

Überblicks-Skript „Verbale Sacherschließung im Kontext“ (Stand: April 2014)

Bisher wird durch RSWK und GND (früher „SWD“) ein erheblicher Mehrwert geschaffen.

- Basis: Autopsie/Inhaltsanalyse von Dokumenten, Erkennen und Herausarbeiten von Themen, deren Wiedergabe in der Dokumentationssprache
- Deutschsprachige verbale Sucheinstiege
- Gebräuchliches, laufend aktualisiertes Vokabular (Nachschlagewerke, Kompromiss zwischen Fachsprachen und Allgemeinsprache)
- kontrollierte Terminologie (Thesaurusstruktur)
- Schlagwörter und ihre einzelnen Bestandteile sind in beliebiger Kombination suchbar
- Zusätzlich präsentieren Schlagwortfolgen die Schlagwörter in geregelter Anordnung (eingängige „syntaktische“ Muster) und damit rasch erfassbare „Kurz-Abstracts“
- Tiefe Durchdringung der deutschen Katalogdaten mit RSWK-Schlagwörtern (Anwendung seit 1986, auch für ältere Bestände, in Universalbibliotheken, aber auch vielen Spezialbibliotheken und -sammlungen)
- Sehr gutes Fremddatenangebot durch kooperative Anwendung im ganzen deutschsprachigen Raum (in jedem Verbund mindestens eine Reihe wichtiger Bibliotheken), DNB für Reihe A der Nationalbibliografie (mit Ausnahmen)
- Die Online-Kommunikation aller Bibliotheken mit der Normdatei schafft eine breitere Basis für die Normdatenverantwortung und -pflege
- Normdaten bieten diverse Sucheinstiege, Selektions- und Navigationsmöglichkeiten (Synonyme, Relationen, Codierungen)
- Identifizieren und eindeutige Beziehungen zwischen Entitäten qualifizieren Normdaten als Bestandteile des semantischen Netzes
- Normdaten werden sehr aktuell gehalten (z.B. rasche Einführung neuer Begriffe und Benennungen, Namensänderungen, Aktualisierung von Lebensdaten und -stationen)

Bestimmend für die aktuellen Veränderungen in der (Sach)Erschließung sind:

- Internationalisierung von Regeln und Formaten (Aufgabe deutscher Sonderlösungen; optimale Interoperabilität von Daten, Anschluss an Weiterentwicklungen mit breiter Relevanz),
- Das Internet ist zunehmend Ort der Publikation und Suchraum; die neue Anforderung lautet, Linked Data für das inhaltliche Retrieval bereitzustellen.
- Veränderte Benutzergewohnheiten
 - o einerseits die von Google und Mobilgeräten geprägten Recherchegewohnheiten der „breiten Masse“ (1. „one-click“, 2. vom reinen Text zu multimedialen Elementen, 3. Möglichkeit zur Personalisierung und Interaktion, Web 2.0-Elemente)
 - o andererseits immer höhere Anforderungen an die wissenschaftliche Information (1. wachsende Informationsmenge, 2. immer differenzierteres Wissen)
- Es gibt neue technische Möglichkeiten (z. B. maschinelle Indexierung, virtuelle Datenclustering, Visualisierungstechniken, tief in die Volltexte reichende Discovery-Suchindizes)

Der heutige Kontext der verbalen Inhaltserschließung ist geprägt von oft unvermeidbarer Heterogenität.

In Bibliothekskatalogen ist Konsistenz noch in gewissem Ausmaß möglich, aber in großen Suchräumen sind Metadaten zunehmend heterogen. Dokumente sind z. T. gar nicht oder unterschiedlich mit Metadaten erschlossen durch

- kontrolliertes Vokabular (Normdateien, Fachthesauri)
- freies Vokabular, „Folksonomies“
- klassifikatorische Erschließung mit verbalen Elementen
- Abstracts

- Kataloganreicherung (v.a. Inhaltsverzeichnisse)

- Maschinell erzeugte Indexate

(dies alles ggf. in verschiedenen Sprachen)

Neben Bibliotheken steuern Metadaten bei: Autoren, Verlage, Fachinformationseinrichtungen, andere Dienstleister und Endnutzer.

