Für eine pan-europäische Lexikologie und Lexikographie mittels des Linked Open Data Frameworks

Thierry Declerck, DFKI GmbH, Saarbrücken, Deutschland

Eveline Wandl-Vogt. ÖAW, Wien, Österreich

Einleitung

Das rasche Wachstum an Webinhalten erfordert innovative Lösungen für die automatische Analyse von solchen Inhalten. Nur mit derartigen Lösungen können Herausforderungen in Szenarien wie die umfangreiche Analyse und Interpretation von heterogenen Datenmengen verschiedener Sprachen, Medien und aus diversen Organisationen bewältigt werden.

Für die inhaltliche Analyse von täglich neu produzierten, heterogenen, mehrsprachigen und multimedialen Inhalten greifen sprachtechnologische Anwendungen immer häufiger auf sprach- und medienunabhängige Datenanalysen und Repräsentationsmethoden wie etwa Linked Data[1] und Semantic Web Technologien zurück.

Notwendig dafür ist die Darstellung von sprach- und medienspezifischen linguistischen Informationen auf einer semantischen Ebene. Nur so wird eine Form von Analytics möglich, welche auf die zunehmende Bandbreite von Medien und menschlichen Sprachen angewendet werden kann, welche sich heute im Web findet.

Dabei spielen lexikalische Ressourcen eine wesentliche Rolle. Und weiterdenkend, stellt sich die Frage inwiefern lexikalische Ressource nicht nur als Basis für die semantische Analyse von Webinhalten sich eignen, sondern ob sie nicht im gleichen Format im Web zu stehen haben als die Wissensobjekte, die sie ja auch beschreiben. Dies ist auch eine zentrale Frage der modernen digitalisierten Lexikographie: wie werden Wörterbücher konzipiert, in einer Zeit in der Sprachdaten unterschiedlichsten Typs in digitaler Form vorliegen und zugreifbar sind? Wie soll ein Wörterbuch in diesem neuen Kontext aussehen?

Wir beschreiben in diesem Beitrag, wie Vertreter von zwei Wissenschaftsgemeinden – Sprach- und Semantic Web Technologie auf der einen Seite, und Lexikologie und Lexikographie auf der anderen Seite, sich diese Fragestellung annehmen, auch im Rahmen von zwei Europäischen Projekten, die wir auch kurz beschreiben.

[1] S. http://linkeddata.org/ für Details.

Das LIDER Projekt

LIDER (http://www.lider-project.eu/) schafft die Grundlage für ein Ökosystem aus frei verfügbaren, verlinkten und semantisch interoperablen Ressourcen. LIDER untersucht, wie Sprache („Linguistic Linked Data"-Repräsentationen[2] von Korpora, Wörterbüchern, lexikalischen und syntaktischer Metadaten etc.) und multimediale Daten (Bild, Video etc.) als Basistechnologie für die Analyse unternehmensweiter, mehrsprachiger und cross-medialer Inhalte im Netz fungieren können.

LIDER hilft dabei, eine Community zur Linguistic Linked Licensed Data (3LD) zu gründen, in der unter Linguistic Linked Data sowohl frei verfügbare linguistische Ressourcen als auch lizensierte linguistische Daten verstanden werden.

LIDER arbeitet auch an einer Referenzarchitektur für das Erstellen von Linguistic Linked Data auf Basis von bereits existierenden und zukünftigen Plattformen sowie von frei verfügbaren Quellen.

Das ENeL Projekt

ENeL (http://www.cost.eu/domains_actions/isch/Actions/IS1305) ist eine so-genannte COST Aktion der Europäischen Union und der Kommission zur Unterstützung von paneuropäischer Forschung. ENeL zielt auf dem Aufbau eines Europäischen Netzwerks für e-Lexikographie (Enel).

