Multilingualität und Lokalisierung zur Wissenserkundung
oder
Vom Nutzen semantischer Netze für das Information Retrieval

Winfried Gödert

1. Ausgangslage

Fasst man die Entwicklung bisheriger und die Vorstellungen zukünftiger Information Retrieval Systeme unter dem Gesichtspunkt zusammen, welche Eigenschaften sie zur Verarbeitung von thematischen Anfragen haben bzw. haben sollen, so lassen sich folgende Generationen von *Retrieval-Paradigmen* angeben:

Matching von Wörtern

Hiermit ist der gegenwärtige technische Mindeststandard beschrieben, über den alle Retrievalsysteme, OPACs oder Suchmaschinen verfügen. Dieser Standard reicht aber weder aus, um Null-Treffer-Mengen zu vermeiden, noch um differenzierte thematische Recherchen durchzuführen oder gar Interessen zu berücksichtigen, die mit Wissenserkundung beschrieben werden.

Begriffliches Suchen

Begriffliches Suchen versucht, die Wortsuche auf das thematisch Gemeinte zu erstrecken, Mehrdeutigkeiten zu vermeiden und insbesondere Null-Treffer-Ergebnisse zu vermeiden. Hierzu werden vorab definierte Synonymie-Relationen aus Normdateien in den Rechercheablauf einbezogen. Neuere Studien haben gezeigt, dass nunmehr auch die Zahl der OPACs steigt, in denen derartige Möglichkeiten angeboten werden.[1]

Berücksichtigung und Exploration des begrifflichen Umfeldes eines Suchbegriffs

Idealerweise besteht das Ziel, das semantische Umfeld eines Suchbegriffs benutzergesteuert für die Bildung von Treffermengen zu berücksichtigen. Derzeit sind zum Erreichen der Zielsetzung zwei Entwicklungslinien zu beobachten. In OPACs oder anderen Recherchesystemen mit offenen Dokumentbeständen werden vorab definierte a priori Beziehungen, die das semantische Umfeld eines Begriffs beschreiben, für die Bildung der Treffermengen berücksichtigt. Ein auf die Dezimalklassifikation gestütztes Verfahren wurde bereits im System *ETHICS* der ETH Zürich eingesetzt.[2] In Recherchesystemen mit abgeschlossenen Dokumentbeständen kommen mit Erfolg Kombinationen aus linguistischen und statistischen Verfahren zum Einsatz, die teilweise bereits Ansprüchen einer begrifflichen Exploration genügen.[3]

---

[1] J. Hubrich: Die Schlagwortrecherche in deutschsprachigen OPACs: Typen der Schlagwortsuche und der Einsatz der Schlagwortnormdatei (SWD) dargelegt unter Rückgriff auf eine empirische Untersuchung. In: Bibliotheksdienst. 39(2005) H.5, 626–653.

[2] Vgl. für das methodische Vorgehen z.B.: H. Funk, K. Loth: Sachabfrage im ETHICS auf der Basis der UDK: ein OPAC. In: Wissensorganisation im Wandel: Dezimalklassifikation - Thesaurusfragen - Warenklassifikation. Proc. 11. Jahrestagung der Gesellschaft zur Klassifikation. Frankfurt 1988. 43–47.

[3] Als besonders gelungenes Beispiel kann auf das sog. „Wissensnetz“ der digitalen Brockhaus-Enzyklopädie verwiesen werden; vgl. C. Rösener: Die Stecknadel im Heuhaufen: Natürlichsprachlicher Zugang zu Volltextdatenbanken. Frankfurt a.M. 2005. X, 243.

Berücksichtigung und Exploration von Themen

Thematische Exploration bricht mit der Vorstellung, dass dem Recherchierenden alle begrifflichen Zusammenhänge des gewünschten Themas zum Zeitpunkt der Recherche bereits bewusst sein müssen und postuliert, dass das Aufzeigen neuer Zusammenhänge die Ergebnismenge positiv beeinflusst. Realisierungen setzen eine Verbindung von begrifflicher Exploration mit Berücksichtigung von dokumentspezifischen a posteriori Relationen durch Boolesche Verknüpfungen oder syntaktischen Operationen voraus. Aus heutiger Sicht kann kein Beispiel angegeben werden, in dem diese Zielsetzung realisiert wäre.

