This is an author's accepted manuscript version of a conference report published in Information - Wissenschaft & Praxis (Wiesbaden), 62 (2011) 2-3, pp. 125-126.
The final publisher's version is available online at:
https://b-i-t-online.de/pdf/iwp/IWP2011-2-3.pdf

Mind the lexical gap – EuroVoc Building Block of the Semantic Web

Ein Konferenzbericht

Axel Huckstorf[1] and Vivien Petras[2]

[1] *Stiftung Wissenschaft und Politik, Berlin*
[2] *Institut für Bibliotheks- und Informationswissenschaft*
Humboldt-Universität zu Berlin

Luxemburg, 18. und 19. November 2010

Ein Konferenzereignis der besonderen Art fand im November letzten Jahres in Luxemburg statt. Initiiert durch das Amt für Veröffentlichungen der Europäischen Union (http://publications.europa.eu) waren Bibliothekare und Information Professionals eingeladen, um über die Zukunft mehrsprachiger kontrollierter Vokabulare in Informationssystemen und insbesondere deren Beitrag zum Semantic Web zu diskutieren. Organisiert wurde die Konferenz durch das EuroVoc Team, dass den Thesaurus der Europäischen Union bearbeitet. Die letzte EuroVoc Konferenz fand im Jahr 2006 statt. In der Zwischenzeit ist EuroVoc zu einem ontologie-basierten Thesaurusmanagementsystem übergegangen und hat systematisch begonnen, Semantic Web Technologien für die Bearbeitung und Repräsentation einzusetzen und sich mit anderen Vokabularen zu vernetzen. Ein produktiver Austausch fand mit den Produzenten anderer europäischer und internationaler Vokabulare (z.B. United Nations oder FAO) sowie Vertretern aus Projekten, die an Themen über automatische Indexierung (hier

insbesondere parlamentarische und rechtliche Dokumente) sowie Interoperabilitiät zwischen Vokabularen arbeiten, statt.

Mehrsprachigkeit, das Semantic Web und die parlamentarische Dokumentation waren die Fokuspunkte der zweitägigen Diskussionen, die sich sogar in der Mehrsprachigkeit der Beiträge ausdrückten (Englisch und Französisch), die von Übersetzern der EU simultan übersetzt wurden.

Giovanni Sartor vom European University Institut präsentierte in seiner Eröffnungspräsentation das Fachgebiet Legal Informatics, wo rechtliche Fragen und Informationstechnologien aufeinandertreffen, insbesondere um – auch mit Semantic Web Technologien – präzise Modellierungen von rechtlichen Zusammenhängen darzustellen und maschinell verarbeitbar zu machen.

Danach präsentierte Christine Laaboudi-Spoiden vom EuroVoc Team die vielen Neuerungen, die der Thesaurus in den letzten Jahren erfahren hat. Mit der gleichzeitigen Veröffentlichung in 22 Sprachen ist EuroVoc ein Standardwerkzeug in der parlamentarischen Dokumentation, dass derzeit in eine neue Semantic Web-kompatible Thesaurus-Management-Umgebung (OWL Datenmodell, SKOS u.a. Thesaurusformate) überführt. Gleichzeitig werden in seit 2008 laufenden Projekten Mappings mit anderen europäischen und internationalen Thesauri erstellt. Ziel ist es, EuroVoc in der Linked Open Data Cloud zu veröffentlichen. Die Methoden und Technologien des Thesaurus-Mappings wurden von Laurent Begin (Mondeca) und Jerôme Euzenat (Inria) vorgestellt. Unterschiedliche Terminologie, nicht nur Mehrsprachigkeit, und unterschiedliche Struktur der Vokabularien stellen die größten Probleme bei automatischen Mapping-Verfahren dar.

Neben diesen Fokus-Themen wurden interessante EU-Projekte vorgestellt, die die parlamentarische Dokumentation automatisierbar, interoperabel und einfacher gestalten sollen, so z.B. das Cellar Projekt für zur Aggregation der E-Publishing-Vorhaben, das Peter Schmitz vom EU-Amt für Veröffentlichungen vorstellte.

Ein Highlight war die Thesaurus Speed Dating Session, in der verschiedene Thesauri in einem Durchlaufverfahren den Konferenzteilnehmern vorgestellt wurden und diskutiert wurden (siehe separater Block).