Die Arbeitsgruppen von ENeL setzen sich mit der Tatsache auseinander, dass Computer und die Verfügbarkeit des World Wide Web (WWW) die Bedingungen für die Produktion und Rezeption von Wörterbüchern deutlich verändert haben. Für Redakteure wissenschaftlicher Wörterbücher ist das WWW nicht nur eine Quelle der Inspiration, sondern auch eine neue und große Herausforderung. Zum Beispiel wenn es darum geht, die Lücke zwischen der Öffentlichkeit und wissenschaftlichen Wörterbüchern zu schließen und dabei den Benutzern einfacher Zugang zu wissenschaftlicher Wörterbüchern zu gewährleisten. Es wird auch versucht, einen breiteren und systematischen Austausch von Know-how und gemeinsamen Standards und Lösungen zu schaffen und ein gemeinsames Konzept zu entwickeln, wie die Grundlage für eine neue Art der Lexikographie auszusehen hat. Darüber hinaus wird der paneuropäische Charakter der lexikographischen Arbeit in Europa in den Mittelpunkt gesetzt. Wie kann man von nationalen Wörterbücherprojekten zu supranationalen Wörterbüchern gelangen? Wie kann die Mehrsprachigkeit des Europäischen Bürgers optimal berücksichtigt werden? Gibt es paneuropäische Wörter oder Begriffe? Gibt es gemeinsame Neologismen, die in Wörterbücher berücksichtigt werden müssen? So dass sich die Frage stellt, ob es neben paneuropäischen Korpora (wie Europarl, s. http://www.statmt.org/europarl/), auch paneuropäische Wörterbücher geben kann, oder soll?

In diesem Beitrag präsentieren wir einen implementierten Ansatz zu einer Linked Data konformen Modellierung von lexikographischen Daten, die eine Vernetzung und

[2] Siehe auch http://linguistics.okfn.org/resources/llod/.

Integration von bestehenden multilingualen lexikalischen und enzyklopädischen Ressourcen erlaubt.

Linked (Open) Data und Linguistic Linked Open Data

Wir geben hier die Definition, die in Wikipedia steht: **„Linked Open Data (LOD)** bezeichnet im World Wide Web frei verfügbare Daten, die per Uniform Resource Identifier (URI) identifiziert sind und darüber direkt per HTTP abgerufen werden können und ebenfalls per URI auf andere Daten verweisen. Idealerweise werden zur Kodierung und Verlinkung der Daten das Resource Description Framework (RDF) und darauf aufbauende Standards wie SPARQL und die Web Ontology Language (OWL) verwendet, so dass Linked Open Data gleichzeitig Teil des Semantic Web ist. Die miteinander verknüpften Daten ergeben ein weltweites Netz, das auch als „Linked [Open] Data Cloud" oder „Giant Global Graph" bezeichnet wird. Dort wo der Schwerpunkt weniger auf der freien Nutzbarkeit der Daten wie bei freien Inhalten liegt (Open Data), ist auch die Bezeichnung **Linked Data** üblich." (http://www.wikiwand.com/de/Linked_Open_Data, konsultiert am 2014-11-10)

Die ersten solchen Datenmengen im LOD waren ursprünglich klassische Wissensobjekte, die enzyklopädischer Natur sind. Aber in den letzten Jahren sind Bestrebungen aktiv gewesen, auch linguistisches Wissen so zu kodieren, und eine „Linguistic Linked Open Data" Gemeinschaft ist entstanden (s. http://linguistics.okfn.org/resources/llod/). Eine graphische Darstellung des aktuellen Standes des „Linguistic Linked Open Data cloud diagram" ist auf einer Extraseite am Ende dieses Beitrages, im Anhang, wiedergegeben. Und genau in diesem Rahmen findet die Kooperation zwischen den ENeL und LIDER Projekten statt. Als Repräsentationsformalismus für die LOD konformen Modellierung von lexikographischen Daten, die ENeL Teilnehmer uns zu Verfügung gestellt haben, verwenden wird das Ontolex Modell, das im nächsten Abschnitt eingeführt wird.

