Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

Doktor-Ingenieur
(Dr.-Ing.)

im Fachgebiet
Informatik

vorgelegt

von **Dipl.-Inf. Sebastian Hellmann**

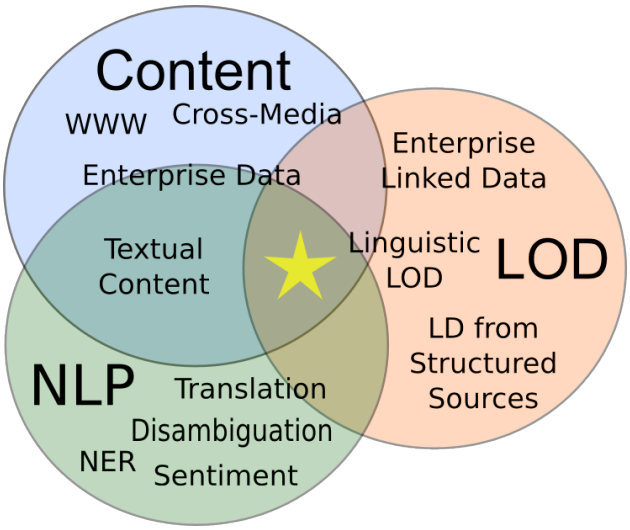
geboren am 14. März 1981 in Göttingen, Deutschland

Die Annahme der Dissertation wurde empfohlen von:
1. Prof. Dr. Klaus-Peter Fähnrich, Universität Leipzig
2. Prof. Dr. Hans Uszkoreit, Universität des Saarlandes

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 01.09.2014 mit dem Gesamtprädikat
magna cum laude.

INTEGRATING NATURAL LANGUAGE PROCESSING (NLP) AND LANGUAGE RESOURCES USING LINKED DATA

SEBASTIAN HELLMANN



Universität Leipzig

January 8, 2015

AUTHOR:
Dipl. Inf. Sebastian Hellmann

TITLE:
Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data

INSTITUTION:
Institut für Informatik, Fakultät für Mathematik und Informatik, Universität Leipzig

BIBLIOGRAPHIC DATA:
2013, XX, 197p., 33 illus. in color., 8 tables

SUPERVISORS:
Prof. Dr. Klaus-Peter Fähnrich
Prof. Dr. Sören Auer
Dr. Jens Lehman

© January 8, 2015

Für Hanne,
meine Eltern Anita und Lothar
und meine Schwester Anna-Maria

THESIS SUMMARY

Title:
Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data
Author:
Sebastian Hellmann
Bib. Data:
2013, XX, 197p. 33 illus. in color. 8 tab. no appendix

A GIGANTIC IDEA RESTING ON THE SHOULDERS OF A LOT OF DWARFS. This thesis is a compendium of scientific works and engineering specifications that have been contributed to a large community of stakeholders to be copied, adapted, mixed, built upon and exploited in any way possible to achieve a common goal: *Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data.*

The explosion of information technology in the last two decades has led to a substantial growth in quantity, diversity and complexity of *web-accessible linguistic data*. These resources become even more useful when linked with each other and the last few years have seen the emergence of numerous approaches in various disciplines concerned with linguistic resources and NLP tools. It is the challenge of our time to *store*, *interlink* and *exploit* this wealth of data accumulated in more than half a century of computational linguistics, of empirical, corpus-based study of language, and of computational lexicography in all its heterogeneity.

The vision of the *Giant Global Graph* (GGG) was conceived by Tim Berners-Lee aiming at connecting all data on the Web and allowing to discover new relations between this openly-accessible data. This vision has been pursued by the *Linked Open Data* (LOD) community, where the cloud of published datasets comprises 295 data repositories and more than 30 billion RDF triples (as of September 2011).

RDF is based on globally unique and accessible URIs and it was specifically designed to establish links between such URIs (or resources). This is captured in the *Linked Data paradigm* that postulates four rules: (1) Referred entities should be designated by URIs, (2) these URIs should be resolvable over HTTP, (3) data should be represented by means of standards such as RDF, (4) and a resource should include links to other resources.

Although it is difficult to precisely identify the reasons for the success of the LOD effort, advocates generally argue that open licenses as well as open access are key enablers for the growth of such a network as they provide a strong incentive for collaboration and contribution by third parties. In his keynote at BNCOD 2011, Chris Bizer argued that with RDF the overall data integration effort can be "split between data publishers, third parties, and the data consumer", a claim that can be substantiated by observing the evolution of many large data sets constituting the LOD cloud.

As written in the acknowledgement section, parts of this thesis has received numerous feedback from other scientists, practitioners and industry in many different ways. The main contributions of this the-

sis are summarized here: Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data

PART I – INTRODUCTION AND BACKGROUND. During his keynote at the Language Resource and Evaluation Conference in 2012, Sören Auer stressed the decentralized, collaborative, interlinked and interoperable nature of the Web of Data. The keynote provides strong evidence that *Semantic Web technologies such as Linked Data are on its way to become main stream for the representation of language resources*. The jointly written companion publication for the keynote was later extended as a book chapter in *The People's Web Meets NLP* and serves as the basis for Chapter 1 "Introduction" and Chapter 2 "Background", outlining some stages of the Linked Data publication and refinement chain. Both chapters stress the importance of open licenses and open access as an enabler for collaboration, the ability to interlink data on the Web as a key feature of RDF as well as provide a discussion about scalability issues and decentralization. Furthermore, we elaborate on how conceptual interoperability can be achieved by (1) re-using vocabularies, (2) agile ontology development, (3) meetings to refine and adapt ontologies and (4) tool support to enrich ontologies and match schemata.

PART II - LANGUAGE RESOURCES AS LINKED DATA. Chapter 3 "Linked Data in Linguistics" and Chapter 6 "NLP & DBpedia, an Upward Knowledge Acquisition Spiral" summarize the results of the Linked Data in Linguistics (LDL) Workshop in 2012 and the NLP & DBpedia Workshop in 2013 and give a preview of the MLOD special issue. In total, five proceedings – three published at CEUR (OKCon 2011, WoLE 2012, NLP & DBpedia 2013), one Springer book (Linked Data in Linguistics, LDL 2012) and one journal special issue (Multilingual Linked Open Data, MLOD to appear) – have been (co-)edited to create incentives for scientists to convert and publish Linked Data and thus *to contribute open and/or linguistic data to the LOD cloud*. Based on the disseminated call for papers, *152 authors contributed one or more accepted submissions* to our venues and 120 reviewers were involved in peer-reviewing.

Chapter 4 "DBpedia as a Multilingual Language Resource" and Chapter 5 "Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Linked Data Cloud" contain this thesis' contribution to the DBpedia Project in order to further increase the size and inter-linkage of the LOD Cloud with lexical-semantic resources. Our contribution comprises extracted data from Wiktionary (an online, collaborative dictionary similar to Wikipedia) in more than four languages (now six) as well as language-specific versions of DBpedia, including a quality assessment of inter-language links between Wikipedia editions and internationalized content negotiation

rules for Linked Data. In particular the work described in Chapter 4 created the foundation for a DBpedia Internationalisation Committee with *members from over 15 different languages with the common goal to push DBpedia as a free and open multilingual language resource*.

PART III - THE NLP INTERCHANGE FORMAT (NIF). Chapter 7 "NIF 2.0 Core Specification", Chapter 8 "NIF 2.0 Resources and Architecture" and Chapter 9 "Evaluation and Related Work" constitute one of the main contribution of this thesis. The *NLP Interchange Format* (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. The core specification is included in Chapter 7 and describes which URI schemes and RDF vocabularies must be used for (parts of) natural language texts and annotations in order to create *an RDF/OWL-based interoperability layer with NIF built upon Unicode Code Points in Normal Form C*. In Chapter 8, classes and properties of the *NIF Core Ontology* are described to formally define the relations between text, substrings and their URI schemes. Chapter 9 contains the evaluation of NIF.

In a questionnaire, we asked questions to 13 developers using NIF. UIMA, GATE and Stanbol are extensible NLP frameworks and NIF was not yet able to provide off-the-shelf NLP domain ontologies for all possible domains, but only for the plugins used in this study. After inspecting the software, the developers agreed however that NIF is adequate enough to provide a generic RDF output based on NIF using literal objects for annotations. All developers were able to map the internal data structure to NIF URIs to serialize RDF output (Adequacy). The development effort in hours (ranging between 3 and 40 hours) as well as the number of code lines (ranging between 110 and 445) suggest, that the implementation of NIF wrappers is easy and fast for an average developer. Furthermore the evaluation contains a comparison to other formats and an evaluation of the available URI schemes for web annotation.

In order to collect input from the wide group of stakeholders, a total of 16 presentations were given with extensive discussions and feedback, which has lead to a constant improvement of NIF from 2010 until 2013. After the release of NIF (Version 1.0) in November 2011, a total of *32 vocabulary employments and implementations for different NLP tools and converters were reported* (8 by the (co-)authors, including Wiki-link corpus (Section 11.1), 13 by people participating in our survey and 11 more, of which we have heard). Several roll-out meetings and tutorials were held (e.g. in Leipzig and Prague in 2013) and are planned (e.g. at LREC 2014).

PART IV - THE NLP INTERCHANGE FORMAT IN USE. Chapter 10 "Use Cases and Applications for NIF" and Chapter 11 "Publication

of Corpora using NIF" describe 8 concrete instances where NIF has been successfully used. One major contribution in Chapter 10 is the usage of NIF as the recommended RDF mapping in the *Internationalization Tag Set* (ITS) 2.0 W3C standard (Section 10.1) and the conversion algorithms from ITS to NIF and back (Section 10.1.1). One outcome of the discussions in the standardization meetings and telephone conferences for ITS 2.0 resulted in the conclusion there was *no alternative RDF format or vocabulary other than NIF* with the required features to fulfill the working group charter. Five further uses of NIF are described for the Ontology of Linguistic Annotations (OLiA), the RDFaCE tool, the Tiger Corpus Navigator, the OntosFeeder and visualisations of NIF using the RelFinder tool. These 8 instances provide an implemented proof-of-concept of the features of NIF.

Chapter 11 starts with describing the conversion and hosting of the huge Google Wikilinks corpus with 40 million annotations for 3 million web sites. The resulting RDF dump contains 477 million triples in a 5.6 GB compressed dump file in turtle syntax. Section 11.2 describes how NIF can be used to publish extracted facts from news feeds in the RDFLiveNews tool as Linked Data.

PART V - CONCLUSIONS. Chapter 12 provides lessons learned for NIF, conclusions and an outlook on future work. Most of the contributions are already summarized above. One particular aspect worth mentioning is the increasing number of NIF-formated corpora for Named Entity Recognition (NER) that have come into existence after the publication of the main NIF paper *Integrating NLP using Linked Data* at ISWC 2013. These include the corpora converted by Steinmetz, Knuth and Sack for the NLP & DBpedia workshop and an OpenNLP-based CoNLL converter by Brümmer. Furthermore, we are aware of three LREC 2014 submissions that leverage NIF: *NIF4OGGD - NLP Interchange Format for Open German Governmental Data*, N^3 *– A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format* and *Global Intelligent Content: Active Curation of Language Resources using Linked Data* as well as an early implementation of a GATE-based NER/NEL evaluation framework by Dojchinovski and Kliegr. Further funding for the maintenance, interlinking and publication of Linguistic Linked Data as well as support and improvements of NIF is available via the expiring LOD2 EU project, as well as the CSA EU project called LIDER (`http://lider-project.eu/`), which started in November 2013. Based on the evidence of successful adoption presented in this thesis, we can expect a decent to high chance of reaching critical mass of Linked Data technology as well as the NIF standard in the field of Natural Language Processing and Language Resources.

PUBLICATIONS

Citations at the margin were the basis for the respective sections or chapters.

This thesis is based on the following publications, books and proceedings, in which I have been author, editor or contributor. *At the respective margin of each chapter and section*, I included the references to the appropriate publications.

STANDARDS

- Section F[1] and G[2] of the W3C standard about the "Internationalization Tag Set (ITS) Version 2.0" are based on my contributions to the W3C Working Group and have been included in this thesis.
- In this thesis, I included parts of the NIF 2.0 standard[3], which was a major result of the work described here.

BOOKS AND JOURNAL SPECIAL ISSUES, (CO)-EDITED

- Linked Data in Linguistics. Representing and connecting language data and language metadata. Chiarcos, Nordhoff, and Hellmann (2012)
- Multilingual Linked Open Data (MLOD) 2012 data post proceedings. Hellmann, Moran, Brümmer, and McCrae (to appear)

PROCEEDINGS, (CO)-EDITED

- Proceedings of the 6th Open Knowledge Conference (OkCon 2011). Hellmann, Frischmuth, Auer, and Dietrich (2011)
- Proceedings of the Web of Linked Entities workshop in conjunction with the 11th International Semantic Web Conference (ISWC 2012). Rizzo, Mendes, Charton, Hellmann, and Kalyanpur (2012)
- Proceedings of the NLP and DBpedia workshop in conjunction with the 12th International Semantic Web Conference (ISWC 2013). Hellmann, Filipowska, Barriere, Mendes, and Kontokostas (2013b)

1 http://www.w3.org/TR/its20/#conversion-to-nif
2 http://www.w3.org/TR/its20/#nif-backconversion
3 http://persistence.uni-leipzig.org/nlp2rdf/specification/core.html

JOURNAL PUBLICATIONS, PEER-REVIEWED

- Internationalization of Linked Data: The case of the Greek DBpedia edition. Kontokostas et al. (2012)
- Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. Chiarcos, Hellmann, and Nordhoff (2011)
- Learning of OWL Class Descriptions on Very Large Knowledge Bases. Hellmann, Lehmann, and Auer (2009)
- DBpedia and the Live Extraction of Structured Data from Wikipedia. Morsey, Lehmann, Auer, Stadler, and Hellmann (2012)
- DBpedia - A Crystallization Point for the Web of Data. Lehmann et al. (2009)

CONFERENCE PUBLICATIONS, PEER-REVIEWED

- NIF Combinator: Combining NLP Tool Output. Hellmann, Lehmann, Auer, and Nitzschke (2012)
- OntosFeeder – A Versatile Semantic Context Provider for Web Content Authoring. Klebeck, Hellmann, Ehrlich, and Auer (2011)
- The Semantic Gap of Formalized Meaning. Hellmann (2010)
- RelFinder: Revealing Relationships in RDF Knowledge Bases. Heim, Hellmann, Lehmann, Lohmann, and Stegemann (2009)
- Integrating NLP using Linked Data. Hellmann, Lehmann, Auer, and Brümmer (2013)
- Real-time RDF extraction from unstructured data streams. Gerber et al. (2013)
- Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud. Hellmann, Brekle, and Auer (2012)
- Linked-Data Aware URI Schemes for Referencing Text Fragments. Hellmann, Lehmann, and Auer (2012)
- The TIGER Corpus Navigator. Hellmann, Unbehauen, Chiarcos, and Ngonga Ngomo (2010)
- NERD meets NIF: Lifting NLP extraction results to the linked data cloud. Rizzo, Troncy, Hellmann, and Brümmer (2012)
- Navigation-induced Knowledge Engineering by Example. Hellmann, Lehmann, Unbehauen, et al. (2012)
- LinkedGeoData - Adding a Spatial Dimension to the Web of Data. Auer, Lehmann, and Hellmann (2009)
- The Web of Data: Decentralized, collaborative, interlinked and interoperable. Auer and Hellmann (2012)
- DBpedia live extraction. Hellmann, Stadler, Lehmann, and Auer (2009)
- Triplify: Light-weight linked data publication from relational databases. Auer, Dietzold, Lehmann, Hellmann, and Aumueller (2009)

- Standardized Multilingual Language Resources for the Web of Data: `http://corpora.uni-leipzig.de/rdf`. Quasthoff, Hellmann, and Höffner (2009)
- The Open Linguistics Working Group. Chiarcos, Hellmann, Nordhoff, Moran, et al. (2012)

BOOK CHAPTERS

- Towards Web-Scale Collaborative Knowledge Extraction. Hellmann and Auer (2013)
- Knowledge Extraction from Structured Sources. Unbehauen, Hellmann, Auer, and Stadler (2012)
- The German DBpedia: A sense repository for linking entities. Hellmann, Stadler, and Lehmann (2012)
- Learning of OWL class expressions on very large knowledge bases and its applications. Hellmann, Lehmann, and Auer (2011)
- The Open Linguistics Working Group of the Open Knowledge Foundation. Chiarcos, Hellmann, and Nordhoff (2012b)

ACKNOWLEDGMENTS

I feel unable to give proper attribution to my scientific colleagues who have contributed to this thesis. Of course, I have cited the relevant work, where appropriate. There have been many other occasions, however, where feedback and guidance have been provided and work has been contributed. Although, I mention some people and also groups of people (e.g. authors, reviewers, community members), I would like to stress that there are many more people behind the scenes who were pulling strings to achieve the common goal of free, open and interoperable data and web services.

I would like to thank all colleagues with whom we jointly organized the following workshops and edited the respective books and proceedings: Philipp Frischmuth, Sören Auer and Daniel (Open Knowledge Conference 2012), Christian Chiarcos, Sebastian Nordhoff (Linked Data in Linguistics 2012), Giuseppe Rizzo, Pablo N. Mendes, Eric Charton, Aditya Kalyanpur (Web of Linked Entities 2012), Steven Moran, Martin Brümmer, John McCrae (MLODE and MLOD 2012 and 2014), Agata Filipowska, Caroline Barriere, Pablo N. Mendes and Dimitris Kontokostas (NLP & DBpedia 2013) for the collaboration on common workshops, proceedings and books. Furthermore, I would like to thank once more the 152 authors who have submitted their work to our venues and 120 reviewers for their valuable help in selecting high quality research contributions.

I would like to be thankful for all the discussions, we had on mailing lists of the Working Groups for Open Data in Linguistics, DBpedia, NLP2RDF and the Open Annotation W3C CG.

Furthermore, I would like to thank Felix Sasaki, Christian Lieske, Dominic Jones and Dave Lewis and the whole W3C Working Group for the discussions and for supporting the adoption of NIF in the W3C recommendation.

I would like to thank our colleagues from the LOD2 project and AKSW research group for their helpful comments during the development of NIF and this thesis. This work was partially supported by a grant from the European Union's 7th Framework Programme provided for the project LOD2 (GA no. 257943). Special thanks go to Martin Brümmer, Jonas Brekle and Dimitris Kontokostas as well as our future AKSW league of 7 post-docs (Martin, Seebi, Axel, Jens, Nadine, Thomas) and its advisor Sören.

I would like to thank Prof. Fähnrich for his scientific experience with the efficient organization of the process of a PhD thesis. In par-

ticular, I would like to thank Dr. Sören Auer and Dr. Jens Lehmann for their continuous help and support.

Additional thanks to Michael Unbehauen for his help with the LaTeX layout, Martin Brümmer for applying the Relfinder on NIF output to create the screenshot in Section 10.6, Dimitris Kontokostas for updating the image in Section 4.1.

CONTENTS

Part I

INTRODUCTION AND BACKGROUND

1 INTRODUCTION

Auer and Hellmann (2012); Chiarcos et al. (2011); Chiarcos, Nordhoff, and Hellmann (2012); Hellmann and Auer (2013); Hellmann, Lehmann, et al. (2013)

The vision of the *Giant Global Graph*[1](GGG) was conceived by Tim Berners-Lee aiming at connecting all data on the Web and allowing to discover new relations between the data. This vision has been pursued by the *Linked Open Data*(LOD) community, where the cloud of published datasets comprises 295 data repositories and more than 30 billion RDF triples.[2] Although it is difficult to precisely identify the reasons for the success of the LOD effort, advocates generally argue that open licenses as well as open access are key enablers for the growth of such a network as they provide a strong incentive for collaboration and contribution by third parties. Bizer (2011) argues that with RDF the overall data integration effort can be "split between data publishers, third parties, and the data consumer", a claim that can be substantiated by looking at the evolution of many large data sets constituting the LOD cloud. We outline some stages of the Linked Data publication and refinement chain (cf. Auer and Lehmann (2010); Berners-Lee (2006); Bizer (2011)) in Figure 1 and discuss these in more detail throughout this thesis.

1.1 NATURAL LANGUAGE PROCESSING

Hellmann, Lehmann, et al. (2013)

In addition to the increasing availability of open, structured and interlinked data, we are currently observing a plethora of *Natural Language Processing* (NLP) tools and services being made available and new ones appearing almost on a weekly basis. Some examples of web services providing just *Named Entity Recognition* (NER) services are

1 http://dig.csail.mit.edu/breadcrumbs/node/215
2 Version 0.3 from Sept. 2011 – http://lod-cloud.net/state/

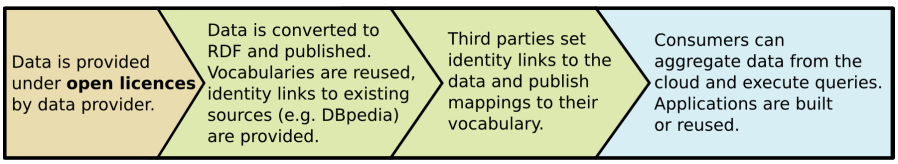


Figure 1: Summary of the above-mentioned methodologies for publishing and exploiting Linked Data (Chiarcos et al., 2011). The data provider is only required to make data available under an open license (left-most step). The remaining, data integration steps can be contributed by third parties and data consumers

Zemanta[3], *OpenCalais*[4], *Ontos*[5], *Enrycher*[6], *Extractiv*[7], *Alchemy API*[8] or *DBpedia Spotlight*[9]. Similarly, there are tools and services for language detection, part-of-speech (POS) tagging, text classification, morphological analysis, relationship extraction, sentiment analysis and many other NLP tasks. Each of the tools and services has its particular strengths and weaknesses, but exploiting the strengths and synergistically combining different tools is currently an extremely cumbersome and time consuming task. The programming interfaces and result formats of the tools have to be analyzed and differ often to a great extend. Also, once a particular set of tools is integrated this integration is *not reusable* by others.

We argue that simplifying the interoperability of different NLP tools performing similar but also complementary tasks will facilitate the comparability of results, the building of sophisticated NLP applications as well as the synergistic combination of tools. Ultimately, this might yield a boost in precision and recall for common NLP tasks. Some first evidence in that direction is provided by tools such as *RDFaCE* (Khalili, Auer, & Hladky, 2012), *Spotlight* (Mendes, Jakob, García-Silva, & Bizer, 2011) and *Fox* (Ngonga Ngomo, Heino, Lyko, Speck, & Kaltenböck, 2011)[10], which already combine the output from several backend services and achieve superior results.

Another important factor for improving the quality of NLP tools is the availability of large quantities of qualitative background knowledge on the currently emerging Web of Linked Data (Auer & Lehmann, 2010). Many NLP tasks can greatly benefit from making use of this wealth of knowledge being available on the Web in structured form as *Linked Open Data* (LOD). The precision and recall of Named Entity Recognition, for example, can be boosted when using background knowledge from *DBpedia*, *Geonames* or other LOD sources as crowdsourced and community-reviewed and timely-updated gazetteers. Figure 2 shows a snapshot of the LOD cloud with highlighted language resources that are relevant for NLP.

Of course the use of gazetteers is a common practice in NLP. However, before the arrival of large amounts of Linked Open Data their creation, curation and maintenance in particular for multi-domain NLP applications was often impractical.

The use of LOD background knowledge in NLP applications poses some particular challenges. These include:

3 http://www.zemanta.com/
4 http://www.opencalais.com/
5 http://www.ontos.com/
6 http://enrycher.ijs.si/
7 http://extractiv.com/
8 http://www.alchemyapi.com/
9 http://spotlight.dbpedia.org
10 http://aksw.org/Projects/FOX

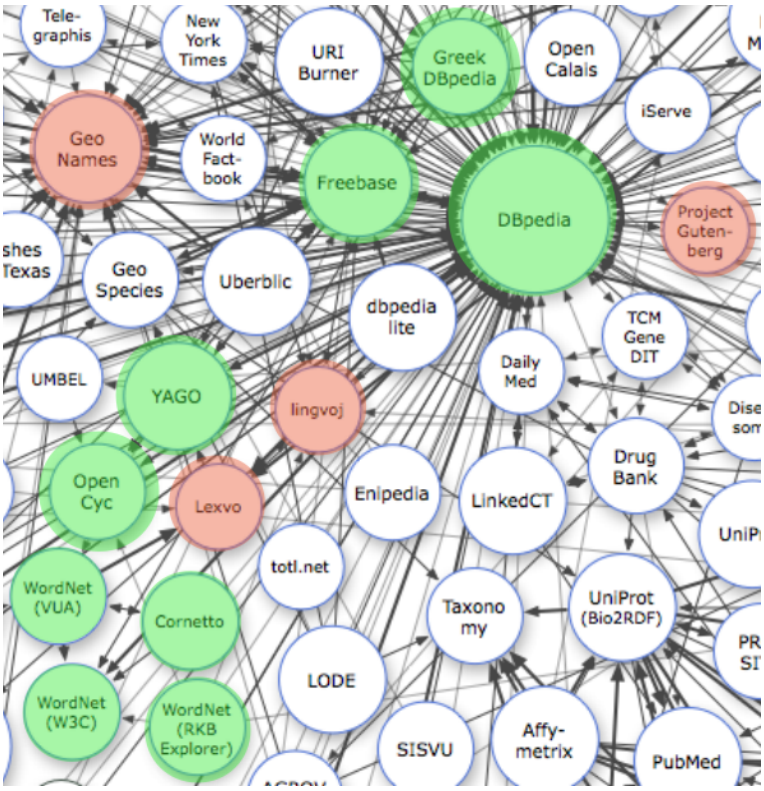


Figure 2: Language resources in the LOD cloud (as of September 2012). Lexical-semantic resources are colored green and linguistic meta data red.

- *identification* – uniquely identifying and reusing identifiers for (parts of) text, entities, relationships, NLP concepts and annotations etc.;
- *provenance* – tracking the lineage of text and annotations across tools, domains and applications;
- *semantic alignment* – tackle the semantic heterogeneity of background knowledge as well as concepts used by different NLP tools and tasks.

