Ernesto W. De Luca, Potsdam, und Ingetraut Dahlberg, Bad König

Die Multilingual Lexical Linked Data Cloud: Eine mögliche Zugangsoptimierung?

Sehr viele Informationen sind bereits im Web verfügbar oder können aus isolierten strukturierten Datenspeichern wie Informationssystemen und sozialen Netzwerken gewonnen werden. Datenintegration durch Nachbearbeitung oder durch Suchmechanismen (z. B. D2R) ist deshalb wichtig, um Informationen allgemein verwendbar zu machen. Semantische Technologien ermöglichen die Verwendung definierter Verbindungen (typisierter Links), durch die ihre Beziehungen zueinander festgehalten werden, was Vorteile für jede Anwendung bietet, die das in Daten enthaltene Wissen wieder verwenden kann. Um eine semantische Daten-Landkarte herzustellen, benötigen wir Wissen über die einzelnen Daten und ihre Beziehung zu anderen Daten. Dieser Beitrag stellt unsere Arbeit zur Benutzung von Lexical Linked Data (LLD) durch ein Meta-Modell vor, das alle Ressourcen enthält und zudem die Möglichkeit bietet sie unter unterschiedlichen Gesichtspunkten aufzufinden. Wir verbinden damit bestehende Arbeiten über Wissensgebiete (basierend auf der Information Coding Classification) mit der Multilingual Lexical Linked Data Cloud (basierend auf der RDF/OWL-Repräsentation von EuroWordNet und den ähnlichen integrierten lexikalischen Ressourcen MultiWordNet, MEMODATA und die Hamburg Metapher DB).

Deskriptoren: Wissensrepräsentation, Datenverknüpfung, Klassifikationssystem, Konkordanz, Semantisches Netzt

The Multilingual Lexical Linked Data Cloud: A possible semantic-based access to the Web?

A lot of information that is already available on the Web or retrieved from local information systems and social networks is structured in data silos that are not semantically related. For this reason, data integration, through reengineering (e.g. triplify), or querying (e.g. D2R) is an important task in order to make information available for everyone. Thus, in order to build a semantic map of the data, we need knowledge about data items itself and the relation between heterogeneous semantic technologies make it emerge that the use of typed links that directly express their relations are an advantage for every application that can reuse the incorporated knowledge about the data items. In this paper, we present our work of providing Lexical Linked Data (LLD) through a meta-model that contains all the resources and gives the possibility to retrieve and navigate them from different perspectives. We combine the existing work done on knowledge domains (based on the Information Coding Classification) within the Multilingual Lexical Linked Data Cloud (based on the RDF/OWL EurowordNet and the related integrated lexical resources MultiWordNet, EuroWordNet, MEMODATA Lexicon, the Hamburg Methaphor Database).

Descriptors: Knowledge representation, data linkage, classification system, concordance, semantic network

Le Multilingual Lexical Linked Data Cloud: possibilité d'optimisation de l'accès?

La plupart des informations sont déjà mises à disposition sur le web ou peuvent être extraites de systèmes de stockages de données structurés tels que les systèmes d'information et les réseaux sociaux. Dès lors l'intégration des données par le travail postérieur ou par le biais de mécanismes de recherche (par exemple D2R) devient capitale, afin de rendre l'information universellement utilisable. Les technologies sémantiques facilitent l'utilisation de liens définis (liens typés), qui retiennent les relations entre eux. Ceci offre des avantages à toutes les applications qui peuvent utiliser les connaissances contenues dans les données. Afin de produire une carte de données sémantique, nous devons connaître les données individuelles et leurs relations avec les autres données. Cet article présente nos travaux sur l'utilisation de Lexical Linked Data (LLD) par un méta-modèle qui contient toutes les ressources et qui fournit également la possibilité de les trouver sous différents aspects. Nous relions ainsi des travaux effectués sur les domaines de connaissances (basés sur l'Information Coding Classification) avec la Multilingual Lexical Linked Data Cloud (basée sur la représentation RDF/OWL de EuroWordNet et les ressources lexicales intégrées similaires MultiWordNet, MEMODATA et la Hamburg Métaphore DB).

Descripteurs: représentation des connaissances, liaison de données, système de classification, concordance, net sémantique

DOI 10.1515/iwp-2014-0040

Viele der im Web verfügbaren oder aus lokalen Systemen und social networks gewonnenen Informationen stammen aus isolierten Silos ohne semantische Verknüpfung untereinander. Die entstehenden semantischen Technologien führen aber dazu, dass Daten, die mittels typisierter Links verknüpft sind und dadurch unmittelbar ihre Beziehungen zueinander ausdrücken, Vorteile für all diejenigen Anwendungen bieten, die das enthaltene Wissen über die Daten nutzen können. Deshalb ist Datenintegration durch Nachbearbeitung oder durch Suchmechanismen (z. B. D2R) eine wichtige Aufgabe, um Information allgemein verfügbar zu machen. Dazu brauchen wir Wissen über die einzelnen Daten und ihre Beziehungen untereinander, um daraus semantische Beziehungsnetze knüpfen zu können. Vorgestellt wird die Bereitstellung von Lexical Linked Data (LLD) mittels eines Meta-Modells, das alle Ressourcen vereinigt und die Möglichkeit bietet, Daten aus unterschiedlichen Perspektiven zu finden oder in ihnen zu navigieren. Wir kombinieren damit existierende Arbeit über Wissensgebiete (basierend auf der Information Coding Classification) mit der Multilingual Lexical Linked Data Cloud (basierend auf der RDF/OWL Repräsentation von EuroWordNet und ähnlichen integrierten lexikalischen Ressourcen (MultiWordNet, MEMODATA und die Hamburg Metapher DB).