Das Ontolex Modell

Das Ontolex Modell wird im Rahmen eines W3C Vorhabens[3], einer so-genannten Community Group, diskutiert und steht kurz vor der offiziellen Veröffentlichung. Es basiert auf dem Modell *lemon* (s. J. McCrae et al., 2012) und auch auf dem ISO Modell Lexical Markup Framework (LMF; http://www.lexicalmarkupframework.org/). Diese Modelle beschreiben einen modularen Ansatz zur Lexikonbeschreibung, und dadurch verliert die traditionelle Sicht an Bedeutung, dass der „Einstieg" zu einem Lexikoneintrag beim sogenannten „Headword" stattzufinden hat. Alle Elemente eines Wörterbucheintrages sind gleichwertig und können unabhängig voneinander beschrieben und durch expliziten Relationsmarker miteinander verbunden werden. Das neue mit *lemon* und Ontolex ist, dass die Komponente eines Lexikoneintrages im Netz verteilt werden können und durch RDF Relationen („properties") miteinander verlinkt werden. Praktisch heißt das, dass ein Wörterbuchautor nicht alle

[3] Siehe http://www.w3.org/community/ontolex/.

Komponente oder Elemente eines Eintrages detailliert beschreiben und an einem „Ort“ halten muss, aber dass sie/er auch auf bestehende Elemente (zum Beispiel die Etymologie eines Wortes) zurückgreifen kann, und einfach darauf verweisen kann. Wir sind überzeugt, dass diese Eigenschaften des Modells die Zusammenarbeit zwischen verschiedenen wissenschaftlichen Lexikographen ermöglichen und unterstützen können, und dass dadurch virtuelle Forschungsumgebungen im lexikographischen Bereich entstehen können.

Die RDF basierte Verlinkung gewährleistet, dass der Eintrag dennoch eine Einheit bleibt, in der Form eines genannten Graphs. Abbildung 1 unten zeigt eine graphische Darstellung von Ontolex.