Mit der Verbreitung der **Volltexterschließung** stellt sich stärker die Frage nach dem besonderen Nutzwert von Metadaten gegenüber der inhaltserschließenden Aufbereitung von Volltexten.

Heterogenitätsbehandlung ist deshalb eine wichtige Aufgabe der künftigen Retrievaltools.

- Künftig wird es mehr Mischformen geben (kontrollierte / freie Terminologie, zugeteilte / extrahierte Indexate, intellektuelle / maschinelle Indexierung). Auch „flache“, oberflächliche und natürlichsprachige Zugriffspunkte verlangen nach Auswertung und müssen mit kontrolliertem Vokabular und Normdaten interoperabel gemacht werden.
- Zur Veranschaulichung dient das Schalenmodell nach Krause.
- Ein qualitativ gut, konsistent erschlossener Kernbereich wertet auch die Erschließung der übrigen Dokumente auf. Er bietet die Basis für statistische und probabilistische Verfahren der maschinellen Erschließung. Die vorhandenen Metadaten nach RSWK/GND (vorwiegend bei Printpublikationen, aber auch elektronischen Ressourcen) sind von Qualität und Menge her sehr geeignet, um in großen Suchräumen mit heterogener (oder fehlender) Inhaltserschließung (vorwiegend dort, wo Internetressourcen einbezogen sind) virtuelle Konsistenz zu schaffen und den Sucherfolg zu verbessern.

Normdaten erleichtern maschinelles Indexieren

- Maschinelles Indexieren = Verzicht auf intellektuelle Inhaltsanalyse, um große Mengen von Dokumenten inhaltlich so gut wie möglich erschließen zu können (vor allem Internetressourcen)
- klassische Verfahren der linguistischen Datenverarbeitung: Parsing, Lemmatisierung usw., Basis sind große Wörterbücher, die gepflegt werden müssen (Mehrsprachigkeit der Dokumente als Problem)
- Beispiele für die „klassische“ Methode: IDX (MILOS), AUTINDEX
- statistische und probabilistische Verfahren: große Mengen guter Erschließungsdaten als Basis, d.h. in der Regel nur sinnvoll, wenn auf intellektueller Erschließung in einem Datenbestand aufgesetzt werden kann. Beste Grundlage im verbalen Bereich: Normdaten
- Projekt PETRUS der DNB, Teilprojekt automatische Schlagwortvergabe (Software Averbis, zunächst für elektronische Dissertationen): noch kein Routineeinsatz
- Intellektuelle Inhaltsanalyse ist vielfach nicht ersetzbar.

Anforderungen an Recherchertools (OPACs usw.)

- Die Techniken zur Unterstützung des inhaltlichen Retrievals haben i.A. mit der Erschließung nicht Schritt gehalten. Mit verbesserten Recherchertools ist eine bessere Auswertung der Normdaten und Schlagwortfolgen möglich.
- Die klassische Schlagwortsuche führt in einen konsistent erschlossenen Kernbereich. Hier war früher häufig eine zweistufige Suche möglich bzw. üblich: Anzeige 1. einer Liste relevanter Indexate zur Vorauswahl, 2. der Dokumente zu den vorausgewählten Indexaten. Dokumente ohne Schlagwörter blieben unentdeckt.
- Daneben tritt zunehmend die Strategie der „schmutzigen Suche“ in einem umfassenden Index im ersten Schritt mit anschließender Suchverfeinerung im zweiten Schritt.
- Nutzer brauchen an wichtigen Stellen im Rechercheprozess eine knappe Inhaltsbeschreibung einzelner Dokumente. Bei Bedarf muss auch eine menüartige Auswahl aus relevanten Themen möglich sein. Die Entitäten, auf die es dabei ankommt, sind die

Themen (die durch Indexate repräsentiert werden), nicht die Formalbeschreibungen einzelner Dokumente.

