Thesen zu Rahmenbedingungen und Zielen der verbalen Sacherschließung für die nächsten 3-5 Jahre

Dr. Gerhard Stumpf (Stand: 12.9.2011)[1]

Konsistenzerhaltung im heterogenen Umfeld

Erschließungs- und Suchräume mit heterogenen Metadaten werden zum Regelfall. Konsistenz und Homogenität der Suchergebnisse sind als Teil des Benutzerkomforts jedoch weiterhin anzustreben, sei es durch Datenaufbereitung, sei es durch virtuelle Konsistenzerhaltung im Suchprozess und bei der Präsentation der Ergebnisse.

Der wichtigste Mehrwert in Metadaten der verbalen Inhaltserschließung liegt – neben dem Eröffnen von Sucheinstiegen - in der Möglichkeit, einerseits **thematisch Gleiches und Ähnliches zusammenzuführen** und andererseits **Unterschiedliches unterscheidbar und selektierbar** zu machen. An den wichtigen Stellen muss auf Qualität und Exaktheit geachtet werden, um diese Anforderungen erfüllen zu können.

Das Ziel der Sacherschließung muss sein, dass Benutzer nicht nur „etwas" finden, sondern auf eine thematische Suche **bereits im ersten Schritt möglichst viele und präzise Ergebnisse** erhalten können. Oft werden nicht nur relevante Dokumente, sondern (annähernd) vollständige und verlässliche Ergebnissets benötigt. Darüber hinaus sind aber Angebote zur **Einschränkung, Erweiterung und zu weiteren Entdeckungen** unverzichtbar.

Zur Beschreibung des heutigen Szenarios der inhaltlichen Erschließung im bibliothekarischen Umfeld eignet sich nach wie vor das **Schalenmodell**[2]: Abnehmend nach Menge und Qualität der inhaltsrelevanten Metadaten gruppieren sich in einem bibliographischen Datenbestand konzentrische Datenschichten um einen relativ am besten erschlossenen Kernbereich. In diesem konzentrieren sich die Dokumente, die mit Autopsie und normdatengestützt intellektuell und relativ konsistent erschlossen wurden. Statistische und probabilistische Verfahren (maschinelle Indexierung, aber auch die virtuelle Clusterung der Metadaten von Dokumenten), die zur Erschließung der übrigen Schalen eingesetzt werden sollen, haben nur dann Aussicht auf Erfolg, wenn sie auf eine größere Menge von Daten aus dem Kernbereich zurückgreifen können.[3] Dabei setzen im verbalen Bereich gute Suchergebnisse mit aktuellem Suchvokabular die laufende Aktualisierung und Anwendung eines kontrollierten Indexierungsvokabulars voraus.

---

[1] Impulse für diese Standortbestimmung kamen u. a. aus Vorträgen des Dt. Bibliothekartags Berlin 2011. Einige weiterführende Links:
Wiesenmüller, Heidrun: Den Kern erhalten - Qualität an der richtigen Stelle (BT Berlin 2011, Vortrag im Block "Die Zukunft der Katalogisierung") http://www.opus-bayern.de/bib-info/volltexte//2011/1000/
Wiesenmüller, Heidrun: Zwischen Wunsch und Wirklichkeit - Bibliotheksdaten und Bibliothekskataloge. 5 Thesen (Impuls-Referat im Rahmen der VDB-Mitgliederversammlung auf dem Bibliothekartag) http://www.vdb-online.org/wordpress/wp-content/uploads/2011/06/Vortrag-Wiesenmueller.pdf
Weitere Erschließungs-Beiträge vom Berliner Bibliothekartag:
http://www.bib-info.de/verband/publikationen/opus/berlin-2011/vortraege-nach-themenfeldern/erschliessung.html

[2] Krause, Jürgen: Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung ("Schalenmodell"). Bonn 1996. (Quelle am 28.6.2011: http://www.gesis.org/publikationen/archiv/iz-arbeitsberichte/abstracts/#c3193). - Überarbeitet in: Krause, J., Shell Model, Semantic Web and Web Information Retrieval. In I. Harms, H.-D. Luckhardt & H. W. Giessen (Eds.), Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. Festschrift für Harald H. Zimmermann (pp. 95–106). München 2006: K. G. Saur.

