Thesauri und Interoperabilität mit anderen Vokabularen: die neue Thesaurusnorm ISO 25964

Kempf, Andreas Oskar

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

 DOI 10.1515/iwp-2013-0050 — Information. Wissenschaft & Praxis 2013; 64(6):365–368

Andreas Oskar Kempf, Köln

Thesauri und Interoperabilität mit anderen Vokabularen

Die neue Thesaurusnorm ISO 25964 – Information and documentation – Thesauri and interoperability with other vocabularies
Part 1: Thesauri for information retrieval
Part 2: Interoperability with other vocabularies

Normen und Standards, insbesondere wenn sie international formuliert sind, bilden die zentrale Grundlage für eine gemeinsame Verständigung. Als Konsensergebnis eines häufig sich über Jahre hinziehenden Diskussionsprozesses liefern sie wichtige State-of-the-Art-Spezifikationen, etwa für Produkte und Dienstleistungen, und geben weiterführende Beispiele guter Anwendungspraxis. Das gilt in besonderer Weise auch für den Bereich Information und Dokumentation. Gerade hier ist ein gemeinsames Verständnis, etwa von Dokumentationssprachen, notwendig, um Informationen und Daten sinnvoll und effizient auszutauschen. Erst die Interoperabilität von Vokabularen und damit die Fähigkeit von Dokumentationssprachen, miteinander zu kommunizieren, ermöglicht beispielsweise die nutzbringende Einbindung unterschiedlicher Vokabulare bei der Suchanfrage. Die Herausbildung immer enger geknüpfter Informationsnetzwerke, wie sie sich seit einigen Jahren besonders prominent in der Entwicklung des Semantic Web beobachten lässt, verdeutlicht die Bedeutung einer gerade die Grundlagen umfassenden Standardisierungsarbeit.

Der zunehmenden Vernetzung von Informationsressourcen trägt die ISO-Norm 25964 zu Thesauri und Interoperabilität mit anderen Vokabularen, deren erster Teil bereits im Jahr 2011 und deren zweiter Teil nun Anfang März dieses Jahres veröffentlich wurde, in besonderer Weise Rechnung. Beide Teile wurden durch eine internationale Arbeitsgruppe, die im Jahr 2008 unter der Projektleiterin Stella Dextre Clarke ihre Arbeit aufnahm, verfasst. Vorgänger der Norm, die von dem erklärten Anspruch geleitet ist, den Übergang zum elektronischen und damit maschinengängigen Informationsmanagement zu vollziehen, bildeten ISO 2788-1986 bzw. der British Standard (BS) 5723:1987 sowie ISO 5964-1985 bzw. BS 6723:1985. Beide ISO-Vorgängernormen wurden einer gründlichen Revision unterzogen. Zusätzlich flossen Inhalte aus BS 8723, der erstmals konkreter auf die Standardisierung von Konkordanzen (Mappings bzw. Mapping-Beziehungen) eingeht und schrittweise in den Jahren 2005 bis 2008 veröffentlich wurde, in beide Teile der neuen ISO-Norm 25964 ein.

Im Fokus der vorliegenden Besprechung stehen die Aspekte Interoperabilität und Standardisierung von Terminologie-Mappings bzw. sog. Terminologieföderationen. Beides bildet gleichzeitig den Themenschwerpunkt der ISO-Norm. Wenngleich insbesondere Teil 2 Richtlinien zur Erstellung von Thesaurusföderationen an die Hand gibt, liefert bereits der erste Teil in Form von Richtlinien zum Thesaurusaufbau wichtige Grundlagen, um Interoperabilität von bzw. zwischen Thesauri zu erzielen. Interoperabilität wird dabei allgemein als Fähigkeit zum Informationsaustausch zwischen zwei oder mehreren Systemen bzw. Systemkomponenten definiert, um die gewonnenen Informationen etwa bei der Recherche in Datenkollektionen nachzunutzen. Sie lasse sich einerseits dadurch, dass Daten in einer normierten Weise erstellt und vorgehalten werden, um sie in andere Systeme zur Nachnutzung zu importieren, andererseits durch Mappings zwischen den Begriffen der beteiligten Vokabulare erreichen. Zum Gebrauch von Thesauri bei der Indexierung werden in dieser Norm keine Aussagen gemacht. Ausführungen hierzu finden sich in der ISO-Norm 999.

