werden. DVD/CD-Datenbanken sind ein sehr heterogenes Medium und unterscheiden sich im Aufbau und der zugrundeliegenden Technik. Dennoch können die Datenbanken grob in Gruppen geteilt werden. Da gibt es zum Beispiel Datenbanken, die überwiegend das PDF-Format einsetzen, andere basieren auf den HTML-Standard. Auch Online-Datenbanken durchlaufen diese Station. Zum Einsatz kommen bei diesen Datenbanken zunächst Webcrawler-Techniken, um die Inhalte aufzubereiten.

Für jede Datenbank ist für die Extraktion ein individueller Job erforderlich. Lediglich Teilaufgaben können nachgenutzt werden. Wichtig ist, dass schon während der Datenextraktion Routinen für ein späteres Update einbezogen werden. Denn der Index sollte immer den aktuellsten Stand der Datenbanken abbilden.

**Maschinenlesbarkeit:** Diese Station kann übersprungen werden, falls die extrahierten Daten bereits in einem Format vorliegen, die ein Rechner weiterver-

arbeiten kann. Anderenfalls werden OCR-Techniken angewandt, um die Maschinenlesbarkeit herzustellen. Im Projekt der Tiefenerschließung kommt die Open Source Software Tesseract zum Einsatz, eine Software, die von Google entwickelt wird und Teil des Digitalisierungsprozesses von Google Books ist.

**Datenverarbeitung:** In Vorbereitung auf die Indexierung werden die gewonnenen Daten durch zusätzliche Metadaten angereichert. Teilweise stammen diese Metadaten aus DBIS, teilweise auch von den Datenbanken selbst. Nur gepflegte Metadaten gewährleisten eine nahtlose Eingliederung in den Katalog durch eine einheitliche Facetten-, Treffer- und Detailanzeige.

**Indexierung:** Das Projekt der Tiefenerschließung verwendet einen einheitlichen Index, auf den alle teilnehmenden Einrichtungen zugreifen. Obwohl lediglich ein Index existiert, werden nur jene Datenbanken bei der Suche mit einbezogen, für die eine Einrichtung eine Lizenz erworben hat. Für jeden Volltext ist ein Einrichtungsbezug hinterlegt, wodurch eine Mandantenfähigkeit ermöglicht wird. Auch für den Index kommt eine Open Source Lösung zur Verwendung. In diesem Fall eine Solr/Lucene-Kombination.