[3] Auf Grund der bisherigen Erfahrungen liefern automatische Indexierungsverfahren nur dann gute Ergebnisse, wenn sie auf inhaltserschließenden Metadaten, besonders Normdaten, aufbauen können, die in der Regel einer intellektuellen Inhaltsanalyse entstammen. „Klassische Erschließungsziele wie Zusammenführung von Gleichem, vollständiger Nachweis, zuverlässige und einheitliche inhaltliche Suche sind nur durch Erschließung, nicht durch automatische Indexierung zu erreichen. - **Allgemein sind die Grenzen automatischer Erschließungsverfahren**

Heterogenitätsbehandlung zielt dort, wo trotz zunehmender Heterogenität große Mengen konsistenter Metadaten vorhanden sind, auf **Konsistenzerhaltung**. Entscheidend ist dabei, im Sinn des Schalenmodells durch die Kombination verschiedener Verfahren für die Suche Konsistenz in dem Maß zu erzeugen, wie sie für den Sucherfolg benötigt wird, und die in den Metadaten teilweise bestehenden Inkonsistenzen zu glätten.

Paradigmenwandel auf dem Weg zum Semantic web

Im Kontext der gesamten Erschließung ist wohl von einer zunehmenden **Konvergenz von Formal- und Sacherschließung** auszugehen, da sie sich verändern wird von der Katalogisierung eher statisch-deskriptiver, konstruierter Container („Titelaufnahmen“) hin zum Datenmanagement für kleine Einheiten zur Adressierung von Dingen der wirklichen Welt („Ressourcen“).

Die Erstellung und Pflege von **Normdaten** rückt also noch stärker in den Fokus. Deren Zweckbestimmung wandelt sich auch. So gibt es hier Akzentverschiebungen, vereinfacht dargestellt etwa diese:

Früher waren am wichtigsten:	**Künftig sind wichtiger:**
Dokumente (Publikationen) als „Vorlagen“, aus denen bestimmte Angaben abgeschrieben werden (FE); Inhaltsanalyse als getrennter Arbeitsgang (SE)	Entitäten (Basis: FRBR), die identifiziert und in Relation gesetzt werden, woraus dann weitgehend die bibliographische Beschreibung samt Inhaltserschließung entsteht
Eintragungen in Indizes für Stichwortsuche und alphabetisches Browsing	Relationen, dargestellt als Links zum Navigieren
Ansetzungsformen (als Anzeigeformen verstanden)	Identifier (als datentechnische Anker für Entitäten verstanden)
Ordnungsprinzip primär alphanumerisch, daher Tendenz zur Zusammenführung des Ähnlichen an einer Alphabetstelle	Ordnung nach Facetten, also Entitäts- und Relationstypen, bzw. nach Benutzerkriterien (Häufigkeit, Relevanz)
Leitform: Katalog, Verzeichnis, Index	Leitform: Netz
Der Bibliothekskatalog als abgeschlossener Referenzrahmen / Suchraum, auf den allein sich Konsistenz bezieht	Bibliothekskataloge sind nur noch Knoten im größeren Netz; Konnektivität tritt als Ziel neben Konsistenz; zur Herstellung von Konsistenz können diverse Ressourcen herangezogen werden.
Vorrangig ist eine menschenlesbare, sequenzielle Darstellung (ISBD)	Vorrangig sind maschineninterpretierbare Daten (Linked data), aus denen eine menschenlesbare Anzeige generiert wird
Die SWD als eine Art Universalthesaurus für den deutschsprachigen Raum, fast autonom geregelt durch die RSWK in Definition, Verwendung und Verknüpfung der	Global gültige Entitäten und Identifier (VIAF) mit im deutschsprachigen Raum gebräuchlichen Vorzugsbenennungen im Rahmen eines international angewandten Rahmenregelwerks

dort erreicht, wo die Intelligenz beginnt. - Eine Entscheidung für automatische Indexierung muss das Wissen um deren Leistungsfähigkeit insb. deren Grenzen der Leistungsfähigkeit berücksichtigen. - Einer Entscheidung für automatische Indexierung sollte daher eine Zielbestimmung hinsichtlich der gewünschten Retrievalmöglichkeiten jetzt und in 10 Jahren vorausgehen. Entscheidungen gegen eine Erschließung sind nur mit erheblichen Konsistenzverlusten umkehrbar.“ (Klaus Lepsky, http://www.iws.fh-koeln.de/institut/personen/lepsky/skript-3d-automatische-indexierung-linguistik-statistik-06.pdf , Folie 12)