This is an author's accepted manuscript version of a conference report published in Information - Wissenschaft & Praxis (Wiesbaden), 62 (2011) 2-3, pp. 125-126.
The final publisher's version is available online at:
https://b-i-t-online.de/pdf/iwp/IWP2011-2-3.pdf

„Thesauri ‚Speed Dating' Workshop" – oder: Die Suche nach dem „idealen" Thesaurus

Zur Auflockerung des üblichen Konferenzablaufs hatten sich die Organisatoren der EuroVoc-Konferenz für den ersten Konferenztag etwas Besonderes ausgedacht. Es handelte sich hierbei um einen Programmpunkt, der mit dem Titel „Thesauri ‚Speed Dating' Workshop" überschrieben war. Während eines zweieinhalbstündigen Workshops sollten die Konferenzteilnehmer/innen sozusagen „im Schnelldurchlauf" neun bei internationalen Einrichtungen/Organisationen in Anwendung befindliche multilinguale Fachthesauri bzw. kontrollierte Schlagwortsysteme kennenlernen und anhand der gewonnen Erkenntnisse Schlussfolgerungen für den aus ihrer Sicht „idealen Thesaurus" ziehen. Dazu waren alle Konferenzteilnehmer/innen bereits bei der Registrierung in sechs Arbeitsgruppen aufgeteilt worden. Außerdem erhielt jede/r Konferenzteilnehmer/in zur Vorbereitung auf den Workshop sehr übersichtlich aufbereitete *Quick Reference Cards*, die zu jedem Thesaurus eine Reihe von Basisinformationen enthielten, anhand derer ein erster Überblick und Vergleich der einzelnen Thesauri möglich war.

Es sei durchaus ein Experiment, auf das man sich mit diesem Veranstaltungsformat einlasse, verriet ein Vertreter des Konferenzveranstalters, dem Amt für Veröffentlichungen der Europäischen Union. Dieses Experiment darf jedoch in der Nachbetrachtung als äußerst gelungen angesehen werden.

Für jede der sechs Arbeitsgruppen war jeweils ein *Date* mit Vertreter/innen folgender Thesauri vorgesehen: *Date* 1: „European Training Thesaurus" (Cedefop - European Centre for the Development of Vocational Training, Thessaloniki), „TESE - Multilingual Thesaurus on Education Systems in Europe" (Europäische Kommission: Executive Agency Education, Audiovisual and Culture, Brüssel), *Date* 2: „Unesco Thesaurus" (Unesco, Paris), *Date* 3: "GEMET – General Multilingual Environmental Thesaurus" (European Environment Agency, Kopenhagen), "Agrovoc – Food and Agricultural Organisation Thesaurus" (Food and Agriculture Organisation, Rom), "Inspire Feature Concept Dictionary" (Europäische Kommission: Joint Research Centre, Brüssel), *Date* 4: "ESCO – European Taxonomy of Skills Competencies and Occupations" (Europäische Kommission: Generaldirektion Beschäftigung, Soziale Angelegenheiten und Chancengleichheit, Brüssel), *Date* 5: "European Thesaurus on International Relations and Area Studies" (European Network on International

Relations and Area Studies), *Date* 6: "UNBIS – United Nations Thesaurus" (Vereinte Nationen: Dag Hammarskjöld Library, New York).

Den Thesaurusvertreter/innen kam die Aufgabe zu, jeder Arbeitsgruppe innerhalb von jeweils zwanzig Minuten Rede und Antwort zu stehen. In der Regel bekamen die Thesaurusvertreter/innen etwa fünf Minuten Zeit für die Vorstellung ihres Thesaurus/ihrer Thesauri; anschließend konnten Fragen durch die Gruppe gestellt werden. Ziel des Workshops sollte es zudem sein, innerhalb der Arbeitsgruppen Diskussionen über die vorgestellten Thesauri anzuregen und die Spezifika bzw. Gemeinsamkeiten der jeweiligen Thesauri herauszustellen, dabei stets den Blick auf die übergeordnete Leitfrage nach dem „idealen Thesaurus" gerichtet. Jeder Arbeitsgruppe war ein „Rapporteur" aus verschiedenen EU-Institutionen zur Seite gestellt worden, die zum einen als Moderatoren fungierten und zum anderen die Aufgabe hatten, die Ergebnisse der Diskussionen in den einzelnen Gruppen nach Beendigung aller sechs *dates* zusammenzufassen. Die Zusammenfassung der Schlussfolgerungen aller sechs Arbeitsgruppen wiederum und deren Präsentation im Plenum übernahm die Thesaurus-Expertin Stella Dextre Clarke, die – in ihrer Funktion als Leiterin der ISO 25964-Arbeitsgruppe – im Anschluss an den Workshop zum aktuellen Entwicklungsstand der neuen ISO-Norm für mono- und multilinguale Thesauri referierte.

Die Diskussionen in den Arbeitsgruppen erbrachten zunächst die eindeutige Erkenntnis, dass die Relevanz von Thesauri als Instrument des Zugangs zu und Managements von (Fach)Informationen auch im Internet-Zeitalter unbestritten ist und sich durch das exponentielle Wachstum von Wissen eine umso gewichtigere Ordnungsfunktion für kontrollierte Vokabulare ergibt („*clients need not recall, but precision!*"). Notwendig sei allerdings eine ständige Weiterentwicklung der kontrollierten Vokabulare – eine systematische Auswertung von Suchanfragen der User kann hier sehr hilfreich sein. Gefordert wurde zudem die Bereitstellung von Austauschformaten (z.B. SKOS-RDF) für die einzelnen Dokumentationssprachen, damit diese für die Wissensverknüpfung im Semantic Web auch eingesetzt werden können. Einige der Thesaurusanbieter fühlten sich durch die Workshop-Diskussionen angeregt, aufgrund der erkannten thematischen Überschneidungen Möglichkeiten der Kooperation in der Terminologiearbeit auf ihrem Fachgebiet zu erwägen.