1 Wissensorganisation und Data Linking

Bei der nationalen ISKO Konferenz in London, 2010, wurde über die Zukunft der Wissensorganisation im Semantic Web diskutiert, insbesondere mit Bezug auf Linked Data. Es wurde zum bisher erfolgreichsten Ereignis in der englischen ISKO-Geschichte als sich einige Vortragende auf Tim Berner-Lees Gedanken über Linked Data Prinzipien bezogen, über die er schon im Juli 2006 geschrieben hatte. Die Idee ist recht einfach, man kann jeder Information einen Wert hinzufügen, wenn man sie auf eine andere bezieht (http://www.iskouk.org/events/linked data sep2010.htm).

Bereits bei der ISKO Konferenz von 2009 hatte David Crystal (Crystal 2009) in seinem Einführungsvortrag „Semantic targeting: past, present and future“ über die Evolution des sprachlichen Ansatzes zur Inhaltsanalyse gesprochen, die er in den letzten 20 Jahren entwickelt hatte. Diese begann mit der Wissensmanagement-Taxonomie, die für die allgemeinen Enzyklopädien der Cambridge Gruppe benötigt wurden und dieser folgte die Transformation in eine Internet-Taxonomie mit Anwendungen in maschineller Klassifikation von Dokumenten, Suchmaschinen-Assistenz, e-Kommerz, Online-Werbung und Internet-Sicherheit. Die neueren Entwicklungen beziehen sich hauptsächlich auf Werbung, einem Gebiet, dessen Ideen sich von einfacher Stichwort-Analyse zu Kontext-Werbung entwickelten und die nun mit semantischen Inhalten angereichert werden können. Diese Begriffe beinhalteten auch die Art und Weise der Behandlung von Empfindungen, Gefühlen, Absichten und kulturellem Standort. Weitere Arbeiten entstanden auf diesem Gebiet im mehrsprachigen Zusammenhang, indem man zwei Thesauri zum Wiederauffinden von Bildern (Ménard 2009) miteinander verglich oder indem man Medline Abstracts semantisch auffüllte zur Erweiterung des Informationszugangs (Ibekwe SanJuan 2009). Die Ausbeutung von Daten in einer Cloud wurde im Beitrag von Paul Miller in „Exploiting Data in the Cloud“ behandelt, wobei das Anwachsen der Datenmenge diskutiert und die Notwendigkeit von Normung (z. B. durch Linked Data) zur Verbesserung des Strukturierungsprozesses erkannt wurden. Systeme der Wissensorganisation sind erforderlich, um besser an Informationen und Wissen heranzukommen, indem die neuen Techniken, wie „RDF representations“ (Michael Smethurst und Tom Scott 2009), „social tagging“ (Bartlomiej Puzon, et al. 2009) „relational databases“ (Susan Lim Lee Hong Singh et al. 2009) oder die Technik der Facettenklassifikation (Elaine Ménard 2009) verwendet werden.

2 Data Linking mit einer Universalklassifikation

Schon vielfach wurde vorgeschlagen, die Strukturen von etablierten universalen Klassifikationssystemen zu verwenden, um an die Daten im WordNet oder anderen Netzen des Semantic Web besser heranzukommen und ihren Zusammenhang zu erkennen, so z. B. durch Winfried Gödert in seiner Einführung in die Vorträge der Kölner Konferenz, über „Concepts in Context“ (Gödert 2010). Eine Gruppe des ITC in Trento (jetzt FBK) hat sich mit diesen Möglichkeiten auseinandergesetzt und auch eine Reihe von entsprechenden Untersuchungen in ihrer Arbeit „Revising the Wordnet Hierarchy: semantics, coverage

Tab. 1: Information Coding Classification. (Englisch).