Schlagwörter. Sukzessive wurde die SWD für Online-Kataloge der 1. Generation tauglich gemacht.	(RDA), wobei nur für deren konkrete Verwendung und Kombination noch spezifische SE-Regeln nötig sind

Die im Aufbau befindlichen internationale Gesamt-Normdatei **VIAF**[4] kann in Zukunft große Bedeutung erlangen, vielleicht größere als vielen Vereinheitlichungen und Internationalisierungen im Bereich Regelwerk und Datenformat.

Ziele

Die von der AG Strategie der Sacherschließung des Standardisierungsausschusses 2008 formulierten Anforderungen legen – neben der Vereinfachung des Regelwerks – den Akzent einerseits auf das quantitative Datenangebot, andererseits auf die bessere Nutzbarmachung der Daten für das Retrieval:

Um das derzeit als unbefriedigend angesehene Kosten-Nutzen-Verhältnis der RSWK-Erschließung zu verbessern, muss auf mehreren Ebenen zugleich angesetzt werden. Anzustreben sind folgende Ziele:

1. Wirtschaftlichere Anwendung der RSWK für die Bibliotheken
2. Erhöhung des Anteils von Titeln mit verbaler Sacherschließung gemäß RSWK
3. Verbesserte Nutzbarmachung der RSWK-Erschließung für das Retrieval und Einbezug nicht verbal sacherschlossener Titel in ein übergreifendes Retrieval
4. Profilierung von RSWK/SWD als Sacherschließungswerkzeug auch außerhalb des bibliothekarischen Bereichs

Um die genannten Ziele umzusetzen, sind folgende Maßnahmen erforderlich:

- Bessere Nutzbarmachung der Schlagwortketten für das Retrieval
- Bessere Nutzbarmachung des strukturierten Suchvokabulars der SWD für das Retrieval
- Vereinfachung und bessere Präsentation des Regelwerks in Hinblick auf die Bedürfnisse der verschiedenen Anwendergruppen
- Vereinheitlichung bzw. Wiederannäherung bisher abweichender Anwendungen („Schlagwortreihen“, Einzelschlagwörter sowie Sonderregeln für Altkarten, VD 17, RSWK Musik)
- Pragmatischer und liberaler Umgang mit Heterogenität, um einen möglichst hohen Nachnutzungsgrad zu gewährleisten
- Werbung und Offenheit für die Anwendung der SWD außerhalb des engeren bibliothekarischen Bereiches[5]

Retrieval / Benutzerkomfort

Allgemein treten neben die kataloginternen Retrievalaspekte zunehmend weiterführende Angebote, z. B. im Bereich „Katalog 2.0“. Zweifellos sind auch die Ansprüche durchschnittlicher Bibliotheksbenutzer an Intuitivität und Komfort der Suchinstrumente gestiegen.

Für die verbale Inhaltserschließung sollten folgende Grundsätze des **Benutzerkomforts** beachtet werden:

- Benutzer benötigen weiterhin verbale Sucheinstiege mit einer **im Deutschen gebräuchlichen Terminologie** (allgemeinsprachlich und fachsprachlich).