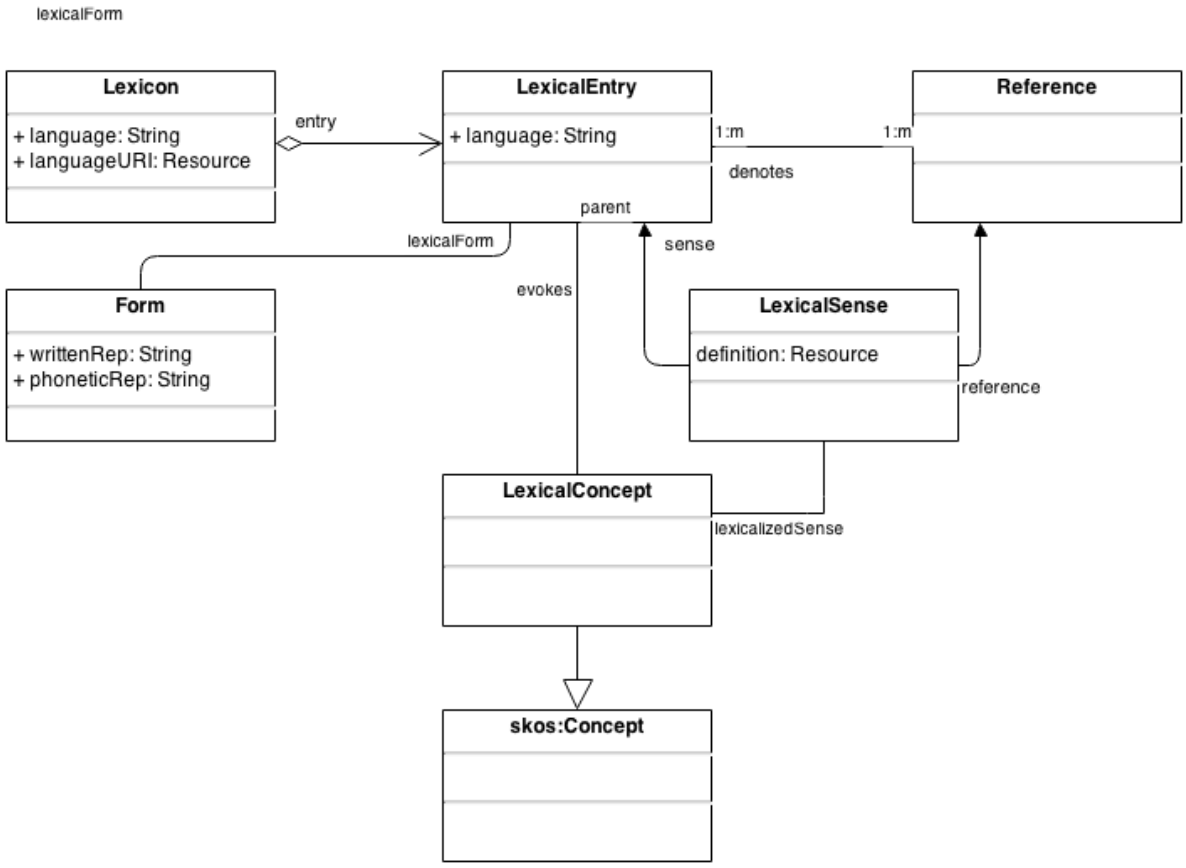

Abbildung 1: Graphische Darstellung des Ontolex Modells

Unsere Experimente

Von ENeL Teilnehmern haben wir lexikographische Daten erhalten, die wir dann in Ontolex umgewandelt haben. Es handelt sich momentan um:

- 2 Österreichische Dialektwörterbücher (in Tustep/XML und Word)
- Beispiele aus einem Slowakischen Wörterbuch (in XML, + PDF/Word)
- Ein Slowenisches Wörterbuch (in XML, basiert auf dem LMF Standard)
- 2 Wörterbücher von Arabischen Dialekten (in TEI kodiert)
- 1 Beispiel aus einem Baskisch-Deutsch Wörterbuch (in XML)
- 1 Beinspiel aus eine Französischen Wiktionary-basierten Wörterbuch

- 1 Konzept-basiertes Wörterbuch des Limburgischen (in Excel)
- 1 Beispiel aus dem multilingualen KDictionary (in XML)
- Beispiele aus dem Digital Scottisch Lexicon (Old Scottisch, html + 1 Beispiel in TEI)

Wir haben alle Daten zunächst “manuell” analysiert und einzelne Einträge manuell in ein Ontologie Editierwerkzeug nach den Ontolex Modell eingetragen. Waren wir mit der Modellierung zufrieden, sind dann Skripte geschrieben worden, die die einzelnen Quellen dann vollständig in das RDF Format von Ontolex übertragen haben. In manchen Fällen haben wir gesehen, dass Lemmata von Einträgen eine Bedeutung mit anderen Lemmata (auch aus anderen Lexikons) teilen, so dass direkte (auch multilinguale) Relationen zwischen solchen Elemente automatisch etabliert werden können.

Für einzelne Fälle haben wir dann (manuell) einen Link zwischen Lexikonelementen und externen enzyklopädischen Quellen eingefügt, um zu zeigen, wie lexikalischen Daten mit Daten eines anderen Typs effizient verlinkt werden können. Abbildung 2 unten zeigt ein Screenshot aus unserem Ontologie Editor, aus dem der Leser einige Elemente des Ontolex Modell sehen kann.