0 ALLGEMEINE FORM-u.SACH-BEGRIFFE	01 THEORIEN, PRINZIPIEN	02 GEGENSTAND, BESTANDTEILE	03 PROZESSE, METHODIK	04 EIGEN-SCHAFTEN od.1.Art Auspr.	05 PERSONEN oder 2.Art d. Ausprägung	06 INSTITUTIONEN oder 3.Art der Ausprägung	07 TECHNIK, HERSTELLUNG	08 ANWENDUNG, DETERMINA-TION	09 VERBREITUNG, SYNTHESE
1 FORM & STRUKTUR-BEREICH	11 Logik	12 Mathematik	13 Statistik	14 Systemforschung	15 Organisation	16 Meßwesen	17 Kybernetik	18 Normung	19 Prüfungs- & Kontrollwesen
2 MATERIE & ENERGIE BEREICH	21 Mechanik	22 Materiephysik	23 Allgemeine & technische Physik	24 Elektronik	25 Physikalische Chemie	26 Chemie der Stoffe	27 Technische Chemie	28 Energiewesen	29 Elektrotechnik
3 KOSMO & GEO-BEREICH	31 Astronomie & Astrophysik	32 Weltraumfor-schung & Tech-nologie	33 Grundlagen-Geowissen-schaften	34 Atmosphären-Wissenschaft & Technik	35 Hydro- & Ozeanologie	36 Spezielle Geo-logie & Minera-logie	37 Bergbau	38 Hüttenwesen & Werkstoffwissen-schaften	39 Geographie
4 BIO-BEREICH	41 Allgemeine Bio-logie & Grund-wissenschaften	42 Mikrobiologie	43 Pflanzenbiologie & Kultivierung	44 Tierbiologie & -haltung	45 Veterinärmedizin	46 Landwirtschaft	47 Forstwirtschaft	48 Lebensmittel-wissenschaft & Technologie	49 Ökologie & Umwelt
5 HUMAN-BEREICH	51 Humanbiologie	52 Gesundheits-wesen	53 Pathologie und Medizin	54 Operative & therapeutische Medizin	55 Psychologie	56 Pädagogik	57 Beruf, Arbeit & Freizeit	58 Sport	59 Haushalt & Häusliches Leben
6 SOZIO-BEREICH	61 Soziologie	62 Staat und Politik	63 Verwaltung	64 Finanzwesen	65 Sozialhilfe / Sozialpolitik	66 Recht	67 Raumplanung & Urbanistik	68 Wehrwesen	69 Geschichte
7 WIRTSCHAFTS- & PRODUK-TIONS-BEREICH	71 Allgemeine Wirt-schaftswissen-schaften	72 Betriebswirt-schaftslehre	73 Allgemeine Technik & Technologie	74 Geräte- & Maschinenbau	75 Bauwesen	76 Warenlehre	77 Fahrzeugwesen	78 Verkehrswesen	79 Versorgung & Dienstleistungen
8 WISSENSCHAFTS- & INFORMATI-ONS-BEREICH	81 Wissenschafts-wissenschaft	82 Informations-wissenschaft	83 Informatik	84 Allgemeines Informations-wesen	85 Kommunika-tionswissenschaften	86 Publizistik	87 Graphische Technik & Wirtschaft	88 Nachrichten-technik	89 Semiotik
9 GEISTESWISSEN-SCHAFT & KUL-TURBEREICH	91 Sprache	92 Literatur und Philologie	93 Musik	94 Bildende Kunst	95 Darstellende Kunst, Theater	96 Kultur-wissenschaften i.e.S.	97 Philosophie	98 Religion	99 Christliche Religion & Theologie

© ICC 1982, rev. 2002 Dr.Ingetraut Dahlberg

and balancing“ genannt (Pianta E. et al.) Sie entschieden sich zur Benutzung der Dewey Decimal Classification (DDC) und ordneten schließlich für 48 „Basic Domains“ aus fünf Wissensbereichen die gefundenen Notationen der Dewey Decimal Classification tabellarisch an (siehe Tab. 2 am Ende, mittlere Spalte). Es erwies sich, dass für die sog. Basic Domains zum Teil mannigfache Notationen (also lange Zahlenreihen) notwendig wurden, was verdeutlicht, dass es unnötig aufwendig erscheint, mit diesem System sinnvoll zu arbeiten (wenn es auch das älteste, am weitesten verbreitetste der gegenwärtig noch benutzten, universalen Klassifikationssysteme ist (von 1876, natürlich in vielen Neuauflagen mit der Zeit aufdatiert, aber in ihrer Grundstruktur beibehalten). Vermutlich hätten sie aber mit der Universal Decimal Classification (UDC) von 1895 ähnliche Erfahrungen gemacht. Es sind dies die beiden einzigen universalen Systeme, die eine durchgehend numerische Notation benutzen, was natürlich für die Abbildung der hierarchischen Struktur von großem Vorteil ist. Andere universale Systeme, wie die Library of Congress Classsification (LCC) von 1903, die Bliss Bibliographic Classification (BBC) von 1935, die Colon Classification des Inders Ranganathan (RCC) von 1933 und die Library Bibliographical Classification der Russen (LBC) von 1956, alle immer wieder in Neuauflagen, haben jedoch alle eine sog. gemischte Notation von Buchstaben und Zahlen. Alle genannten Systeme benutzen Disziplinen als Hauptklassen, was freilich dazu führt, dass Schwierigkeiten entstehen, wenn sich im Laufe der Zeit neue Disziplinen entwickeln und eine entsprechend sinnvolle Stelle im System benötigen.

Die Erfahrung der o. g. ITC-Gruppe mit der DDC brachte De Luca zu dem Entschluss, erstens einen Vergleich mit den Notationen der neuen Information Coding Classification (ICC) vorzunehmen (siehe Tab. 1) und zweitens ein erstes Mapping vorzunehmen, um die Möglichkeiten aufzuzeigen, die diese Erweiterung für eine detailliertere Strukturierung der Benennungen von Wissensgebieten/Domains haben könnte. Die ICC existiert auch auf Deutsch, so dass ihre Nutzung einen deutlichen

Vorteil zur Domäne-Beschreibung deutscher SynSets gegeben ist. Die „WordNet Domains“ existieren hingegen nur auf Englisch.