1.2 OPEN LICENSES, OPEN ACCESS AND COLLABORATION

Chiarcos et al. (2011)

DBpedia, FlickrWrappr, 2000 U.S. Census, LinkedGeoData, LinkedMDB are some prominent examples of LOD data sets, where the conversion, interlinking, as well as the hosting of the links and the converted RDF data has been completely provided by third parties with no effort and cost for the original data providers.[11] DBpedia (Lehmann et al., 2009), for example, was initially converted to RDF solely from the openly licensed database dumps provided by Wikipedia. With

11 More data sets can be explored here: http://thedatahub.org/tag/published-by-third-party

Openlink Software a company supported the project by providing hosting infrastructure and a community evolved, which created links and applications. Although it is difficult to determine whether open licenses are a necessary or sufficient condition for the collaborative evolution of a data set, the opposite is quite obvious: *Closed* licenses or *unclearly licensed* data are an impediment to an architecture which is focused on (re-)publishing and linking of data. Several data sets, which were converted to RDF could not be re-published due to licensing issues. Especially, these include the Leipzig Corpora Collection (LCC) (Quasthoff et al., 2009) and the RDF data used in the TIGER Corpus Navigator (Hellmann et al., 2010) in Section 10.4. Very often (as it is the case for the previous two examples), the reason for closed licenses is the strict copyright of the primary data (such as newspaper texts) and researchers are unable to publish their annotations and resulting data. The open part of the American National Corpus (OANC[12]) on the other hand has been converted to RDF and was republished successfully using the POWLA ontology (Chiarcos, 2012c). Thus, the work contributed to OANC was directly reusable by other scientists and likewise the same accounts for the RDF conversion.

Note that the *Open* in Linked Open Data refers mainly to *open access*, i.e. retrievable using the HTTP protocol.[13] Only around 18% of the data sets of the LOD cloud provide clear licensing information at all.[14] Of these 18% an even smaller amount is considered *open* in the sense of the open definition[15] coined by the Open Knowledge Foundation. One further important criteria for the success of a collaboration chain is whether the data set explicitly allows to redistribute data. While often self-made licenses allow scientific and non-commercial use, they are incomplete and do not specify how redistribution is handled.

1.3 LINKED DATA IN LINGUISTICS

Chiarcos, Nordhoff, and Hellmann (2012)

The explosion of information technology in the last two decades has led to a substantial growth in quantity, diversity and complexity of web-accessible linguistic data. These resources become even more useful when linked with each other, and the last few years have seen the emergence of numerous approaches in various disciplines concerned with linguistic resources.

It is the challenge of our time to store, interlink and exploit this wealth of data accumulated in more than half a century of computational linguistics (Dostert, 1955), of empirical, corpus-based study of

12 http://www.anc.org/OANC/
13 http://richard.cyganiak.de/2007/10/lod/#open
14 http://www4.wiwiss.fu-berlin.de/lodcloud/state/#license
15 http://opendefinition.org/

language (Francis & Kucera, 1964), and of computational lexicography (Morris, 1969) in all its heterogeneity.

A crucial question involved here is the *interoperability* of the language resources, actively addressed by the community since the late 1980s (Text Encoding Initiative, 1990), but still a problem that is partially solved at best (Ide & Pustejovsky, 2010). A closely related challenge is *information integration*, i.e., how heterogeneous information from different sources can be retrieved and combined in an efficient way.

With the rise of the Semantic Web, new representation formalisms and novel technologies have become available, and, independently from each other, researchers in different communities have recognized the potential of these developments with respect to the challenges posited by the heterogeneity and multitude of linguistic resources available today. Many of these approaches follow the *Linked Data paradigm* (Berners-Lee, 2006, Section 2.2) that postulates rules for the publication and representation of web resources. If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects.

This thesis provides an excerpt of the broad variety of approaches towards the application of the Linked Data paradigm to linguistic resources in Chapter 3. It assembles the contributions of the workshop on Linked Data in Linguistics (LDL-2012), held at the 34th Annual Meeting of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS), March 7th-9th, 2012, in Frankfurt/M., Germany, organized by the Open Linguistics Working Group (OWLG, cf. Section 2.1) of the Open Knowledge Foundation (OKFN),[16] an initiative of experts from different fields concerned with linguistic data, including academic linguists (e.g., typology, corpus linguistics), applied linguistics (e.g., computational linguistics, lexicography and language documentation), and NLP engineers (e.g., from the Semantic Web community). The primary goal of the working group is to promote the idea of open linguistic resources, to develop means for their representation, and to encourage the exchange of ideas across different disciplines. Accordingly, the chapter represents a great bandwidth of contributions from various fields, representing principles, use cases, and best practices for using the Linked Data paradigm to represent, exploit, store, and connect different types of linguistic data collections.

One goal of the book accompanying the workshop on Linked Data in Linguistics (Chiarcos, Nordhoff, & Hellmann, 2012, LDL-2012) is to document and to summarize these developments, and to serve as a point of orientation in the emerging domain of research on Linked Data in Linguistics. This documentary goal is complemented by so-

16 http://okfn.org

cial goals: (a) to facilitate the communication between researchers from different fields who work on linguistic data within the Linked Data paradigm; and (b) to explore possible synergies and to build bridges between the respective communities, ranging from academic research in the fields of language documentation, typology, translation studies, digital humanities in general, corpus linguistics, computational lexicography and computational linguistics, and computational lexicography to concrete applications in Information Technology, e.g., machine translation, or localization.

1.4 NLP FOR AND BY THE SEMANTIC WEB – THE NLP INTERCHANGE FORMAT (NIF)

Chiarcos, Nordhoff, and Hellmann (2012); Hellmann, Lehmann, et al. (2013)

In recent years, the interoperability of linguistic resources and NLP tools has become a major topic in the fields of computational linguistics and Natural Language Processing (Ide & Pustejovsky, 2010). The technologies developed in the Semantic Web during the last decade have produced formalisms and methods that push the envelop further in terms of expressivity and features, while still trying to have implementations that scale on large data. Some of the major current projects in the NLP area seem to follow the same approach such as the graph-based formalism GrAF developed in the ISO TC37/SC4 group (Ide & Suderman, 2007) and the ISOcat data registry (Windhouwer & Wright, 2012), which can benefit directly by the widely available tool support, once converted to RDF. Note that it is the declared goal of GrAF to be a pivot format for supporting conversion between other formats and not designed to be used directly and the ISOcat project already provides a Linked Data interface. In addition, other data sets have already converted to RDF such as the typological data in Glottolog/Langdoc (Nordhoff, 2012), language-specific Wikipedia versions (cf. Chapter 4), Wiktionary (cf. Chapter 5). An overview can be found in Chapter 3.

The recently published NLP Interchange Format (NIF)[17] aims to achieve interoperability for the output of NLP tools, linguistic data and language resources in RDF, documents on the WWW and the Web of Data (LOD cloud).

NIF addresses the interoperability problem on three layers: the *structural*, *conceptual* and *access* layer. NIF is based on a Linked Data enabled URI scheme for identifying elements in (hyper-)texts (structural layer) and a comprehensive ontology for describing common NLP terms and concepts (conceptual layer). NIF-aware applications will produce output (and possibly also consume input) adhering to the NIF Core ontology as REST services (access layer). Other than more centralized solutions such as UIMA (Ferrucci & Lally, 2004) and GATE (Cunningham, Maynard, Bontcheva, & Tablan, 2002), NIF

17 http://persistence.uni-leipzig.org/nlp2rdf/

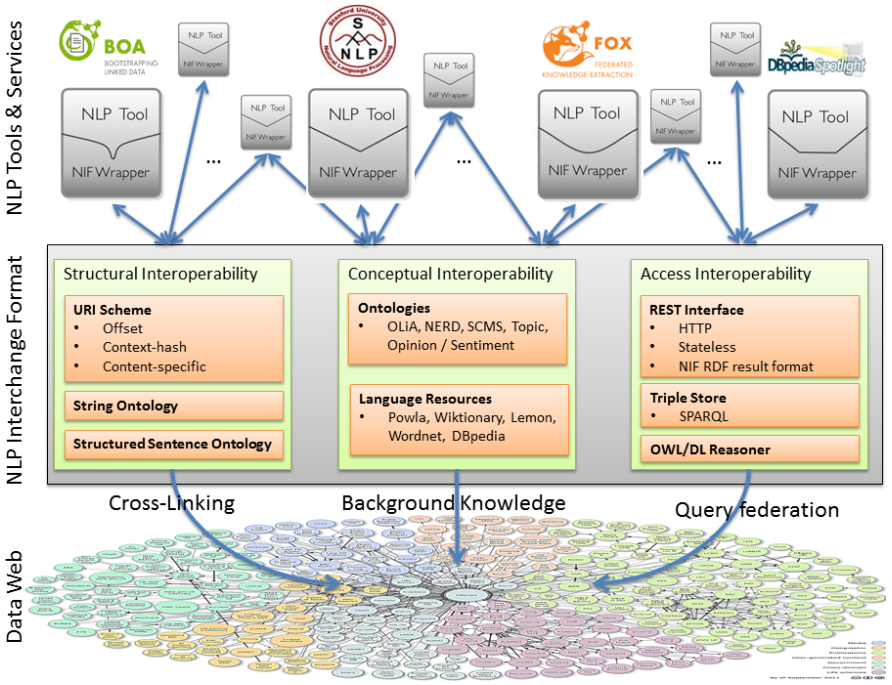


Figure 3: NIF architecture aiming at establishing a distributed ecosystem of heterogeneous NLP tools and services by means of structural, conceptual and access interoperability employing background knowledge from the Web of Data (Auer & Hellmann, 2012).

enables the creation of heterogeneous, distributed and loosely coupled NLP applications, which use the Web as an integration platform. Another benefit is, that a NIF wrapper has to be only created once for a particular tool, but enables the tool to interoperate with a potentially large number of other tools without additional adaptations. NIF can be partly compared to LAF and its extension GrAF (Ide & Pustejovsky, 2010) as LAF is similar to the proposed URI schemes and the NIF Core Ontology[18], while other (already existing) ontologies are re-used for the different annotation layers of NLP (cf. Section 7.3). Furthermore, NIF utilizes the advantages of RDF and uses the Web as an integration and collaboration platform. Extensions for NIF can be created in a decentralized and agile process, as has been done in the NERD extension for NIF (Rizzo et al., 2012). Named Entity Recognition and Disambiguation (NERD)[19] provides an ontology, which maps the types used by web services such as *Zemanta*, *OpenCalais*, *Ontos*, *Evri*, *Extractiv*, *Alchemy API* and *DBpedia Spotlight* to a common taxonomy. Ultimately, we envision an ecosystem of NLP tools and services to emerge using NIF for exchanging and integrating rich annotations. Figure 3 gives an overview over the architecture of NIF, connecting tools, language resources and the Web of Data.

18 http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
19 http://nerd.eurecom.fr

1.5 REQUIREMENTS FOR NLP INTEGRATION

Hellmann, Lehmann, et al. (2013)

In this section, we will give a list of requirements, we elicited within the LOD2 EU project[20], which influenced the design of NIF. The LOD2 project develops the *LOD2 stack*[21], which integrates a wide range of RDF tools, including a *Virtuoso* triple store as well as Linked Data interlinking and OWL enrichment tools.

COMPATIBILITY WITH RDF. One of the main requirements driving the development of NIF, was the need to convert any NLP tool output to RDF as virtually all software developed within the LOD2 project is based on RDF and the underlying triple store.

COVERAGE. The wide range of potential NLP tools requires that the produced format and ontology is sufficiently general to cover all or most annotations.

STRUCTURAL INTEROPERABILITY. NLP tools with a NIF wrapper should produce unanimous output, which allows to merge annotations from different tools consistently. Here structural interoperability refers to the way *how* annotations are represented.

CONCEPTUAL INTEROPERABILITY. In addition to structural interoperability, tools should use the same vocabularies for the same kind of annotations. This refers to *what* annotations are used.

GRANULARITY. The ontology is supposed to handle different granularity not limited to the document level, which can be considered to be very coarse-grained. As basic units we identified a document collection, the document, the paragraph and the sentence. A keyword search, for example, might rank a document higher, where the keywords appear in the same paragraph.

PROVENANCE AND CONFIDENCE. For all annotations we would like to track, where they come from and how confident the annotating tool was about correctness of the annotation.

SIMPLICITY. We intend to encourage third parties to contribute their NLP tools to the LOD2 Stack and the NLP2RDF platform. Therefore, the format should be as simple as possible to ease integration and adoption.

SCALABILITY. An especially important requirement is imposed on the format with regard to scalability in two dimensions: Firstly, the triple count is required to be as low as possible to reduce the overall memory and index footprint (URI to id look-up tables). Secondly, the complexity of OWL axioms should be low or modularised to allow fast reasoning.

20 http://lod2.eu
21 http://stack.linkeddata.org

1.6 OVERVIEW AND CONTRIBUTIONS

PART I – INTRODUCTION AND BACKGROUND. During his keynote at the Language Resource and Evaluation Conference in 2012, Sören Auer stressed the decentralized, collaborative, interlinked and interoperable nature of the Web of Data. The keynote provides strong evidence that *Semantic Web technologies such as Linked Data are on its way to become main stream for the representation of language resources*. The jointly written companion publication for the keynote was later extended as a book chapter in *The People's Web Meets NLP* and serves as the basis for Chapter 1 "Introduction" and Chapter 2 "Background", outlining some stages of the Linked Data publication and refinement chain. Both chapters stress the importance of open licenses and open access as an enabler for collaboration, the ability to interlink data on the Web as a key feature of RDF as well as provide a discussion about scalability issues and decentralization. Furthermore, we elaborate on how conceptual interoperability can be achieved by (1) re-using vocabularies, (2) agile ontology development, (3) meetings to refine and adapt ontologies and (4) tool support to enrich ontologies and match schemata.

PART II - LANGUAGE RESOURCES AS LINKED DATA. Chapter 3 "Linked Data in Linguistics" and Chapter 6 "NLP & DBpedia, an Upward Knowledge Acquisition Spiral" summarize the results of the Linked Data in Linguistics (LDL) Workshop in 2012 and the NLP & DBpedia Workshop in 2013 and give a preview of the MLOD special issue. In total, five proceedings – three published at CEUR (OKCon 2011, WoLE 2012, NLP & DBpedia 2013), one Springer book (Linked Data in Linguistics, LDL 2012) and one journal special issue (Multilingual Linked Open Data, MLOD to appear) – have been (co-)edited to create incentives for scientists to convert and publish Linked Data and thus *to contribute open and/or linguistic data to the LOD cloud*. Based on the disseminated call for papers, *152 authors contributed one or more accepted submissions* to our venues and 120 reviewers were involved in peer-reviewing.

Chapter 4 "DBpedia as a Multilingual Language Resource" and Chapter 5 "Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Linked Data Cloud" contain this thesis' contribution to the DBpedia Project in order to further increase the size and inter-linkage of the LOD Cloud with lexical-semantic resources. Our contribution comprises extracted data from Wiktionary (an online, collaborative dictionary similar to Wikipedia) in more than four languages (now six) as well as language-specific versions of DBpedia, including a quality assessment of inter-language links between Wikipedia editions and internationalized content negotiation rules for Linked Data. In particular the work described in Chapter 4

created the foundation for a DBpedia Internationalisation Committee with *members from over 15 different languages with the common goal to push DBpedia as a free and open multilingual language resource*.

PART III - THE NLP INTERCHANGE FORMAT (NIF). Chapter 7 "NIF 2.0 Core Specification", Chapter 8 "NIF 2.0 Resources and Architecture" and Chapter 9 "Evaluation and Related Work" constitute one of the main contribution of this thesis. The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. The core specification is included in Chapter 7 and describes which URI schemes and RDF vocabularies must be used for (parts of) natural language texts and annotations in order to create *an RDF/OWL-based interoperability layer with NIF built upon Unicode Code Points in Normal Form C*. In Chapter 8, classes and properties of the *NIF Core Ontology* are described to formally define the relations between text, substrings and their URI schemes. Chapter 9 contains the evaluation of NIF.

In a questionnaire, we asked questions to 13 developers using NIF. UIMA, GATE and Stanbol are extensible NLP frameworks and NIF was yet not able to provide off-the-shelf NLP domain ontologies for all possible domains, but only for the plugins used in this study. After inspecting the software, the developers agreed however that NIF is general enough and adequate to provide a generic RDF output based on NIF using literal objects for annotations. All developers were able to map the internal data structure to NIF URIs to serialize RDF output (Adequacy). The development effort in hours (ranging between 3 and 40 hours) as well as the number of code lines (ranging between 110 and 445) suggest, that the implementation of NIF wrappers is easy and fast for an average developer. Furthermore the evaluation contains a comparison to other formats and an evaluation of the available URI schemes for web annotation.

In order to collect input from the wide group of stakeholders, a total of 16 presentations were given with extensive discussions and feedback, which has lead to a constant improvement of NIF from 2010 until 2013. After the release of NIF (Version 1.0) in November 2011, a total of *32 vocabulary employments and implementations for different NLP tools and converters were reported* (8 by the (co-)authors, including Wiki-link corpus (Section 11.1), 13 by people participating in our survey and 11 more, of which we have heard). Several roll-out meetings and tutorials were held (e.g. in Leipzig and Prague in 2013) and are planned (e.g. at LREC 2014).

PART IV - THE NLP INTERCHANGE FORMAT IN USE. Chapter 10 "Use Cases and Applications for NIF" and Chapter 11 "Publication of Corpora using NIF" describe 8 concrete instances where NIF has

been successfully used. One major contribution in Chapter 10 is the usage of NIF as the recommended RDF mapping in the Internationalization Tag Set 2.0 W3C standard (Section 10.1) and the conversion algorithms from ITS to NIF and back (Section 10.1.1). One outcome of the discussions in the standardization meetings and telephone conferences for ITS 2.0 resulted in the conclusion that there was *no alternative RDF format or vocabulary other than NIF* with the required features to fulfill the working group charter. Five further uses of NIF are described for the Ontology of Linguistic Annotations (OLiA), the RDFaCE tool, the Tiger Corpus Navigator, the OntosFeeder and visualisations of NIF using the RelFinder tool. Theses 8 instances provide an implemented proof-of-concept of the features of NIF.

Chapter 11 starts with describing the conversion and hosting of the huge Google Wikilinks corpus with 40 million annotations for 3 million web sites. The resulting RDF dump contains 477 million triples in a 5.6 GB compressed dump file in turtle syntax. Section 11.2 describes how NIF can be used to publish extracted facts from news feeds in the RDFLiveNews tool as Linked Data.

PART V - CONCLUSIONS. Chapter 12 provides lessons learned for NIF, conclusions and an outlook on future work.

2 BACKGROUND

2.1 THE WORKING GROUP ON OPEN DATA IN LINGUISTICS (OWLG)

Chiarcos, Hellmann, and Nordhoff (2012b)
Chiarcos et al. (2011)
Chiarcos, Hellmann, and Nordhoff (2012a)
Chiarcos, Hellmann, and Nordhoff (2012b)

2.1.1 *The Open Knowledge Foundation*

The Open Knowledge Foundation (OKFN) is a nonprofit organisation aiming to promote the use, reuse and distribution of open knowledge. Activities of the OKFN include the development of standards (Open Definition), tools (CKAN) and support for working groups and events.

The *Open Definition* sets out principles to define "openness" in relation to content and data: "A piece of content or data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and share-alike."[1]

The OKFN provides a catalog system for open datasets, *CKAN*[2]. CKAN is an open-source data portal software developed to publish, to find and to reuse open content and data easily, especially in ways that are machine automatable.

The OKFN also serves as host for various working groups addressing problems of open data in different domains. At the time of writing, there are 19 OKFN *working groups* covering fields as different as government data, economics, archeology, open text books or cultural heritage.[3] The OKFN organizes various events such as the Open Knowledge Conference (OKCon), and facilitates the communication between different working groups.

In late 2010, the *OKFN Working Group on Open Linguistic Data* (OWLG) was founded. Since its formation, the Open Linguistics Working Group has been steadily growing, we have identified goals and problems that are to be addressed, and directions that are to be pursued in the future. Preliminary results of this ongoing discussion process were summarized in this section: Section 2.1.2 specifies the goals of the working group; Section 2.1.3 identifies four major problems and challenges of the work with linguistic data; Section 2.1.4 gives an overview of recent activities and the current status of the group.

1 http://www.opendefinition.org
2 http://ckan.org/
3 For a complete overview see http://okfn.org/wg.

2.1.2 *Goals of the Open Linguistics Working Group*

As a result of discussions with interested linguists, NLP engineers, and information technology experts, we identified seven open problems for our respective communities and their ways to use, to access, and to share linguistic data. These represent the challenges to be addressed by the working group, and the role that it is going to fulfill:

1. promote the idea of open data in linguistics and in relation to language data;
2. act as a central point of reference and support for people interested in open linguistic data;
3. provide guidance on legal issues surrounding linguistic data to the community;
4. build an index of indexes of open linguistic data sources and tools and link existing resources;
5. facilitate communication between existing groups;
6. serve as a mediator between providers and users of technical infrastructure;
7. assemble best-practice guidelines and use cases to create, use and distribute data.

In many aspects, the OWLG is not unique with respect to these goals. Indeed, there are numerous initiatives with similar motivation and overlapping goals, e.g. the Cyberling blog,[4] the ACL Special Interest Group for Annotation (SIGANN),[5] and large multi-national initiatives such as the ISO initiative on Language Resources Management (ISO TC37/SC4),[6] the American initiative on Sustainable Interoperability of Language Technology (SILT),[7] or European projects such as the initiative on Common Language Resources and Technology Infrastructure (CLARIN),[8] the Fostering Language Resources Network (FLaReNet),[9] and the Multilingual Europe Technology Alliance (META).[10]

The key difference between these and the OWLG is that we are not grounded within a *single* community, or even restricted to a hand-picked set of collaborating partners, but that our members represent

4 http://cyberling.org/
5 http://www.cs.vassar.edu/sigann/
6 http://www.tc37sc4.org
7 http://www.anc.org/SILT
8 http://www.clarin.eu
9 http://www.flarenet.eu
10 http://www.meta-net.eu

the whole band-width from academic linguistics over applied linguistics and human language technology to NLP and information technology. We do not consider ourselves to be in competition with any existing organization or initiative, but we hope to establish new links and further synergies between these. The following section summarizes typical and concrete scenarios where such an interdisciplinary community may help to resolve problems observed (or, sometimes, overlooked) in the daily practice of working with linguistic resources.

2.1.3 *Open linguistics resources, problems and challenges*

Among the broad range of problems associated with linguistic resources, we identified four major classes of problems and challenges that may be addressed by the OWLG:

LEGAL QUESTIONS Often, researchers are uncertain with respect to legal aspects of creating and distributing linguistic data. The OWLG can represent a platform to discuss such problems, experiences and to develop recommendations, e.g. with respect to the publication of linguistic resources under open licenses.

TECHNICAL PROBLEMS Often, researchers come up with questions regarding the choice of tools, representation formats and metadata standards for different types of linguistic annotation. These problems are currently addressed in the OWLG, proposals for the interoperable representation of linguistic resources and NLP analyses by means of W3C standards such as RDF are actively explored, and laid out with greater level of detail in this article.

REPOSITORY OF OPEN LINGUISTIC RESOURCES So far, the communities involved have not yet established a common point of reference for existing open linguistic resources, at the moment there are multiple metadata collections. The OWLG works to extend CKAN with respect to open resources from linguistics. CKAN differs qualitatively from other metadata repositories:[11] (a) CKAN focuses on the license status of the resources and it encourages the use of *open* licenses; (b) CKAN is not specifically restricted to linguistic resources, but rather, it is used by all working groups, as well as interested individuals outside these working groups.[12]

11 For example, the metadata repositories maintained by META-NET (http://www.meta-net.eu), FLaReNet (http://www.flarenet.eu/?q=Documentation_about_Individual_Resources) or CLARIN (http://catalog.clarin.eu/ds/vlo).

12 Example resources of potential relevance to linguists but created outside the linguistic community include collections of open textbooks (http://wiki.okfn.org/Wg/opentextbooks), the complete works of Shakespeare (http://openshakespeare.org), and the Open Richly Annotated Cuneiform Corpus (http://oracc.museum.upenn.edu).

SPREAD THE WORD Finally, there is an agitation challenge for open data in linguistics, i.e. how we can best convince our collaborators to release their data under open licenses.

2.1.4 *Recent activities and on-going developments*

In the first year of its existence, the OWLG focused on the task to delineate what questions we may address, to formulate general goals and identify potentially fruitful application scenarios. At the moment, we have reached a critical step in the formation process of the working group: having defined a (preliminary) set of goals and principles, we can now concentrate on the tasks at hand, e.g. to collect resources and to attract interested people in order to address the challenges identified above.

The Working Group maintains a home page,[13] a mailing list[14], a wiki,[15] and a blog.[16] We conduct regular meetings and organize regular workshops at selected conferences.

A number of possible community projects have been proposed, including the documentation of workflows, documenting best practice guidelines and use cases with respect to legal issues of linguistic resources, and the creation of a Linguistic Linked Open Data (LLOD) cloud, which is one of the main topic of this thesis.[17]

2.2 TECHNOLOGICAL BACKGROUND

Chiarcos, Hellmann, and Nordhoff (2012a)

Several standards developed by different initiatives are referenced or used throughout this work. One is the *Extensible Markup Language* (XML, Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 1997) and its predecessor, the Standard Generalized Markup Language (SGML, Goldfarb & Rubinsky, 1990). These are text-based formats that allow to encode documents in an appropriate way for representing and transmitting machine-readable information.

XML and SGML have been the basis for most proposals for *interoperable representation formalisms specifically for linguistic resources*, for example the Corpus Encoding Standard (CES, Ide, 1998) developed by the Text Encoding Initiative (TEI[18]), or the Graph Annotation Format (GrAF, Ide & Suderman, 2007) developed in the context of the Linguistic Annotation Framework (LAF) by ISO TC37/SC4[19]. Earlier standards for linguistic corpora used XML data structures (i.e.,

13 http://linguistics.okfn.org
14 http://lists.okfn.org/mailman/listinfo/open-linguistics
15 http://wiki.okfn.org/Wg/linguistics
16 http://blog.okfn.org/category/working-groups/wg-linguistics
17 Details on these can be found on the OWLG wiki, http://wiki.okfn.org/Wg/linguistics.
18 http://www.tei-c.org
19 http://www.tc37sc4.org

trees) directly, but since Bird and Liberman (2001), it is generally accepted that generic formats to represent linguistic annotations should be based on graphs. State-of-the-art formalisms for linguistic corpora follow this assumption, and represent linguistic annotations in XML standoff formats, i.e., as bundles of XML files that are interlinked with cross-references, e.g., with formats like ATLAS (Bird & Liberman, 2001), PAULA XML (Dipper, 2005), or GrAF (Ide & Suderman, 2007).