Der erste Teil der Thesaurusnorm mit dem Untertitel Thesauri für Information Retrieval behandelt in den beiden ersten Dritteln den Aufbau, die Pflege und das Management von mono- und multilingualen Thesauri. Die Autorinnen und Autoren der Norm reagieren damit auf die zunehmende Mehrsprachigkeit von Thesauri, wie sie sich etwa im Zuge europäischer Informationsinfrastrukturprojekte zeigt. Es werden die besonderen Herausforderungen bei der Thesaurusübersetzung, wie etwa unterschiedliche Stufen von Äquivalenzbeziehungen, benannt und anhand von Beispielen illustriert. Ebenso werden besondere Problemstellungen, wie etwa das Fehlen von

Äquivalenzbezeichnungen in einer der Übersetzungssprachen hervorgehoben und unterschiedliche mit Beispielen versehene Lösungsansätze angeführt.

Im letzten Drittel von Teil 1 folgen Richtlinien und Empfehlungen zur Präsentation von Daten in einer standardisierten Form für den Import und den Gebrauch von Daten in anderen Systemen. Die Norm trägt damit der Entwicklung Rechnung, dass Thesauri mittlerweile zumeist unter Anwendung spezieller Software-Lösungen aufgebaut und gepflegt sowie mitunter in weitere Software-Produkte, wie Content-Management-Systeme, integriert werden. Mit dem Ziel, Interoperabilität zu erleichtern, wird die Verwendung spezieller Austauschformate (u. a. MARC, SKOS) und -protokolle (u. a. SWADE-E SKOS API und ADL Thesaurus Protocol) für einen effizienten Datenaustausch besonders hervorgehoben. Dem gleichen Ziel folgt schließlich auch das in der Modellierungssprache UML (Unified Modeling Language) präsentierte Datenmodell, das die einem Thesaurus zugrundeliegende Datenstruktur spezifiziert. Anwendbar auf alle Arten von Thesauri, erleichtert es die einheitliche und konsistente Erstellung eines Thesaurus und den Datenaustausch mit anderen kontrollierten Vokabularen. Neben den Thesaurusbegriffen, den traditionellen Relationstypen und Datenfeldern, etwa für Scope Notes, Erläuterungen oder redaktionelle Anmerkungen, sowie der Unterscheidung zwischen Deskriptoren und Synonymverweisen enthält es die Möglichkeit, anwendungsspezifisch typisierbare Begriffsgruppen zu modellieren. Im Internet frei verfügbar[1] und mit dem Datenmodell des W3C-Standards SKOS zur einheitlichen Publikation von Thesauri im Web kompatibel, enthält es darüber hinaus zusätzliche Erweiterungen.[2]

Teil 2 der ISO-Norm 25964 trägt den Untertitel Thesauri und Interoperabilität mit anderen Vokabularen. Noch deutlicher trägt er der Anforderung an Thesauri Rechnung, den Austausch von Informationen über einzelne Datenkollektionen hinweg zu ermöglichen. Entsprechend beinhaltet der zweite Teil im Kern Richtlinien zum Aufbau von Mappings bzw. von Beziehungen zwischen Konzepten zweier oder mehrerer Vokabulare. Erklärter Anspruch ist es, irreführende Schlussfolgerungen beim Aufbau von Mappings zu verhindern.