Doch wie sieht er nun aus, der „ideale Thesaurus"? Zu den Grunderkenntnissen des Workshops gehörte die Feststellung, dass es *den* idealen Thesaurus nicht gibt. Vielmehr

ist derjenige Thesaurus als ideal anzusehen, der auf die spezifischen Interessen der Nutzergemeinschaft eines Thesaurus zugeschnitten ist und sich damit zwangsläufig von jedem anderen Thesaurus unterscheidet. Auch politische, institutionelle oder finanzielle Rahmenbedingungen können die Gestaltung eines Thesaurus beeinflussen. Dabei sind Normen für kontrollierte Vokabulare keinesfalls unwichtig, aber sie dürfen der beabsichtigten Anwendung nicht entgegenstehen. Die Anpassung eines Thesaurus an die Bedürfnisse der Nutzer korreliert stark mit seiner Popularität unter den Anwendern. Diese Nutzergemeinschaft ist derzeit im Wandel begriffen, was eine Herausforderung für die Thesaurushersteller darstellt. Der ideale Thesaurus sollte daher nicht sofort als solcher zu erkennen sein, so eine Anregung, sondern eher googleähnlich anmuten und Möglichkeiten zur Einbeziehung der Nutzer bieten. Überhaupt sollte der Nutzer bei der Konzeption jedes Thesaurus an erster Stelle stehen. Als Erfolgsfaktoren für „ideale" Thesauri sahen die Workshopteilnehmer/innen außerdem an: Linked Data-Fähigkeit, Multilingualität, regelmäßige Aktualisierung der Thesaurusbegriffe, Sorgfalt beim Erstellen der semantischen Beziehungen zwischen den Begriffen sowie beim Mapping verschiedener kontrollierter Vokabulare. Da die künftig maßgebliche Nutzergemeinschaft für Thesauri stark durch das Semantic Web bestimmt sein wird, müssen Thesaurusanbieter – auch angesichts sinkender Budgets – verstärkt miteinander kooperieren, um die Vorzüge von Thesauri für die Wissensverknüpfung im Semantic Web vollends zur Geltung kommen zu lassen, insbesondere auch in Verbindung mit Ontologien. Bis zur Umsetzung der Vision eines weltweiten semantischen Pools, in dem alles mit allem verlinkt ist, wird es jedoch noch ein weiter Weg sein.

Link zu den stichwortartigen Workshop Conclusions:

http://eurovoc.europa.eu/drupal/sites/all/files/conference2010/EuroVocConference_Speeddatingconclusions.ppt

Am zweiten Konferenztag wurde die automatische Indexierung parlamentarischer Information mit EuroVoc in mehreren Themen vorgestellt. Gleichzeitig wurde in der Session über Metadaten und Interoperability über Projekte refereriert, die großangelegte Mapping- u.a. Interoperabilitätsvorhaben durchführen: MACS (Mapping von LCSH – Rameau – SWD), Europeana (mehrsprachiges Alignment von Vokabularen im Europeana Semantic Data Layer), das Vocabulary Mapping Framework der XXX, in dem 53 Vokabulare gemappt wurden, und das Star Project in Großbritannien, die semantische Technologien zum Mapping von digitalen Archäologie-Ressourcen einsetzen. Wiederkehrende Diskussionsthemen waren hier die Lizensierung bzw. offene Veröffentlichung von Terminologien und eine damit verbundene Verlinkung mit der Linked Open Data Cloud, was wiederum Fragen der Vertrauenswürdigkeit und Autorität von Quellen aufwarf: einerseits trägt die offene Verlinkung von Terminologie-Ressourcen zur weltweiten Vernetzung und Wissensverbreitung bei, andererseits wird damit die Kontrollfähigkeit bei qualitativ hochwertigen Quellen verringert und die Unterscheidbarkeit zwischen vertrauenswürdigen bzw. professionell erstellten Ressourcen und anderen erschwert.

In ihrer Abschlussrede fasste die Generaldirektorin des Amtes für Veröffentlichungen, Martine Reicherts, die Themen der Konferenz noch einmal zusammen: die Probleme des Informationsaustausches in einer heterogenen Daten- und Formatwelt (insbes. Verlinkung, Mehrsprachigkeit und Kosten) machen die Zusammenarbeit im Semantic Web eine Notwendigkeit, allerdings sind Fragen der Qualitätskontrolle und Finanzierung noch längst nicht geklärt. Auf die nächste EuroVoc Konferenz darf man also gespannt sein.

Alle Präsentationen der Konferenz sind auf der EuroVoc Webseite einzusehen:

http://eurovoc.europa.eu/drupal/?q=node/936