In parallel to these formalisms, which are specific to linguistic resources, other communities have developed the *Resource Description Framework* (RDF, Lassila & Swick, 1999). Although RDF was originally invented to provide formal means to describe resources, e.g. books in a library or in an electronic archive (hence its name), its data structures were so general that its use has extended far beyond the original application scenario. RDF is based on the notion of *triples* (or 'statements'), consisting of a *predicate* that links a *subject* to an *object*. In other words, RDF formalizes relations between resources as labeled edges in a directed graph. Subjects are represented using globally unique Uniform Resource identifiers (URIs) and point (via the predicate) to another URI, the object part, to form a graph. (Alternatively, triples can have simple strings in the object part that annotate the subject resource.) At the moment, RDF represents the primary data structure of the Semantic Web, and is maintained by a comparably large and active community. Further, it provides crucial advantages for the publication of linguistic resources in particular: RDF provides a graph-based data model as required by state-of-the-art approaches on generic formats for linguistic corpora, and several RDF extensions were specifically designed with the goal to formalize knowledge bases like terminology data bases and lexical-semantic resources. For resources published under open licenses, an RDF representation yields the additional advantage that resources can be interlinked, and it is to be expected that an additional gain of information arises from the resulting network of resources. If modeled with RDF, linguistic resources are thus not only *structurally interoperable* (using RDF as representation formalism), but also *conceptually interoperable* (with metadata and annotations are modeled in RDF, different resources can be directly linked to a single repository). Further, concrete applications using linguistic resources can be build on the basis of the rich ecosystem of format extensions and technologies that has evolved around RDF, including APIs, RDF databases (triple stores), the query language SPARQL, data browsing and visualization tools, etc.

For the formalization of knowledge bases, several RDF extensions have been provided, for example the *Simple Knowledge Organization System* (SKOS, Miles & Bechhofer, 2009), which is naturally applicable to lexical-semantic resources, e.g., thesauri. A thorough logi-

cal modeling can be achieved by formalizing linguistic resources as ontologies, using the *Web Ontology Language* (OWL, McGuinness & Van Harmelen, 2004), another RDF extension. OWL comes in several dialects (profiles), the most important being OWL/DL and its sub-languages (e.g. OWL/Lite, OWL/EL, etc.) that have been designed to balance expressiveness and reasoning complexity (McGuinness & Van Harmelen, 2004; W3C OWL Working Group, 2009) OWL/DL is based on Description Logics (DL, Baader, Horrocks, & Sattler, 2005) and thus corresponds to a *decidable* fragment of first-order predicate logic. A number of reasoners exist that can draw inferences from an OWL/DL ontology and verify consistency constraints. Primary entities of OWL Ontologies are *concepts* that correspond to classes of objects, *individuals* that represent instances of these concepts, and *properties* that describe relations between individuals. Ontologies further support *class operators* (e.g. intersection, join, complement, instanceOf, subClassOf), as well as the specification of *axioms* that constrain the relations between individuals, properties and classes (e.g. for property P, an individual of class A may only be assigned an individual of class B). As OWL is an extension of RDF, every OWL construct can be represented as a set of RDF triples.

RDF is based on globally unique and accessible URIs and it was specifically designed to establish links between such URIs (or resources). This is captured in the *Linked Data paradigm* (Berners-Lee, 2006) that postulates four rules:

1. Referred entities should be designated by URIs,
2. these URIs should be resolvable over HTTP,
3. data should be represented by means of standards such as RDF,
4. and a resource should include links to other resources.

With these rules, it is possible to follow links between existing resources to find other, related, data and exploit network effects. The *Linked Open Data (LOD) cloud*[20] represents the resulting set of resources. If published as Linked Data, linguistic resources represented in RDF can be linked with resources already available in the Linked Open Data cloud. At the moment, the LOD cloud covers a number of lexico-semantic resources, including the Open Data Thesaurus,[21] WordNet,[22] Cornetto (Dutch WordNet),[23] DBpedia (machine-readable version of the Wikipedia),[24] Freebase (an entity database),[25] OpenCyc

20 http://lod-cloud.net
21 http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData
22 http://semanticweb.cs.vu.nl/lod/wn30, http://www.w3.org/TR/wordnet-rdf, http://wordnet.rkbexplorer.com
23 http://www2.let.vu.nl/oz/cltl/cornetto
24 http://www.dbpedia.org
25 http://freebase.com

(database of real-world concepts),[26] and YAGO (a semantic knowledge base).[27] Additionally, the LOD cloud includes knowledge bases of information about languages and bibliographical information that are relevant for here, e.g., Lexvo (metadata about languages),[28] lingvoj (metadata about language in general),[29] Project Gutenberg (bibliographical data base)[30] and the OpenLibrary (bibliographical data base).[31] Given the interest that researchers take in representing linguistic resources as Linked Data, continuing growth of this set of resources seems to be assured. Several contributions assembled in this volume discuss the linking of their resources with the Linked Open Data cloud, thereby supporting the overarching vision of a Linguistic Open Data (sub-) cloud of linguistic resources, a *Linguistic Linked Open Data cloud* (LLOD).

2.3 RDF AS A DATA MODEL

Chiarcos et al. (2011)

RDF as a data model has distinctive features, when compared to its alternatives. Conceptually, RDF is close to the widely used Entity-Relationship Diagrams (ERD) or the Unified Modeling Language (UML) and allows to model entities and their relationships. XML is a serialization format, that is useful to (de-)serialize data models such as RDF. Major drawbacks of XML and relational databases are the lack of (1) global identifiers such as URIs, (2) standardized formalisms to explicitly express links and mappings between these entities and (3) mechanisms to publicly access, query and aggregate data. Note that (2) can not be supplemented by transformations such as XSLT, because the linking and mappings are implicit. All three aspects are important to enable ad-hoc collaboration. The resulting technology mix provided by RDF allows any collaborator to join her data into the decentralized data network employing the HTTP protocol which immediate benefits herself and others. In addition, features of OWL can be used for inferencing and consistency checking. OWL – as a modelling language – allows, for example, to model transitive properties, which can be queried on demand, without expanding the size of the data via backward-chaining reasoning. While XML can only check for validity, i.e. the occurrence and order of data items (elements and attributes), consistency checking allows to verify, whether a data set adheres to the semantics imposed by the formal definitions of the used ontologies.

26 http://sw.opencyc.org
27 http://mpii.de/yago
28 http://www.lexvo.org
29 http://www.lingvoj.org
30 http://www4.wiwiss.fu-berlin.de/gutendata
31 http://openlibrary.org

2.4 PERFORMANCE AND SCALABILITY

Chiarcos et al. (2011); Hellmann and Auer (2013)

RDF, its query language SPARQL and its logical extension OWL provide features and expressivity that go beyond relational databases and simple graph-based representation strategies. This expressivity poses a performance challenge to query answering by RDF triple stores, inferencing by OWL reasoners and of course the combination thereof. Although the scalability is a constant focus of RDF data management research[32], the primary strength of RDF is its flexibility and suitability for data integration and not superior performance for specific use cases. Many RDF-based systems are designed to be deployed in parallel to existing high-performance systems and not as a replacement. An overview over approaches that provide Linked Data and SPARQL on top of relational database systems, for example, can be found in Auer, Dietzold, et al. (2009). The NLP Interchange Format (cf. Chapter 7) allows to express the output of highly optimized NLP systems (e.g. UIMA) as RDF/OWL. The architecture of the Data Web, however, is able to scale in the same manner as the traditional WWW as the nodes are kept in a de-centralized way and new nodes can join the network any time and establish links to existing data. Data Web search engines such as *Swoogle*[33] or *Sindice*[34] index the available structured data in a similar way as Google does with the text documents on the Web and provide keyword-based query interfaces.

2.5 CONCEPTUAL INTEROPERABILITY

Chiarcos et al. (2011); Hellmann and Auer (2013)

While RDF and OWL as a standard for a common data format provide structural (or syntactical) interoperability, conceptual interoperability is achieved by globally unique identifiers for entities, properties and classes, that have a fixed meaning. These unique identifiers can be interlinked via `owl:sameAs` on the entity-level, re-used as properties on the vocabulary level and extended or set equivalent via `rdfs:subClassOf` or `owl:equivalentClass` on the schema-level. Following the ontology definition of Gruber (1993), the aspect that ontologies are a "shared conceptualization" stresses the need to collaborate to achieve agreement. On the class and property level RDF and OWL give users the freedom to reuse, extend and relate to other work in their own conceptualization. Very often, however, it is the case that groups of stakeholders actively discuss and collaborate in order to form some kind of agreement on the meaning of identifiers as has been described in Hepp, Siorpaes, and Bachlechner (2007). In

32 `http://factforge.net` or `http://lod.openlinksw.com` provide SPARQL interfaces to query billions of aggregated facts.
33 `http://swoogle.umbc.edu`
34 `http://sindice.com`

the following, we will give four examples to elaborate how conceptual interoperability is achieved:

- In a knowledge extraction process (e.g. when converting relational databases to RDF) vocabulary identifiers can be reused during the extraction process. Especially community-accepted vocabularies such as FOAF, SIOC, Dublin Core and the DBpedia Ontology are suitable candidates for reuse as this leads to conceptual interoperability with all applications and databases that also use the same vocabularies. This aspect was the rationale for designing Triplify (Auer, Dietzold, Lehmann, Hellmann, & Aumueller, 2009), where the SQL syntax was extended to map query results to existing RDF vocabularies.
- During the creation process of ontologies, direct collaboration can be facilitated with tools that allow agile ontology development such as *OntoWiki*, *Semantic Mediawiki* or the *DBpedia Mappings Wiki*[35]. This way, conceptual interoperability is achieved by a distributed group of stakeholders, who work together over the Internet. The created ontology can be published and new collaborators can register and get involved to further improve the ontology and tailor it to their needs.
- In some cases, real life meetings are established, e.g. in the form of Vo(cabulary) Camps, where interested people meet to discuss and refine vocabularies. VoCamps can be found and registered on `http://vocamp.org`.
- A variety of RDF tools exists, which aid users in creating links between individual data records as well as in mapping ontologies.
- Semi-automatic enrichment tools such as ORE (Bühmann & Lehmann, 2012) allow to extend ontologies based on the entity-level data .

35 `http://mappings.dbpedia.org`

Part II

LANGUAGE RESOURCES AS LINKED DATA

3

LINKED DATA IN LINGUISTICS

Chiarcos, Hellmann, and Nordhoff (2012a); Chiarcos, Nordhoff, and Hellmann (2012); Hellmann, Brekle, and Auer (2012); Hellmann, Filipowska, et al. (2013b, 2013a); Hellmann et al. (to appear); Kontokostas et al. (2012); Lehmann et al. (2009)

Researchers in NLP and Linguistics are currently discovering Semantic Web technologies and employing them to answer novel research questions. Through the use of Linked Data, there is the potential to solve many issues currently faced by the language resources community. In particular, there is significant evidence that RDF allows better data integration than existing formats (Chiarcos, Nordhoff, & Hellmann, 2012), in part through a rich ecosystem of tools provided by the Semantic Web, such as query (Garlik, Seaborne, & Prud'hommeaux, 2013) and federation (Quilitz & Leser, 2008). In addition, the Semantic Web has already been used by several authors (Windhouwer & Wright, 2012) to define data categories and enable better resource interoperability. The utility of this method of publishing language resources has lead to the interest of a significant sub-community in linguistics (Chiarcos, Hellmann, Nordhoff, Moran, et al., 2012).

Language resources include language data such as written or spoken corpora and lexica, multimodal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, multimedia databases, etc.

For this thesis, we are especially interested in resources used to assist and augment language processing applications, even if the nature of the resource is not deeply entrenched in Linguistics, but only as long as the usefulness is well motivated (DBpedia redirects and disambiguation pages are one example (Mendes, Jakob, & Bizer, 2012)). The focus of this chapter is on language resources that were published as Linked Data using appropriate technologies such as RDF and OWL. Figure 4 displays the state of the LLOD cloud after the MLODE Workshop 2012 in Leipzig, organized by organized by Hellmann, Moran, Brümmer and Konkokostas.[1]

For the book "*Linked Data in Linguistics 2012*", we were happy to have attracted a large number of high quality contributions from very different domains for the workshop on Linked Data in Linguistics (LDL-2012) held March 7^{th} - 9^{th}, 2012, as part of the 34^{th} Annual Meeting of the German Linguistics Society (DGfS) in Frankfurt a. M., Germany. The set of subdisciplines included in this volume is diverse; the goal is the same: provide scientific data in an open format which permits integration with other data repositories.

The book is organized in four parts: Parts I, II and III describe applications of the Linked Data paradigm to major types of linguistic resources, i.e., **lexical-semantic resources**, **linguistic corpora** and

1 http://sabre2012.infai.org/mlode

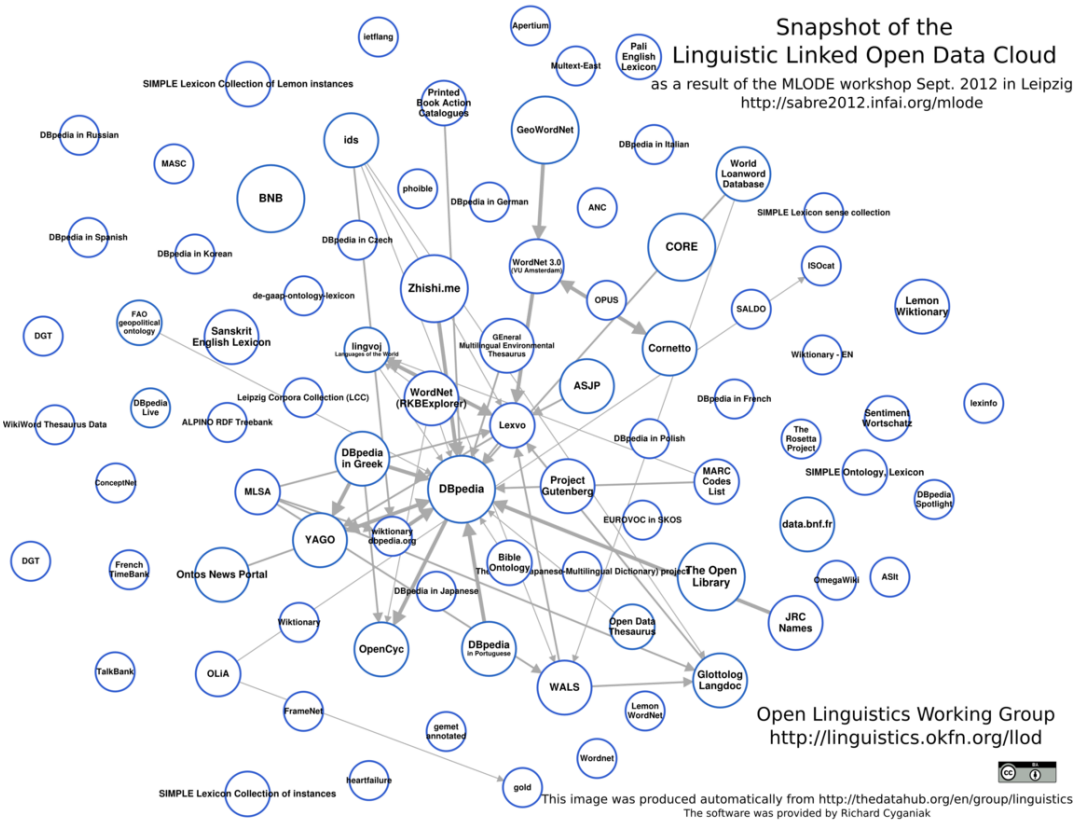


Figure 4: The Linguistic Linked Open Data Cloud as a result of the MLODE Workshop 2012 in Leipzig

other knowledge bases, respectively. These parts represent the contributions of the participants of the Workshop Linked Data in Linguistics (LDL-2012). In Part IV, the editors describe recent efforts to **link linguistic resources** – and thus to create a Linked Open Data (sub-)cloud of linguistic resources – in the context of the Open Linguistic Working Group (OWLG) of the Open Knowledge Foundation (OKFN). They illustrate how lexical-semantic resources, corpora and other linguistic knowledge bases can be interlinked and what possible gains of information are to be expected, using representative examples for the respective classes of linguistic resources.

As we are interested in linking different language resources, it should be noted that there is a natural overlap between these categories, and therefore, many contributions could be classified under more than one category. Bouda and Cysouw (2012), for example, discuss not only lexical resources, but also corpus representation, and knowledge bases for linguistic metadata; Schalley (2012) and Declerck, Lendvai, Mörth, Budin, and Váradi (2012) describe not only linguistic knowledge bases, but also corpus data and multi-layer annotations; and the contributions by Chiarcos (2012a), Hellmann, Stadler, and Lehmann (2012), and Nordhoff (2012) that are presented in the context of linking linguistic resources, could also have been pre-

sented in the respective parts on linguistic corpora, lexical-semantic resources and other (linguistic) knowledgebases.

3.1 LEXICAL RESOURCES

Chiarcos, Hellmann, and Nordhoff (2012a)

Part I describes the modeling of various lexical-semantic resources as illustrated for lexical-semantic resources.

Bouda and Cysouw (2012) describe the digitization of dictionaries, and how the elements (head words, translations, annotations) found in there can be served in a Linked Data way while at the same time maintaining access to the document in its original form. To this end, they use standoff markup, which furthermore allows the third-party annotation of their data. They also explore how these third-party annotations could be shared in novel ways beyond the normal scope of normal academic distribution channels, e.g. Twitter.

McCrae, Montiel-Ponsoda, and Cimiano (2012) describe the *lemon* format that has been developed for the sharing of lexica and machine readable dictionaries. They consider two resources that seem ideal candidates for the Linked Data cloud, namely WordNet 3.0 and Wiktionary, a large document based dictionary. The authors discuss the challenges of converting both resources to *lemon*, and in particular for Wiktionary, the challenge of processing the mark-up, and handling inconsistencies and underspecification in the source material. Finally, they turn to the task of creating links between the two resources and present a novel algorithm for linking lexica as lexical Linked Data.

Herold, Lemnitzer, and Geyken (2012) report on the lexical resources of the long-term project 'Digitales Wörterbuch der deutschen Sprache' (DWDS) which aims at the integration of several lexical and textual resources in order to document the German language and its use at several stages. They describe the explicit linking of four lexical resources on the level of individual articles which is achieved via a common meta-index. The authors present strategies for the actual dictionary alignment as well as a discussion of models that can adequately describe complex relations between entries of different dictionaries.

Lewis et al. (2012) describe perspectives of Linked Data in the fields of software localisation and translation. They present a platform architecture for sharing, searching and interlinking of Linked Localisation and Language Data on the web. This architecture rests upon a semantic schema for the respective resources that is compatible with existing localisation data exchange standards and can be used to support the round-trip sharing of language resources. The paper describes the development of the schema and data management processes, web-based tools and data sharing infrastructure that use it. An initial proof of concept prototype is presented which implements a web application that segments and machine translates content for crowd-sourced post-editing and rating.

3.2 LINGUISTIC CORPORA

Chiarcos, Hellmann, and Nordhoff (2012a)

Part II deals with problems to create, to maintain and to evaluate linguistic corpora and other collections of linguistically annotated data. Previous research indicates that formalisms such as RDF and OWL are suitable to represent linguistic annotations Burchardt, Padó, Spohr, Frank, and Heid (2008); Cassidy (2010) and to build NLP architectures on this basis Hellmann (2010); Wilcock (2007), yet so far, it has rarely been applied to this type of linguistic resource.

van Erp (2012) describes interoperability problems of linguistic resources, in particular corpora, and develops a vision to apply the Linked Data approach to these issues. In her contribution, the constraints for linguistic resource reuse and the tasks are detailed, accompanied by a Linked Data approach to standardise and reconcile concepts and representations used in linguistic annotations.

As mentioned above, these problems are addressed in the NLP community by generic data models for linguistic corpora that are based on directed graphs.

Eckart, Riester, and Schweitzer (2012) describe such a state-of-the-art approach on the task of resource integration for multiple independent layers of annotation in a multi-layer annotated corpus that is based on a graph-based data model, although not on RDF, but an XML standoff format and a relational database management system. They present an annotated corpus of German radio news including syntactic information from a parser, as well as manually annotated information status labels and prosodic labels. They describe each annotation layer and focus on the linking of the data from both layers of annotation, and show how the resource can support data extraction on both annotation layers. Although they do not directly make use of the Linked Data paradigm, the problems identified and the data model employed represent important steps towards the development of representation formalisms for multi-layer corpora by means of RDF and as Linked Data, see, for example, Chiarcos (2012a).

Carl and Hœg Müller (2012) describe a fascinating intersection between pure structural syntactic data and human-machine interaction in translation processes. Human behaviour while translating on a computer can be recorded with eye trackers and capturing of user input (mouse, keyboard). This behavioural data can then be linked to syntactic data extracted from the sentence translated (constituency, dependency). The intuition is that syntactically complicated sentences will have a repercussion in the user behaviour (longer gaze, slower input, more corrections). Carl and Müller, just like Bouda and Cysouw, and Eckart et al., use standoff annotation to allow for overlapping annotations. Their use of structural data on the one hand and behavioural data from a novel domain on the other hand shows the benefits the provision of data as Linked Data can have.

Blume, Flynn, and Lust (2012) describe DTA, an online tool for the study of language acquisition. DTA allows for data creation, data management and collaborative use of child language data from a variety of languages (Spanish, French, English, Sinhala). Language Acquisition is a relative newcomer to the area of Linked Data, and it is exciting to see that areas somewhat distant from the NLP origins of Linked Data are beginning to join the movement.

3.3 LINGUISTIC KNOWLEDGEBASES

Chiarcos, Hellmann, and Nordhoff (2012a)

While Part II focused on annotated linguistic data, Part III presents a number of repositories of knowledge about languages and linguistic terminology that can be used, for example, for annotating linguistic data with linguistic analyses and metadata.

Windhouwer and Wright (2012) describe the linking from language resources to linguistic data categories in ISOcat, a repository of linguistic terminology developed to foster semantic interoperability of linguistic resources. This registry follows a grass roots approach, which means that any linguist can add the data categories (s)he needs. However, the goal of improving semantic interoperability can only be met if the data categories are reused by a wide variety of linguistic resource types. A resource indicates its usage of data categories by linking to them, this paper describes the technical prerequisites to achieve this in an RDF-based approach.

Declerck, Lendvai, Mörth, Budin, and Váradi (2012) describe strategies for exploiting the large set of dynamically increasing, freely available language data incorporated in the Linked Open Data (LOD) framework. Such language data currently mostly exist in the form of raw, unstructured textual expressions within RDF labels or comments. Incorporating them as structured language data within the LOD leads to a linguistic enrichment of the data sets that express linked (domain) knowledge resources, and this will enable the creation of more accurate, knowledge-aware NLP applications. This integration of linguistic information in knowledge representation systems should be done in compliance with both ISO (multi-layer linguistic annotation and data categories) and W3C (RDF, SKOS) standards. By this, new linguistically enriched datasets can also be more easily ported into the LOD format: e.g., repositories in the field of Digital Humanities often hold language data in taxonomical structures. The potential of linked language data for digital humanities is illustrated here for the detection of motifs in literary texts. For this purpose, a formal representation of the taxonomical structure of the Thompson Motif-Index of folk-literature (Thompson, 1955-58, TMI) is presented.

In a similar vein, Pareja-Lora (2012) reports on the development of a concept taxonomy for a different type of linguistic annotation,

namely pragmatic annotations. Pragmatics has to deal with a real mix of different linguistic topics, such as (i) speech acts, (ii) deixis, presuppositions and implicatures; or (iii) pragmatic coherence relations, which traditionally have been tackled following several fragmentary and/or partial approaches. Pareja-Lora describes an approach to specify formally the different elements that a pragmatic annotation scheme should contemplate and make explicit with the goal to facilitate the interoperability of linguistic annotations up to the pragmatic level.

While the terminology repositories and taxonomies described in this part so far have been developed for interoperability of NLP tools and linguistic annotations, the remaining chapters of this part deal with typological databases that provide information about languages from a slightly different angle of research.

Moran (2012) tackles the very basic unit of linguistics, the phoneme, and shows how heterogeneous data bases of phoneme inventories found in the worlds languages can be integrated with a Linked Data approach via mapping of the relations found in the original data bases to his ontology. His system is in production stage, and Moran shows how a number of phonological hypotheses can be confirmed or refuted using his PHOIBLE database. Moran furthermore explores the difference between queries in traditional relational databases and SPARQL queries.

Schalley (2012) Andrea Schalley casts a wide net and lists the criteria a typological knowledge base would have to respond to in an ideal world. She then discusses challenges for the realization and sketches the development of a computational tool that utilises Semantic Web technologies in order to provide novel ways to process, integrate, and query cross-linguistic data. Its data store incorporates a set of ontologies (comprising linguistic examples, annotations, language background information, and metadata) backed by a software logic reasoner. This allows for highly targeted querying and answers on rather specific questions such as (i) which size (in terms of speaker count) do languages have that have kin-sensitive pronouns?, or (ii) which languages code joint attention in their grammar, and if so, where in the grammar do they do it?

3.4 TOWARDS A LINGUISTIC LINKED OPEN DATA CLOUD

Chiarcos, Hellmann, and Nordhoff (2012a)

The last part describes joint activities of different members of the Open Linguistics Working Group (OWLG) aiming to develop a Linked Open Data (sub-)cloud of linguistic resources.

Chiarcos, Hellmann, and Nordhoff (2012b) describe the Open Linguistics Working Group (OWLG), its goals, addressed problems, recent activities and on-going developments.

Chiarcos (2012a) describes the formalization of annotated linguistic corpora by means of OWL/DL with a focus on genericity and

interoperability. Structural interoperability of linguistic corpora is addressed with POWLA, an OWL/DL formalization of a data model designed to represent any kind of linguistic annotation assigned to textual data; conceptual interoperability between annotations of different corpora can be established using the OLiA ontologies, an architecture of modular OWL/DL ontologies that formalize the linking of annotation schemes with community-maintained terminology repositories.

Hellmann, Stadler, and Lehmann (2012) describe the DBpedia, one of the major free data sets in the Web of Data, as an example of a lexical-semantic resource. In particular, the internationalization of the DBpedia is addressed – including the development of a German DBpedia. The authors further describes the NLP Interchange Format (NIF), that can be used, for example, to develop NLP pipelines that perform the task to assign words the corresponding DBpedia concept (entity linking). NIF represents the output of NLP tools in RDF, and thus, it is possible to integrate this data into an existing Linked Data infrastructure.