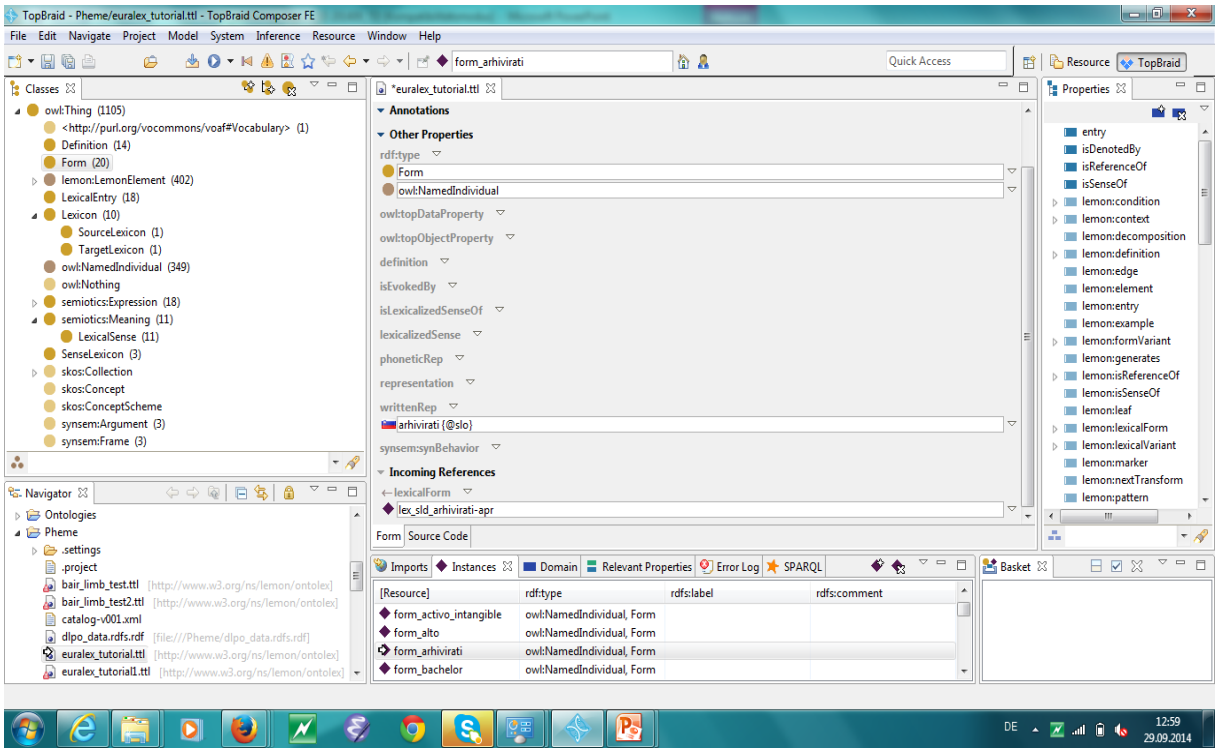

Abbildung 2: Ein Schnappschuss aus dem Ontologie Editor

Referenzen

Thierry Declerck, Eveline Wandl-Vogt. Cross-linking Austrian dialectal Dictionaries through formalized Meanings. in: Andrea Abel, Chiara Vettori, Natascia Ralli (eds.): *Proceedings of the XVI EURALEX International Congress, Pages 329-343.*

Thierry Declerck, Eveline Wandl-Vogt. How to semantically relate dialectal Dictionaries in the Linked Data Framework. *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014), Gothenburg, Sweden, ACL, 4/2014*

Maud Ehrmann, Francesca Cecconi, Daniele Vannella, John P. McCrae, Philipp Cimiano, and Roberto Navigli. A Multilingual Semantic Network as Linked Data: lemon-BabelNet**.** *Proceedings of the 3rd Workshop on Linked Data in Linguistics*

Philipp Cimiano and Christina Unger**. Multilingualität und Linked Data.** *In: Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien* (Editors: Tassilo Pellegrini, Harald Sack, and Sören Auer)

J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. Interchanging lexical resources on the Semantic Web**.** *Language Resources and Evaluation* (2012).

