

Annabella Schmitz

RDF-LIFTING VON OntoLex-Lemon AUS DEM DIGITALEN FAMILIENNAMENWÖRTERBUCH DEUTSCHLANDS MIT XTriples

Keywords Linked Open Data; RDF; Digital Lexicography; XML; Digital Humanities

Das digitale Familiennamenwörterbuch Deutschlands (DFD) ist ein Forschungsprojekt zwischen der Akademie der Wissenschaften und der Literatur Mainz, der Technischen Universität Darmstadt und der Johannes Gutenberg-Universität Mainz. Anhand der Telefonanschlüsse der Telekom aus dem Jahr 2005 sollen in Deutschland vorkommende Familiennamen digital lexikographisch erfasst werden. Berücksichtigt werden hier Bedeutungen der Familiennamen, Zuordnung zu festgelegten Kategorien, Häufigkeit der Vorkommen eines Namens in Deutschland und, wenn vorhanden, im Ausland, Sprachvorkommen sowie die Kartierung der geographischen Verbreitung der Familiennamen in Deutschland.

Um die Bestandteile des Familiennamenwörterbuchs als Linked Open Data bereitzustellen und eine Vernetzung im Semantic Web zu etablieren, bietet sich eine Modellierung mit der OntoLex-Lemon Ontologie (Cimiano/McCrae/Buitelaar (Hg.) 2016) an. OntoLex-Lemon ist ein Vokabular zur Darstellung von lexikographischen Ressourcen als RDF (Resource Descriptive Framework). Es besteht aus mehreren Modulen, die sich je nach Intention des Lexikons wählen lassen. Für das Digitale Familiennamenwörterbuch eigneten sich das lexikographische Modul (Bosque-Gil/Gracia (Hg.) 2019) und das Kernmodul am besten. Das lexikographische Modul, da hiermit die Struktur der ursprünglichen Ressource beibehalten werden kann, und das Kernmodul zur Darstellung von weiteren linguistischen Eigenschaften und semantischen Verlinkungen mit anderen Webseiten, die nicht der genauen Struktur der Webseite des Digitalen Familiennamenwörterbuchs entsprechen. Zusätzlich wurden für das RDF-Modell mit OntoLex noch weitere Vokabulare herangezogen, um die Darstellung des Wörterbuchs vollständig zu machen. Letztendlich können mit OntoLex-Lemon semantische Aussagen über das Familiennamenwörterbuch und dessen lexikalische Einträge getroffen werden. So ist eine Modellierung mit OntoLex-Lemon sehr hilfreich, um die semantischen Aussagen des Digitalen Familiennamenwörterbuchs mit anderen lexikalischen Ressourcen vernetzbar und vergleichbar zu machen. Eine beispielhafte Modellierung eines Familiennamenartikels ist in der Abbildung 1 anhand des Familiennamens „Eis“ zu sehen.

Damit dann schließlich das OntoLex-Lemon-Modell für das Digitale Familiennamenwörterbuch in Turtle (Terse RDF Triple Language), einer Syntax für RDF, für alle publizierten Familiennamenartikel umgesetzt werden kann, kann der Webservice XTriples herangezogen werden.

Mit diesem Webservice lassen sich RDF-Triples (semantische Aussagen, bestehend aus Subjekt, Prädikat und Objekt) aus XML-Daten extrahieren, die dann in verschiedenen Output-Formaten verfügbar gemacht werden. Neben Turtle, was für das DFD generiert werden soll, stehen noch RDF/XML, N-Triples, N-Quads, TriX, JSON und SVG Graph als Output zur Verfügung. Das RDF-Modell kann mit einer einfachen Konfiguration erstellt werden, indem die XML-Daten der Familiennamenartikel des Digitalen Familiennamenwörterbuchs

als Rest-Schnittstelle in der Konfiguration hinterlegt werden. Die Konfiguration für die XTriples besteht aus einer XML-ähnlichen Struktur und es wird für jede RDF-Aussage des gewünschten Ziel-Modells ein Subjekt, ein Prädikat und ein Objekt in der Konfiguration definiert. Mögliche semantische Aussagen eines Artikels des Digitalen Familiennamenwörterbuchs mit ihrer Umsetzung in Turtle wären:

- 1) Das DFD ist eine lexikographische Ressource.
 - :dfd a lexicog:LexicographicResource.
- 2) Das DFD hat den Eintrag „Eis“.
 - :dfd lexicog:entry lexicog:LexicographicResource.
- 3) „Eis“ ist ein lexikographischer Eintrag.
 - <http://www.namenforschung.net/id/name/9815/1> a lexicog:Entry.

So wird dann jeder einzelne hinterlegte Artikel auf dieser Schnittstelle gecrawlt und mittels XPath-Ausdrücken werden Informationen für die einzelnen RDF-Statements in der Konfiguration aus den XML-Daten gelesen und von einer XML-Struktur in einen RDF-Graphen bzw. ein anderes mögliches Ausgabeformat abgeleitet (RDF-Lifting).