Sieht man die Unterstützung von Vorgängen der Wissenserkundung als generelle Zielsetzung für die Entwicklung von Rechercheumgebungen, so stellt sich die Frage nach den Anforderungen an die hierfür einzusetzenden Instrumente.

2. Multilingualität und Lokalisierung

Multilinguale Erschließung wird gerne unter der Zielsetzung gesehen, Beziehungen zwischen Entitäten normierten Vokabulars in verschiedenen Sprachen herzustellen und so idealerweise eine Verbindung von der begrifflichen Recherche in der einen Sprache zur Recherche in einer anderen Sprache zu schaffen (Crosswalks), ohne dass dabei ontologische Unterschiede berücksichtigt werden.

Ein solches Verständnis muss immer dann an Grenzen stoßen, wenn in den beteiligten Ordnungssystemen nicht nur allgemein verbindliche („universale“) Strukturen abgebildet sind, sondern Teilbereiche sozialer Wirklichkeitskonstruktionen. Solche spezifischen Wirklichkeitskonstruktionen, die nicht einem strikten universalen Bezugssystem zugeordnet werden können, sollen nachfolgend als *Lokalisierung* verstanden werden. Hierbei kann es sich handeln um:

- Historische Entwicklungen und Zusammenhänge
- Ethnische Themen
- Religiöse Themen
- Juristische Themen
- Nationale Organisationsformen
- Politische Strukturen
- Erziehungs- und Bildungssystem
- Alltagskulturelle Themen (Sport, Haushalt, Hobby, Brauchtum, …)
- Fauna und Flora

Besondere Aufmerksamkeit verdienen Lokalisierungs-Überlegungen in multilingualen Erschließungs- und Retrievalkontexten, da dort in der Regel jede Sprache Beziehung zu einem Bezugssystem haben wird, in dem derartige spezifische Wirklichkeitskonstruktionen vorkommen.

Zur Realisierung multilingualer Erschließung sind verschiedene Wege vorgeschlagen worden.

Für Klassifikationen ist ein multilingualer Zugang über Erweiterungen des Registervokabulars in mehreren Sprachen möglich, unabhängig davon, ob das System mit seinen Benennungen selbst in eine andere Sprache übersetzt oder um Klassen erweitert wurde, die

durch den Bezugsraum der Übersetzung erforderlich oder als wünschenswert angesehen wurden.[4]

Für multilinguale Thesauri werden bislang folgende Vorgehensweisen empfohlen:

1. Benutzung einer Leitsprache, die Ausgangspunkt für die begriffliche Strukturierung ist. Die anderen Sprachen werden in Form einer Art (Quasi-) Synonymie-Relation angebunden. Die begriffliche Deckungsgleichheit zwischen den Deskriptoren kann bei dieser Vorgehensweise nicht immer gewährleistet werden.

   Im Grundsatz handelt es sich bei dieser Vorgehensweise um einen monolingualen Thesaurus (soweit es die begriffliche Struktur betrifft) mit einem multilingualen Zugangsvokabular.

2. Jede berücksichtigte Sprache wird für den Aufbau der Struktur gleich behandelt. Die Deskriptoren der verschiedenen Sprachen werden wie vor in unterschiedlicher begrifflicher Deckungsgleichheit aufeinander abgebildet. Zusätzlich wird versucht, die jeweiligen Strukturen ebenfalls aufeinander zu beziehen.[5]

   Diese Vorgehensweise setzt als Idealbild die Strukturgleichheit der zu verwendenden Begriffe in den verschiedenen Sprachen voraus. Selbst wenn diese Erwartung realistisch wäre – mit der Hinzunahme weiterer Sprachen wird sie immer fragwürdiger – eine solche Vorgehensweise beraubt sich der Chance, die in den Bezugsräumen vorhandenen Wirklichkeitskonstruktionen (Lokalisierungen) spezifisch abzubilden, aufeinander zu beziehen und Crosswalks dazwischen herzustellen.