Eingangs trifft der zweite Teil allgemeine Aussagen zu Thesaurus-Mappings. Er liefert Definitionen und Symbole und unterscheidet zwischen drei verschiedenen Strukturmodellen für Mappings. Das erste Strukturmodell steht für den Fall einer sog. strukturellen Einheit der am Mapping beteiligten Vokabulare. Sämtliche Vokabulare verfügen über exakt dieselbe Struktur von hierarchischen und assoziativen Beziehungen. Als Beispiel werden unterschiedliche Übersetzungen eines Thesaurus genannt, die innerhalb eines Managementsystems verwaltet werden. Das zweite Strukturmodell zeichnet sich dadurch aus, dass sämtliche Vokabulare eines Mappings, die in diesem Fall auf der Strukturebene durchaus unterschiedlich aufgebaut sein können, direkt miteinander verlinkt werden. Von jedem einzelnen Vokabular können bilaterale Beziehungen zu jedem anderen Vokabular ausgehen. Um den Mapping-Aufwand gering zu halten, können die Relationen allerdings auch auf Beziehungen lediglich in eine Richtung beschränkt werden. Das letzte Strukturmodell sieht vor, dass Mappings lediglich von einem zentralen Vokabular zu sämtlichen anderen an dem Mapping-Prozess beteiligten Vokabularen aufgebaut werden. Entsprechend können einzig von diesem zentralen Vokabular, das gleichsam als Drehscheibe dient, und nicht mehr von jedem einzelnen Vokabular aus Ressourcen, die mittels eines der anderen Vokabulare erschlossen wurden, abgesucht werden. Dieses Strukturmodell eignet sich insbesondere, wenn viele verschiedene Vokabulare miteinander verbunden werden oder ein Vokabular einen besonderen Status besitzt.

Es folgen Empfehlungen zu unterschiedlichen, zwischen den an einem Mapping beteiligten Vokabularen möglichen Typen von Verbindungen. Analog zu den innerhalb eines Thesaurus enthaltenen Beziehungen wird zwischen Äquivalenz-, hierarchischen und assoziativen Mapping-Relationen unterschieden. Weitere Relationstypen, etwa im Fall von Mappings zu Ontologien, so die Autorinnen und Autoren, seien allerdings denkbar. Da Äquivalenzbeziehungen, die für die Verwendung von Mappings einen besonderen Mehrwert darstellen, häufig relativ unsauber ausfallen, wird für sie noch einmal zwischen drei unterschiedlichen Formen von Äquivalenzbeziehung differenziert. Neben exakten und sog. nicht-exakten sind demnach auch sog. partielle Äquivalenzbeziehungen denkbar. Für letztere, so die Empfehlungen, gelte es, auf der Grundlage subjektiver Entscheidungen diese, sofern möglich, als hierarchische, nicht-exakte oder aber als sog. zusammengesetzte Äquivalenz-Mappings durch die Kombination von Begriffen zu modellieren. In diesem Fall kann die Äquivalenz durch die Schnittmenge der kombinierten Begriffe oder kumulativ ausgedrückt

1 http://niso.org/schemas/iso25964/Model_2011-06-02.jpg

2 Eine Tabelle zu den Entsprechungen zwischen dem ISO- und dem SKOS- bzw. SKOS-XL-Datenmodell findet sich als Beispiel sehr guter Zusammenarbeit zwischen den beiden daran beteiligten Arbeitsgruppen unter http://www.niso.org/apps/group_public/download.php/9627/Correspondence%20ISO25964-SKOSXL-MADS-2012-10-21.pdf.

werden. Auch in Fällen von Mappings zu Vokabularen, die sich durch Präkoordination auszeichnen, wird der Aufbau von Äquivalenzbeziehungen durch die Kombination von Begriffen empfohlen.