Nordhoff (2012) presents a knowledge base that conveys information *about* linguistic resources, it thus exemplifies how metadata can be provided within the Linguistic Linked Open Data cloud: Sebastian Nordhoff describes how existing work on language classification can interface with bibliographical work based on standards like TEI and Dublin Core in the Glottolog/Langdoc project. His work affords links to the vast amounts of bibliographical data contained in the LOD cloud on the one hand, and language classification and language history on the other. Further, he illustrates the linking between LOD resources for the example of Glottlog/Langdoc and ASJP online, which measures the lexical distance between languages.

Using POWLA, the DBpedia, OLiA and Glottolog/Langdoc as examples, the final contribution by Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff describes how corpora, lexical-semantic resources, and other linguistic knowledge bases can be interlinked, and how additional information can be obtained by building a Linked Open Data (sub-)cloud of linguistic resources.

3.5 STATE OF THE LINGUISTIC LINKED OPEN DATA CLOUD IN 2012

Chiarcos et al. (2011)

The idea of Linked Open Data is gaining ground: data sets from different subdisciplines of linguistics and neighboring fields are currently prepared. Related efforts, e.g. those summarized in the previous section, include fields as diverse as language acquisition, the study of folk motifs, phonological typology, translation studies, pragmatics, comparative lexicography. The coverage of the LLOD cloud is thus increasing, a major aspect of on-going work is to increase the

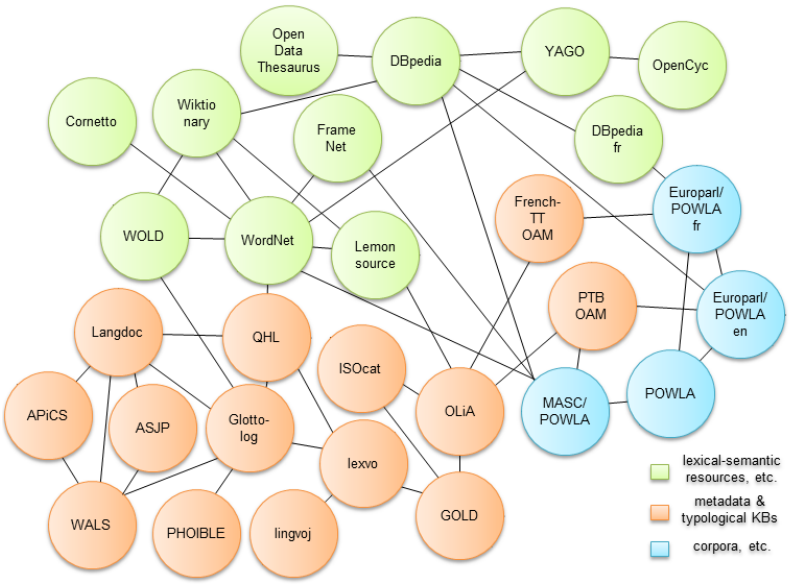


Figure 5: Draft for the Linguistics Linked Open Data (LLOD) cloud in 2012. Source: http://linguistics.okfn.org/resources/llod

density of the graph, as well. Figure 5 shows a current sketch of the LLOD cloud.[2]

The colors in the diagram correspond to different types of resources, lexical-semantic resources and general-purpose knowledge bases are shown in green, metadata repositories and typological databases in orange and corpora in blue. **Corpora** are illustrated with selected examples only, the English and French versions of Europarl v.3 as described in this article, and the Manually Annotated Subcorpus (MASC) of the American National Corpus Ide, Fellbaum, Baker, and Passonneau (2010). Like these, other corpora with comparable annotations can be represented in RDF/OWL using the POWLA scheme.

Using tools like DBpedia Spotlight (Mendes et al., 2011) and formats such as NIF, these corpora can be easily linked with **lexical-semantic resources** such as DBpedia and its language-specific instantiations. (In the diagram, only the French version is shown, further language-specific DBpedia instantiations are available and described in the Chapter 4) Other general knowledge bases that are available in the LOD have been included in the diagram besides DBpedia: YAGO,[3] OpenCyc,[4] the Open Data Thesaurus,[5] different ver-

2 It should be noted that the LLOD cloud is still work in progress. The resources in Figure 5 are available, albeit not all of them have already been converted to RDF, and not every linking has already been implemented. The diagram is inspired by the LOD diagram by Richard Cyganiak and Anja Jentzsch (http://lod-cloud.net).

3 http://www.mpi-inf.mpg.de/yago-naga/yago

4 http://www.opencyc.org

5 http://thedatahub.org/dataset/open-data-thesaurus

sions of the English WordNet[6] and the Dutch WordNet Cornetto.[7] Lemon is a formalism to publish lexical resources as Linked Data and has been applied to WordNet, Wiktionary and other resources (McCrae, Spohr, & Cimiano, 2011), and it builds on earlier models such as LingInfo (Buitelaar et al., 2006) and LexOnto (Cimiano, Haase, Herold, Mantel, & Buitelaar, 2007). Other groups are actively working on further Wiktionary instantiations that may be integrated into the LLOD, e.g. (C. Meyer & Gurevych, 2010) and the appraoch described in Chapter 5. RDF versions of FrameNet are also developed, but have not yet been publicly released (Picca, Gliozzo, & Gangemi, 2008; Scheffczyk, Pease, & Ellsworth, 2006).

In the group of lexical-semantic resources, the World Loanword Database (WOLD)[8] has a special status, because it combines characteristics of a lexical-semantic resource with those of a **typological database**. Another typological project dealing with lexical-semantic resources is Quantitative Modeling of Historical-comparative Linguistics (QHL), which digitizes dictionaries of South American languages and provides the data as RDF (Bouda & Cysouw, 2012). Besides Glottolog and Langdoc, other typological databases in the diagram include the World Atlas of Language Structures (WALS);[9] the Atlas of Pidgin and Creole Language Structures (Michaelis, Maurer, Haspelmath, & Huber, in preparation, APiCS); the Phonetics Information Base and Lexicon (PHOIBLE),[10] containing phoneme inventories from over 1,000 languages Moran (2012); and the Automated Similarity Judgment Program (Brown, Holman, Wichmann, & Velupillai., 2008, ASJP),[11] which provides word lists for over 5,000 languages as well as standardized aggregated lexical distances between language pairs computed from those word lists.

The same group of resources also includes **metadata repositories**: Lexvo and lingvoj[12] are repositories that provide terminology to describe languages; GOLD, ISOcat and the OLiA Reference Model provide information about linguistic categories and phenomena, and various OLiA Annotation Models (OAMs, illustrated only with the examples discussed in this article) formalize annotation schemes.

approach to specify formal consistency conditions (i.e. OWL, or, for other use cases, SKOS and related RDF-based formats) allows us to be open to novel, unforeseen use cases.

From the perspective of the OWLG, where different researchers with different agendas are involved, it is not possible to define a concrete application that unites all our efforts. Instead, we have come to

6 http://ckan.net/dataset/w3c-wordnet, http://ckan.net/dataset/vu-wordnet, http://ckan.net/dataset/rkb-explorer-wordnet
7 http://ckan.net/dataset/cornetto
8 http://wold.livingsources.org
9 http://www.wals.info
10 http://phoible.org
11 http://cldbs.eva.mpg.de/asjp
12 http://ckan.net/dataset/lexvo, http://ckan.net/dataset/lingvoj

the insight that RDF and Linked Data may be appropriate solutions for our different, community-specific problems, and cooperate in the development and the linking of resources according to this premise. The development of the LLOD is therefore not guided by a particular application we have in mind, but by the premise to publish data. To put it bluntly, the publication of data precedes the creation of (further) applications as Figure 1 shows. The members of the OWLG are convinced that cross-disciplinary research is an important goal and therefore strive for maximum interoperability between different tools and resources, and RDF represents the most promising foundation for this purpose.[13]

3.6 QUERYING LINKED RESOURCES IN THE LLOD

Chiarcos et al. (2011)

The LLOD cloud does not only provide us with interoperable representations of language resources, but also with the possibility to conduct queries across different resources. Integrating information from various sources allows us to enrich resources, to validate their information and thereby to achieve an improvement in terms of information quality and quantity.

For the special case of parallel corpora, Chiarcos (2012a) gives an example for the querying of multiple interlinked resources, where utterances from word-aligned French and English Europarl corpora and their alignment were modeled in RDF and queried with SPARQL. Similar applications for other complex corpora, especially multi-layer corpora, are possible. This example showed how modeling language resources in RDF can contribute to their **structural interoperability**. Section 10.2 and Chiarcos (2012b) provided another example, where information from terminology repositories was used to formulate a query on the basis of well-defined concepts rather than resource-specific tags. Using interlinked language resources thus improved the **conceptual interoperability** of linguistic annotations and corpus queries. Here, we give two other examples, concerned with the **enrichment** of language resources by information from the LLOD.

3.6.1 *Enriching metadata repositories with linguistic features (Glottolog ↦ OLiA)*

If linguistic corpora are annotated with languoids as defined in Glottlog, it is possible to identify which languoid makes use of which

13 Research on interoperability of linguistic resources so far has concentrated in different resource types, e.g. XML-standoff formats such as GrAF Ide and Suderman (2007) for linguistic corpora, or special-purpose XML formats such as the Lexical Markup Framework LMF Francopoulo et al. (2006). These efforts provide interoperability within their particular domain, but only with an RDF linearization (which exists for both formats), interoperability *between* corpora and lexical-semantic resources can be achieved.

linguistic categories and features and to use this information in typological research.

On the basis of the resources described before, this can be extrapolated from annotations that occur in the respective corpus.[14] The following query retrieves all syntactic categories that are used for a particular Glottolog languoid (given a set of corpora to which this query is applied):

```
PREFIX dcterms:  <http://purl.org/dc/terms/>.
PREFIX powla:    <http://purl.org/powla/powla.owl#>.
PREFIX olia:     <http://purl.org/olia/olia.owl#>.
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema#>.
CONSTRUCT { ?languoid <#uses> ?syntacticCategory }
WHERE {
  ?node dcterms:language ?languoid
  FILTER(regex(str(?languoid),"http://glottolog.livingsources.org/resource/languoid/id/.*")).
  ?node a powla:Node.
  ?node a ?syntacticCategory
  FILTER(regex(str(?syntacticCategory), "http://purl.org/olia/olia.owl#.*")).
  ?syntacticCategory rdfs:subClassOf olia:SyntacticCategory.
}
```

On this basis, then, one may study to what extent genealogical relationships correspond to certain syntactic features (as far as reflected in the underlying resources). For instance, one might formulate a rule which asserts the existence of a grammatical category to a `glottolog:superlanguoid` if all its sublanguoids happen to have this particular property. To give an example, the category "Preposition" is found in corpora of German, Dutch, English, and all other Germanic languages. Such a category can therefore be posited on the family level. Postpositions on the other hand are only found in a subset of the Germanic languages and thus do not "climb up the tree" as high as their prenominal brethren.

If knowledge bases with other metrics of language relatedness (e.g. ASJP, Brown et al. (2008)) are included, one can test whether these metrics correspond to the occurrence of similar grammatical features. The Linked Data approach furthermore allows to map nodes of different trees to each other. Computation of consensus trees from trees based on different datasets is another possibility.

14 It should be noted that this approach is *approximative* only, because it considers only information expressed in annotations. It is possible that the underlying schemes make a number of simplifying assumptions, e.g. not to distinguish two functionally different categories that appear superficially and that cannot be unambiguously distinguished by NLP tools or human annotators. Greater precision can probably be achieved if such queries are applied to language-annotated lexicons that make use of a standard vocabulary to represent detailed grammatical information, as created, for example, in the context of the LEGO project (Poornima & Good, 2010) whose lexicons are linked to the GOLD ontology (Farrar & Langendoen, 2003). The queries necessary for this purpose would be, however, almost identical.

3.6.2 *Enriching lexical-semantic resources with linguistic information (DBpedia (↦ POWLA) ↦ OLiA)*

Unlike classical lexical-semantic resources, DBpedia offers almost no information about the linguistic realization of the entities it contains. Using corpora with entity links and syntactic annotation, however, this information can be easily obtained. The following SPARQL query identifies possible syntactic realizations for concepts in a given corpus:

```
PREFIX powla: <http://purl.org/powla/powla.owl#>
PREFIX olia:  <http://purl.org/olia/olia.owl#>
PREFIX itsrdf: <http://www.w3.org/2005/11/its/rdf#>

CONSTRUCT { ?semClass <#realizedAs> ?syntClass }
WHERE {
  ?x a powla:Node.
  ?x itsrdf:taIdentRef ?semClass.
  ?x a ?syntClass
  FILTER(regex(str(?syntClass),"http://purl.org/olia/olia.owl#")).
  ?syntClass rdfs:subClassOf olia:MorphosyntacticCategory.
}
```

The newly generated triples can then be added to DBpedia, and provide us with information about possible grammatical realizations of an entity. A practical application of such information can be seen, for example, in the improvement of entity-linking algorithms with linguistic filters.

4

DBPEDIA AS A MULTILINGUAL LANGUAGE RESOURCE: THE CASE OF THE GREEK DBPEDIA EDITION.

Kontokostas et al. (2012)

DBpedia (Lehmann et al., 2009) is one of the most prominent LD examples. It is an effort to extract knowledge represented as RDF from Wikipedia as well as to publish and interlink the extracted knowledge according to the Linked Data principles. DBpedia is presently the largest hub on the Web of Linked Data (Kobilarov, Bizer, Auer, & Lehmann, 2009).

The early versions of the *DBpedia Information Extraction Framework* (DIEF) used only the English Wikipedia as sole source. Since the beginning, the focus of DBpedia has been to build a fused, integrated dataset by integrating information from many different Wikipedia editions. The emphasis of this fused DBpedia was still on the English Wikipedia as it is the most abundant language edition. During the fusion process, however, language-specific information was lost or ignored. The aim of approach described in this section is to establish best practices (complemented by software) that allow the DBpedia community[1] to easily generate, maintain and properly interlink language-specific DBpedia editions. We realized this best practice using the Greek Wikipedia as a basis and prototype and contributed this work back to the original DIEF. We furthermore envision the Greek DBpedia to serve as a hub for an emerging*Greek Linked Data*(GLD) Cloud (Bratsas et al., 2011).

The Greek Wikipedia is, when compared to other Wikipedia language editions, still relatively small – 66th in article count[2] – with around 65,000 articles. Although the Greek Wikipedia is presently not as well organized – regarding infobox usage and other aspects – as the English one there is a strong support action by the Greek government[3] foreseeing Wikipedia's educational value to promote article authoring in schools, universities and by everyday users. This action is thus quickly enriching the GLD cloud. In addition, the Greek government, following the initiative of open access of all public data, initiated the geodata project,[4] which is publishing data from the public sector. The Greek DBpedia will not only become the core where all these datasets will be interlinked, but also provides guidelines

1 The authors established the *DBpedia Internationalization Committee* to gather other interested community members aiming to create a network of internationalized DBpedia editions (http://dbpedia.org/internationalization).

2 Accessed on 20/10/2011: http://stats.wikimedia.org/EN/Sitemap.htm

3 http://advisory.ellak.gr/?p=12

4 http://geodata.gov.gr

```
{{TemplateMapping
| mapToClass = Actor
| mappings =
   {{ PropertyMapping | templateProperty
      = name | ontologyProperty = foaf:name }}
   {{ PropertyMapping | templateProperty
      = birth_place | ontologyProperty = birthPlace }}
}}
```

Listing 1: Example infobox-to-ontology mapping.

on how they could be published, how non-Latin characters can be handled and how the Transparent Content Negotiations (TCN) rules (RFC 2295) (Holtman & Mutz, 1998) for de-referencing can be implemented. article

4.1 CURRENT STATE OF THE INTERNATIONALIZATION EFFORT

The introduction of the Mapping-Based Infobox Extractor in Lehmann et al. (2009) alongside crowd-sourcing approaches in Hellmann, Stadler, et al. (2009) allowed the international DBpedia community to easily define infobox-to-ontology mappings using a relatively simple syntax[5] (cf. Listing 1). As a result of this development, there are presently mappings for 15 languages[6] defined in addition to English.

The DBpedia 3.7 release[7] was the first DBpedia release to use the new I18n-DIEF. The 15 languages that had infobox-to-ontology mappings defined, were extracted as localized datasets. The extensions to the DIEF, that are described in this chapter and Kontokostas et al. (2012), were generic enough to create the other localized editions. This indicates that the internationalized DIEF is sufficiently configurable to cope with the extensive amount of languages provided by Wikipedia.

Shortly after the DIEF was extended, five DBpedia chapters, apart from the English one, existed: German, Greek, Korean, Portuguese and Russian.[8] The Portuguese and Russian DBpedia editions were created right from the start using the new I18n-DIEF. Although the German and Korean chapters existed before the Greek chapter, they faced the challenges that are addressed in this section: Both chapters had developed their own approaches independently and while the German chapter used percent-encoded URIs the Korean chapter

5 http://mappings.dbpedia.org
6 ca, de, el, es, fr, ga, hr, hu, it, nl, pl, pt, ru, sl, tr.
7 http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/
8 Accessed on 25/10/2011: http://wiki.dbpedia.org/Internationalization/Chapters

also made an effort to export their datasets with localized IRIs (Kim, Weidl, Choi, & Auer, 2010).

In particular the German chapter had used the DIEF, thus discarding all articles without an English interlanguage link, which resulted in a fragmented graph structure (targets of links were not contained in the knowledge base). Furthermore, both the German and the Korean DBpedias did not incorporate the DBpedia Ontology and the mapping rules (Kontokostas et al., 2012) which resulted in a low data quality and a scheme, which was not synchronized with the fused DBpedia. Recently the German and Korean chapters have adopted the best practices we describe in this chapter and Kontokostas et al. (2012), i.e. they configured the DIEF to their language, created mappings and realized the TCN rules for IRI de-referencing.

4.2 LANGUAGE-SPECIFIC DESIGN OF DBPEDIA RESOURCE IDENTIFIERS

Earlier versions of the fused DBpedia extracted non-English Wikipedia articles only when they provided an English interlanguage link and the created resources use the default DBpedia namespace. Although this approach minimizes the use of non-Latin characters in resource identifiers, it has a the following drawbacks:

1. The merging is solely based on the link from the non-English resource to the English article. It has been shown that such links are more appropriate, if the interlanguage links go in both directions (Erdmann, K.Nakayama, Hara, , & Nishio, 2008). Because we introduce `owl:sameAs` (a transitive property) to link between language editions, we especially conducted measurements to test the integrity of our design (cf. Section 4.3).

2. A large number of articles, without an English translation link, are discarded. For instance, the DIEF produces 30% less triples for the Greek DBpedia than the I18n-DIEF (Kontokostas et al., 2012).

3. The extracted non-English articles cannot provide information other than their abstract and label, as everything else either conflicts with an English definition or creates multiple definitions.

4. The English Wikipedia is treated as the authority, which may not be the case for language-specific articles. For instance, the article about the Eiffel Tower in the French Wikipedia[9] contains more detailed information. Up to now, though, the English version of DBpedia was the only available option.

9 http://fr.wikipedia.org/wiki/Tour_Eiffel [accessed on 2011/11/07].

It is more appropriate that resources in non-English languages are published according to the Wikipedia's naming strategy, i.e. with the original article name, using a language-specific namespace (e.g. `http://el.dbpedia.org/` for Greek). As new languages are publishing their data, the English DBpedia might be transferred into `http://en.dbpedia.org/` and the default namespace could be used solely for the "Cross-language knowledge fusion" (Lehmann et al., 2009, p. 164).

4.3 INTER-DBPEDIA LINKING

Using the language-specific resource naming approach, an interlanguage link (ILL) can be utilized to connect resources across different DBpedia language editions and thus creates a multilingual semantic space. To accomplish this, a new extractor was developed for the I18n-DIEF, called *Interlanguage-Link (ILL) Extractor*. It extracts ILLs and generates RDF triples using the `dbpedia-owl:interlanguageLink` predicate. Using these links as a raw dataset, we examine whether they can be used to generate `owl:sameAs` links between resources extracted from different Wikipedia language editions. The ILL correspondence is not always reliable since by following ILLs across different languages, conflicts may appear Bolikowski (2009), as the following example illustrates:

$$\texttt{en:Tap (valve)} \mapsto \texttt{it:Rubinetto} \mapsto \texttt{es:Grifo} \mapsto \texttt{en:Griffin}$$

We performed an analysis on the ILLs, which form a directed graph (V, E), where V is the set of Wikipedia pages as nodes and E is the set of ILLs between two pages which define the edges. Wikipedia mentions the following editor guideline: "An interlanguage link is mainly suitable for linking to the most closely corresponding page in another language".[10] Thus, each concept, represented as a set of Wikipedia pages, can be defined as a subgraph consisting of the corresponding pages in each language. When this subgraph contains *at most* one article from each language, the correspondence is *consistent*, otherwise we consider it a *conflict situation*. Using the simplified dataset provided by Bolikowski (2009),[11] the graph properties were re-calculated and presented in Table 1. From the results, we can estimate the extent of conflicts and whether the conflicts are reduced, if the ILL graph is restricted to two-way links only. The conflict analysis was performed using the English articles as starting point for the measurements since it is the largest dataset.

We observe that the relative error is very small: 0.21% of the total number of English articles are participating in conflicts, creating

10 http://meta.wikimedia.org/wiki/Help:Interwiki_linking#Interlanguage_links

11 urlhttp://wikitools.icm.edu.pl/m/dumps/

Property	**Graph with all links**	**Graph restricted to two-way links**
Graph type	Directed	Undirected
Graph order (number of nodes)	878,333	825,764
Graph size (number of Links)	47,487,880	(2×) 23,001,554
Connected components (weak)	34,623	34,412
Conflicts (paths between two English articles)	380,902	(2×) 16,063
Different (English) *articles in conflicts*	5,400 (0.21%)	1,900 (0.07%)
Total English articles (as of August 2008)	2.5M	

Table 1: The ILL Graph Properties for all edges and for the subgraph of two-way edges. The calculations were performed with the open-source *R Project for Statistical Computing* (`http://www.r-project.org/`).

a total number of 380,902 conflicts. By restricting the graph to two-way links, 52,569 nodes and 1,484,772 edges are discarded. However, the discarded nodes (5.9%) are responsible for 65% of different English articles participating in conflicts, and the discarded links (3.1%) are responsible for 91.6% of the conflicts in English articles. The fact that the connected components are reduced from 34,623 to 34,412, i.e. 0.61%, is an indication that the graph structure does not change significantly, if the graph is restricted to two-way edges. Even though the relative error is small, by removing the one-way edges from ILLs, the conflicts are further reduced to 0.07% of the total number of English articles and the conflicts are reduced to 8.4%. The reason why we conducted the measurements are the strong semantical implications of `owl:sameAs`, as it produces equivalence classes in a multilingual network of language-specific DBpedia. This is why it is necessary to reduce any errors to a minimum. Our analysis indicates that the created conflicts are not significant if the `owl:sameAs` triples are considered only for two-way ILL edges.

In order to implement this analysis a new tool was created, that utilizes the ILL Extractor output from two languages, and generates `owl:sameAs` triples only for two-way edges. An example of a link extracted this way is:

$$\textbf{dbp-el}{:}\Theta\varepsilon\sigma\sigma\alpha\lambda o\nu\grave{\iota}\chi\eta \ \textbf{owl}{:}\text{sameAs} \ \textbf{dbp}{:}\text{Thessaloniki}$$

In addition to the inter-DBpedia linking, another tool was developed that transitively links a non-English DBpedia to all the external LOD datasets that are linked to the English DBpedia. Even though this could be accomplished using a SPARQL query (Garlik et al.,

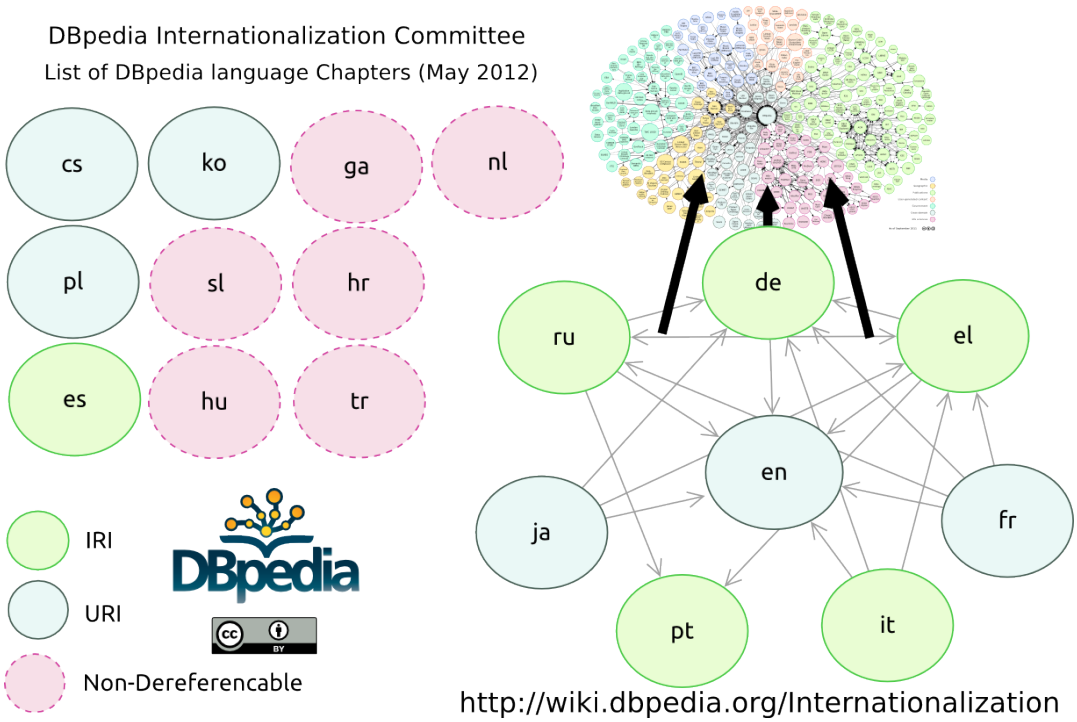


Figure 6: DBpedia language editions, manually drawn in 2012

2013), this procedure would consume substantial server resources in querying and loading all the datasets. Our tool does not have this problem because the triples are created offline, directly in the N-Triples format.