Im Zusammenhang mit der Erstellung der deutschen Ausgabe der *Dewey Decimal Classification* wurde durch einen neuen Vorschlag Multilingualität und Lokalisierung verbunden.[6] Das Ergebnis besteht aus der Schlussfolgerung, dass eine Klassifikation mit jeder Übersetzung in andere Sprachen eine jeweils neue Sichtweise gegenüber den vorherigen Ausgaben eröffnet: Je mehr bei der Strukturierung des Systems darauf geachtet wird, nur „universale“ Strukturen zu berücksichtigen, desto mehr können bei der Gestaltung
des Zugangsvokabulars oder bei Erweiterung der Klassenstruktur Gesichtspunkte einer Lokalisierung eingebracht werden. Ergebnis wäre ein entlokalisiertes universales Kernsystem mit einem Kranz lokalisierter Systeme, die nicht allein Übersetzung des Kernsystems sind, sondern in ihrer Struktur das Lokalisierungsgebiet der jeweiligen Sprache berücksichtigen (vgl. Abb.1 und Abb.2).

Abb.1: Struktur des globalen Kernsystems mit lokalisierten Erweiterungen und Registern

Abb.2: Die Zielprojektion: Entlokalisierte DDC_oglobal mit Lokalisierungen

3. Das Projekt CrissCross

[4] Vgl. beispielhaft die schon erwähnte Realisierung des Systems ETHICS der ETH Zürich.

[5] Vgl. z.B.: G.J.A. Riesthuis: Information languages and multilingual subject access. In: Subject retrieval in a networked environment: Proceedings of the IFLA Satellite Meeting held in Dublin, OH […]. Hg.: I.C. McIlwaine. München 2003. 11–17.

[6] W. Gödert, M. Preuss: Anforderungen an ein Klassifikationssystem in einer globalisierten Welt. Vortrag anlässlich des DDC Workshops in Frankfurt a.M., 20. April 2005. Folien der Präsentation unter: http://www.ddc-deutsch.de/publikationen/pdf/workshop2005-goedert-preuss.pdf.

Das Projekt hat die Zielsetzung[7], ein erweitertes multilinguales und thesaurusbasiertes Registervokabular zur *Dewey-Dezimalklassifikation* (*DDC Deutsch*) zu erstellen, das als Recherchevokabulars zu heterogen erschlossenen Dokumenten verwendet werden kann. Dazu soll eine Verbindung zur *Schlagwortnormdatei (SWD)* hergestellt werden, indem jedes Sachschlagwort der *SWD* eine *DDC*-Notation erhält und schließlich sollen die im Projekt *MACS*[8] begonnen Arbeiten fortgesetzt werden, Links zwischen den Schlagwörtern der *Schlagwortnormdatei (SWD),* der *Library of Congress Subject Headings (LCSH)* und *Répertoire d'autorité-matière encyclopédique et alphabétique unifé (Rameau)* herzustellen.

Damit steht das Projekt in der Tradition der zuvor charakterisierten Überlegungen zur Erstellung multilingualer Thesauri. Denkt man an die Berücksichtigung weiterer Sprachen, so ergibt sich ein zusätzliches Argument, über die Art der semantischen Brücken zwischen den Begriffen nachzudenken.

4. Multilingualität, Lokalisierung und semantische Netze

Konsequent gedachte Lokalisierung erfordert eine größere semantische Ausdrucksvielfalt, als sie in der Regel in den klassischen Dokumentationssprachen mit Begrenzung auf Äquivalenzen (Synonyme, Quasi-Synonyme), Hierarchien, Assoziationen oder genetische Zusammenhänge vorhanden ist.

Der Wunsch nach Verfeinerung und Anreicherung semantischer Relationen im normierten Vokabular großer Normdateien wird zunehmend auch aus Retrievalsicht geäußert.[9] Dieses Interesse wird im Kontext des Semantic Web gestützt durch die Diskussion um maschinelle Interpretation und Verwertbarkeit des Vokabulars und der Relationen. Zielsetzung ist hierbei, geeignete Repräsentationsformen für Ontologien zu finden, um ihren semantischen Gehalt mit anderen Web-Anwendungen verbinden zu können.