- Die Indexate der verbalen Sacherschließung müssen also weiterhin eine konzise Inhaltsbeschreibung (Präsentation) ermöglichen und einzelne Sucheinstiege bieten, aber auch für die Suchverfeinerung (Drill down) geeignet sein.
- Noch zu wenig beachtet: das Navigieren in einem thematischen (nicht bloß sprachlich-begrifflichen) Umfeld
- Eine Herausforderung ist es, für den Benutzer Transparenz zu schaffen (wie kommen Suchergebnisse zustande?), ohne den Recherchevorgang zu komplizieren und zu hemmen.

Wirtschaftlichkeit

- Die RSWK-Fremddatennutzung wurde / wird im Titelbereich optimiert. Es ist nur noch wenig Eigenleistung erforderlich (zumindest bei Universalbibliotheken).
- Die Normdatei ist kooperativ geführt, elaboriert und sehr aktuell. Auch hier fällt immer weniger Eigenleistung an (zumindest bei Universalbibliotheken).
- Die Konzentration auf inhaltliche Schwerpunkte von lokalem Interesse ist möglich (Sammelschwerpunkte, regionale Literatur); dort kann mehr Aufwand investiert und ggf. tiefer erschlossen werden. Bei den DFG-Fachinformationsdiensten (ehem. SSG) wird eine gute Inhaltserschließung erwartet.
- Die Vereinfachung und Liberalisierung der RSWK ab 2010 bedeutet einen deutlichen Einschnitt. Heterogenität innerhalb der RSWK-Daten wird in gewissem Maß akzeptiert, auf Ausnahmen wird zu Gunsten einfacher Grundregeln weitgehend verzichtet, der Erfassungsaufwand sinkt.

RSWK und GND werden zunehmend im internationalen Kontext gesehen: LCSH, RDA

- o Library of Congress Subject Headings (LCSH): Präkombination in z.T. komplexen Headings = starr, oft zu grob, Ergänzung/Aktualisierung umständlich, keine Thesaurusstruktur, kein konsequent kontrolliertes Vokabular, Synonyme und Relationen nur in Browsinglisten darstellbar
- o RSWK / GND: Präkoordination (Schlagwortfolgen aus kleinteiligem Vokabular), Verknüpfung der Einzelelemente aus der Normdatei, Zerlegungskontrolle, vielfältige Kombinationsmöglichkeit einzelner Schlagwörter, Thesaurusstruktur, Linked-data-Eignung = eindeutig das fortschrittlichere System.
- o Konvergenz? Es herrscht durchaus Unzufriedenheit mit den LCSH in den USA → evtl. bewegt man sich hin zu unserem System (aber gebremst durch die vielen vorhandenen LCSH-Daten in vielen Katalogen)
- Im Regelwerk RDA sind auch die Entitäten definiert, die hauptsächlich für die Sacherschließung relevant sind (object, concept, event, place; Regeln noch nicht ausgeführt). Bisher (Stand 2014) ist nicht daran gedacht, die LCSH oder andere SE-Verfahren durch etwas Anderes abzulösen.

Es gibt ein bisher noch wenig genutztes Potenzial der Normdaten. Im Kontext der Linked data-Ansätze dürfte sich hier aber noch einiges bewegen.

- Auswertung der Codierungen (z.B. Zeitcode → Recherchekriterium „behandelte Zeit“)
- Georeferenzierung mittels Koordinaten, Einbindung von Kartendarstellungen
- Crosskonkordanzen zu Fachthesauri und zu Klassifikationen (DDC, Projekt CrissCross); auch an die Nutzung der Normdatei für Konkordanzen zwischen Klassifikationen ist zu denken.
- Die multilinguale Verlinkung mit anderssprachigen Normdaten (MACS-Projekt) wird weiter ein Weg sein, um einen eigensprachlichen Zugang zu fremdsprachlich erschlossenen Dokumenten zu schaffen.