Da alle Familiennamenartikel für das Digitale Familiennamenwörterbuch als TEI-XML-Artikel erstellt werden, sind alle dazu benötigten Daten schon in einer durchsuchbaren Struktur vorhanden und können anhand des Modells in ihrer Gesamtheit mithilfe einer beispielhaften Konfiguration extrahiert werden, sodass das beispielhafte Modell (Abb. 1) für jeden Familiennamenartikel in der gezeigten Struktur umgesetzt werden kann.

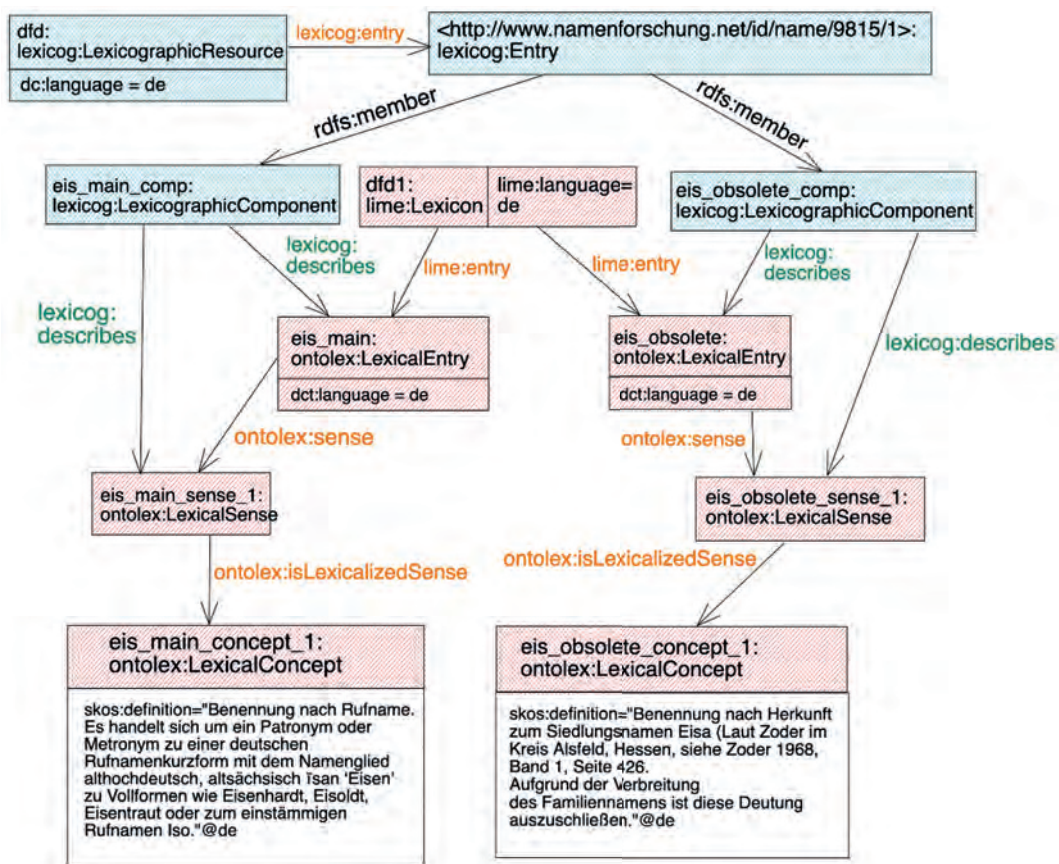


Abb. 1: OntoLex-Kernmodell des Familiennamens „Eis“ (Griebel 2022)

Folglich kann mithilfe von XTriples und mit dem RDF-Vokabular OntoLex ein Wörterbuch wie das Digitale Familiennamenwörterbuch, welches die Artikel in XML-Daten vorliegen hat, schnell, einfach und effizient als RDF/Turtle und damit auch als Linked Open Data bereitgestellt werden.

Literatur

Bosque-Gil, J./Gracia, J. (Hg.) (2019): The OntoLex Lemon Lexicography Module: Final Community Group Report 17 September 2019. <https://www.w3.org/2019/09/lexicog/> (Stand: 25.03.22).

Cimiano, P./McCrae, J./Buitelaar, P. (Hg.) (2016): Lexicon model for ontologies: community report. <https://www.w3.org/2016/05/ontolex/> (Stand: 24.03.22).

Digitales Familiennamenwörterbuch Deutschlands:
<https://www.namenforschung.net/dfd/woerterbuch/liste/> (Stand: 24.03.22).

Griebel, J. (2022): Eis. Digitales Familiennamenwörterbuch Deutschlands.
<http://www.namenforschung.net/id/name/9815/1> (Stand: 25.03.22).

XTriples: <https://xtriples.lod.academy/index.html> (Stand: 25.03.22).

Kontaktinformationen

Annabella Schmitz

Akademie der Wissenschaften und der Literatur Mainz
annabella.schmitz@adwmainz.de