Aus dokumentationssprachlicher Sicht können *semantische Netze* als eine Verallgemeinerung bisheriger Ansätze angesehen werden, indem eine stärkere Differenzierung des verwendeten Relationeninventars bei genauerer Bestimmung des Typs und seiner formalen Eigenschaften vorgenommen wird. Primäres Ziel ist eine Form der Wissensrepräsentation, die beispielsweise über den Weg von Inferenzschlüssen entlang der Relationenpfade im Netz vorhandene, aber nicht explizit ausgewiesene Beziehungen zwischen Netz-Entitäten abzuleiten gestatten. Die Art der zu berücksichtigenden Relationen bestimmt sich dabei häufiger aus Nutzen- und Zweckorientierungen, als dies aus klassischen Dokumentationssprachen mit universaler Ausrichtung bekannt ist.

In der Übertragung dieser Ideen auf dokumentationssprachliche Kontexte kann ein großes Potenzial für Gestaltung zukünftiger Retrievalumgebungen gesehen werden, sofern entsprechend aussagekräftige Netze mit Bezug zu erschlossenen Dokumenten zur Verfügung stehen.

Es stellt sich die Frage, ob die Erstellung derartiger Netze von Grund auf neu geschehen sollte (naturgemäß muss dann in der Regel auch die Erschließung erneut durchgeführt werden), oder ob nicht die Weiterentwicklung vorhandener großer Ordnungsstrukturen mit bereits umfangreichen erschlossenen Dokumentbeständen in geeignete Ontologiemodelle mehr Erfolg versprechende Ergebnisse liefern könnte.

---

[7] Vgl. zum Projekt: http://www.d-nb.de/wir/projekte/crisscross.htm.

[8] Vgl. zum Projekt: http://www.d-nb.de/wir/projekte/macs.htm und https://ilmacs.uvt.nl/pub/.

[9] D. Tudhope, H. Alani u. C. Jones: Augmenting thesaurus relationships: possibilities for retrieval. In: Journal of digital information. 1(2001) no.8. [http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Tudhope/].

Die Prüfung dieser Frage erfordert eine genauere Kenntnis der formalen Eigenschaften der vorhandenen Dokumentationssprachen. Interessanterweise existieren bislang kaum quantitativ orientierte Studien zur Frage der Relationen-Zahl – sowohl absolut als auch differenziert nach Relationstypen – oder zur Relationen-Dichte, -Homogenität oder -Zuverlässigkeit.

In einer Studie aus dem Jahr 2004[10] konnte für den Sachschlagwortbestand der *Schlagwortnormdatei (SWD)* – das größte normierte deutschsprachige Vokabular (mit ca. 160.000 Sachschlagwörtern, 140.000 Synonymen) – ermittelt werden, dass 18% der Schlagwörter über gar keine Relation verfügten und 34 % ohne Ober-, Unter- oder Verwandte Begriffe waren. Ein Eindruck zur Relationendichte in der *SWD* vermittelt ein Vergleich mit dem *Standardthesaurus Wirtschaft*: Auf der ersten Hierarchieebene befinden sich in der *SWD* 45% aller Schlagwörter (im *STW*: 20%), auf der zweiten Ebene 27% (*STW*: 22%), auf der dritten Ebene 13% (*STW*: 23%), auf der vierten Ebene 7% (*STW*: 16%) und auf der fünften Ebene 4% (*STW*: 10%).

Analysen der drei genannten Normdateien lassen zusammenfassend erkennen:

1. Die Relationierung der Begriffe ist unterschiedlich umfangreich und dicht ausgeprägt (allein wegen der Bindung an erschlossene Bestände). Für die *LCSH* und für *Rameau* können zwar keine der *SWD* vergleichbaren Daten angegeben werden[11], Stichproben deuten jedoch auf vergleichbare Verhältnisse hin.