Für den praktischen Einsatz von Mappings besonders hilfreich ist eine Übersicht über die unterschiedlichen Mapping-Typen und die Auswirkungen der Mappings auf das Retrieval. Differenziert nach den beiden zentralen Anwendungsfällen von Mappings, dem Indexierungsprozess und der Informationsrecherche, werden anhand von Beispielen die Auswirkungen auf Recall und Precision erläutert. Aus diesem praxisnahen Einblick in den Mehrwert von Mappings werden Schlussfolgerungen und Empfehlungen abgeleitet, die in fünf zentrale Fragen münden, die nach Meinung der Autorinnen und Autoren zu Beginn eines jeden neuen Mapping-Projekts stehen sollten. Dazu zählt etwa die Frage, wie differenziert die Mappings aufgebaut werden sollen. Auch die nachfolgenden, mit Beispielen versehenen Ausführungen zum generellen Vorgehen beim Aufbau von Mappings und zu den unterschiedlichen Mapping-Techniken eignen sich sehr gut für die Praxis. Daneben wird von den Autorinnen und Autoren der Norm auch die Verwendung computerunterstützter Verfahren beim Mapping-Aufbau behandelt und in einzelnen Arbeitsschritten erläutert. Vor dem Hintergrund der zunehmenden Entwicklung derartiger Verfahren wird gleichwohl ausdrücklich darauf hingewiesen, dass die von einem Matching-Algorithmus als geeignet identifizierten Mapping-Kandidaten abschließend stets von einem Experten zu prüfen sind.

Es folgen Ausführungen zu Datenmanagement und Austauschformaten sowie zur Pflege der Mappings. Vor dem Hintergrund, dass es noch kein Standardschema zum Management von Mappings gibt, das komplett den Richtlinien der Thesaurusnorm entspricht, wird die Datenvorhaltung in einer XML-Datenbank oder in einem sog. RDF-Store empfohlen. Für die Publikation der Daten im Semantic Web wird explizit dazu geraten, ein SKOS-kompatibles Datenformat zu verwenden. So sind in SKOS selbst unterschiedliche Mapping-Relationen vorgesehen. Für die Pflege der Mappings wird hervorgehoben, dass sämtliche Änderungen, etwa in Bezug auf die Reichweite eines Konzepts, sowohl im Start- als auch im Endvokabular eines Mappings die Kontrolle und ggf. Modifikation der daran geknüpften Mappings zur Folge haben. Differenziert danach, ob die Änderungen das Start- oder das Zielvokabular eines Mappings betreffen, findet sich im Anschluss auch hier eine Übersicht darüber, welche praktischen Arbeitsschritte für das Mapping damit verbunden sind. Es folgen mit Beispielen versehene Empfehlungen zur Darstellung der Mappings, die zwischen Darstellungsmöglichkeiten in Thesaurusverwaltungssystemen für ein übersichtliches Datenmanagement bei der Pflege und Weiterentwicklung von Mappings und endnutzerbezogenen Anforderungen an Retrievalsysteme unterscheiden. Je nach Nutzungsszenario, so die Autorinnen und Autoren, könne die Art der Darstellung sehr stark variieren. Im Falle einer automatischen Verarbeitung der Mappings bei einer Suchanfrage könne etwa auf die Darstellung komplett verzichtet werden.

In der zweiten Hälfte von Teil 2 der neuen Thesaurusnorm werden unterschiedliche Arten von Vokabular unter dem Gesichtspunkt der Interoperabilität bzw. möglicher Mappings mit Thesauri dargestellt. Die Ausführungen haben zumeist beschreibenden und empfehlenden statt normativen Charakter. Neben Klassifikationen zählen hierzu etwa Taxonomien und Terminologien sowie Synonymringe und Normdaten zu Personen oder Körperschaften, deren zentrale Eigenschaften und semantische Komponenten ebenso wie deren Verwendungsformen jeweils zu Beginn kurz beschrieben werden. Grund dafür ist, dass es zum einen zu diesen Typen von Vokabular, mit Ausnahme von Terminologien, keine eigenen ISO-Normen gibt. Zum anderen können sie die Recherche sehr gut unterstützen, indem sie sich zur Anreicherung des Suchvokabulars eignen.