In total, 33,148 `owl:sameAs` links to the English DBpedia were established (2339 links were only one-way and have been removed (6.59%)). As a result of our inter-DBpedia linking, a total of 101,976 additional `owl:sameAs` links were created, linking the Greek DBpedia with 20 external LOD datasets.[12]

4.4 OUTLOOK ON DBPEDIA INTERNATIONALIZATION

With the maturing of Semantic Web technologies proper support for internationalization is a crucial issue. This particularly involves the internationalization of resource identifiers, RDF serializations and corresponding tool support. The Greek DBpedia is the first step towards Linked Data internationalization and the first successful attempt to serve Linked Data with de-referencable IRIs that also serves as a guide for LOD publishing in non-Latin languages. Apart from the de-referencable IRI solution, this work provides the tools for a truly international DBpedia, as Greek is a comparatively complex language with non-Latin characters and non-standard punctuation. Since the Greek DBpedia provides qualitative information comparable to the

12 http://el.dbpedia.org/en/datasets

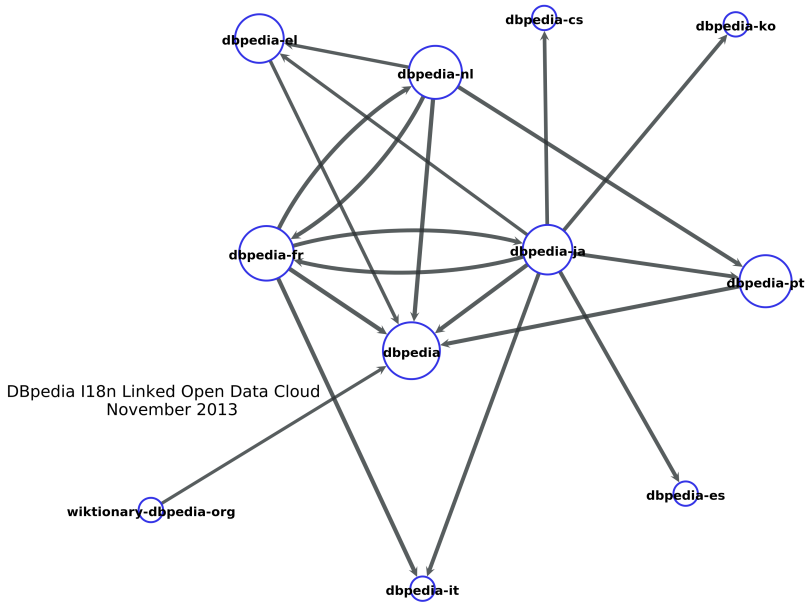


Figure 7: DBpedia language editions, generated from thedatahub.io in 2013

English DBpedia, our I18n-DIEF can be easily transferred to other non-Latin Wikipedia editions and can (with slight language specific adoptions) be expected to give similar qualitative results.

As a result of our findings, the main DBpedia edition can also significantly contribute towards the IRI adoption. The switch of the English DBpedia edition as one major Linked Data hub to use IRIs will encourage other Linked Data providers to follow. Already 17.8% (i.e. 1,679,124 out of 9,485,630 in the 3.6 release) of all resources contain the % escape character and can therefore be simpler written as IRIs.

A follow up of this work is the institutionalization of the Internationalization Committee[13] (IC). The IC could play a stronger role in the coordination and management of the various language editions of DBpedia as well as maintain and revise the best-practices laid out in this article. The committee should establish a common platform for sharing, hosting and collaboratively integrating various language editions. It should also agree upon the required technical specifications for I18n DBpedia editions as well as provide the appropriate documentation (guidelines and support documents) for the I18n operation. We included the manually drawn Figure 6 and Figure 7 to visualize the progress of the DBpedia I18n effort. Note that collecting the metadata in RDF for Figure 7 is still work in progress at the time of writing this thesis.

Another area of research is the more efficient utilization of the Wikipedia interlanguage links. The approach discussed in Section 4.3 was safe and straightforward. A further analysis of the *conflict situations* and how they could be resolved will be of great importance

13 http://dbpedia.org/Internationalization

both for Wikipedia and the internationalization of the Semantic Web. The *conflict situations* analysis could also provide new data and make us re-examine the use of `owl:sameAs` – as a too strong semantic implication – with other vocabularies (i.e. SKOS). We could also utilize the *conflicts*, which are now discarded, by adding `rdfs:seeAlso` links.

Infobox Mappings will play a central role in the integration and evolution of international DBpedia editions. Developing better mapping tools is a crucial strategy to facilitate this process.

DBpedia's goal is to "make it easier for the amazing amount of information in Wikipedia to be used in new and interesting ways" and to "inspire new mechanisms for navigating, linking and improving the encyclopedia itself".[14] Our work and the remaining work in Kontokostas et al. (2012) provides new tools for improving Wikipedia, because DBpedia may serve as an important statistical diagnostic tool for Wikipedia that helps to identify and resolve existing and emerging issues.

14 http://dbpedia.org

5

LEVERAGING THE CROWDSOURCING OF LEXICAL RESOURCES FOR BOOTSTRAPPING A LINGUISTIC LINKED DATA CLOUD

Hellmann, Brekle, and Auer (2012)

The exploitation of community-built lexical resources has been discussed repeatedly. *Wiktionary* is one of the biggest collaboratively created lexical-semantic and linguistic resources available, written in 171 languages of which approximately 147 can be considered active[1], containing information about hundreds of spoken and even ancient languages. For example, the English *Wiktionary* contains nearly 3 million words[2]. A *Wiktionary* page provides for a lexical word a hierarchical disambiguation to its language, part of speech, sometimes etymologies and most prominently senses. Within this tree numerous kinds of linguistic properties are given, including synonyms, hyponyms, hyperonyms, example sentences, links to Wikipedia and many more. C. M. Meyer and Gurevych (2011) gave a comprehensive overview on why this dataset is so promising and how the extracted data can be automatically enriched and consolidated. Aside from building an upper-level ontology, one can use the data to improve NLP solutions, using it as comprehensive background knowledge. The noise should be lower when compared to other automatic generated text copora (e.g. by web crawling) as all information in *Wiktionary* is entered and curated by humans. Opposed to expert-built resources, the openness attracts a huge number of editors and thus enables a faster adaption to changes within the language.

The fast changing nature together with the fragmentation of the project into *Wiktionary language editions* (*WLE*) with independent layout rules, called *ELE guidelines* (Entry Layout Explained, see Section 5.2.2) poses the biggest problem to the automated transformation into a structured knowledge base. We identified this as a serious problem: Although the value of *Wiktionary* is known and usage scenarios are obvious, only some rudimentary tools exist to extract data from it. Either they focus on a specific subset of the data or they only cover one or two WLE. The development of a flexible and powerful tool is challenging to be accommodated in a mature software architecture and has been neglected in the past. Existing tools can be seen as adapters to single WLE — they are hard to maintain and there are too many languages, that constantly change. Each change in the *Wiktionary* layout requires a programmer to refactor complex code. The last years showed, that only a fraction of the available data is ex-

1 http://s23.org/wikistats/wiktionaries_html.php

2 See http://en.wiktionary.org/wiki/semantic for a simple example page

tracted and there is no comprehensive RDF dataset available yet. The key question is: Can the lessons learned by the successful DBpedia project be applied to *Wiktionary*, although it is fundamentally different from Wikipedia? The critical difference is that only word forms are formatted in infobox-like structures (e.g. tables). Most information is formatted covering the complete page with custom headings and often lists. Even the infoboxes itself are not easily extractable by default DBpedia mechanisms, because in contrast to DBpedias *one entity per page* paradigm, *Wiktionary* pages contain information about *several* entities forming a complex graph, i.e. the pages describe the lexical word, which occurs in several languages with different senses per part of speech and most properties are defined *in context* of such child entities. Opposed to the currently employed classic and straightforward approach (implementing software adapters for scraping), we propose a declarative mediator/wrapper pattern. The aim is to enable non-programmers (the community of adopters and domain experts) to tailor and maintain the WLE wrappers themselves. We created a simple XML dialect to encode the ELE guidelines and declare triple patterns, that define how the resulting RDF should be built. This configuration is interpreted and run against *Wiktionary* dumps. The resulting dataset is open in every aspect and hosted as Linked Data[3]. Furthermore the presented approach can be extended easily to interpret or *triplify* other MediaWiki installations or even general document collections, if they follow a global layout.

5.1 RELATED WORK

In the last five years, the importance of *Wiktionary* as a lexical-semantic resource has been examined by multiple studies. C. Meyer and Gurevych (2010); C. M. Meyer and Gurevych (2010) presented an impressive overview on the importance and richness of *Wiktionary*. In Zesch, Müller, and Gurevych (2008) the authors presented the *JWKTL* framework to access *Wiktionary* dumps via a Java API. In C. M. Meyer and Gurevych (2011) this JWKTL framework was used to construct an upper ontology called *OntoWiktionary*. The framework is reused within the *UBY project* (Gurevych et al., 2012), an effort to integrate multiple lexical resources (besides *Wiktionary* also *WordNet*, *GermaNet*, *OmegaWiki*, *FrameNet*, *VerbNet* and *Wikipedia*). The resulting dataset is modelled according to the *LMF ISO standard*(Francopoulo et al., 2006). K. Moerth and T. Declerck and P. Lendvai and T. Váradi (2011) and Declerck et al. (2012) discussed the use of *Wiktionary* to canonicalize annotations on cultural heritage texts, namely the Thompson Motif-index. Zesch, Müller, and Gurevych (2008) and Weale, Brew, and Fosler-Lussier (2009) also showed, that *Wiktionary* is suitable for calculating semantic relatedness and synonym detection; and it out-

3 http://wiktionary.dbpedia.org/

name	active	available	RDF	#triples	ld	languages
JWKTL	✓	dumps	✗	-	✗	en, de
wikokit	✓	source + dumps	✓	n/a	✗	en, ru
texai	✗	dumps	✓	~ 2.7 million	✗	en
lemon scraper	✓	dumps	✓	~16k per lang	✗	6
blexisma	✗	source	✗	-	✗	en
WISIGOTH	✗	dumps	✗	-	✗	en, fr
lexvo.org	✓	dumps	✓	~353k	✓	en

Table 2: Comparison of existing Wiktionary approaches (ld = Linked Data hosting). None of the above include any crowd-sourcing approaches for data extraction. The wikokit dump is not in RDF.

performs classic approaches. Furthermore, other NLP tasks such as sentiment analysis have been conducted with the help of *Wiktionary* (Chesley, Vincent, Xu, & Srihari, 2006).

Several questions arise, when evaluating the above approaches: Why are there not more NLP tools reusing the free *Wiktionary* data? Why are there no web mashups of the data[4]? Why has *Wiktionary* not become the central linking hub of lexical-semantic resources, yet?

From our point of view, the answer lies in the fact, that although the above papers presented various desirable properties and many use cases, they did not solve the underlying knowledge extraction and data integration task sufficiently in terms of coverage, precision and flexibility. Each of the approaches presented in Table 2 relies on tools to extract machine-readable data in the first place. In our opinion these tools should be seen independent from their respective usage and it is not our intention to comment on the scientific projects built upon them in any way here. We will show the state of the art and which open questions they raise.

JWKTL is used as data backend of *OntoWiktionary* as well as UBY[5] and features a modular architecture, which allows the easy addition of new extractors (for example *wikokit* (Krizhanovsky, 2010) is incorporated). The Java binaries and the data dumps in LMF are publicly available. Among other things, the dump also contains a mapping from concepts to lexicalizations as well as properties for part of speech, definitions, synonyms and subsumption relations. The available languages are English, German (both natively) and Russian through *wikokit*. According to our judgement, *JWKTL* can be considered the most mature approach regarding software architecture

4 For example in an online dictionary from http://en.wikipedia.org/wiki/List_of_online_dictionaries

5 http://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/, http://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/

and coverage and is the current state of the art. *Texai*[6] and *Blexisma*[7] are also Java based APIs, but are not maintained anymore and were most probably made obsolete by changes to the *Wiktionary* layout since 2009. There is no documentation available regarding scope or intended granularity. A very fine grained extraction was conducted using WISIGOTH (Sajous, Navarro, Gaume, Prévot, & Chudy, 2010), but unfortunately there are no sources available and the project is unmaintained since 2010. Two newer approaches are the *lexvo.org* service and the algorithm presented in McCrae et al. (2012). The *lexvo.org* service offers a Linked Data representation of *Wiktionary* with a limited granularity, namely it does not disambiguate on sense level. The source code is not available and only the English *Wiktionary* is parsed. As part of the Monnet project[8], McCrae et al. (2012) presented a simple scraper to transform *Wiktionary* to the *lemon* RDF model (McCrae et al., 2011). The algorithm (like many others) makes assumptions about the used page schema and omits details about solving common difficulties as shown in the next section. At the point of writing, the sources are not available, but they are expected to be published in the future. Although this approach appears to be the state of the art regarding RDF modelling and linking, the described algorithm will *not scale to the community-driven heterogeneity* as to be defined in Section 5.2. All in all, there exist various tools that implement extraction approaches at various levels of granularity or output format. In the next section, we will show several challenges that in our opinion are insufficiently tackled by the presented approaches. Note that this claim is not meant to diminish the contribution of the other approaches as they were mostly created for solving a single research challenge instead of aiming to establish *Wiktionary* as a stable point of reference in computational linguistics using Linked Data.

5.2 PROBLEM DESCRIPTION

5.2.1 *Processing Wiki Syntax*

Pages in *Wiktionary* are formatted using the *wikitext* markup language[9]. Operating on the parsed HTML pages, rendered by the *MediaWiki engine*, does not provide any significant benefit, because the rendered HTML does not add any valuable information for extraction. Processing the database backup XML dumps[10] instead, is con-

6 http://sourceforge.net/projects/texai/
7 http://blexisma.ligforge.imag.fr/index.html
8 See http://www.monnet-project.eu/. A list of the adopted languages and dump files can be found at http://monnetproject.deri.ie/lemonsource/Special:PublicLexica
9 http://www.mediawiki.org/wiki/Markup_spec
10 http://dumps.wikimedia.org/backup-index.html

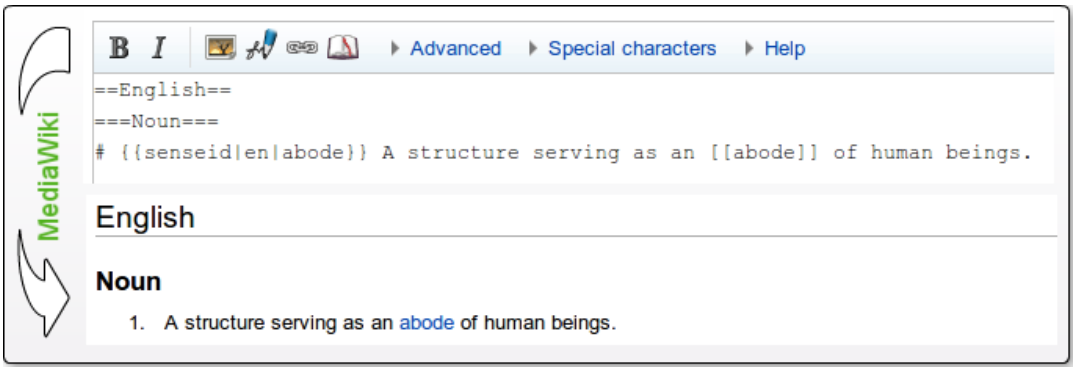


Figure 8: An excerpt of the *Wiktionary* page *house* with the rendered HTML.

venient as we could reuse the DBpedia extraction framework[11] in our implementation. The framework mainly provides input and output handling and also has built-in multi-threading by design. Actual features of the wikitext syntax are not notably relevant for the extraction approach, but we will give a brief introduction to the reader, to get familiar with the topic. A wiki page is formatted using the lightweight (easy to learn, quick to write) markup language *wikitext*. Upon request of a page, the MediaWiki engine renders this to an HTML page and sends it to the user's browser. An excerpt of the *Wiktionary* page *house* and the resulting rendered page are shown in Figure 8.

The markup == is used to denote headings, # denotes a numbered list with * for bullets, `[[link label]]` denotes links and `{{}}` calls a template. Templates are user-defined rendering functions that provide shortcuts aiming to simplify manual editing and ensuring consistency among similarly structured content elements. In MediaWiki, they are defined on special pages in the `Template:` namespace. Templates can contain any wikitext expansion, HTML rendering instructions and placeholders for arguments. In the example page in Figure 8, the `senseid` template[12] is used, which does nothing being visible on the rendered page, but adds an id attribute to the HTML `li`-tag, which is created by using #. If the English *Wiktionary* community decides to change the layout of senseid definitions at some point in the future, only a single change to the template definition is required. Templates are used heavily throughout *Wiktionary*, because they substantially increase maintainability and consistency. But they also pose a problem to extraction: on the unparsed page only the template name and its arguments are available. Mostly this is sufficient, but if the template adds static information or conducts complex operations on the arguments, which is fortunately rare, the template result can only be obtained by a running MediaWiki installation hosting the pages. The resolution of template calls at extraction time slows the process down notably and adds additional uncertainty.

11 http://wiki.dbpedia.org/Documentation
12 http://en.wiktionary.org/wiki/Template:senseid

5.2.2 *Wiktionary*

Wiktionary has some unique and valuable properties:

- **Crowd-sourced**
  Wiktionary is community edited, instead of expert-built or automatically generated from text corpora. Depending on the activeness of its community, it is up-to-date to recent changes in the language, changing perspectives or new research. The editors are mostly semi-professionals (or guided by one) and enforce a strict editing policy. Vandalism is reverted quickly and bots support editors by fixing simple mistakes and adding automatically generated content. The community is smaller than Wikipedia's but still quite vital (between 50 and 80 very active editors with more than 100 edits per month for the English *Wiktionary* in 2012[13]).
- **Multilingual**
  The data is split into different Wiktionary Language Editions (WLE, one for each language). This enables the independent administration by communities and leaves the possibility to have different perspectives, focus and localization. Simultaneously one WLE describes multiple languages; only the representation language is restricted. For example, the German *Wiktionary* contains German description of German words **as well as** German descriptions for English, Spanish or Chinese words. Particularly the linking across languages shapes the unique value of *Wiktionary* as a rich multi-lingual linguistic resource. Especially the WLE for not widely spread languages are valuable, as corpora might be rare and experts are hard to find.
- **Feature rich**
  As stated before, *Wiktionary* contains for each lexical word –A lexical word is just a string of characters and has no disambiguated meaning yet– a disambiguation regarding language, part of speech, etymology and senses. Numerous additional linguistic properties exist normally for each part of speech. Such properties include word forms, taxonomies (hyponyms, hyperonyms, synonyms, antonyms) and translations. Well maintained pages (e.g. frequent words) often have more sophisticated properties such as derived terms, related terms and anagrams.
- **Open license**
  All the content is dual-licensed under both the *Creative Commons CC-BY-SA 3.0 Unported License*[14] as well as the *GNU Free Documentation License (GFDL)*.[15] All the data extracted by our approach falls under the same licences.

13 http://stats.wikimedia.org/wiktionary/EN/TablesWikipediaEN.htm
14 http://en.wiktionary.org/wiki/Wiktionary:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License
15 http://en.wiktionary.org/wiki/Wiktionary:GNU_Free_Documentation_License

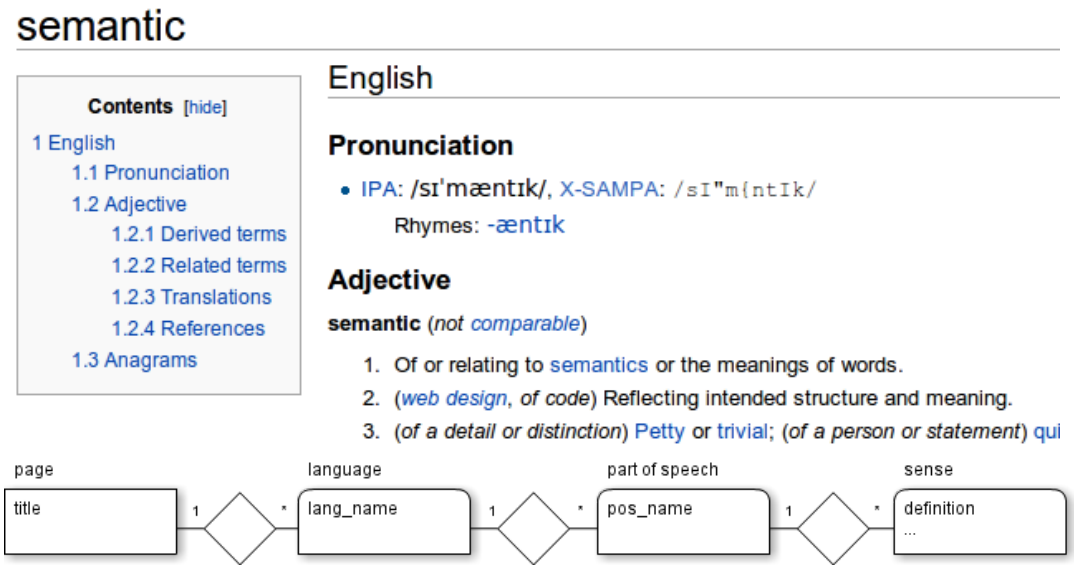


Figure 9: Example page `http://en.wiktionary.org/wiki/semantic` and underlying schema, only valid for the English *Wiktionary*, as other WLE might look very different.

- **Big and growing**
  English contains 2,9M pages, French 2,1M, Chinese 1,2M, German 0,2 M. The overall size (12M pages) of *Wiktionary* is in the same order of magnitude as Wikipedia's size (20M pages)[16]. The number of edits per month in the English *Wiktionary* varies between 100k and 1M — with an average of 200k for 2012 so far. The number of pages grows — in the English *Wiktionary* with approx. 1k per day in 2012.[17]

The most important resource to understand how *Wiktionary* is organized are the *Entry Layout Explained* (ELE) help pages. As described above, a page is divided into sections that separate languages, part of speech etc. The table of content on the top of each page also gives an overview of the hierarchical structure. This hierarchy is already very valuable as it can be used to disambiguate a lexical word. The schema for this tree is restricted by the ELE guidelines[18]. The entities illustrated in Figure 9 of the ER diagram will be called *block* from now on. The schema can differ between WLEs and normally evolves over time.

5.2.3 *Wiki-scale Data Extraction*

The above listed properties that make *Wiktionary* so valuable, unfortunately pose a serious challenge to extraction and data integration ef-

16 http://meta.wikimedia.org/wiki/Template:Wikimedia_Growth
17 http://stats.wikimedia.org/wiktionary/EN/TablesWikipediaEN.htm
18 For English see http://en.wiktionary.org/wiki/Wiktionary:ELE

forts. Conducting an extraction for specific languages at a fixed point in time is indeed easy, but it eliminates some of the main features of the source. To fully synchronize a knowledge base with a community-driven source, one needs to make distinct design choices to fully capture all desired benefits. MediaWiki was designed to appeal to non-technical editors and abstains from intensive error checking as well as formally following a grammar — the community gives itself just layout guidelines. One will encounter fuzzy modelling and unexpected information. Editors often see no problem with such "noise" as long as the page's visual rendering is acceptable. Overall, the main challenges can be summed up as (1) the constant and frequent changes to data *and schema*, (2) the heterogeneity in WLE schemas and (3) the human-centric nature of a wiki.

5.3 DESIGN AND IMPLEMENTATION

Existing extractors as presented in Section 5.1 mostly suffer from their *inflexible* nature resulting from their narrow use cases at development time. Very often approaches were only implemented to accomplish a short term goal (e.g. prove a scientific claim) and only the needed data was extracted in an *ad-hoc* manner. Such evolutionary development generally makes it difficult to generalize the implementation to heterogeneous schemas of different WLE. Most importantly, however, they ignore the community nature of *Wiktionary*. Fast changes of the data require ongoing maintenance, ideally by the wiki editors from the community itself or at least in tight collaboration with them. These circumstances pose serious requirements to software design choices and should not be neglected. All existing tools are rather monolithic, hard-coded black boxes. Implementing a new WLE or making a major change in the WLE's ELE guidelines will require a programmer to refactor most of its application logic. Even small changes like new properties or naming conventions will require software engineers to align settings. The amount of maintenance work necessary for the extraction correlates with change frequency in the source. Following this argumentation, a community-built resource can only be efficiently extracted by a community-configured extractor. This argument is supported by the successful crowd-sourcing of DBpedia's internationalization as described in Chapter 4 and (Kontokostas et al., 2012) and the non-existence of *open* alternatives with equal extensiveness.
Given these findings, we can now conclude four high-level requirements:

- declarative description of the page schema;
- declarative information/token extraction, using a terse syntax, maintainable by non-programmers;

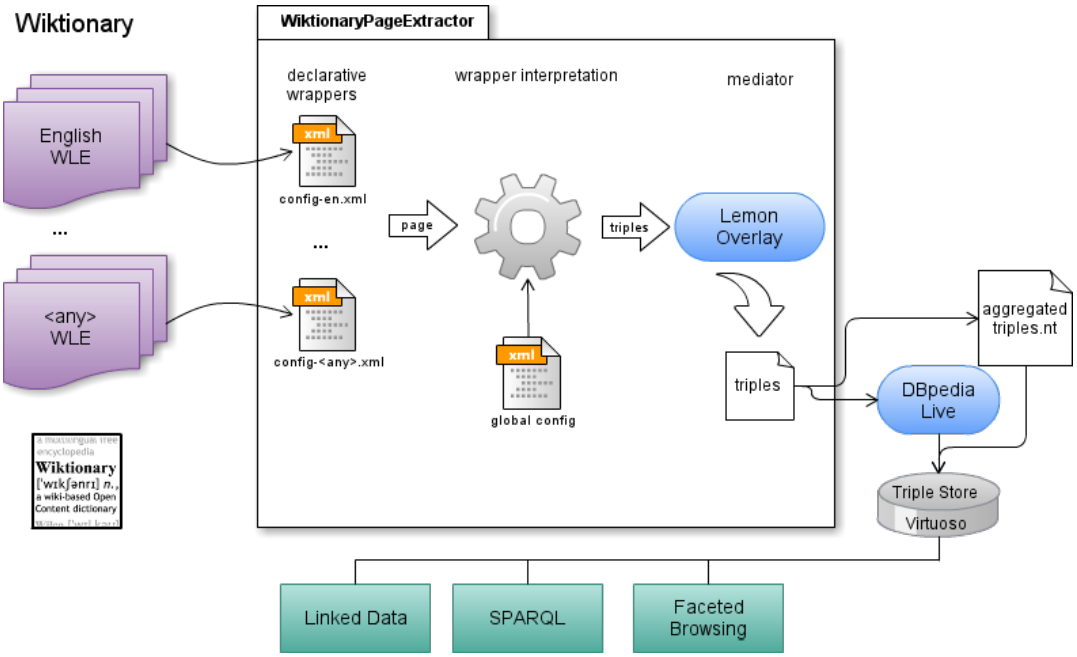


Figure 10: Architecture for extracting semantics from Wiktionary leveraging the DBpedia framework.