2. Die vorhandene Relationierung folgt über mehrere Stufen selten homogenen Gesichtspunkten[12] und bietet somit keinerlei Voraussetzungen für logische Inferenzprozesse.

3. Es gibt viele Beispiele semantischer Cluster, die in jeder der drei Dateien dem Gedanken der Lokalisierung Rechnung tragen. Bei geeigneter Strukturierung sind so gute Voraussetzungen gegeben, das Verständnis der jeweiligen Wirklichkeitskonstruktion sichtbar zu machen und für begriffliche Crosswalks zwischen semantischen Strukturen zur Verfügung zu stellen. Als Beispiel denke man etwa an die Thematik „Regierungssysteme“ und ihre begriffliche wie strukturelle Repräsentation.[13]

Der Vorschlag lautet nun:

1. Benutzung einer Kern-Ontologie mit universalen Relationen.

2. Die Lokalisierung erfolgt über sprachspezifische semantische Netze, die nach Festlegung eines geeigneten Relationeninventars aus den vorhandenen Dateien entwickelt und an die Kernontologie angeschlossen werden.

---

[10] Unveröffentlichte Studie, die im Auftrag der Deutschen Bibliothek an der FH Köln, Institut für Informationswissenschaft durchgeführt wurde.

[11] Beide stehen nicht unmittelbar als maschinenlesbare – und damit statistisch auswertbare – Dateien zur Verfügung.

[12] Vgl. die Studie zu den LCSH: ALA / Subcommittee on Subject Relationships/Reference Structures: Final Report to the ALCTS/CCS Subject Analysis Committee. June 1997. In: http://www.ala.org/ala/alctscontent/catalogingsection/catcommittees/subjectanalysis/subjectrelations/finalreport.htm.

[13] Vgl. die Beispiele in der zum Vortrag verwendeten Präsentation: http://www.bibliothekartag.at/bibliotag2006/Vortraege/VortraegePDF/Goedert_multilingualitaet_lokalisierung.pdf.

Für die Bestimmung des Relationeninventars sind möglicherweise Anlehnungen an bekannte Vorarbeiten zur Erstellung universaler facettierter Ordnungsstrukturen nützlich.

Derartige Netze können insbesondere die wirklichkeitskonstruierenden Teile einer speziellen Lokalisierung durch ein eigenes Set von Relationen flexibler abbilden, als es bei der Integration in die Struktur der Systematik machbar wäre. Die Verbindung zwischen den semantischen Netzen muss dann nicht mehr einer Philosophie kontextfreier semantischer Übereinstimmung folgen, sondern hat „nur“ noch eine Brückenfunktion zwischen verschiedenen Lokalisierungen. Jede der beteiligten Dateien kann dabei autonom weiter gepflegt und entwickelt werden, ohne dass Änderungen für den semantischen Gehalt der Verlinkungen berücksichtigt werden müssten (vgl. Abb.3).

Abb.3 Lokalisierte semantische Netze mit Kernontologie

5. Konsequenzen für das Retrieval

Ein Retrievalmodell kann für diesen Vorschlag nur in ganz groben *funktionalen* Umrissen gegeben und nicht im Sinne einer Benutzersicht gegeben werden (vgl. Abb.4). Im Vordergrund dieses Modells steht die Nutzung der verschiedenen hinterlegten Relationsarten für Navigations- und Retrievalzwecke sowie der bedarfsorientierte Überstieg mittels der Kernontologie aus einem lokalisierten Netz in ein anderes. Man denke als Beispiel wieder an das Thema „Regierungssysteme“ und seine begriffliche wie strukturelle Repräsentation. Die in der Abbildung angedeutete Recherche nach Einzelbegriffen muss um die Möglichkeiten thematischer Recherchen durch postkoordinierende Verknüpfungen ergänzt gedacht werden.