[4] http://viaf.org/

[5] https://wiki.bsz-bw.de/lib/exe/fetch.php?media=v-team:katalogisierung:normdaten:sacherschliessung:ag_strategie_sacherschliessung_abschlussbericht.pdf

- Normdaten einerseits, nicht normierte Indexate sowie aus Volltextindexierung (auch von Abstracts) gewonnene Sucheinstiege andererseits sind **unterschiedliche Datenschichten** und dürfen nicht so vermengt werden, dass der spezifische Mehrwert von Normdaten verloren geht. Die Verlässlichkeit kontrollierten Vokabulars darf nicht durch andere verbale Elemente beeinträchtigt werden. Für ein Ranking sollen Normdaten immer größeres Gewicht haben.
- Die funktionalen Unterschiede zwischen bibliothekarischen u. a. Metadaten einerseits und den aus Volltexten (Abstracts, Kataloganreicherung) gewonnenen Indexaten andererseits sind umso stärker zu berücksichtigen, je mehr die Recherche auf Volltextindexierung zurückgreifen kann.
- Sucheinstiege über fremdsprachige, semantisch zutreffende Äquivalente sollen möglich werden durch **Verlinkung mit anderssprachigem Normvokabular**.
- **Crosskonkordanzen** sollen u. a. die Sucherweiterung auf nur klassifizierte Ressourcen und die Nutzung der Benennungen aus Klassifikationen als Sucheinstieg ermöglichen.
- Wichtig sind **optimierte Anzeigeformen und eine intuitive Benutzerführung** (Klartext, Menüs; Navigationsschritte verständlich anbieten). Ein „Google-like"-Design ermöglicht einen niedrigschwelligen Erstzugang, ist aber bei weitem nicht ausreichend, damit der gebotene Mehrwert für das Retrieval zum Tragen kommt.
- Es sollte möglich sein, bei Bedarf auch komplexe Suchanfragen zu stellen.
- Die **facettierte Navigation** (Drill down) des Suchmaschinenkatalogs stellt ein intuitives und interaktives Werkzeug dar, das den Benutzererwartungen entspricht und deshalb in effektiver und kreativer Weise ausgebaut werden sollte. Der iterative Prozess der Suchverfeinerung geht mit einer gedanklichen Analyse des Rechercheproblems und einer Grobanalyse der verfügbaren Ressourcen durch den Benutzer einher. Es ist daher wichtig, die facettierte Navigation auf kontrollierten Metadaten aufzubauen und die Navigatoren von störenden Anzeigen frei zu halten.
- Die **Anzeige von Datenelementen aus Normdatensätzen** auch für Benutzer ist sinnvoller denn je. Zum Benutzerkomfort gehört auch die optionale Möglichkeit, durch Einblick in Normdaten die erzielten Rechercheergebnisse transparenter und künftige voraussehbarer zu machen.
- Die auch in der GND verankerten **Schlagwort-Hinweissätze** [die keine eigenständigen Elemente des semantischen Netzes sind] haben den Zweck, Benutzer darauf hinzuweisen, dass bestimmte Suchanfragen nicht direkt zum Ziel führen, ohne dass eine automatische Umlenkung der Suche möglich wäre. Aus diesem Grund müssen sie beim Retrieval über Schlagwörter auch angezeigt werden (ein uraltes Desiderat; neue Techniken ermöglichen evtl. neue Lösungsansätze).
- Eine konsequentere Normdatennutzung ermöglicht **neue Techniken des Browsing und des entdeckenden Suchens** (stärker kontextsensitiv, Suchraumerweiterung, Expansion). Den heutigen Erkenntnissen über kognitive Prozesse und den nachhaltigen Wissenserwerb entsprechend, sollen dem Benutzer über die mit jeweils in einem Datenbestand vorhandenen Titelsätzen verknüpften Normdaten hinaus weitere Metadaten, das ganze semantische Netz oder Teile davon zugänglich werden.
- **Neue Features**, die mit relativ wenig Aufwand zur besseren Nutzung der vorhandenen Daten führen, könnten z.B. sein:
    - Autovervollständigen im Suchfeld auf der Basis von Schlagwort-Ansetzungsformen bzw. individualisierten Personennamen
    - In Suchmaschinenkatalogen: Beim Drill down neue Facetten auf der Basis von Zeitcode (thematisierte Zeit) und Ländercode (thematisierte geographische Räume)
    - In Suchmaschinenkatalogen: Mehrfach- statt Einfachauswahl beim Drill down
    - Beim Ranking gezieltere Berücksichtigung SE-spezifischer Parameter