#### Projektstand

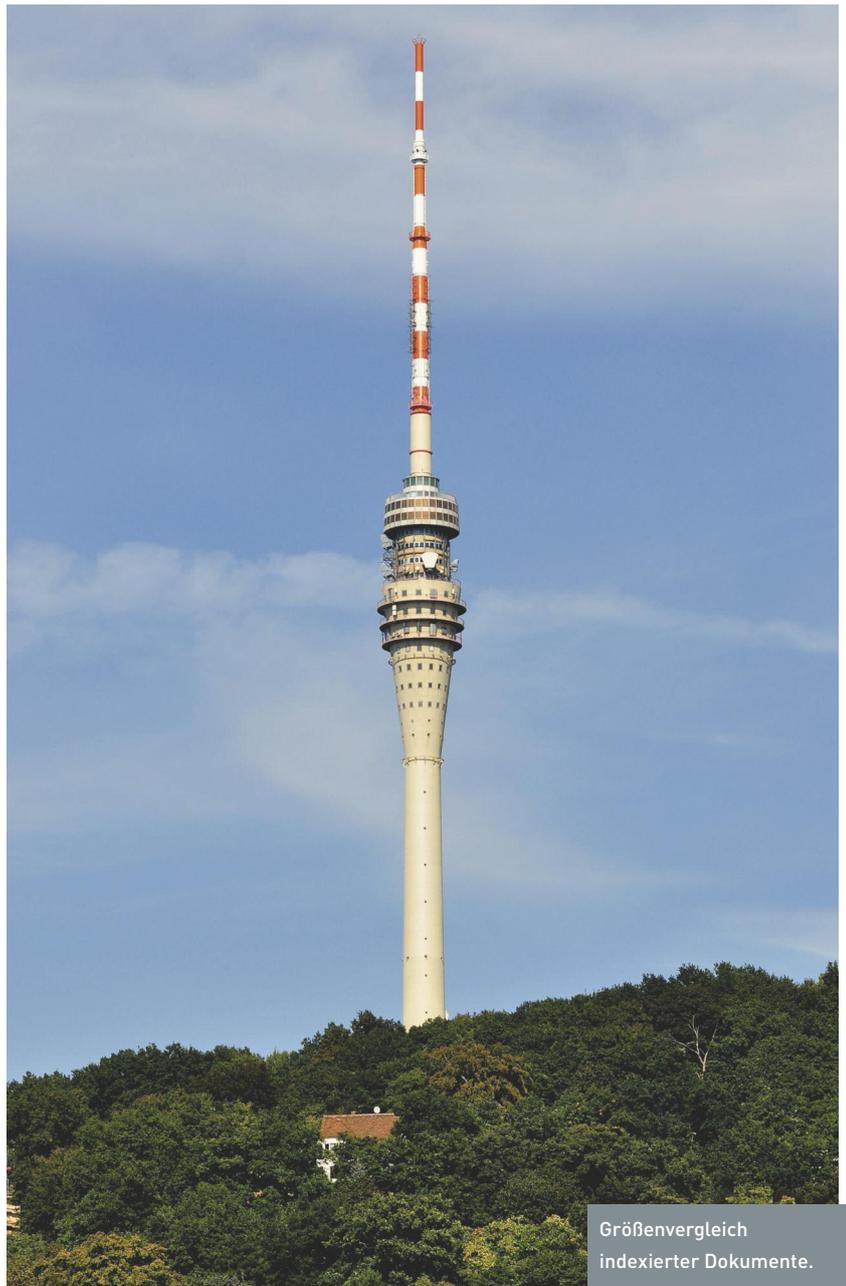
Über 350 Datenbanken konnten bisher erschlossen werden und sind nun über den Index abrufbar. Derzeit hervorzuheben ist die schon mehrfach erwähnte Normen-Datenbank Perinorm, in Umfang und Nutzung die Top-Datenbank auf DBoD. 23 von 26 teilnehmenden Bibliotheken lizenzieren die Datenbank und können damit schon heute die Vorteile der Tiefenerschließung nutzen.

Ein Größenvergleich soll veranschaulichen, welches Ausmaß die schon indexierten Volltexte einnehmen. Würden alle indexierten Volltexte einseitig auf normales A4-Papier gedruckt, entstünde ein Stapel von 250 Metern Höhe, etwa die Höhe des Dresdner Fernsehturms. Die Zahl zeigt eindrucksvoll, wie viel Inhalt in den DBoD-Datenbanken steckt und durch die Tiefenerschließung bereits ans Licht gebracht wurde. Faszinierend ist, dass es die Rechentechnik ermöglicht, ein beliebiges Wort, in diesem riesigen Stapel, in weniger als einer Sekunde zu finden.

Ein Meilenstein in diesem Jahr war die Integration der Tiefenerschließung in den SLUB-Katalog. Ein weiterer Meilenstein wird erreicht, wenn andere sächsische Hochschulbibliotheken jeweils ebenfalls direkt mit ihren Katalogsystemen angeschlossen werden. Eine Schnittstelle für das Discovery-Frontend „finc“ soll daher als nächstes folgen.

#### Nachhaltigkeit

Trotz aller Unkenrufe ist die DVD/CD im Datenbanksegment noch lange kein Auslaufmodell. Zwar werden immer mehr Datenbanken online publiziert,



Größenvergleich  
indexierter Dokumente.

jedoch vollzieht sich dieser Prozess langsamer als erwartet. Wie gezeigt, eignet sich das Konzept der maschinellen Tiefenerschließung zudem für die Adaption beliebiger digitaler Ressourcen. Das Ziel bleibt immer dasselbe: Bereitstellung einer Volltextsuche im Katalog.

Sowie alle wesentlichen Datenbanken bearbeitet sind, ist vorgesehen, die Open access publizierten Dokumente aus dem sächsischen Hochschulschriftenserver Qucosa (<http://www.qucosa.de>) zu indexieren. Soweit die Dokumente der retrodigitalen Sammlungen der SLUB mittels Texterkennungssoftware qualitätsvolle maschinenlesbare Texte liefern, sollen auch diese indexiert werden.

Fragen zu DBoD oder der Maschinellen Tiefenerschließung sind unter [info@dbod.de](mailto:info@dbod.de) stets willkommen. Auch über Hinweise und Anregungen freuen wir uns.



FALK  
NIEDERLEIN