Abb.4: Vereinfachtes funktionales Retrievalmodell zur semantischen Navigation

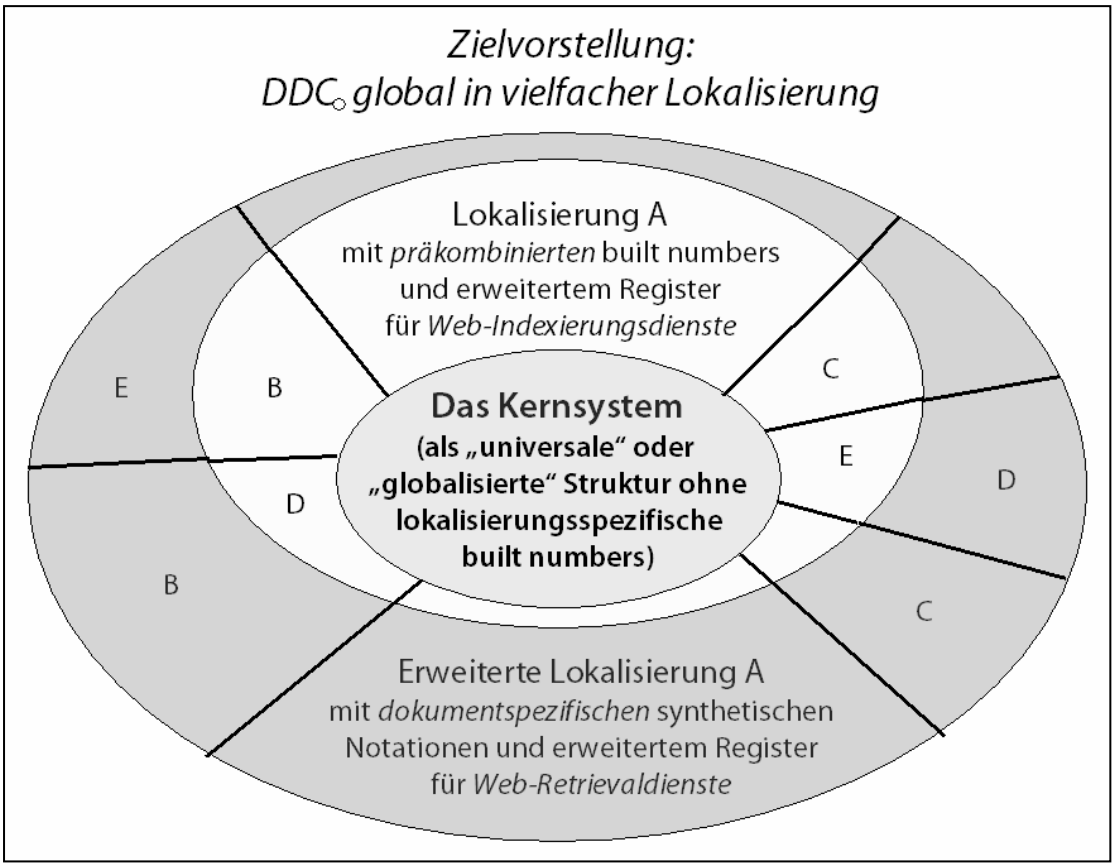

Abb.1: Struktur des globalen Kernsystems mit lokalisierten Erweiterungen und Registern