- Im Kommen: Erschließung webbasierter Nachschlagewerke durch Normdaten-Permalinks (Beispiele: Wikipedia, digitale Lexika der BSB)
- Wichtig für Aktivitäten auf der Basis von Normdaten ist die Vereinheitlichung der Ansetzungsformen für Formal- und Sacherschließung in der Gemeinsamen Normdatei (GND). Die konsequente Relationierung der Datensätze für bestimmte Entitäten ertüchtigt die GND, um eine tragende Stütze des Semantic Web zu werden.
- Normdaten fügen sich in die Datenmodelle FRBR (für bibliographische Daten) und FRSAID (für themenbezogene Normdaten) ein.
- VIAF (viaf.org): Virtuelle internationale Normdatei (bisher Personennamen und Geographika). Eine Entität mit weltweit gültiger ID-Nummer hat verschiedensprachige Benennungen, über die jeweils ein und derselbe Normsatz angesteuert und die bei der Präsentation selektiv gezeigt werden können. VIAF ist eines der zukunftsträchtigsten Projekte der Internationalisierung, jenseits aller Bemühungen um „internationale Regelwerke“. Seit 2012 im Rahmen von OCLC institutionalisiert.

Nutzung von Normdaten für das Semantic Web (Linked Data)

- Es lohnt sich, die vorhandenen semantischen Strukturen für die inhaltliche Erschließung von Ressourcen im Internet nutzbar zu machen. Wegen der hervorragenden Ausgangsbasis kann damit ein stabiler und qualitätvoller Beitrag zur „Linked Data Cloud“ geleistet werden.
- Dies ermöglicht die Etablierung neuer Dienstleistungen und kann zugleich ein wirkungsvolles Instrument der Heterogenitätsbehandlung sein.
- Notwendig ist die Veränderung textueller Strukturen in der Normdatei in eine Relationierung (maschinelle Auswertbarkeit von Entitätsbeziehungen) mit der Umsetzung in RDF.
- Dies betrifft nicht nur Normdaten, sondern auch die damit verknüpften Titeldaten. Deshalb besteht eine enge Verbindung mit der Open Data -Bewegung. Ein ähnlicher Effekt wird mit der Teilnahme von Bibliotheken am WorldCat erzielt. Bibliothekarische Metadaten werden im Web sichtbar, auch in neuer Gestalt und in nicht-bibliothekarischen Kontext
 - o Zumindest die Normdaten lösen sich allmählich vom primär bibliographischen Kontext und Zweck. Bibliothekarische Metadaten werden auch für neue Zielgruppen (nicht nur für Bibliotheksbenutzer) interessant.
 - o Barrieren für eine breite Mitarbeit an den Normdaten müssen weiter abgebaut werden, ohne qualitative Standards in Frage zu stellen.
- Reich- und Tragweite der bibliothekarischen Arbeit des einzelnen Indexierers werden immer größer. Der kooperative Gedanke und das Bewusstsein, mit der früher nur dem eigenen Bestand gewidmeten Inhaltserschließungsarbeit jetzt der Vernetzung des Wissens auf der ganzen Welt zu dienen, müssen gefördert werden.

Fazit

Die intellektuelle verbale Inhaltserschließung, gezielt weiterentwickelt und eingesetzt, kann somit
 → einen wesentlichen Beitrag zur künftigen Informations-Infrastruktur und
 → einen wesentlichen Beitrag zur Positionierung der Bibliotheken in der Wissensgesellschaft leisten.

Bibliothekarinnen und Bibliothekare mit Fachwissen sind dabei mehr denn je gefragt.

Dr. Gerhard Stumpf
 Stand: April 2014