Vor dem Hintergrund einer starken Verwendung des Ontologiebegriffs in den Forschungsbereichen künstliche Intelligenz und Knowledge Engineering in der Informatik sowie in der Semantic-Web-Literatur wird in diesem Abschnitt der ISO-Norm auch auf Ontologien eingegangen. Dabei beziehen sich die Ausführungen ausschließlich auf sog. domainspezifischen heavyweight-Ontologien, die sich in Abgrenzung zu sog. lightweight-Ontologien, wie sie für eine Sammelbezeichnung für sämtliche Formen von Wissensorganisationssystemen stehen, durch die für Ontologien spezifischen Aussagen, sog. Axiome, auszeichnen. Mit Blick auf die unterschiedlichen Anwendungskontexte beider Vokabulartypen wird deutlich hervorgehoben, dass Ontologien, anders als Thesauri, weder zum Zweck des Information Retrieval anhand von Schlagwörtern oder Notationen noch zur terminologischen Kontrolle in Form von Disambiguierung gedacht sind, sondern in erster Linie dazu dienen, logische maschinenlesbare Schlussfolgerungen zu ermöglichen. Entsprechend werden auch keine Empfehlungen zum Mapping zwischen Thesauri und Ontologien gegeben. Stattdessen folgen relativ allgemein gehaltene Ausführungen zur möglichen Neumodellierung eines Thesaurus zu einer Ontologie zum einen und zur komplementären Verwendung von Thesaurus und Ontologie zum anderen.

Zusammenfassend bilden beide Teile der neuen ISO-Norm, die einen wichtigen Beitrag zur weiteren Standardisierung der Arbeit mit Thesauri sowie angrenzender Vokabulartypen leisten, eine sehr gute Handreichung für Praktiker im Umgang mit Thesauri. Mit dem Schwerpunkt auf Interoperabilität von Thesauri tragen beide Teile der wachsenden Vernetzung von Informationsressourcen Rechnung. Durch die große Bandbreite der behandelten Themenbereiche vom Thesaurusaufbau über Austauschformate bis hin zum Management von Mappings bieten sie sowohl für Einsteiger, die über keinen informationswisschenschaftlichen Hintergrund verfügen, als auch für Experten, die das eigene Erschließungsvokabular mit anderen Thesauri und weiteren Formen von Vokabular verknüpfen möchten, substantielle Richtlinien. Als besonders benutzerfreundlich erweisen sich dabei neben den zahlreichen praxisnahen Beispielen in den einzelnen Unterkapiteln die umfangreichen Register am Ende beider Teile. Haben einige Abschnitte, insbesondere des zweiten Teils, weniger normativen als vielmehr beschreibenden und empfehlenden Charakter, so zeigt dies lediglich, wie sehr dieser Bereich keinesfalls als abgeschlossenen gelten kann, sondern sich durch ein hohes Maß an Neu- und Weiterentwicklungen auszeichnet. Gerade vor dem Hintergrund der besonderen Entwicklungsdynamik und eigener Praxiserfahrungen ist der ausdrückliche Appell der Autorinnen und Autoren, bei der Verwendung prozessunterstützter Verfahren zum Aufbau von Mappings abschließend stets eine Qualitätskontrolle durch Fachleute vornehmen zu lassen, noch einmal besonders hervorzuheben.

Beide Teile der ISO-Norm 25964 sind direkt über die Internationale Organisation für Normung sowie in Deutschland zusätzlich über den Beuth-Verlag zum Preis von um die 200,– Euro erhältlich.[3]

Dr. Andreas Oskar Kempf
GESIS – Leibniz-Institut für Sozialwissenschaften
Computational Social Science
Unter Sachsenhausen 6–8
50667 Köln
andreas.kempf@gesis.org
www.gesis.org

Dr. Andreas Oskar Kempf ist wissenschaftlicher Mitarbeiter bei GESIS – Leibniz-Institut für Sozialwissenschaften. Verantwortlich für den Thesaurus und die Klassifikation Sozialwissenschaften zählen zu seinen Forschungsgebieten die Interoperabilität von Wissensorganisationssystemen und ihre Einbindung in Anwendungen des Semantic Web.

3 Zu Teil 1 siehe http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53657; zu Teil 2 siehe http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53658