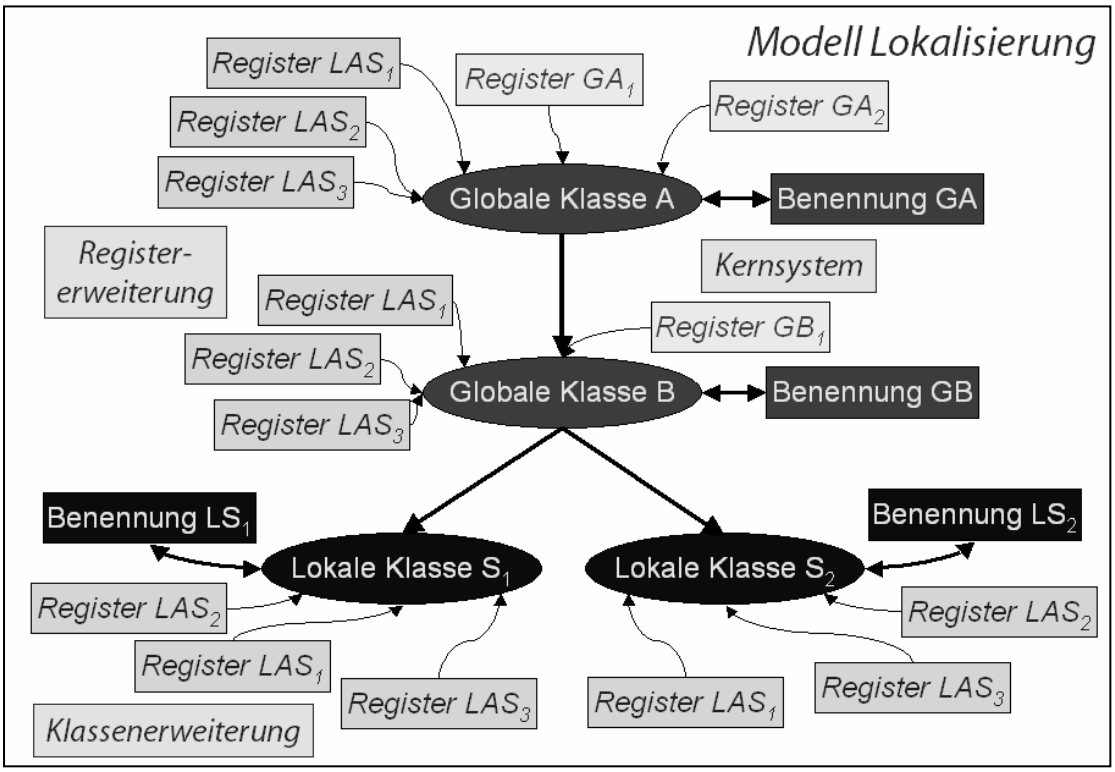

Abb.2: Die Zielprojektion: Entlokalisierte DDC_oglobal mit Lokalisierungen

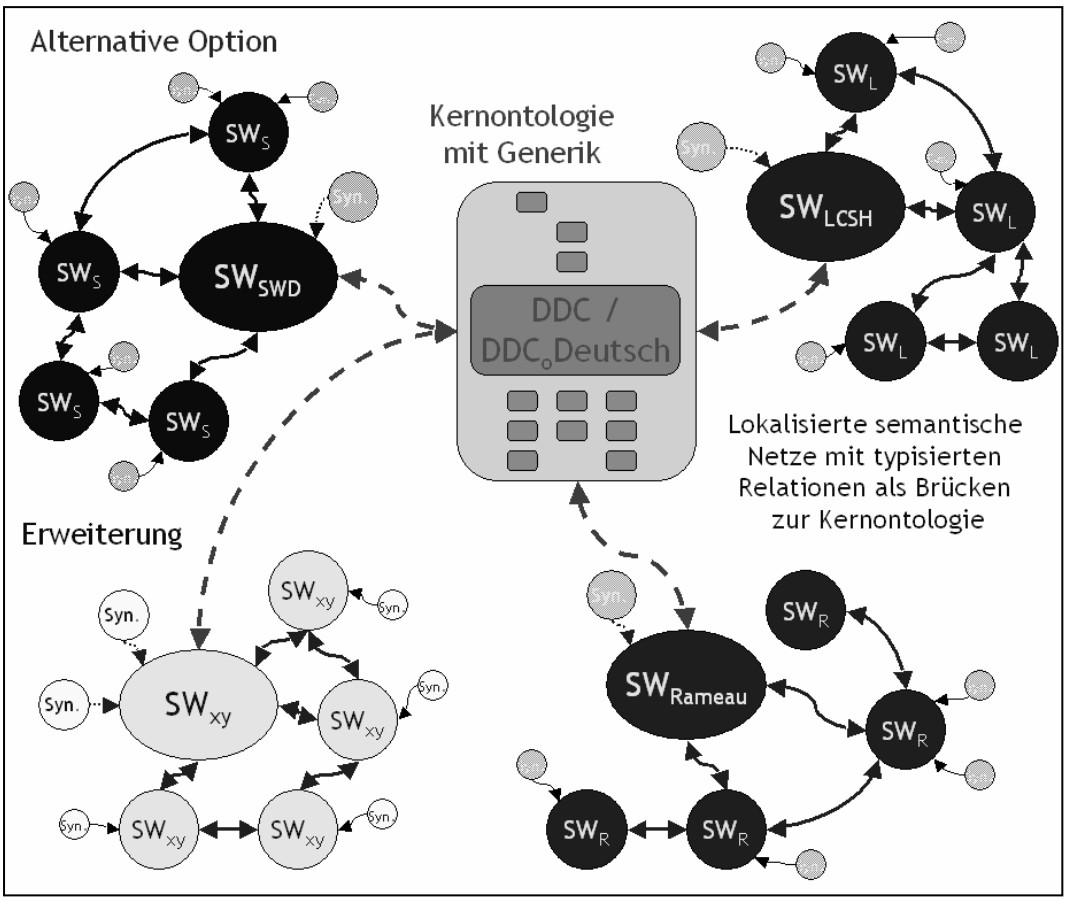