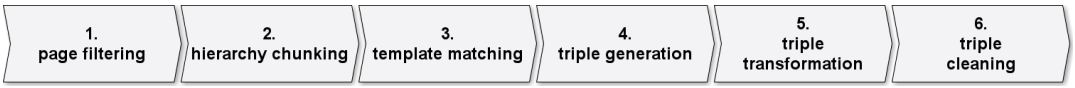


Figure 11: Overview of the extractor workflow.

- configurable mapping from language-specific tokens to a global vocabulary;
- fault tolerance (uninterpretable data is skipped).

We solve the above requirements by proposing an extension to the DBpedia framework (in fact an additional extractor), which follows a rather sophisticated workflow, shown in Figure 10.

The *Wiktionary* extractor is invoked by the DBpedia framework to handle a page. It therefore uses a language-specific configuration file, that has to be tailored to match the WLE's ELE guidelines to interpret the page. At first, the resulting triples still adhere to a language-specific schema, that directly reflects the assumed layout of the WLE. A generic lossless transformation and annotation using the *lemon* vocabulary is then applied to enforce a global schema and reduce semantic heterogeneity. Afterwards the triples are returned to the DBpedia frameworks, which takes care of the serialization and (optionally) the synchronization with a triple store via DBpedia Live (Hellmann, Stadler, et al., 2009). The process of interpreting the declarative wrapper is explained in more detailed in Figure 11.

5.3.1 *Extraction Templates*

As mentioned in Section 5.2.2, we define *block* as the part of the hierarchical page that is responsible for a certain entity in the extracted RDF graph. For each *block*, there can be declarations on how to process the page on that level. This is done by so called *extraction templates* (ET, not to be confused with the templates of *wikitext*). Each possible section in the *Wiktionary* page layout (i.e. each linguistic property) has an ET configured (explained in detail below). The idea is to provide a declarative and intuitive way to encode *what to extract*. For example consider the following page snippet:

```
===Synonyms===
* [[building]]
* [[company]]
```

Since the goal is to emit a link to each resource per line, we can write the ET in the following style, using the popular scraping paradigms such as regular expressions:

```
===Synonyms===
(* [[\$target]]
)+
```

Some simple constructs for variables "`$target`" and loops "`(*`", "`)+`" are defined for the ET syntax. If they are *matched against* an actual wiki page, *bindings* are extracted by a matching algorithm. We omit a low-level, technical description of the algorithm — one can think of it like a Regular Expression *Named Capturing Group*. The found *variable bindings* for the above example are `{(target->building), (target->company)}`. The triple generation rule encoded in XML looks like:

```
<triple s="http://some.ns/$entityId" p="http://some.ns/hasSynonym" o="http://some.ns/$target" />
```

Notice the reuse of the `$target` variable: The data extracted from the page is inserted into a triple. The variable `$entityId` is a reserved global variable, that holds the page name e.g. the word. The created triples in N-Triples syntax are:

```
<http://some.ns/house>   <http://some.ns/hasSynonym>   <http://some.ns/building> .
<http://some.ns/house>   <http://some.ns/hasSynonym>   <http://some.ns/company> .
```

The actual patterns are more complex, but the mechanism is consistently used throughout the system.

5.3.2 *Algorithm*

The algorithm of processing a page works as follows:

Input: Parsed page obtained from the DBpedia Framework (essentially a lexer is used to split the Wiki Syntax into tokens)

1. Filter irrelevant pages (user/admin pages, statistics, list of things, files, templates, etc.) by applying string comparisons on the page title. Return an empty result on that condition.

2. Build a finite state automaton[19] from the page layout encoded in the WLE specific XML configuration. This schema also contains so called *indicator templates* for each *block*, that — if they match at the current page token — indicate that their respective block starts. So they trigger state transitions. In this respect the mechanism is similar to McCrae et al. (2012), but in contrast our approach is declarative — the automaton is constructed *on-the-fly* and not hard-coded. The current state represents the current position in the disambiguation tree.

3. The page is processed token by token:
   a) Check if *indicator templates* match. If yes, the corresponding block is entered. The *indicator templates* also emit triples like in the *extraction template* step below. These triples represent the block in RDF – for example the resource `http://wiktionary.dbpedia.org/resource/semantic-English` represents the English block of the page "semantic".
   b) Check if any *extraction template* of the current block match.

      If yes, transform the variable bindings to triples.[20] Localization specific tokens are replaced as configured in the so called *language mapping* (explained in detail in section 5.3.3).

4. The triples are then *transformed*. In our implementation *transformation* means, that all triples are handed to a static function, which return a set of triples again. One could easily load the triples into a triple store like JENA and apply arbitrary SPARQL Construct and Update transformations. This step basically allows post-processing, e.g. consolidation, enrichment or annotation. In our case, we apply the schema transformation (by the mediator) explained in detail in Section 5.3.4).

5. The triples are sorted and de-duplicated to remove redundancy in the RDF dumps.

Output: Set of triples (handed back to the DBpedia Framework).

19 Actually a finite state transducer, most similar to the Mealy-Model.

20 In our implementation: Either declarative rules are given in the XML config or alternatively static methods are invoked on user-defined classes (implementing a special interface) for an imperative transformation. This can greatly simplify the writing of complex transformation.

5.3.3 *Language Mapping*

The language mappings are a very simple way to translate and normalize tokens, that appear in a WLE. In the German WLE, for example, a noun is described with the German word "*Substantiv*". Those tokens are translated to a shared vocabulary, before emitting them (as URIs for example). The configuration is also done within the language specific XML configuration:

```
<mapping from="Substantiv" to="Noun">
<mapping from="Deutsch" to="German">
...
```

5.3.4 *Schema Mediation by Annotation with* lemon

The last step of the data integration process is the schema normalization. The global schema of all WLE is not constructed in a centralized fashion — instead we found a way to both making the data globally navigable and keeping the heterogeneous schema without loosing information. *lemon* (McCrae et al., 2011) is an RDF model for representing lexical information (with links to ontologies — possibly DBpedia). We use part of that model to encode the relation between *lexical entries* and *lexical senses*. *lemon* has great potential of becoming the *de facto* standard for representing dictionaries and lexica in RDF and is currently the topic of the OntoLex W3C Community group[21]. The rationale is to add *shortcuts* from *lexical entities* to *senses* and propagate properties that are along the intermediate nodes down to the senses. This can be accomplished with a generic algorithm (a generic tree transformation, regardless of the depth of the tree and used links). Applications assuming only a *lemon* model, can operate on the shortcuts and if applied as an overlay — leaving the original tree intact — this still allows applications, to also operate on the actual tree layout. The (simplified) procedure is presented in Figure 12[22]. The use of the *lemon* vocabulary and model as an additional schema layer can be seen as our mediator. This approach is both lightweight and effective as it takes advantage of *multi-schema modelling*.

5.4 RESULTING DATA

The extraction has been conducted as a proof-of-concept on four major WLE: The English, French, German and Russian *Wiktionary*. The datasets combined contain more than 80 million facts. The data

21 http://www.w3.org/community/ontolex/

22 Note, that in the illustration it could seem like the information about part-of-speech would be missing in the *lemon* model. This in not the case. Actually from the part-of-speech nodes, there is a link to corresponding language nodes. These links are also propagated down the tree.

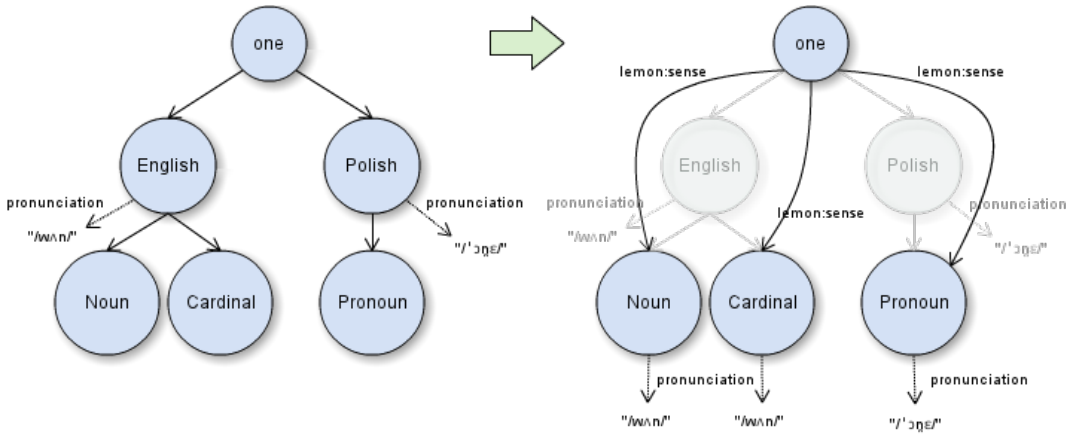


Figure 12: Schema normalization.

language	*#words*	*#triples*	*#resources*	*#predicates*	*#senses*	*XML lines*
en	2,142,237	28,593,364	11,804,039	28	424,386	930
fr	4,657,817	35,032,121	20,462,349	22	592,351	490
ru	1,080,156	12,813,437	5,994,560	17	149,859	1449
de	701,739	5,618,508	2,966,867	16	122,362	671

Table 3: Statistical comparison of extractions for different languages. XML lines measures the number of lines of the XML configuration files

is available as N-Triples dumps[23], Linked Data[24], via the *Virtuoso Faceted Browser*[25] or a SPARQL endpoint[26]. Table 3 compares the size of the datasets from a quantitative perspective.

The statistics show, that the extraction produces a vast amount of data with broad coverage, thus resulting in the largest lexical linked data resource. There might be partially data quality issues with regard to missing information (for example the number of *words with senses* seems to be relatively low intuitively), but detailed quality analysis has yet to be done. Instead we defined some simple quality measures that can be automatically computed.

Table 4 gives an assessment of the quality of the language configuration independent from the quality of the underlying source data:

t/w: *Triples per word.* The simplest measure of information density. *#wws*:*Words with senses.* The number of words, that have at least one sense extracted. An indicator for the ratio of pages for which valuable information could be extracted, but consider stub pages, that are actually empty. *s/wws*: *Senses per word with sense.* Gives an idea of the average senses per word while ignoring unmaintained pages.

23 http://downloads.dbpedia.org/wiktionary
24 for example http://wiktionary.dbpedia.org/resource/dog
25 http://wiktionary.dbpedia.org/fct
26 http://wiktionary.dbpedia.org/sparql

language	*t/w*	*#wws*	*s/wws*	*t/l*
en	13.35	591,073	1.39	2.70
fr	7.52	750,206	1.26	1.73
ru	11.86	211,195	1.40	2.25
de	8.01	176,122	1.43	1.06

Table 4: Statistical quality comparison.

t/l:*Triples per line.* The number of triples divided by the number of line breaks in the page source (plus one). Averaged across all pages.

5.5 LESSONS LEARNED

MAKING UNSTRUCTURED SOURCES MACHINE-READABLE CREATES FEEDBACK LOOPS Although this is not yet proven by empirical data, the argument that extracting structured data from an open data source and making it freely available in turn encourages users of the extracted data to contribute to the source, seems reasonable. The clear incentive is to *get the data out again*. This increase in participation besides improving the source, also illustrates the advantages of machine readable data to common Wiktionarians. Such a positive effect from DBpedia supported the current *Wikidata*[27] project.

SUGGESTED CHANGES TO WIKTIONARY Although it's hard to persuade the community of far-reaching changes, we want to conclude how *Wiktionary* can increase its data quality and enable better extraction.

- **Homogenize Entry Layout across all WLE's.**
- **Use anchors to markup senses:** This implies creating URIs for senses. These can then be used to be more specific when referencing a *word* from another article. This would greatly benefit the evaluation of automatic anchoring approaches like in C. M. Meyer and Gurevych (2011).
- **Word forms:** The notion of word forms (e.g. declensions or conjugations) is not consistent across articles. They are hard to extract and often not given.

5.6 DISCUSSION AND FUTURE WORK

Our main contributions in this section are an extremely flexible extraction from *Wiktionary*, with simple adaption to new Wiktionaries and changes via a declarative configuration. By doing so, we are provisioning a linguistic knowledge base with unprecedented detail and

27 http://meta.wikimedia.org/wiki/Wikidata

coverage. The DBpedia project provides a mature, reusable infrastructure including a public Linked Data service and SPARQL endpoint. All resources related to our *Wiktionary* extraction, such as source-code, extraction results, pointers to applications etc. are available from our project page.[28] As a result, we hope it will evolve into a central resource and interlinking hub on the currently emerging Web of Linguistic Data.

5.6.1 *Next Steps*

Wiktionary Live: Users constantly revise articles. Hence, data can quickly become outdated, and articles need to be re-extracted. DBpedia-Live enables such a continuous synchronization between DBpedia and Wikipedia. The WikiMedia foundation kindly provided us access to their update stream, the Wikipedia OAI-PMH[29] live feed. The approach is equally applicable to *Wiktionary*. The *Wiktionary* Live extraction will enable users for the first time ever to query *Wiktionary* like a database in real-time and receive up-to-date data in a machine-readable format. This will strengthen *Wiktionary* as a central resource and allow it to extend its coverage and quality even more.
Wiki based UI for the WLE configurations: To enable the crowd-sourcing of the extractor configuration, an intuitive web interface is desirable. Analogue to the mappings wiki[30] of DBpedia, a wiki could help to hide the technical details of the configuration even more. Therefore a JavaScript based WYSIWYG XML editor seems useful. There are various implementations, which can be easily adapted.
Linking: Finally, an alignment with existing linguistic resources like WordNet and general ontologies like YAGO or DBpedia is essential. That way *Wiktionary* will allow for the interoperability across a multilingual semantic web.

5.6.2 *Open Research Questions*

Publishing Lexica as Linked Data: The need to publish lexical resources as Linked Data has been recognized recently (Nuzzolese, Gangemi, & Presutti, Submitted). Although principles for publishing RDF as Linked Data are already well established (Auer & Lehmann, 2010), the choice of identifiers and first-class objects is crucial for any linking approach. A number of questions need to be clarified, such as which entities in the lexicon can be linked to others. Obvious candidates are entries, senses, synsets, lexical forms, languages, ontology instances and classes, but different levels of granularity have to

28 http://wiktionary.dbpedia.org
29 Open Archives Initiative Protocol for Metadata Harvesting, cf. http://www.mediawiki.org/wiki/Extension:OAIRepository
30 http://mappings.dbpedia.org/

be considered and a standard linking relation such as `owl:sameAs` will not be sufficient. Linking across data sources is at the heart of Linked Data. An open question is how lexical resources with differing schemata can be linked and how are linguistic entities to be linked with ontological ones. There is most certainly an impedance mismatch to bridge.

The success of DBpedia as a "crystallization point for the Web of Data" is predicated on the stable identifiers provided by Wikipedia and are an obvious prerequisite for any data authority. Our approach has the potential to drive this process by providing best practices and live showcases and data in the same way DBpedia has provided it for the LOD cloud. Especially, our work has to be seen in the context of the recently published Linguistic Linked Data Cloud (Chiarcos, Hellmann, Nordhoff, Moran, et al., 2012) and the community effort around the Open Linguistics Working Group (OWLG)[31] and NIF (Chapter 7). Our Wiktionary conversion project provides valuable data dumps and Linked Data services to further fuel development in this area.

Algorithms and methods to bootstrap and maintain a Lexical Linked Data Web: State-of-the-art approaches for interlinking instances in RDF knowledge bases are mainly build upon similarity metrics (Ngonga Ngomo & Auer, 2011; Volz, Bizer, Gaedke, & Kobilarov, 2009) to find duplicates in the data, linkable via `owl:sameAs`. Such approaches are not directly applicable to lexical data. Existing linking properties either carry strong formal implications (e.g. `owl:sameAs`) or do not carry sufficient domain-specific information for modeling semantic relations between lexical knowledge bases.

31 http://linguistics.okfn.org

6

NLP & DBPEDIA, AN UPWARD KNOWLEDGE ACQUISITION SPIRAL

Hellmann, Filipowska, et al. (2013b, 2013a)

This chapter summarizes the workshop NLP & DBpedia 2013. Communities interested in Natural Language Processing (NLP) and in the Semantic Web, in particular DBpedia, come together to explore different ways of collaborating, and helping each other, toward a common goal of understanding and representing information.

Resources such as DBpedia are a step toward a solution to the knowledge acquisition bottleneck, so often mentioned in earlier days of NLP (Gale, Church, & Yarowsky, 1992). A prerequisite of text processing and understanding is the availability of knowledge about words, concepts and ways of expressing information. But then, to acquire such knowledge, we are required to automatically process text or immerse in costly and error-prone manual knowledge engineering.

Where formerly, there was a chicken and egg problem with a serious bootstrapping issue, we now have structured data in DBpedia, which is readily available to turn the bottleneck into an *upward knowledge acquisition spiral* – a small amount of general knowledge allowing to process text, create more knowledge, validate this knowledge and improve text processing for more acquisition (and so on).

The recent years have seen a major change, mostly through crowdsourcing for the construction of the largest encyclopaedic resource, Wikipedia. Although first, mainly made of unstructured data (paragraphs), the addition of infoboxes, and the expansion of interest toward the Semantic Web, have led to DBpedia – one of the largest openly shared structured resource available today.

However, any resource not curated nor scrutinized by experts will be prone to noise, and that becomes a new and different challenge for NLP. Also, any resource, even as large as DBpedia, is not complete. So far, mainly the infoboxes, which are already semi-structured, are used to build the RDF repository. But even then, Aprosio, Giuliano, and Alberto Lavelli (2013) mention that more than 50% of Wikipedia articles do not include an infobox. So if the article text is analysed, the spiral can turn further, using DBpedia as input for the NLP process and then create more RDF triples to add and integrate into DBpedia (Héder, Mihály and Mendes, Pablo N., 2012).

This workshop's aim is right in the knowledge acquisition spiral, bringing together researchers in both areas to see how NLP can benefit DBpedia and how DBpedia can benefit NLP. The contributions in the workshop allow to highlight multiple facets of this duality.

In the remainder of this article, we discuss the contributions to the NLP&DBpedia workshop. Our main interest, however, are the challenges that the readers can expect to stay unresolved, that is the many interesting underlying issues brought forward by these articles. Another goal of this workshop was to present existing research, systems and resources to allow discussion about different points of convergence and divergence of the NLP and DBpedia community. It is also interesting to illustrate when both communities actually tackle very similar problems, with different approaches.

6.1 KNOWLEDGE ACQUISITION AND STRUCTURING

To some extent Dutta, Meilicke, Niepert, and Ponzetto (2013) explore the problem of the above-mentioned knowledge acquisition bottleneck, by comparing information extraction systems, in particular NELL (Carlson et al., 2010), which spirals on the large corpus ClueWeb09[1] to acquire more and more knowledge, with database extraction approaches based on crowd-sourcing resources such as DBpedia.

While the main focus of Dutta et al. (2013) is more about how to structure the acquired knowledge than on the acquisition method itself, their work raises an important question: To what extent can we (or should we) use Wikipedia and DBpedia to structure and organize data extracted from text? This relates to a known issue in NLP, computational terminology and even more in library science – the debate between classifying (finding which terms in a thesaurus to associate to a document) and free-characterisation (extracting any terms from the text for its representation). The former obliges a thesaurus-like structure to be built before the text is analysed. But then many questions of how such structure was made arise. The latter allows the structure (or none) to emerge from the analysed text, but makes it difficult to compare information extracted from different texts, as there is no agreed-upon schema and synonyms stay unresolved.

The proposal of Paulheim and Ponzetto (2013), is clearly on the acquisition of knowledge to be "fitted" into a known schema, that of the DBpedia ontology. Their proposal suggests the extension of DBpedia through Wikipedia list pages. The main problem is the actual matching between the extracted knowledge and the ontology. Knowledge sharing and matching is always problematic because of two main issues in semantics, that of polysemy (multiple concepts for a word) and synonymy (multiple words for a concept). Furthermore, there are also two main issues in ontology design and knowledge structuring, that of purpose-based versus non-purposed based ontologies, and that of the granularity of the information represented. All those issues combined make it quite difficult to attempt any kind of ontology expansion.

1 http://lemurproject.org/clueweb09/

6.2 REPRESENTATION OF KNOWLEDGE

As we look at NLP and DBpedia, we see that NLP requires knowledge about words, not only about concepts. Obviously the notion of labels exists in DBpedia, but there is more to language than labels. Should this lexical information be represented the same ways as conceptual information is?

The separation between lexical, conceptual, terminological, encyclopaedic, and other kind of knowledge has been a debate for years. Can a single schema allow all types of knowledge? Lexical approaches usually start from words, going from a word to all its senses, and sometimes terminological approaches will start from concepts, and defining all the words that illustrate such concept. If DBpedia is more concept-based, we can then wonder how lexical information would be attached to it, or a more general question of how lexical knowledge has its place within the Semantic Web?

Unger, Mccrae, Walter, Winter, and Cimiano (2013) present a *lemon* lexicon for DBpedia and discuss different issues of lexicalization of conceptual structures.

The BabelNet (Navigli & Ponzetto, 2012) resource, resulting from a merge of WordNet (Fellbaum, 1998) (a widely-used lexical resource in NLP) and Wikipedia, is an example of mixed-level representation in which lexical, conceptual and encyclopaedic knowledge is combined. BabelNet is used in the work of Elbedweihy, Wrigley, and Ciravegna (2013) for the task of QALD (Question Answering over Linked Data) as we will see in the next section. Also Uszkoreit and Xu (2013) talk of developing their own representation, SAR-Graphs (Semantically Associated Relations Graphs) to express not only lexical knowledge, but sentence-based knowledge, that is useful for verbalizing simple predicates but also combined predicates (child of child, for example). These three contributions stimulate a debate on the granularity of the representation of any language resource. Such debate is present in corpus studies, where experts study the value of not only terms, but also phrases (phraseology) in the understanding of language use (Stubbs, 2007).

6.3 NLP TASKS AND APPLICATIONS

Although different tasks are mentioned in our workshop's contributions, three of them are more prominent, that of NER (Named Entity Recognition), Relation Extraction, and Question Answering over Linked Data (QALD).

6.3.1 *Named Entity Recognition*

Named Entity Recognition is defined as the task of assigning a class to entities found in a text, such as person, location, organization, date, etc. NER is a well-recognized task in the NLP community since the beginning of the Message Understanding Conferences (MUC) in 1987 (see Grishman (1997) for a good overview of information extraction and the early MUC conferences). Although not called as such at the time, early work on information extraction looked at text to find *Who did What When How* discovering entities such as places, people and dates. Extracted entities were not necessarily typed, or classified, but as information extraction templates were used, such types were implicitly given by the roles the entities filled (Agent, Place, Date).

Later on, researchers, such as Sekine and Nobata (2004) defined a hierarchical schema of classes for the NER task. Although, the more fine-grained the classes are, however, the more difficult it is to obtain (or even measure) classification results. Obviously, integration and comparison of these hierarchies can have high complexity, if no reference hierarchy is agreed upon. One such reference hierarchy is the recently created NERD ontology (Rizzo et al., 2012), however, containing only 84 types[2] which is coarse grained when compared to the over 500 DBpedia Ontology classes[3], which are used in Dojchinovski and Kliegr (2013).

As mentioned in Steinmetz, Knuth, and Sack (2013) Named Entity Disambiguation is a further step toward identifying not only that an entity is a Person, but who this person actually is by establishing a link toward a more specific reference id or URI in a knowledge base. New names are given to the NED or NERD task, that of Entity linking and "wikifiers" (Dojchinovski & Kliegr, 2013) and the list of emerging tools, which belong to this class of wikifiers is quite huge and growing steadily: Zemanta, OpenCalais, Ontos, Evri, Extractiv, Alchemy API and many more[4].

Wikipedia (and therefore DBpedia) is limited to encyclopaedic knowledge, but often terminological knowledge (how different terms describe different domain specific concepts) as well as lexical knowledge (common words) are available for interlinking with text, thus resembling Word-Sense Disambiguation (WSD), i.e. taking any word in a text and being able to connect the appropriate URI. In Elbedweihy et al. (2013), both tasks (NED and WSD) are tackled using BabelNet.

2 accessed Oct. 10th, 2013 http://nerd.eurecom.fr/ontology

3 An up to date version can be downloaded fromhttp://mappings.dbpedia.org/server/ontology/dbpedia.owl

4 http://en.wikipedia.org/wiki/Knowledge_extraction#Tools contains an up-to-date overview

6.3.2 *Relation extraction*

The task of relation extraction is sometimes seen as a step following that of NER. After entities are extracted, it would be interesting to see how they are related. But sometimes a more "template-like" strategy, as was suggested in early Information Extraction is done. For example, a system would look for "merger" relations between companies, to find out which companies merged. In such case, the relation is known in advance, and we look in text for both the relation and the participants in such relation.

Different types of relations have been investigated over the years, and as NLP and DBpedia come closer, relations found in DBpedia tend to be used. Nebhi (2013) focus on ten different relations found in DBpedia. They identify such relations in text through developed lexical extraction rules. The work of Aprosio et al. (2013) focuses on seven different properties found in DBpedia. By properties, they mean relations for which the subject is most likely a named entity, but the object could be a literal, such as the property *populationTotal*. The line is fuzzy between properties and relations (for example, both contributions mentioned above use the birthDate as a relation to extract in text), and could bring an interesting discussion and debate about this topic. The work of Uszkoreit and Xu (2013) does not target any specific relation and is mostly about the development of a representational schema (as mentioned before) for the English expression of relations.

The explicit expression of relations in text is a topic of interest in the NLP community for a while. Different methods, either statistical (Turney & Littman, 2005) or pattern-based are developed and experimented on (Auger & Barrière, 2010). This is an interesting place for NLP and the Semantic Web to meet as both communities are interested in finding links between concepts and extract facts.