    - Wahrnehmungsfreundliche Visualisierung, z. B. Darstellung verwandter bzw. im Kontext stehender Schlagwörter als Tag clouds
    - Mashups bzw. die Einbindung externer Webdienste mit APIs, um raschen Zugriff auf weiterführende Informationen oder die Weiternutzung von Ergebnissen in anderen Diensten zu ermöglichen
- Inhaltserschließende Daten müssen künftig von Menschen und von Maschinen interpretiert werden können (**Linked data** als Ziel). Dabei ist nicht alles Maschinenlesbare unmittelbar als Anzeigeinformation bestimmt, kann aber indirekt Sucherfolg und Benutzerkomfort steigern.
- Die deskriptive Funktion sequenziell lesbarer Indexate bleibt wichtig; diese sind jedoch nicht mehr statisch zu verstehen; letztlich wird alles verteilt sein und für die Präsentation zur (menschlich-kognitiven) Relevanzbeurteilung dynamisch zusammengezogen (aggregiert) werden können/müssen.
- Die tiefe Durchdringung des Kernbereichs unserer Datenbanken mit RSWK-Daten ist ein enormer Mehrwert, auch wenn diese einer Zeit entstammen, als Linked data noch kein Ziel bibliothekarischer Erschließung waren. Wo ein Thema nur durch mehrere auf einander folgende Schlagwörter repräsentiert und vom Benutzer erkannt werden kann, ist dafür zu sorgen, dass die **Schlagwortfolge** in den hierfür geeigneten Kontexten (Kurztitelanzeige, Detailanzeige, Auswahlmenüs, Listen, Drill downs) in regelmäßiger und verständlicher Form dargestellt wird.

Verbesserung der Datenbasis

Viele dieser Verbesserungen auf der Retrieval-/Outputseite setzen die weitere Verbesserung des Inputs bzw. des Datenangebots voraus. Hierbei müssen vor allem maschinelle **Verfahren zur Anreicherung der Daten** genutzt werden.

Priorität hat der konsequente Datenaustausch zwischen den Verbünden; dringend notwendig sind Vereinbarungen über den zuverlässigen Austausch bestimmter Metadaten in bestimmten Feldern des Datenformats. Insbesondere müssen die Felder für RSWK-Daten sowie für bestimmte Klassifikationen konsequent für diese reserviert bleiben. Erkennbare Tendenzen zur Datenschlamperei sind kontraproduktiv.

Die Inhaltserschließung einer Manifestation eines Werks in einer Datenbank sollte durch geeignete Verfahren auf alle anderen Titelsätze übertragen werden, die sich auf das gleiche Werk beziehen, zumindest für wissenschaftliche Literatur (Methode Pfeffer / SWB-HEBIS). Für Parallelpublikationen gedruckt – elektronisch gilt dies besonders, gerade im Hinblick auf Neuerscheinungen.

Besonderes Augenmerk verdienen die Werke, die noch in keiner Ausgabe inhaltlich erschlossen wurden. Im B3Kat ist die Erschließungssituation befriedigend bis gut bei

- Buchhandelsliteratur der letzten Jahrzehnte,
- Beständen der Spezialbibliotheken im Verbund,
- SSG-Literatur,
- grauer Literatur und Aufsatzliteratur der Regionalbibliographie.

Die quantitative Vermehrung der Inhaltserschließung im B3Kat sollte insbesondere auch einbeziehen:

- Altbestände, v. a. im Blick auf die im B3Kat üblichen Hybridaufnahmen (Massendigitalisate) und im Zusammenhang mit der retrospektiven Nationalbibliographie (VD…)

- Einzeltitel aus Nationallizenzen
- E-Dissertationen, die seit Jahren weder durch die DNB noch durch die Herkunftsbibliotheken ausreichend inhaltlich erschlossen werden (vgl. auch das PETRUS-Projekt der DNB für die Reihe O).

Klassifikation bzw. verbale Erschließung müssen differenziert für die einzelnen Segmente Priorität erhalten. Vereinfachungen der Erschließung sowie der Einsatz automatischer Verfahren sind hier besonders zu erwägen.