Abb.3 Lokalisierte semantische Netze mit Kernontologie

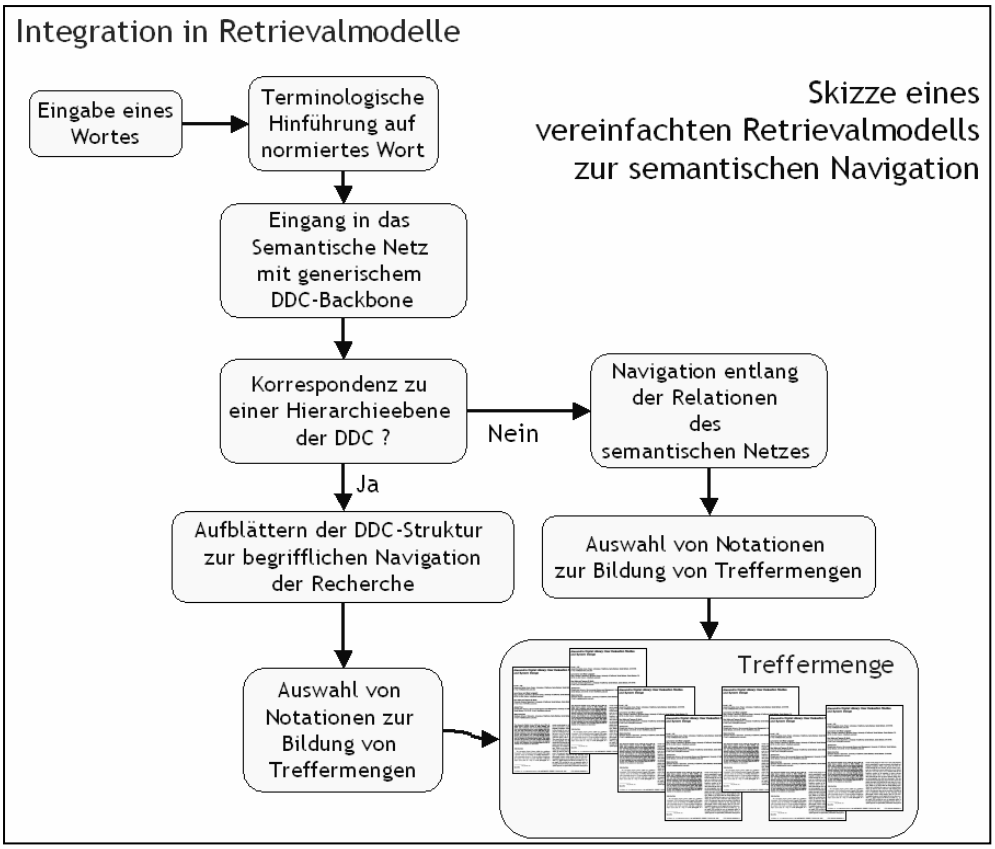

Abb.4: Vereinfachtes funktionales Retrievalmodell zur semantischen Navigation