3 Wissensgebiete in der ICC

Zum Verständnis der ICC kann auf den Artikel von I. Dahlberg, „Information Coding Classification. Geschichtliches, Prinzipien, Inhaltliches“ in dieser Zeitschrift, hingewiesen werden (8/2010), so dass hier nur in aller Kürze die zwei wesentlichen Strukturelemente dieser allgemein noch nicht so sehr bekannten, universalen Facettenklassifikation genannt werden sollen, durch welche Sachgruppen und Wissensgebiete ihre Stellen im System finden. Sie geht also, wie oben implizit angedeutet, in ihren Hauptklassen nicht von Disziplinen aus, sondern von Seinsbereichen und orientiert sich an der Schichtentheorie des Seins, wie sie von den Philosophen J. K. Feibleman (1954) und Nicolai Hartmann (1964) entwickelt wurden waren (teilweise übersetzt in Dahlberg 1974). Außerdem ist sie deshalb eine Facettenklassifikation in hierarchischer Gliederung, weil sie Hauptklassen und Unterklassen durchgängig (mit allerdings zwei Ausnahmen im Bereich 9) nach kategorial definierten Aspekten untergliedert, wo es also für alle Positionen nur eine einzige Stelle gibt, mit der dann alle anderen Stellen des Systems kombiniert werden können, was eine unendliche Flexibilität ermöglicht.

Es werden dazu neun Seinsschichten in drei Stufen eingeführt

Schichten für vorbereitendes Sein:
1 Formen und Strukturen
2 Materie und Energie
3 Kosmos und Erde
Schichten für lebendes Sein:
4 Biologisches Sein
5 Menschliches Sein
6 Soziales Sein
Schichten von produziertem Sein:
7 Materielle Produkte
8 Intellektuelle Produkte
9 Geistige Produkte

Abb. 1: Seinsschichten als Hauptklassen der ICC.

Diese Schichten setzen einander jeweils voraus, sie bedingen sich gegenseitig. Jede Schicht wird durch jeweils neun Aspekte gegliedert, sodass damit für die jeweilige Hauptklasse Sachgruppen (SG) und für deren Unterglie-derung Wissensgebiete (WG) entstehen, die zum Teil auch Disziplinen sind oder mehrere SG eine große, wie z. B. Physik und Chemie oder Medizin, zum anderen Teil Phänomene, wie Luft, Wasser, feste Erde oder Organismen, wie Mikroorganismen, Pflanzen, Tiere.

Die neun Aspekte erscheinen ebenfalls wieder in Dreiergruppen, und ihre Anordnung als Elementstellenplan des Systems wurde „Systematifikator“ genannt, somit gibt es

Konstituierende Aspekte einer SG oder eines WG:
1 Theorie und allg. Grundlagen
2 Objekte u. ihre Bestandteile u Eigenschaften
3 Methoden und Prozesse an u. für d. Objekte
Charakteristische Ausprägungen
4–6 Meist drei Arten von Ausprägungen
Einflüsse von und nach außen
7 Einfluss von außen auf das betr. Gebiet
8 Anwendung von drei (Methoden) auf andere WGs
9 Informationen über ein WG für Außenstehende

Abb. 2: Der Elementstellenplan oder Systematifikator der ICC.

Wer diese Strukturprinzipien in ihrer Anwendung versteht, kann sehr leicht mit dem Gesamtsystem umgehen, Memorierbarkeit ist damit einfach geworden. Des Weiteren ist mit diesen Ein- und Unterteilungen von 9×9, also 81 Sachgruppen, die Möglichkeit der tabellarischen Übersicht gegeben (siehe Tab. 1). Erwähnt werden sollte, dass für alle Sachgruppen und Wissensgebiete Definitionen (aus Brockhaus, Duden, Wahrig, etc.) existieren. Die Systematik wurde in den Jahren 1971 bis 1977 aufgrund einer umfangreichen Sammlung der Benennungen von Wissensgebieten entwickelt und in den letzten Jahren teilweise auf den neuesten Stand gebracht. Insgesamt handelt es sich – nach Bereinigung tausender Synonyme – um etwa 6500 Wissensgebiete, bis zur 3. Hierarchieebene erstmalig publiziert (in Dahlberg 1982).

Es sei noch erwähnt, dass in den letzten Jahren eine Excel-Datei dieser drei Hierarchie-Ebenen erstellt wurde mit den Definitionen aller darin erfassten Begriffe, wobei jedoch die Definitionen der 4. bis 6. Ebene, durchaus vorhanden, noch eingearbeitet werden müssten. Auch existiert zu allen Begriffen der Hierarchie eine englische Übersetzung.

Bisher ist die ICC in Publikationen des INDEKS Verlags verwendet worden, vor allem in der Bibliograhie der Zeitschrift *Knowledge Organization* (jetzt im Internet) und in den drei bisher publizierten Bänden der *International Classification and Indexing Bibliography* (Dahlberg 1982)

für die Gliederung und für jeden einzelnen der 11.092 Einträge.

4 RDF/OWL EuroWordNet und die Multilingual Lexical Linked Data Cloud

Um zu verstehen, wie und warum die Wissensgebiete der ICC in die mehrsprachige lexikalische Linked Data Cloud integriert werden sollten, erscheint es notwendig, zunächst die Struktur dieser Ressource zu erläutern. An erster Stelle möchten wir den Übertragungsprozess der Daten des WordNet und des EuroWordNet beschreiben, um den Umfang innerhalb der ICC-Klassen zu begründen.

Das Princeton WordNet (Fellbaum 1998) wurde bereits in ein OWL-Format umgeschrieben, wie dies durch (van Assem et al 2004) beschrieben wurde, indem sie die OWL-DL-Repräsentation benutzten. Diese Form der Darstellung in RDF/OWL basiert auf dem WordNet-Datenmodell wie es in Abb. 3 zu sehen ist.