Georg Rehm and Felix Sasaki**.** Semantische Technologien und Standards für das mehrsprachige Europa (Editors: B. Humm Ege, B. and A. Reibold eds. Corporate Semantic Web)

Anhang

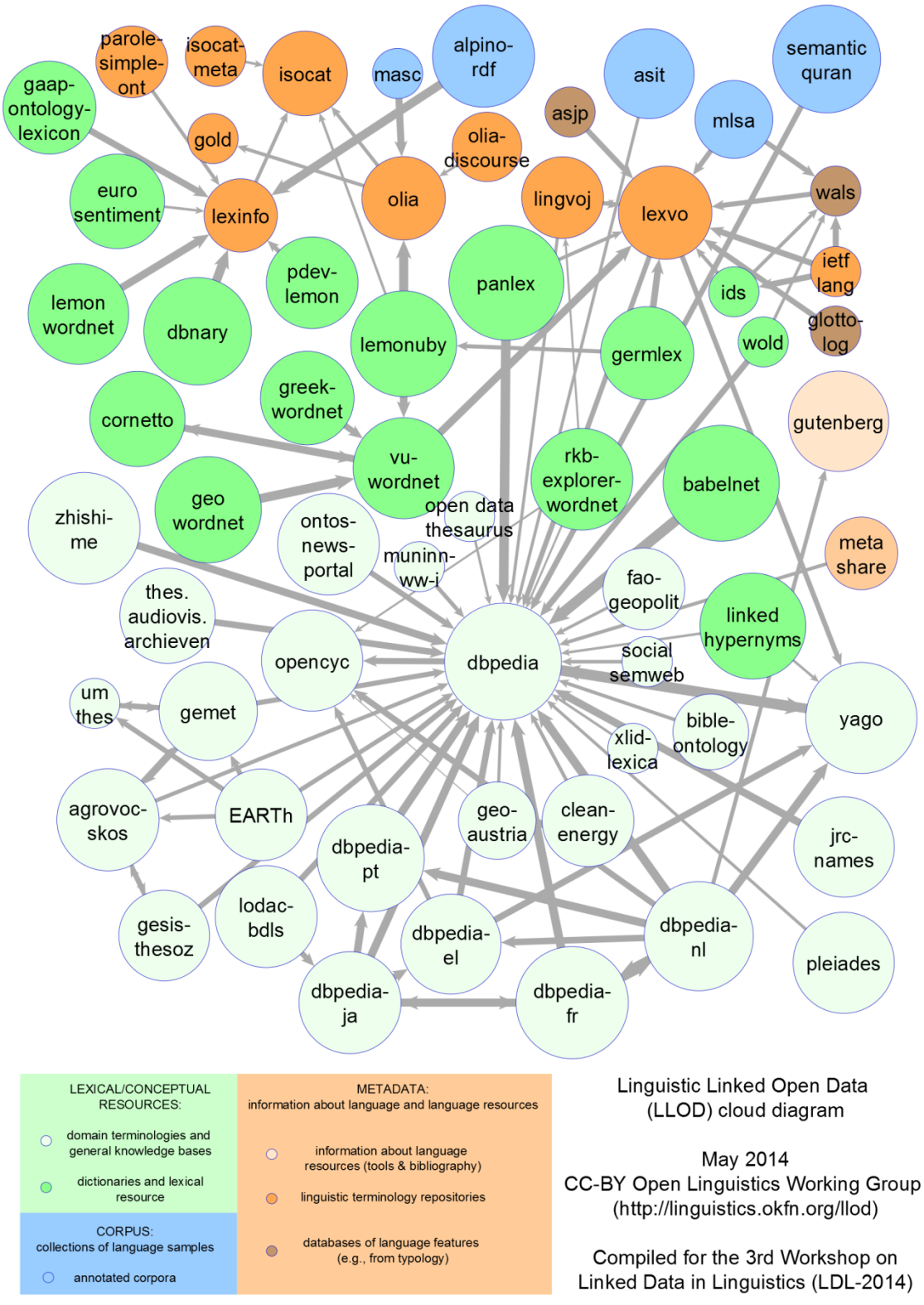

Abbildung 3: Die graphisch Darstellung des Linguistic Linked Open Data clouds