Thematisches Clustering kann auch als Teil eines Retrievalsystems eingesetzt werden. Hierbei werden z. B. Dokumente wahrscheinlich gleichen Inhalts im Rahmen des Suchprozesses gebündelt und dieses ganze Bündel mit einem inhaltlichen Sucheinstieg angesprochen, der lediglich mit Metadaten eines oder weniger Dokumente übereinstimmen muss. Angesichts der aktuellen Datenhaltung in Verbundkatalog und Lokalsystemen verspricht jedoch die einmalige physische Metadatenanreicherung aller Dokumente eines potenziellen Clusters mehr Effekt, weil sie dann für die Suche in allen Lokal- und Fremdsystemen für die dort jeweils vorhandenen Dokumente, unabhängig von den Suchsystemen, jedenfalls bereitstehen.

Situation der Sacherschließung / Hauptaufgabe der nächsten 3 Jahre (nach Einführung der GND)

Es besteht zweifellos ein gewisser Widerspruch zwischen den Anforderungen der Linked-data-Welt (Auflösung der spezifischen Informationen über FRBR-konforme Entitäten in redundanzfreie RDF-Tripel) und den Anforderungen der Suchmaschinentechnologie, insbesondere der Discovery Servives (z.T. flächige, volltextorientierte, auf Masse angelegte Indexierung auf der Basis durchaus redundanter, dokumentbeschreibender Datensätze), wobei bis auf Weiteres auch noch konventionelle OPACs auf relationalen Datenbanken zu versorgen sind.

Eine Herausforderung für die nächsten 3-5 Jahre besteht darin, dass die heute dominierenden Suchmaschinenkataloge mit den gleichen (Norm)Daten bedient werden müssen, die auch für Semantic-web-Anwendungen geeignet sein sollen.

Eine gewisse Redundanz und Kompromisslösungen werden nötig sein, um nicht durch eine zu „fortschrittliche“ Datenmodellierung, die sich in den existierenden Retrievaltools nicht darstellen lässt, Einbußen beim Benutzerkomfort zu riskieren[6].

Andererseits ist es dringend geboten, dass die Suchmaschinenkataloge mit Normdatenrelationen besser umgehen lernen. Mit der Anwendung der GND müssen Relationen auch in aktuellen OPACs in Form verständlicher Navigationslinks abgebildet werden können. Teilweise können Permalinks zu den Normdaten, zur Wikipedia und zu VIAF heute schon nützlich sein. Dies ist im Hinblick auf die GND-Relationstypen für das semantische Netz auszubauen. Übergangsweise ist über andere, „flachere“ Möglichkeiten nachzudenken, wie die in Relationen dargestellten Informationen ansatzweise genützt werden können.

Von den **Discovery Services** ist zu fordern, dass sie die Nutzung der vorhandenen bibliothekarischen Inhaltserschließung bestmöglich unterstützen. Diese muss in einem alle Ressourcentypen umfassenden

[6] Ein Problem wird z. B. die Indexierung der GND-Relationsfelder sein.

Index das Rückgrat für das Retrieval und das Ranking bilden; die Volltextindexierung (insbesondere für Aufsätze) hat ergänzenden Charakter.

Verfahren zur Heterogenitätsbehandlung:

- Die sorgfältige Kategorisierung der Metadaten zwecks Trennung der Datenschichten ist die Grundlage für alle weiteren Verfahren
- Für die nächsten 3-5 Jahre sind die wichtigsten Techniken zum Umgang mit Heterogenität auf der Retrievalebene:
    - Facettierung / Drill down
    - Ranking (möglichst auch nach SE-spezifischen Algorithmen)
- Wesentliche Beiträge zur Reduzierung der Heterogenität können aber auch durch Verfahren der Datenanreicherung, der (virtuellen) Clusterung und (physischen) Dublettenbereinigung geleistet werden.

Neben der Verbesserung der Retrievaltools ist es dringend notwendig, an einem besseren quantitativen Datenangebot zu arbeiten, wobei den qualitativ hochwertigen Daten Vorrang gebührt. In den heutigen bibliographischen Datenbanken gibt es noch riesige Mengen an Titeldaten – ältere und neu produzierte -, die keine bzw. nur spärliche Angaben zum Dokumentinhalt enthalten, obwohl die beschriebenen Ressourcen großenteils wertvolle und spezifische Inhalte haben.