Synset
- NounSynset
- VerbSynset
- Adjective Synset
  - AdjectiveSatelliteSynset
- AdverbSynset

WordSense
- NounWordSense
- VerbWordSense
- AdjectiveWordSense
- AdjectiveSatelliteWordSense
- AdverbWordSense

Word
- Collocation

Abb. 3: OWLHierarchie von PrincetonWordNet.

Beim Vergleich des ursprünglichen Princeton WordNet Synset (das nur Wortbedeutungen zulässt) mit der OWL-Darstellung wird deutlich, dass das RDF/OWL-Schema (in seiner Gesamt-Version) drei Hauptklassen hat, nämlich Synset, WordSense und Word. Die ursprüngliche WordNet Version enthält nur die Synset Klasse. Die zwei Klassen Synset und WordSense haben weitere vier Unterklassen, die auf der Unterscheidung lexikaler Gruppen basieren, es sind NounSynset, VerbSynset, AdjectiveSynset (mit einer zusätzlichen Unterklasse AdjectiveSatelliteSynset) und AdverbSynset. Die Klasse für das Wort hat die Unterklasse Collocation welche Benennungen bezeichnet, die aus zwei oder mehr Wörtern bestehen. Um die Bedeutungen jedes Bestandteils eines Synsets eindeutig zu machen, haben WordSense und Word eine einzige URI, die zum Wiederauffinden von Wörtern und Wortbedeutungen benutzt werden kann, unabhängig von ihren Synsets. Diese Eigenschaft war in der ursprünglichen Version von WordNet nicht vorhanden. Die URI vermitteln Information über die Bedeutung einer Einheit und werden nach folgendem Muster gebildet:

wn20instances: + synset + lexical form + type +
sense number en?)

Wenn wir z. B. die vierte Wortbedeutung des Wortes „Bank“ finden möchten, erhielten wir einen URI wie folgt: http://w3.org/2006/03/wn/wn20/instances/synsetbank-noun-4

Die Eigenschaften des RDF Systems werden nach drei Arten von Relationen unterteilt:

1. solche, die zwei Synsets miteinander verbinden (z. B. Hyponyme Of)
2. solche, die zwei Wortbedeutungen miteinander verbinden (z. B. Antonyme Of)
3. solche, die eine Anzahl von Eigenschaften mit Informationen über Einheiten (z. B.XML Schema Datentypen wie xsd:string) haben, wie man es bei synsetId benutzt.

Um Redundanz zu vermeiden werden nur Relationen in einer transitiven Richtung (z. B. Hyponym Of und nicht Hypernym Of) repräsentiert, die über die OWL-inverseOf-Eigenschaft in das RDF Schema eingesetzt werden können. Insgesamt gibt es 27 Relationen, die in der RDF/OWL-Darstellung des WordNet eingesetzt wurden. Die Fälle aller Klassen und Eigenschaften sind voneinander in verschiedene Dateien getrennt, eine für die Synsets, eine für die WordSenses und Words und eine für jede Relation. Obwohl das RDF-Schema zur Beschreibung der meisten Klassen- und Eigenschafts-Definitionen benutzt wird, sind einige OWL-Aussagen in das Schema integriert worden, um bessere semantische Beschreibungen zu gewährleisten, wie z. B. dem Überprüfen auf Korrektheit der Daten oder der Definition inverser Relationen. Für diese Aussagen muss ein geeignetes Programm die OWL. DL-Norm unterstützen, um die Daten speichern und abfragen zu können.

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;syset-bank-noun-1" rdfs:label="bank">
    <ewn20schema:synsetId>102690337</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Word rdf:about="&ewn20instances;word-bank" ewn20schema:lexicalForm="bank"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-bank-noun-1" rdfs:label="bankv>
    <ewn20schema:word rdf:resource="&ewn20instances:word-bank"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank-noun-1"/>
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank_building-noun-1"/>
</rdf:Description>
```

Abb. 4: RDF/OWL-EuroWordnet Synset Beispiel.

4.1 RDF/OWL EuroWordNet

Wegen der verschiedenen Probleme bezüglich WordNet und seinen Varianten hielten wir es für angebracht, sie in eine RDF/OWL-Repräsentation zu übertragen (siehe unten) um die Entwicklung flexiblerer Revisionsmethoden zu ermöglichen.

Im EuroWordNet enthält ein Synset alle verwandten Wortbedeutungen, Synonyme und Beziehungen zu anderen Synsets sowie auch zum Zwischensprachlichen Register (Inter-Lingual-Index). Diese Information diente der Vorbereitung der Eingliederung in das angemessene RDF-Schema und der Reorganisierung für eine neue Daten-Repräsentation.

Die Entscheidung zur Umwandlung des EuroWordNet basierte auch auf der Notwendigkeit, seinen Umfang auszuweiten, weil nicht alle Bedeutungen durch andere Ressourcen abgedeckt sind. Außerdem, da die meisten Ontologien für Sachgebiete in OWL dargestellt sind und eine einsprachige WordNet RDF/OWL-Repräsentanz bereits implementiert wurde, würde ein EuroWordNet-Austausch diesen „resources“ mehrsprachige Fähigkeiten verleihen. Daher haben wir das WordNet RDF in die RDF/OWL-Repräsentation aufgrund der Arbeit von van Assem et al. (2004). konvertiert.

Da das EuroWordNet verschiedene Relationen besitzt und eine Struktur, die sich vom Princeton WordNet unterscheidet, wurden einige Schritte notwendig, um die Daten an das RDF/OWL-Schema anzugleichen und das RDF-Schema mit den neuen Relationen zu erweitern. So haben wir zuerst die Bedürfnisse für das EuroWordNet analysiert und an das WordNet RDF-Schema für eine mehrsprachige Darstellung des EuroWordNet angeglichen. Danach haben wir die EuroWordNet-Relationen in OWL-Eigenschaften umgewandelt und die Ontologie mit zwei Ontologien von Sachgebieten erweitert. (De Luca et al. 2007)

4.2 Umwandlung von ICC-Wissensgebieten in die RDF/OWL-EuroWordNet-Repräsentation

RDF/OWL-EuroWordNet kann für verschiedene natürliche Sprachverarbeitungsverfahren und maschinelle Lernverfahren eingesetzt werden. De Luca (2008) wertete verschiedene sprachliche Parameter aus, die in dieser Ressource enthalten waren und fand heraus, dass das Wissen über Wissensgebiete (das durch die WordNet Domains) gegeben ist, sehr hilfreich für Retrieval-Fälle sein kann. Mit diesen Ergebnissen entschlossen wir uns, die Synsets mit den ICC-Wisssensgebieten anzureichern und sie in die WordNet Domain Information für jedes einzelne Synset zu integrieren.

Die einzelnen Schritte zur Konvertierung der ICC in die RDF/OWL-EuroWordNet-Repräsentation können folgendermaßen gegliedert werden:

- Analyse der Erfordernisse für die ICC
- Abbilden der ICC Wissensgebiete auf die WordNet Domains
- Adaptieren der EuroWordNet RDF-Schemata auf die ICC
- Mehrsprachigkeit
- OWL Eigenschafts-Konversion
- OWL Domain Erweiterung

van Assem et al. (2004) unterscheiden Word und WordSense in ihrem Datenmodell aus zwei Gründen: Erstens werden verschiedene Relationen bezüglich der Wortbedeutung definiert. Synsets und WordNet gebrauchen diese Unterscheidung in ihrer Datei. Zweitens wird aus Gründen ontologischer Klarheit angenommen, dass Synsets Wortbedeutungen enthalten, um den logischen Ort des Lexikons zu teilen (Wörter als Formen oder Bedeutungen, Synsets als Cluster von Wortbedeutungen durch

Abstraktion ihres Verteilungsbereichs). Indem wir dieses Modell akzeptierten, adaptierten wir ihr Schema, um EuroWordNet zu konvertieren indem wir diese Annahme auch auf eine mehrsprachige Aufgabe anwandten. Ein Beispiel eines OWL-EuroWordNet Synsets wird in Abbildung 4 gegeben. Hier ist die Wortbedeutung von „bank“ in seinem Synset (und synsetId) gezeigt, WordSense, Word und Synonyme (enthält WordSense).

5 Vernetzung der ICC mit dem RDF/OWL-EuroWordNet

Nach der Umwandlung von EuroWordNet in eine OWL-Repräsentation (siehe DeLuca et al., 2007), haben wir uns entschieden, eine Vernetzung mit der ICC-Hierarchie herzustellen. Zur Integration der ICC in die EuroWordNet OWL-Repräsentation analysierten wir die RDF/OWL-EuroWordNet-Hierarchie der Klassen, die auch in OWLDL eingebaut waren. Jede Benennung wurde als eine Klasse erklärt (owl:Class) und jede untergeordnete Benennung als eine Unterklasse (rdfs:subClassOf). Dabei gab es noch zusätzliche Einschränkungen, z. B. owl:disjointW- oder owl:someValuesFrom-Aussagen. Doch gab es keine zusätzlichen Eigenschaften (außer den definierten OWLDL-Aussagen).

Um die ICC mit dem RDF/OWL EuroWordNet zu erweitern haben wir einen zweistufigen Ansatz benutzt:

1. Konvertierung des ICC-Format in das EuroWordNet-OWL-Format
2. Integration der konvertierten Daten in die EuroWordNet-OWL-Hierarchie

Vor der Konvertierung der Wissensgebiete der ICC (Tab. 1) haben wir jedes einzelne Wissensgebiet der ICC mit der entsprechenden WordNet Domäne, wie sie in Tabelle 2 dargestellt wird, verglichen. Diese Vergleiche gaben uns die Möglichkeit, das neue Wissen über die EuroWordNet Hierarchie hinzuzufügen.

Alle ICC-Klassen haben wir zunächst in RDF/OWL-Synset-Klassen konvertiert (z. B. ewn20Schema:NounSynset), sodass sie leichter in die OWL-EuroWordNet-Hierarchie eingefügt werden können. Wir versuchten die Wortbedeutungen jedes ICC-Wissensgebiets eindeutig zu machen, um einen korrekten Vergleich mit dem EuroWordNet Synset und der WordNet-Domain zu finden. Danach erweiterten wir die EuroWordNet Anzahl mit den ICC-Oberbegriffen. Jeder ICC-Oberbegriff wird als eine Klasse deklariert (owl:Class) und jeder Unterbegriff als eine Unterklasse (rdfs:subClassOf).

Die daraus hervorgehende ICC-Top-Ontologie wird ergänzt durch Sprachbeschreibung für jede Klasse (z. B. xml:lang=“en“), so dass wir damit eine korrekte Anzahl von Sprachdateien erhalten, falls vorhanden. Dieselbe Prozedur wurde bei der Pizza.owl und Reise.owl-Ontologie verwendet. Eine mehr ins Einzelne gehende Beschreibung dieser Ergänzungsarbeit ist in De Luca et al. (2007) enthalten.

6 Schlussbemerkungen

Das Daten-Web mit seinen kanonischen Datensätzen (wie z. B. DBpedia, geographische und biologische Daten, Daten der Social Networks, bibliographische, Musik- und Multimediadaten) sowie den Daten aus dem Einsatz von RDFa, Mikroformaten usw. hat schließlich zu einer empirischen Basis für das semantische Netz geführt und damit indirekt auch zur Wissenstechnologie beigetragen. Einerseits wurde damit in den letzten Jahren viel Wert auf die Darstellung lexikalischer Bedeutung gelegt und andererseits hat die Entwicklung lexikalisch-semantischer Ressourcen eine Fülle von Daten mit ihren entsprechenden Implikationen hervorgebracht.

Viele der produzierten Daten bergen allerdings Probleme in ihrer eigenen Struktur und zweitens sind große Datensammlungen bezüglich ihrer Einsatzmöglichkeiten schwer zu beschreiben. Was wird durch bestimmte Daten eigentlich beschrieben, wie sind sie typischerweise angeordnet?

Wörtersammlungen helfen da nicht viel, sie liefern eine Menge von Aussagen und Axiomen, die der Größe und der Gestaltung von Daten im Allgemeinen nicht angepasst sind, denn Größe und Gestalt können nur empirisch ermittelt werden.

Dagegen ist die Anwendung des Wissens über ein Wissensgebiet sehr hilfreich, weil dies eine Hierarchisierung der verschiedenen Begriffe erlaubt, die in die verfügbare lexikalisch vernetzte Data Cloud integriert werden kann.

Dieser Beitrag befasst sich mit diesen Möglichkeiten und zeigt, wie eine Art Hybrid-Forschung dazu führt, die Wissensgebiete der ICC in die LOD Cloud (siehe Abb. 4) einbringen zu können. Dieser Beitrag erbringt daher Zweierlei: (1) die Produktion und Veröffentlichung von lexikalischen LOD-Datensätzen und (2) die Beschreibung einer Methode zur Herstellung einer gemeinsamen lexikalisch-vernetzten Wissensdatenbank mit entsprechender Anreicherung innerhalb der Wissensgebiete der ICC.

Tab. 2: Mapping zwischen WordNet-Domains, DDC und ICC-Codes.

TOP-LEVEL	BASIC DOMAINS	DDC CODES	ICC CODES
Humanities			
	History	[920–990]	[69]
	Linguistics	410	[91]
	Literature	[800, 400]	[92]
	Philosophy	[100–(130.150.176)]	[97]
	Psychology	150	[55]
	Art	[700–(710, 720, 745.5, 790–(791.43, 792, 793.3))]	[94]
	Paranormal	130	[557]
	Religion	200	[98–99]
Free_Time		[790–(791.43, 792,793.3)]	[578]
	Radio-TV	[791.44,791.45]	[88, 866]
	Play	[793.4 : 795–794.6]	[578]
	Sport	[794.6,796 : 799]	[58]
Applied_Science		600	[..8]
	Agriculture	[338.1, 630]	[47]
	Food	[613.2, 613.3, 641, 642]	[48]
	Home	[640–(641, 642, 645)]	[49]
	Architecture	[645, 690, 710, 720]	[756, 946]
	Computer_Science	[004 : 006]	[83]
	Engineering	620	[73]
	Telecommunications	[383, 384]	[88]
	Medicine	[610–(611, 612–612.6)]	[53–54]
Pure_Science		500	[2, 21–29]
	Astronomy	520	[31]
	Biology	[570–577, 611, 612–612.6]	[41–44]
	Animals	590	[44–45]
	Plants	580	[42–43]
	Environment	577	[49]
	Chemistry	540	[25–27]
	Earth	[550, 560, 910–(910.4, 910.202)]	[32–39]
	Mathematics	510	[12]
	Physics	530	[21–23]
Social_Science		[300.1 : 300.9]	[6]
	Anthropology	[301 : 307, 395, 398]	[511]
	Health	[613–(613.2, 613.3, 613.8, 613.9)]	[52]
	Military	[355 : 359]	[68]
	Pedagogy	370	[56]
	Publishing	70	[87]
	Sociology	[301 : 319–(305.8, 306.7), 360–(363.4, 368)]	[61]
	Artisanship	[338.642, 745.5]	[795–797]
	Commerce	[381, 382]	[716]
	Industry	[338–(338.1, 338.642), 660, 670,680]	[715]
	Transport	[385 : 389]	[78]
	Economy	[330–(334, 338), 368, 650]	[71]
	Administration	[351 : 354]	[63]
	Law	340	[66]
	Politics	320	[62]
	Tourism	[910.202, 910.4]	[799]
	Fashion	[390–(392.6, 395, 398), 687]	[761]
		[155.3, 176, 306.7, 363.4, 392.6, 612.6, 613.96]	[516]
	Factotum		not existing

Literatur

Assem van M., Gangemi A., and Schreiber G. (2004). Wordnet in RDFS and OWL. Technical report, W3C.

Crystal, D. (2009). Semantic targeting: past, present and future. In Content Architecture, exploiting and managing diverse resources. ISKO 2009. Aslib, 2009.

Dahlberg, I. (1974). Grundlagen universaler Wissensordnung. Pullach bei München: Verlag Dokumentation, 366 S. DGD-Schriftenreihe, Bd. 3.

Dahlberg, I. (1982). ICC – Information Coding Classification – Principles, Structure and Application Possibilities. Intern.Classif. 9 (1982)2, p. 87–93.

Dahlberg, I., Ed: (1982). International Classification and Indexing Bibliography. Vol.1: Classification systems and thesauri 1950–1982. 141p. Vol. 2: Reference tools and conferences in classification and indexing. 140 p. Vol. 3: Classification and Indexing Systems, theory, structure, methodology. 211p. Vol. 1 contains also the tables of the ICC in 3 hierarchical levels.

Dahlberg, I. (2010). Information Coding Classification. Geschichtliches, Prinzipien, Inhaltliches. Information. Wiss. & Praxis 61 (2010)8, S. 449–454.

De Luca, E. W., Eul, M., Nürnberger, A. (2007). Converting EuroWordNet in OWL and Extending It with Domain Ontologies. In: Proc.Workshop on Lexical-Semantic and Ontological Resources. In Conjunction with the GLDV Conference (GLDV 2007).

De Luca, E.W. (2008). Semantic Support in Multilingual Text Retrieval. Aachen: Shaker Verlag. ISBN 978-3-8322-7489-4.

Feibleman, J.K. (1954). The Integrative Levels in Nature. British J. Philosophy of Science 17 (1954)5, S. 59–66.

Fellbaum, C. (1998). WordNet, an electronic lexical database. MIT Press.

Gödert, W. (2010). Programmatic issues and introduction. Concepts in Context. Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization July 19th–20th, 2010. Edited by Felix Boteram, Winfried Gödert, Jessica Hubrich. Würzburg: Ergon, 2011. (Bibliotheca Academica – Reihe Informations- und Bibliothekswissenschaften ;1), S. 8–12.

Hartmann, N. (1964). Der Aufbau der realen Welt. Grundriss einer allgemeinen Kategorienlehre. Berlin: W. de Gruyter. 579 S.

Ibekwe-SanJuan, I. (2009). Semantic metadata annotation: tagging medline abstracts for enhanced information access. In: Content Architecture, exploiting and managing diverse resources. ISKO. Aslib.

Lim, Lee Hong Susan, Merican, Amir Feisal, Singh, Sarinder Kaur Kashmir, Dimyati, Kaharuddin (2009) Biodiversity information retrieval across networked data sets. In: Content Architecture, exploiting and managing diverse resources. ISKO Aslib.

Matthews, Brian; Jones, Catherine ; Puzon, Bartłomiej; Moon, Jim; Tudhope, Douglas; Golub, Koraljka; Lykke, Marianne (2009). An evaluation of enhancing social tagging with a knowledge organization system. In Content Architecture, exploiting and managing diverse resources. ISKO UK Conference, London, 22–23 June 2009. In: Content Architecture, exploiting and managing diverse resources. ISKO Aslib. http://www.iskouk.org/conf2009/papers/matthews_ISKOUK2009.pdf [1.7.20014]

Ménard, E. (2009). Ordinary image retrieval in a multilingual context: a comparison of two indexing vocabularies. In: Content Architecture, exploiting and managing diverse resources. ISKO. Aslib.

Pianta E., Bentivogli L. Girardi C. (2002) Multiwordnet: developing an aligned multilingual database. In: First International Conference on Global WordNet, Mysore, India.

Smethurst, M., Scott, T. (2009): Building coherence at bbc.co.uk. In: Content Architecture, exploiting and managing diverse resources. ISKO. Aslib.

Prof. Dr.-Ing. Ernesto William De Luca
Fachbereich Informationswissenschaften
Friedrich-Ebert-Straße 4
14467 Potsdam
Telefon 0331 580 1520
Fax 0331 580 1599
deluca@fh-potsdam.de
http://informationswissenschaften.fh-potsdam.de

Prof. Dr.-Ing. Ernesto William De Luca ist Professor für Informationswissenschaft und Direktor des Instituts für Information und Dokumentation an der Fachhochschule Potsdam. Zu seinen Forschungsgebieten gehören Semantic Web Technologien, Empfehlungssysteme und Information Retrieval. Er hat über 80 Beiträge für nationale und internationale Konferenzen, Bücher und Zeitschriften in diesen Fachgebieten verfasst, zahlreiche Workshops organisiert sowie in Programmkomitees hochrangiger Konferenzen mitgewirkt.

Dr. Ingetraut Dahlberg
Am Hirtenberg 13
64732 Bad König
Telefon 06063 913857
Fax 06063 577607
ingetraut.dahlberg@t-online.de

Promotion in Philosophie, Sprachwissenschaft und Wissenschaftsgeschichte. Leiterin der DGD-Bibliothek und Dokumentationsstelle 1965 bis 1970. Gründerin und Vorsitzende der Gesellschaft für Klassifikation (1977 bis 1986) sowie Präsidentin der Internationalen Gesellschaft für Wissensorganisation (ISKO) (1989 bis 1996). Lehraufträge an den Universitäten Mainz und Saarbrücken sowie den Fachhochschulen Hannover und Darmstadt. Verlegerin (INDEKS Verl. 1979 bis 1996). Schriftleitung: International Classification/ Knowledge Organisation (1974 bis 1996).