6.3.3 *Question Answering over Linked Data*

The tasks of Information Retrieval and Question Answering, within the NLP community, provided some of the early attempts toward a more systematized approach to making the field of NLP grow. Those tasks encouraged the development of challenges and competitions with common data (Sparck Jones, 2000, TREC) which we discuss in the next section. The more recent task of Question Answering over Linked Data[5] is a very interesting task, certainly promoting a communication and shared interest between the NLP and the Semantic Web community, and also providing some early attempts within the Semantic Web community at sharing data and evaluation standards.

5 The first challenge started in 2011, and information can be found at http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/

Three contributions look into QALD. The work of Elbedweihy et al. (2013), addresses the task of QALD, with a particular strategy which involves NED and word sense disambiguation, as we mentioned above. In Cabrio, Cojan, Villata, and Gandon (2013), the QALD task is not just tackled, but they go further into the study of inconsistency detection when gathering knowledge to answer questions. They look into English, German, French and Italian chapters of DBpedia, and try to detect inconsistencies and supporting evidence among the different answers. In Unger et al. (2013) the task of QALD is not performed in itself, but it is mentioned as an extrinsic evaluation of the coverage of the *lemon* lexicon, saying that the verbalizations found in the lexicon cover many of the questions.

6.4 RESOURCES

As most workshop contributions combine some techniques from NLP with the Semantic Web, they talk about different resources that would be useful to the community. We don't want to reinvent the wheel. Obviously, even if alternative Semantic Web resources, such as Yago (`http://www.mpi-inf.mpg.de/yago-naga/yago/`) and Freebase (`http://www.freebase.com`) exist, this workshop focuses on DBpedia, which therefore is the Semantic Web resource most referred to in the different contributions.

On the NLP side, many frameworks and typical resources exist as well. Wordnet (`http://wordnet.princeton.edu/`) for example, has been a resource much used in the community for English. More recently, Babelnet (`http://babelnet.org`), mentioned earlier, has been developed to merge Wikipedia and Wordnet. Also GATE, an open source development framework (`http://gate.ac.uk`), is used in Dojchinovski and Kliegr (2013).

We can think that the primary resource for NLP is text, but which text? There has been work in NLP on different types of texts, from news articles to scientific articles, to blogs, to web data. In the present day, textual content is abundant, and the appropriateness of which text should be analysed for which purpose is a pertinent question. In fact, if we see NLP for DBpedia, at the service of expanding DBpedia, then the chosen text should be informative, factual, accurate. As we saw above, mining Wikipedia for more information is an interesting direction, it is not the only one. We also saw (with NELL) that a large crawled Web corpus is a possibility, as it brings large coverage, but it can also bring noise.

Different ways of filtering noise exists, either by trying to evaluate the source of information (trust), or by looking at how consistent or inconsistent different information is, looking at redundancy and conflicts. In Cabrio et al. (2013), the general problem of inconsistent information is tackled.

If we reverse our point of view and see DBpedia at the service of NLP, then the text on which NLP techniques are used is quite arbitrary and depends on further purposes and applications. For example, in Dojchinovski and Kliegr (2013), both news articles and tweets are explored, which are two very different types of texts.

The question of language is valid whether we are looking at "NLP for DBpedia" or "DBpedia for NLP". In Nebhi (2013), French text is analysed, and in Cabrio et al. (2013), four different language chapters of DBpedia are used. This is a minority of contributions exploring other languages than English. As always, work on English is more prominent than that on other language, and it brings awareness that it would be interesting for both communities to work on different languages.

6.4.1 *Gold and silver standards*

The topic of evaluation is both an important one, and a much debated one. In NLP, there has been a tendency in the past 15 years to perform experiments for which there are well defined gold standards and datasets. There has been an increase in the number of competitions and challenges in many sub-fields of NLP, such as automatic summarization (Okumura, Manabu Fukusima & Nanba, 2003), word-sense disambiguation (Navigli, Jurgens, & Vannella, 2013), textual entailment (Cristea, 2009), etc.

In the Semantic Web community, there is less of such rigid evaluation, as the field is younger than NLP, and is still looking at pushing the field with different ideas and concepts without imposing rigid evaluations. Certainly, one of the purpose of this workshop was to start discussion toward bringing more of gold standards and evaluation datasets into the community. Although there are some competitions in other areas, such as the OAEI (Ontology Alignment Evaluation Initiative[6]) which has been happening for a few years now, as well as the QALD (see above) and the plethora of benchmarks for triplestores such as the DBPSB (DBepdia SPARQL Benchmark (Morsey, Lehmann, Auer, & Ngonga Ngomo, 2011)). In the field of NER/NED, however, there are not many datasets or gold standards and only few challenges. The work of Dojchinovski and Kliegr (2013) works towards the standardization of NER and NED benchmarking in an implemented benchmarking system.

As a first important step to develop such a gold standard, it is also good to review and question existing work. The work of Steinmetz et al. (2013) is an extensive comparison of NED benchmarks and characterizes them to see if they could be biased for particular types of algorithms, or types of test data. The contribution therefore opens the

6 http://oaei.ontologymatching.org/

debate as to how we should develop such benchmarks and provides a solid foundation to built upon.

When gold standards are hard (costly, time-consuming) to develop, it can be interesting to develop silver standards that are the results of well-known methods, or the combined results of different methods. Such standards do not replace gold standards, but they at least give an indication of the direction of progress for particular algorithms. One possibility when two communities come together is to take the results of one to become the "silver standard" of the other. Paulheim (2013) describes such a silver standard and discusses its benefits as well as its limitations.

In some work, such as Nebhi (2013) and Aprosio et al. (2013), DBpedia's network of relations is used as a gold standard in relation extraction. Also Wikipedia/DBpedia entities have become the most predominant link targets in NED. Rizzo et al. (2012) report of 7 out of 10 tools that attach Wikipedia/DBpedia URLs as annotations (3 out of 10 for the DBpedia Ontology). Although this is an interesting way to proceed, we can debate whether we are using gold or silver standards and how to unify benchmarks for comparison.

6.5 SUMMARY

We conclude by highlighting a few issues brought forward by the contributions in this workshop. First, the selected papers discuss many problems that have been recognized within the NLP community for a long time, but have only recently been introduced to Semantic Web researchers. The main challenges here concern:

- consensus upon annotation guidelines
- development of extraction rules and agreed upon hierarchies that may be used to unify semantic enrichment and benchmarks
- identification of well-defined tasks and problem classes
- transferability of NLP tasks, resources and tools to other research communities (e.g. library and life sciences) as well as other languages and application areas
- building practical resources and infrastructures, which do not target one single research question, but can be exploited in a more universal manner by NLP tools
- unlock higher layers of semantic annotation to enable state-of-the art OWL-based reasoning on a combination of noisy NLP data and LOD and DBpedia based knowledge structures

Second, and perhaps more importantly, new possibilities emerge from the combination of the communities, and we hope to further

push such possibilities to have more NLP for DBpedia and more DBpedia for NLP, continuing the knowledge spiral, and fighting together to open the knowledge acquisition bottleneck. We hope that the readers of the proceedings (Hellmann, Filipowska, et al., 2013b) will find all papers interesting.

Part III

THE NLP INTERCHANGE FORMAT (NIF)

7

NIF 2.0 CORE SPECIFICATION

Hellmann, Lehmann, et al. (2013)

The idea behind NIF is to allow NLP tools to exchange annotations about text in RDF. Hence, the main prerequisite is that text becomes referenceable by URIs, so that they can be used as resources in RDF statements. In NIF, we distinguish between the *document* d, the *text* t contained in the document and possible *substrings* s_t of this text. Such a substring s_t can also consist of several non-adjacent characters within t, but for the sake of simplicity, we will assume that they are adjacent for this introduction. We call an algorithm to systematically create identifiers for t and s_t a *URI Scheme*. To create URIs, the URI scheme requires a document URI du, a separator sep and the character indices (begin and end index) of s_t in t to uniquely identify the position of the substring. The canonical URI scheme of NIF is based on RFC 5147 [1], which standardizes fragment ids for the text/plain media type. According to RFC 5147, the following URI can address the first occurrence of the substring "Semantic Web" in the text (26610 characters) of the document `http://www.w3.org/DesignIssues/LinkedData.html` with the separator #: `http://www.w3.org/DesignIssues/LinkedData.html#char=717,729` The whole text contained in the document is addressed by "`#char=0,26610`" or just "`#char=0,`". NIF offers several such URI schemes which can be selected according to the requirements of the use case. Their advantages and disadvantages are described in Section 8.4, Section 9.3, Section 9.4 and Hellmann, Lehmann, and Auer (2012) and we will limit ourselves to RFC 5147 in this chapter. For practical reasons, the document URI and the separator are henceforth called the `prefix` part of the URI scheme and the remainder (i.e. "`char=717,729`") will be called the `identifier` part. NIF recommends the prefix to end on slash (/), hash ('#') or on a query component (e.g. `?nif-id=`).

The following specification describes the technical requirements for interoperability, while the NIF Core Ontology provides a formalization in Section 8.1. The technical specification is publicly available under the following URL `http://persistence.uni-leipzig.org/nlp2rdf/specification/core.html`

7.1 CONFORMANCE CHECKLIST

The keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described

1 `http://tools.ietf.org/html/rfc5147`

in RFC 2119[2]. In order to produce an interoperable implementation the following checklist MUST be followed:

1. All texts in NIF MUST be Unicode strings. Furthermore, these unicode strings SHOULD be in Unicode Normal Form C (NFC), which is the recommendation for RDF Literals. In some fringe cases, other normal forms of unicode are legit.
2. All strings MUST be counted in Unicode Code Points. We provide detailed information below.
3. All NIF implementations that expose their interfaces via web service or command line MUST implement the Public API Specification of NIF 2.0 [3].
4. All NIF implementations MUST validate their RDF output with the provided Validator.
5. For each nif:Context, taken out of another nif:Context, implementers must provide a `nif:wasConvertedFrom` provenance link.

7.2 CREATION

NIF is primarily designed to store and transfer text and text annotations. In order to enter the NIF and RDF world, the text, also called the primary data, must be (1) converted to an RDF literal as an object of the `nif:isString` property and (2) we require a way to programatically mint URIs to add annotations to the text. In the example below annotations can be added to the <SubjectURI> which serves as the context, i.e. a representative for the string in `nif:isString`.

```
<SubjectURI> nif:isString "Your text, e.g. a single sentence or
    the content of a whole document; bascially any sequence of
    characters." .
```

In the following, we will use two running examples throughout this specification: a simple sentence and a more complex .txt document as primary data.

Example 1: Web service

The primary use case of NIF is to work as an input and output format for web services. The simple sentence “My favourite actress is Natalie Portman.” serves as an example.

```
curl --data-urlencode input="My favourite actress is Natalie
    Portman."   -d informat=text "http://nlp2rdf.lod2.eu/nif-ws.
    php"
```

2 http://www.ietf.org/rfc/rfc2119.txt
3 http://persistence.uni-leipzig.org/nlp2rdf/specification/api.html

generates the following output:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#> .
<http://nlp2rdf.lod2.eu/nif-ws.php#char=0,40>
    rdf:type nif:RFC5147String , nif:Context ;
    nif:beginIndex "0" ;
    nif:endIndex "40" ;
    nif:isString "My favourite actress is Natalie Portman." .
```

Example 2: Hosting a language resource as linked data

We will give a running example here on how text and annotations can be published as Linked Data. Given a published text available on the web under the following URL `http://persistence.uni-leipzig.org/nlp2rdf/specification/example/david_lynch_dune_quoteid_124.txt`. we can create a new Linked Data URI `http://persistence.uni-leipzig.org/nlp2rdf/specification/example/david_lynch_dune_quoteid_124` as non-information resource URI (a global identifier independent of the data and representation). A web server such as Apache can now be configured to return various information resources via content negotiation (HTTP "Accept:" header) and "303 - See Other" redirects as is common practice in Linked Data:

- `text/plain` 303-redirects to david_lynch_dune_quoteid_124.txt
- `text/html` 303-redirects to an HTML visualization: david_lynch_dune_quoteid_124.php
- `text/turtle` 303-redirects to RDF in Turtle: david_lynch_dune_quoteid_124.ttl
- `application/ld+json` or `application/json` 303-redirects to RDF in JSON-LD: david_lynch_dune_quoteid_124.json
- `application/rdf+xml` 303-redirects to RDF in Json-LD: david_lynch_dune_quoteid_124.owl

Note that the returned turtle data looks like this:

```
<http://persistence.uni-leipzig.org/nlp2rdf/specification/example
    /david_lynch_dune_quoteid_124#char=0,600>
    rdf:type nif:RFC5147String , nif:Context ;
    nif:sourceUrl <http://persistence.uni-leipzig.org/nlp2rdf/
        specification/example/david_lynch_dune_quoteid_124.txt>
    nif:beginIndex "0" ;
    nif:endIndex "600" ;
    nif:isString """# Quote 124 from David Lynch's Dune ...
```

The turtle file was created by this script: [4]:

```
curl --data-urlencode input@david_lynch_dune_quoteid_124.txt --
    data-urlencode prefix="http://persistence.uni-leipzig.org/
    nlp2rdf/specification/example/david_lynch_dune_quoteid_124#"
     -d informat=text  "http://nlp2rdf.lod2.eu/nif-ws.php" >
    david_lynch_dune_quoteid_124.ttl
```

Furthermore, we provide the used .htaccess file in the GitHub repository[5] and an alternative tutorial in the NLP2RDF Wiki[6].

7.2.1 *Definition of Strings*

Achieving interoperability starts at the lowest level. In the following, we will define conventions that technically define strings in a reproducible and therefore interoperable way.

7.2.1.1 *Unicode Normalization Forms, Encoding*

According to the RDF 1.1 specification (Section 3.3 Literals)[7], RDF literals are Unicode strings, which should be in Normal Form C (NFC)[8]. In NIF, we will follow this recommendation in general. There are, however, circumstances which require the use of Normal Form D (NFD) or even NFKC or NFKD. Therefore NIF allows NFD, NFKC and NFKD, if the use case justifies the usage.

One such use case is, if a linguistic annotator wishes to annotate individual diacritics or parts of precomposed characters and syllables. For linguists with this use case, using NFD is obvious and well-justified. We refer the interested reader to these three documents, which give an introduction to this topic: Gernot Katzer's page about the Korean Writing system[9], Wikipedia article about the Korean Hangul[10], Unicode Normal Form specification[11].

7.2.1.2 *String Counting and Determination of Length*

NIF builds on the current best practices for counting strings and creating offsets. The relevant documents are:

1. *The Unicode Standard Version 6.2 - Core Specification*, Section 2.4, Code Points and Characters[12]

4 http://persistence.uni-leipzig.org/nlp2rdf/specification/example/david_lynch_dune_quoteid_124.ttl
5 https://github.com/NLP2RDF/specification/tree/master/example
6 https://github.com/NLP2RDF/software/wiki/How-to-publish-a-txt-corpora-with-NIF-as-Linked-Data
7 http://www.w3.org/TR/rdf11-concepts/#section-Graph-Literal
8 http://unicode.org/reports/tr15/#Norm_Forms
9 http://gernot-katzers-spice-pages.com/var/korean_hangul_unicode.html#
10 https://en.wikipedia.org/wiki/Hangul
11 http://unicode.org/reports/tr15/#Norm_Forms
12 http://www.unicode.org/versions/Unicode6.2.0/ch02.pdf#G25564

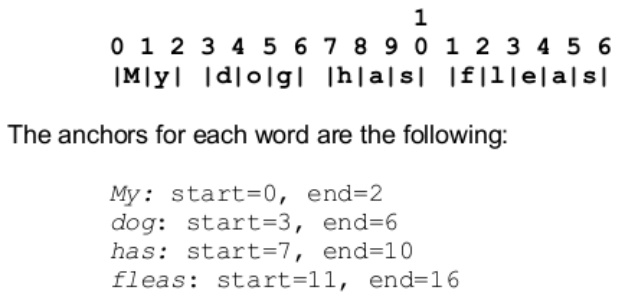


Figure 13: String counting and indexes in ISO 24612:2012

2. Section 2 of RFC 5147[13] (with the exception, that for newlines all code points must be counted in NIF)
3. ISO 24612:2012 - Language resource management – Linguistic annotation framework (LAF), see Figure 13.

NIF REQUIREMENTS

1. Begin and end offsets MUST always count the gaps between characters starting from 0 before the first character as specified in the three standards above.
2. Strings MUST always be counted in Unicode code points. NIF provides a text document to test implementations. The document is available at `http://persistence.uni-leipzig.org/nlp2rdf/specification/example/david_lynch_dune_quoteid_124.txt` and the content of it consists of exactly **600** characters

The following functions (or equivalent) are safe to use:

- Java length()[14]: `"ä".length() == 1`
- PHP utf8_decode()[15]: `strlen(utf8_decode("ä"))===1`
- Python len()[16] in combination with decode() [17]: `len("ä".decode("UTF-8"))`

Below we list some examples, which are not compatible:

- Unix wc[18]: `echo -n "ä" | wc` is 2
- PHP `strlen("ä")===2`
- Python `len("ä")===2`

13 http://tools.ietf.org/html/rfc5147#section-2
14 http://docs.oracle.com/javase/7/docs/api/java/lang/String.html#length%28%29
15 http://php.net/manual/en/function.utf8-decode.php
16 http://docs.python.org/2/library/functions.html#len
17 http://docs.python.org/2/library/stdtypes.html#str.decode
18 https://en.wikipedia.org/wiki/Wc_%28Unix%29

7.2.2 *Representation of Document Content with the nif:Context Class*

In NIF, we consider the definition of "document" as too ambiguous and not practical for NLP purposes. As soon as we start using the term "document" we are suddenly facing many modelling problems, which are relevant for area of *document management*, but only of minor interest for NLP. The biggest modelling problems are the well-known "Theseus's paradox" problem of abstract identity as well as versioning, retrieval, authorship, etc. We would like to state some pertinent examples here:

- **Theseus's paradox**: Tim Berners-Lee web publication about Design Issues for Linked Data[19] was edited several times since it's creation, but was always published under the same URL. For each change, an NLP engine would receive different textual input, but the document URI and therefore the abstract identity would remain the same, regardless of versioning and string changes.
- **Authorship**: The Wikipedia page of George W. Bush[20] has been edited over 45 thousand times. Who is the author of the document? What about user contributions that were deleted (e.g. due to vandalism)? What about the software developer who created the boilerplate HTML such as the navigation bar? Note that authorship on the string and content-level is much easier to trace.
- **Equivalence of redundant documents**: When the content of a document is copied to another URL, both exist in parallel. The content of the document is obviously the same and an NLP engine will (given it is deterministic) produce the same annotations. We are easily able to determine content equality via a string comparison, however, judging whether the documents are equivalent is difficult. The new document has a different URI, was copied there by an activity and therefore has many other properties which are different such as the HTTP headers upon a retrieval action (last-modified, e-tag). Identity according to Leibniz does not hold automatically; we would require an explicit statement that sets both documents equal.

For these reasons, we define that instances of `nif:Context` always refer to the content of the `nif:isString` property. One of the topics, during the creation of the RDF specification, was to allow literals as subjects in RDF statements (Discussion summary)[21]. The discussion

19 http://www.w3.org/DesignIssues/LinkedData.html
20 https://en.wikipedia.org/wiki/George_W._Bush
21 http://www.w3.org/2001/sw/wiki/Literals_as_Subjects

concluded that in principle, there were no predominant technical reasons to deem this approach infeasible. Notation 3 even permits literals as subjects of statements[22]. Therefore instances of nif:Context could be considered as:

```
<http://example.com/demo?cid=83848#char=0,40> owl:sameAs "My
    favourite actress is Natalie Portman." .
```

or alternatively

```
"My favourite actress is Natalie Portman." rdf:type nif:Context .
```

NIF allows the following linking between contexts and document, as well as between two NIF URIs.

7.2.2.1 *Linking to the document*

We can use `nif:sourceUrl`, which is a subproperty of `prov:hadPrimarySource` to link nif:Context to documents.

```
<http://persistence.uni-leipzig.org/nlp2rdf/specification/example
    /david_lynch_dune_quoteid_124#char=0,600>
      rdf:type nif:RFC5147String , nif:Context ;
      nif:beginIndex "0" ;
      nif:endIndex "600" ;
      nif:sourceUrl  <http://persistence.uni-leipzig.org/nlp2rdf/
          specification/example/david_lynch_dune_quoteid_124.txt>
      nif:isString "# Quote 124 from David Lynch's Dune ...
```

7.2.2.2 *Further partitioning of a context*

Some use cases require to have Linked Data URIs per paragraph or per sentence. Then they must use NIF in a way so that the orginal context can be reconstructed or traced with `nif:wasConvertedFrom` which is a subproperty of `prov:wasDerivedFrom`. For each nif:Context, taken out of another nif:Context, implementers MUST provide a `nif:wasConvertedFrom` provenance link between these contexts. Note the change of the prefix in the following example.

```
<http://persistence.uni-leipzig.org/nlp2rdf/specification/example
    /david_lynch_dune_quoteid_124_sentence1#char=0,44>
      rdf:type nif:RFC5147String , nif:Context ;
      nif:beginIndex "0" ;
      nif:endIndex "44" ;
      nif:wasConvertedFrom  <http://persistence.uni-leipzig.org/
          nlp2rdf/specification/example/
          david_lynch_dune_quoteid_124#char=47,91>
      nif:isString "It is by will alone I set my mind in motion.
          """ .
```

22 http://lists.w3.org/Archives/Public/www-rdf-comments/2002JanMar/0127.html

7.3 EXTENSION OF NIF

Hellmann and Auer (2013)

In this section, we describe the extension mechanisms used to achieve interoperability between different annotation layers using RDF and the NIF URI schemes. Several vocabularies (or ontologies) were developed and published by the Semantic Web community, where each one describes one or more layers of annotations. The current best practice to achieve interoperability on the Semantic Web is to re-use the provided identifiers. Therefore, it is straightforward to generate one or more RDF properties for each vocabulary and thus connect the identifiers to NIF. We call such an extension a *Vocabulary Module* and include the following ontologies:

- OLiA[23] for POS tagging and other morpho-syntactical annotations.
- ITS 2.0 RDF[24] for ITS 2.0 related use cases such as entity linking, localization and machine translation.
- lemon[25] for connecting lexical resource.
- MARL[26] for sentiment analysis.
- NERD[27] for class linking.

We introduce three generic properties called `annotation` (for URIs as object), `literalAnnotation` (for literals as object) and `classAnnotation` (for OWL classes as object), which are made available in the NIF Core Ontology (cf. Section 8.1). The third one is typed as OWL annotation property in order to stay within the OWL DL language profile. All further properties used for annotation should be either modelled as a subproperty (via `rdfs:subPropertyOf`) of `annotation`, `literalAnnotation` or `classAnnotation` or left underspecified by using the `annotation`, `literalAnnotation` or `classAnnotation` property directly. This guarantees that on the one hand OWL conventions are followed for uniform processing, while developers, on the other hand, can still use their own annotations via the extension mechanism. The distinction between `annotation`, `literalAnnotation` and `classAnnotation` guarantees that each vocabulary module will still be valid OWL/DL, which is essential for standard OWL reasoners.

When modeling an extension of NIF via a vocabulary module, vocabulary providers can use the full expressiveness of OWL/DL. In the following, we will present several vocabulary modules, including design choices, so they can serve as templates for adaptation and further extensions.

23 http://purl.org/olia
24 http://www.w3.org/2005/11/its/rdf
25 http://lemon-model.net/
26 http://marl.gi2mo.org/
27 http://nerd.eurecom.fr/

7.3.1 *Part of Speech Tagging with OLiA*

The *Ontologies of Linguistic Annotation* (Chiarcos, 2012b, OLiA)[28] provide stable identifiers for morpho-syntactical annotation tag sets, so that NLP applications can use these identifiers as an interface for interoperability. OLiA provides *Annotation Models* for the most frequently used tag sets, such as Penn[29]. These annotation models are then linked to a *Reference Model*, which provides the interface for applications. Consequently, queries such as 'Return all Strings that are annotated (i.e. typed) as `olia:PersonalPronoun` are possible, regardless of the underlying tag set. In the following example, we show how *Penn Tag Set*[30] identifiers are combined with NIF:

```
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#> .
# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
<http://nlp2rdf.lod2.eu/nif-ws.php#char=13,20>
   rdf:type nif:RFC5147String ;
   nif:anchorOf "actress" ;
   nif:beginIndex "13" ;
   nif:endIndex "20" ;
   nif:referenceContext <http://nlp2rdf.lod2.eu/nif-ws.php#char
       =0,40>  ;
   # provenance link
   nif:oliaProv <http://nlp.stanford.edu:8080/parser/index.jsp> ;
   # the tag
   nif:oliaLink <http://purl.org/olia/penn.owl#NN> ;
   # mappings can be found here: https://github.com/NLP2RDF/
       software
   nif:oliaCategory <http://purl.org/olia/olia.owl#Noun> ,
         <http://purl.org/olia/olia.owl#CommonNoun> .
```

`nif:oliaLink` and `nif:oliaCategory` are subproperties of `annotation` and `classAnnotation` respectively and link to the tag set specific annotation model of OLiA as well as to the tag set independent reference ontology. The main purpose of OLiA is not to dictate a certain meaning for linguistic features, but to provide a description of existing annotation sets and a mapping for data integration. OLiA can be extended by third-parties easily to accommodate more tag sets currently not included. Furthermore, all the ontologies are available under an open license[31]. An overview can be found at `http://purl.org/olia`

28 http://purl.org/olia
29 http://purl.org/olia/penn.owl
30 http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html
31 http://sourceforge.net/projects/olia/

7.3.2 *Named Entity Recognition with ITS 2.0, DBpedia and NERD*

The *MultilingualWeb-LT Working Group*[32] has published a standard for the *Internationalization Tag Set* (ITS) Version 2.0[33], which will allow to include coarse-grained NLP annotations into XML and HTML via custom attributes. This specification also includes a link to the ITS RDF Ontology and NIF as the recommended RDF mapping. Complementary to the ITS standardization effort, the *Named Entity Recognition and Disambiguation* (NERD) project (Rizzo et al., 2012) has created mappings between different existing entity type hierarchies to normalize named entity recognition tags.

We will introduce here the most important properties and define how they must be used:

7.3.2.1 *Entity linking*

This section describes the extension of NIF, which can be used to link to link data entities such as the ones provided by DBpedia.

All entities must be attached via the functional OWL property `itsrdf:taIdentRef`:

```
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#>.

# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  rdf:type nif:NamedEntity ;
  nif:anchorOf ''Natalie Portman'' ;
  nif:beginIndex ''24'';
  nif:endIndex ''33'' ;
  itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman>
       .
```

Note that the functionality of OWL properties allows to infer that, if the same subject has two different objects, then these can be considered identical:

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
  itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman>
       .
  itsrdf:taIdentRef <http://rdf.freebase.com/ns/m.09l3p> .
  # entails that
<http://dbpedia.org/resource/Natalie_Portman>
    owl:sameAs  <http://rdf.freebase.com/ns/m.09l3p> .
```

32 http://www.w3.org/International/multilingualweb/lt/
33 http://www.w3.org/TR/its20/

7.3.2.2 *Class linking*

Although the task of Named Entity Recognition (NER) or Class linking is quite old, we were astonished that the modelling in OWL is not straightforward at all and opens many questions. During the discussions, we had, while creating ITS 2.0 [34] and in the last session of the NLP & DBpedia workshop in Sydney in 2013, we came up with several ways to model different semantics.

For the general case, it was agreed to create an underspecified OWL annotation property called `itsrdf:taClassRef` to attach any types to the string, with the simple meaning that this string is now *annotated with the respective type*. This modelling allows virtually any tool to use this free property in its output and guarantees a broad coverage.

Several such classes can be assigned without limitation in RDF:

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
   itsrdf:taClassRef <http://dbpedia.org/ontology/Actor> ;
   itsrdf:taClassRef <http://dbpedia.org/ontology/Artist> ;
   itsrdf:taClassRef <http://dbpedia.org/ontology/Person> ;
   itsrdf:taClassRef <http://xmlns.com/foaf/0.1/Person> .
```

This underspecified property can be refined or converted into these more specific OWL properties:

COARSE-GRAINED, NORMALIZED NERD CLASS As written in Rizzo et al. (2012), we can use the NERD ontology to map many of the other type hierarchies to one of the NERD core classes: `Amount, Animal, Event, Function, Location, Organization, Person, Product, Time`. NIF requires that only one of these classes must be assigned, as they are disjoint annotations. The assignment must be done via the `nif:taNerdCoreClassRef` OWL annotation property.

MOST-SPECIFIC CLASS Often it is practical to directly state, that one of the assigned classes is the most-specific one. While a logical definition of most-specific can have several variants (i.e. the lowest class in a hierarchy that covers the individual or the conjunction of all such classes across hierarchies plus the distinction between assigned and potentially existing or inferable classes), feedback suggested that a solution that does not require a reasoner or external knowledge is preferred. Therefore, the OWL annotation property `nif:taMsClassRef` must link only to classes from the set of all classes of `itsrdf:taClassRef`, which do not have a subclass in the set itself.

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
   itsrdf:taClassRef <http://dbpedia.org/ontology/Actor> ;
   itsrdf:taClassRef <http://dbpedia.org/ontology/Artist> ;
   itsrdf:taClassRef <http://dbpedia.org/ontology/Person> ;
   itsrdf:taClassRef <http://xmlns.com/foaf/0.1/Person> ;
```

34 the discussion is documented on the mailing list archive of the working group

```
# Actor is the only class that does not have a subclass in the
    set of 4.
   nif:taMsClassRef <http://dbpedia.org/ontology/Actor> .
```

DIFFERENTIATION BETWEEN CONTEXT-DEPENDENT AND UNIVERSAL CLASSES Note that all the above mentioned properties do not distinguish between context-dependent meaning and context-independent knowledge as stored in an RDF/OWL knowledge base.

In our example sentence, the entity "Natalie Portman" is mentioned. From the context, i.e. the surrounding sentence, however, we can only learn that she is an actor/actress. A knowledge base might hold additional information such as the fact the she is an Israeli cititzen[35].

Following the original model in Rizzo et al. (2012), NIF allows to make the transition between context-dependent and universal class linking. In the context of the sentence "My favorite actress is Natalie Portman" an NLP engine might assign a class called `Favorite_Actress`, which is definitely true in this context, but does not hold for the real-life actress Natalie Portman in all cases. The `itsrdf:taClassRef` property can be used for this assigment. If the NLP engine, however, aims to extract context-independent types and facts, which have a broader applicability, the output must look as follows:

```
<http://nlp2rdf.lod2.eu/nif-ws.php#char=24,39>
   itsrdf:taIdentRef <http://dbpedia.org/resource/Natalie_Portman
       > ;
   itsrdf:taClassRef <http://dbpedia.org/ontology/Actor> .
# the tool states that the type found is a general property of
    the found entity beyond the context of the current nif:
    Context
<http://dbpedia.org/resource/Natalie_Portman>
   rdf:type <http://dbpedia.org/ontology/Actor> .
```

```
<char=24,39>
   itsrdf:taIdentRef dbpedia:Natalie_Portman ;
   # class assignment valid in context
   itsrdf:taClassRef dbo:Actor ;
   nif:taNerdClassRef nerd:Person .
   # generalized class of the entity Natalie Portman
dbpedia:Natalie_Portman
   rdf:type dbo:Actor .
```

7.3.3 *lemon and Wiktionary2RDF*

URIs of RDF datasets using lemon (McCrae et al., 2011) can be attached to NIF URIs employing the ITS RDF property `itsrdf:termInfoRef`,

35 http://dbpedia.org/class/yago/PeopleFromJerusalem

which can link to lexical entries or senses contained in a lemon lexicon.

```
@prefix wiktionary: <http://wiktionary.dbpedia.org/resource/> .
<http://nlp2rdf.lod2.eu/nif-ws.php#char=13,20>
    nif:anchorOf "actress" ;
    its:termInfoRef wiktionary:actress-English-Noun-1en .
```

8

NIF 2.0 RESOURCES AND ARCHITECTURE

Hellmann, Lehmann, et al. (2013)

8.1 NIF CORE ONTOLOGY

The NIF Core Ontology[1] provides classes and properties to describe and formally define the relations between text, substrings and their URI schemes. We will give a brief summary here and give core definitions where they can not be captured with the expressiveness of OWL DL (description logics are a decidable fragment of predicate logic). The remaining formal definitions are published online as OWL Ontology and not repeated here in detail.

The main class in the ontology is `nif:String`, which is the class of all **words over the alphabet A_U of Unicode characters**. In NIF we define *word*[2] as an arbitrary sequence of Unicode characters not distinguishing between whitespace or line separators and "visible" characters. Even more technical our definition of character comes down to the 1,112,064 code points in the Unicode character set and therefore $|A_U| = 1,112,064$

In the literature this is often called Σ^* for formal languages. We follow the definition of Heyer, Quasthoff, and Wittig (2006, p.325): If A_U^n is the n-ary Cartesian product of the Alphabet A_U of Unicode characters (i.e. all character sequences of length n), then $A^* = \bigcup_{n>=0} A_U^n$ is the set of possible meanings (i.e. the universe of discourse) for individuals of type `nif:String` (we do not exclude the empty word – a zero-length string).

We built NIF upon the Unicode Normalization Form C, as this follows the recommendation of the RDF standard[3] for `rdf:Literal`. Indices are to be counted in-between code points. Another class in the ontology is `nif:URIScheme`. Each `nif:URIScheme` is a subclass of `nif:String` and puts further restrictions over the syntax of the URIs of its members. For example, instances of type `nif:RFC5147String` have to adhere to the NIF URI scheme based on RFC 5147. Users of NIF can create their own URI schemes by subclassing `nif:String` and providing documentation on the Web in the `rdfs:comment` field.

Another very important subclass of `nif:String` is the `nif:Context` OWL class. This class is assigned to the whole string of the text (i.e. all characters in the current viewpoint or interpretation frame). The purpose of an individual of this class is central, because the string of this individual is used to calculate the indices for all substrings. Therefore,

1 http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
2 Not to be confused with the OWL class `nif:Word` for tokenization
3 http://www.w3.org/TR/rdf-concepts/#section-Literals

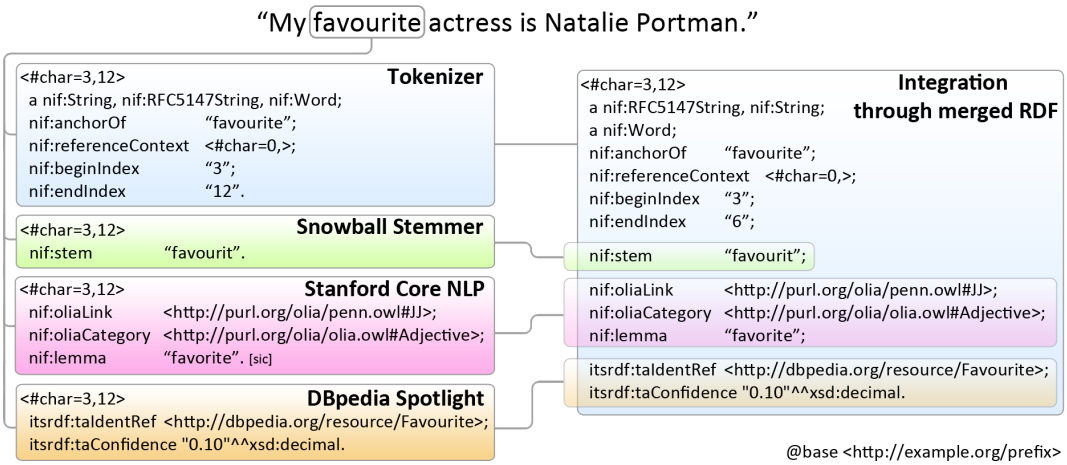


Figure 14: An example of NIF integration. Tool output from four tools is merged via URLs. Reproducible at the NIF demo site: http://nlp2rdf.lod2.eu/demo.php

all substrings have to have a relation `nif:referenceContext` pointing to an instance of `nif:Context`. Furthermore, the datatype property `nif:isString` can be used to include the reference text as a literal within the RDF as is required for the web service scenario. An example of NIF Core can be seen on the top left of Figure 14 which shows a typical merging operation for NIF.

We did not repeat all the classes of the ontology in this thesis. There is an extensive online documentation available at http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html. Figure 15 gives an overview of the ontology.

8.1.1 *Logical Modules*

The NIF ontology[4] is split in three parts: The *terminological model* is lightweight in terms of expressivity and contains the core classes and properties. Overall, it has 98 axioms, 20 classes, 14 data properties and 31 object properties. The *inference model* contains further axioms, which are typically used to infer additional knowledge, such as transitive property axioms. The *validation model* contains axioms, which are usually relevant for consistency checking or constraint validation[5], for instance class disjointness and functional properties. Depending on the use case, the inference and validation model can optionally be loaded. Overall, all three NIF models consist of 177 axioms and can be expressed in the description logic $\mathcal{SHIF}(\mathcal{D})$ with exponential reasoning time complexity (Tobies, 2001).

4 Available at http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/version-1.0/.

5 See e.g. http://clarkparsia.com/pellet/icv/.

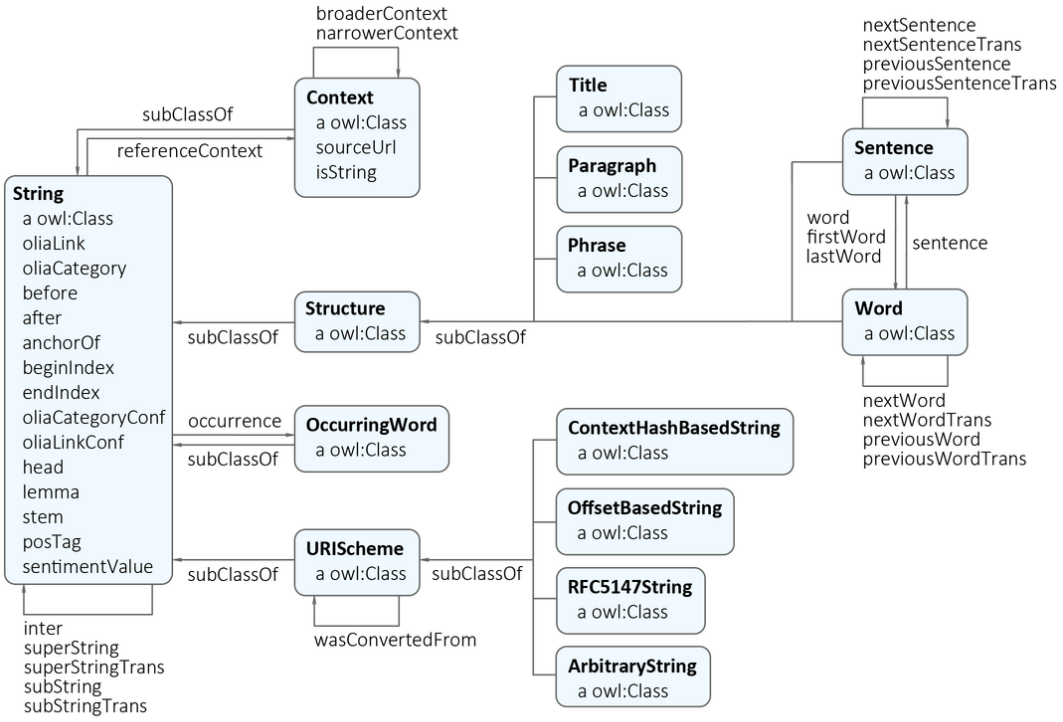


Figure 15: Overview of the NIF Core Ontology

8.2 WORKFLOWS

Hellmann, Lehmann, et al. (2013)

NIF web services are loosely coupled and can receive either text or RDF. To allow seamless NLP integration, clients should create work flows where the text is normalized (Unicode) at the beginning and tokenization is provided. Figure 16 shows one of the possible workflows that uses an NLP tokenizer in a preprocessing step (Hellmann, Lehmann, Auer, & Nitzschke, 2012). Based on the normalization and tokenization, the combined RDF of several tools merges naturally based on the subject URIs as shown in Figure 14. Tokenization con-

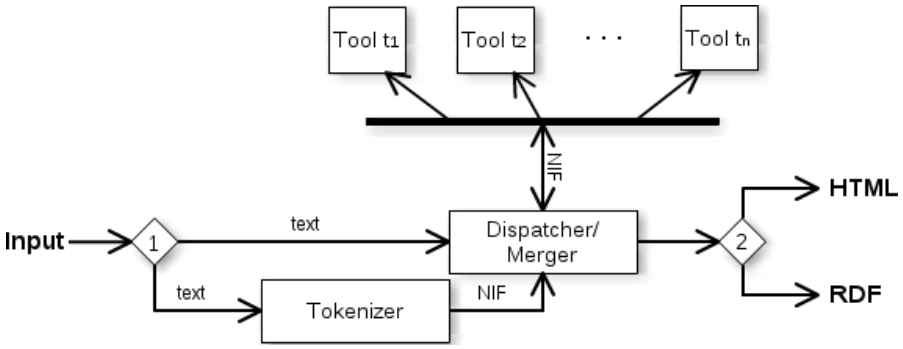


Figure 16: Workflow implemented by the NIF Combinator (Hellmann, Lehmann, Auer, & Nitzschke, 2012)

flicts are a recognized problem in NLP; other algorithms are applicable such as the ones mentioned in Chiarcos, Ritz, and Stede (2012), if no *a priori* resolution is applied.

8.2.1 *Access via REST Services*

The structural and conceptual interoperability layers of NIF are built upon the RDF standard, Linked Data principles and existing ontologies such as OLiA. To improve interoperability and accessibility of NIF components, NIF provides a normative *access layer*, which facilitates easier integration and off-the-shelf solutions by specifying REST parameters. Of special importance is the *prefix* parameter as it enables the client to influence the RDF output. The RDF in Figure 14 is produced by different tools, but can be merged directly under the condition that the URI prefixes and offsets are the same.

NIF can be used for import and export of data from and to NLP tools. Therefore, NIF enables to create ad-hoc **workflows** following a client-server model or the SOA principle. Following such an approach, clients are responsible for implementing the workflow. The NIF Combinator shows one possible implementation of such a workflow. The client sends requests to the different tools either as text or RDF and then receives responses in RDF. This RDF can be aggregated into a local RDF model. Transparently, external data in RDF can also be requested and added without using additional formalisms. For acquiring and merging external data from knowledge bases, existing Semantic Web tools can be used.

The main **interface** are wrappers that provide NIF web services. A NIF web service must be *stateless*, *HTTP method agnostic respective POST and GET*.

NIF 2.0 provides another specification that defines seven parameters, which NIF web services must implement. The specification is not repeated here in detail and can be viewed at `http://persistence.uni-leipzig.org/nlp2rdf/specification/api.html`.

8.2.2 *NIF Combinator Demo*

Figure 16 describes the workflow of the NIF Combinator. The given input (normally text) can be forwarded directly to the dispatcher or optionally prepared by a tokenizer (see Diamond 1 in Figure 16). The tokenizer already outputs the results in NIF and provides tokenization for the remaining components. The dispatcher then calls the selected NLP tools (see checkboxes in Figure 17) which can read as well as write NIF. The NIF output from all the tools is then merged (see Figure 18). Merged results can be shown in HTML format for users. Another option (Diamond 2) is to output RDF directly. This way, the NIF Combinator can be used as an aggregator web service itself, by

Figure 17: Screenshot of the NIF Combinator user interface.

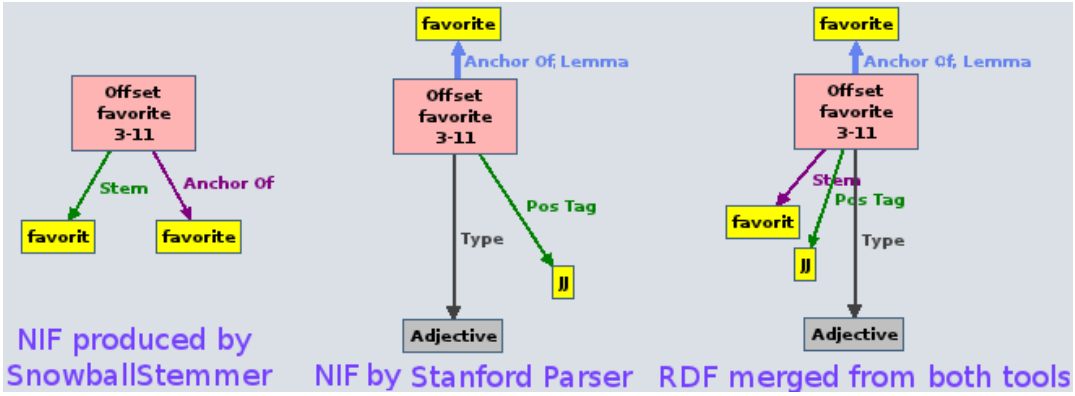


Figure 18: Example of merged RDF from two NLP tools.

simply executing a GET/POST request with the parameters of the HTML forms.

8.3 GRANULARITY PROFILES

Hellmann, Lehmann, et al. (2013)

We will give a brief technical introduction into the four different granularities, which are shown in Figure 19.

NIF SIMPLE. Basic properties describe the strings and the reference text unambiguously. NIF simple allows to express *the best estimate* of an NLP tool in a flat data model. The profile is sufficient for most use cases including simple NLP tasks, such as POS tagging or NER. The client is responsible to resolve any inconsistencies and merge the data retrieved in a web service context. Most properties such as `itsrdf:taIdentRef` and `nif:oliaLink` are functional and enforce (if validated) at most one annotation of a certain type per string. Confidence can be encoded for each annotation, though no alternatives can be included. Provenance can only be encoded for one tool, which is sufficient in the context of a single web service request.

NIF Simple Underspecified. A variant of the above this profile may only be applied, iff the *prefix* equals the annotated information resource. Other information (especially the reference context) may be omitted and later recreated from the identifier part of the URI scheme. In our running example, the file LinkedData.txt can be retrieved from the Web and the identifier `<char=333,345>` would be enough to explicate the remaining triples on the client side. The profile has the lowest triple count (*one triple per annotation*), but can not be queried effectively with SPARQL and has the risk of running out of sync with the primary data.

NIF Stanbol. Alternative annotations with different confidence as well as provenance information (i.e. which NLP engine produced which annotation) can be attached to the additionally created URN for each annotation. The NIF Stanbol profile is complementary to NIF simple, transformation is lossless, except, of course, for the alternatives and the provenance information. The model is interesting for creating algorithms that try to optimize output from different engines and require the detailed NLP graph.

NIF OA (Open Annotation). Open Annotation provides the most expressive model, but requires more triples and creates up to four new URNs per annotation.

Apache Stanbol[6] is a Java framework, that provides a set of reusable components for semantic content management. One component is the content enhancer that serves as an abstraction for entity linking engines. For Stanbol's use case, the NLP graph is required, including provenance, confidence of annotations as well as full information about alternative annotations (often ranked by confidence) and not only the best estimate. The FISE ontology[7] is integrated into NIF as a vocabulary module and a NIF implementation is provided by the project(cf. Section 9.1).

Open Annotation Data Model (OA[8], formerly the annotation ontology (Ciccarese, Ocana, Garcia Castro, Das, & Clark, 2011)) was originally devised as an 'open ontology in OWL-DL for annotating scientific documents on the web' and is now advanced by the Open Annotation W3C Community Group. OA provides structural mechanisms to annotate arbitrary electronic artifacts and resources (including images, websites, audio and video). OA is a generic approach that succeeds in creating an annotation framework for a plethora of use cases and distinguishes between the *body*, the *target* and the *annotation* itself by creating URNs for each of the parts. As NLP has special requirements regarding scalability, NIF offers two more granularities targeting reduced overhead and three different levels of reasoning.

6 http://stanbol.apache.org
7 http://fise.iks-project.eu/ontology/
8 http://www.openannotation.org

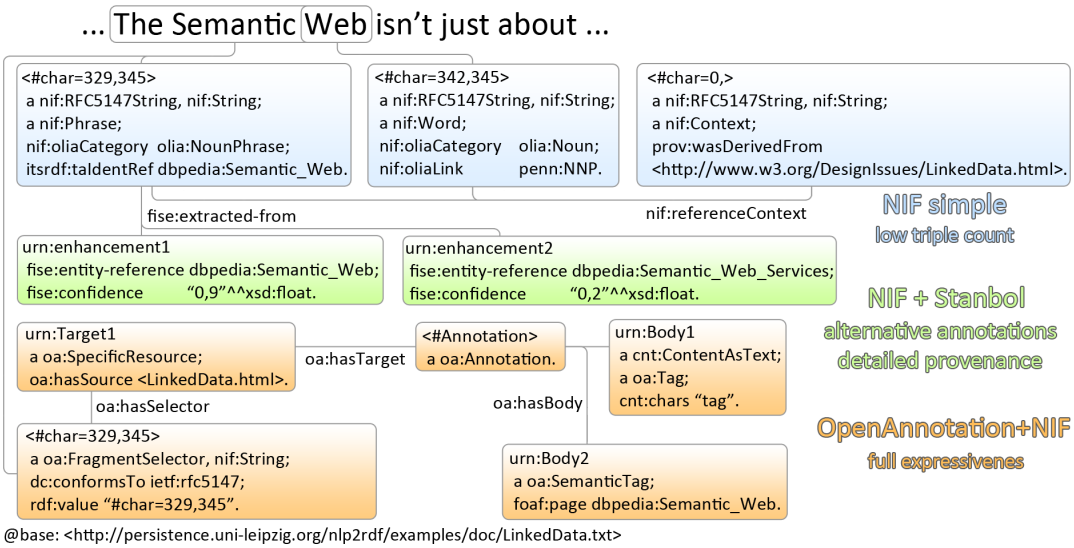


Figure 19: Three of the four granularity profiles of NIF. Open annotation is able to use NIF identifiers as `oa:Selector`.

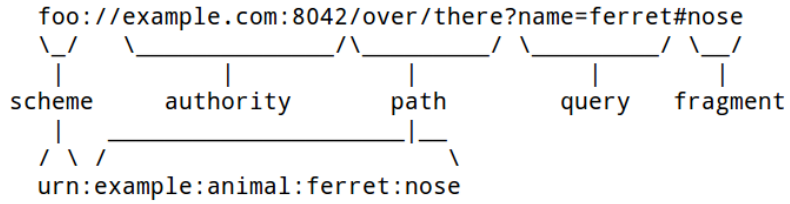


Figure 20: Two example URIs and their component parts, taken from RFC 3986

Furthermore, OA is domain-agnostic, while NIF defines best practices for annotations as well as a community infrastructure to agree on common domain annotations and reference ontologies to create interoperability in the NLP domain.

Especially noticeable is the fact that all three main granularities are complementary and can be kept together. A client could keep token and POS tags in NIF simple to reduce triple count, encode entity linking in NIF Stanbol to keep the alternatives and then have user tags and comments in NIF OA, because OA allows to reply to previous comments (annotations on annotations). An implementation is for example provided in the OpenPHACTS system.[9]

8.4 FURTHER URI SCHEMES FOR NIF

Hellmann, Lehmann, and Auer (2012)

RFC 3986 [10] defines the Generic Syntax for Uniform Resource Identifier (URI). This generic URI syntax consists of a hierarchical sequence of components referred to as the scheme, authority, path, query, and

9 http://ubo.openphacts.org/index.php?id=4684
10 http://tools.ietf.org/html/rfc3986

fragment and can be seen in Figure 20. For our purposes of defining an RDF-based format for representing text and annotations on text, we are especially interested in scheme, path, query and fragment components, as well as retrieval actions resulting in their resolution (e.g. for Linked Data) and the semantics of URIs in RDF. We omit everything unimportant here such as the syntactical conversion of relative URI reference to URIs and refer the reader to the respective RFC's and the W3C recommendations. In order to achieve a certain degree of self-containedness, we will summarize or quote the relevant parts, whenever necessary. To simplify explanation, we group URIs into three main parts: the scheme-part, the scheme-dependent part (authority, path, query) and the fragment identifier part.

RFC 3986 is focusing on the generic syntax of URIs to provide a way to interpret and validate URIs independent of their scheme. It is thus mainly concerned with the syntax or URIs and gives the following Augmented Backus-Naur Form (ABNF) rules[11]:

```
URI           = scheme ":" hier-part [ "?" query ] [ "#" fragment ]
hier-part     = "//" authority path-abempty /
                path-absolute / path-rootless / path-empty
path-abempty  = *( "/" segment )
path-absolute = "/" [ segment-nz *( "/" segment ) ]
path-rootless = segment-nz *( "/" segment )
path-empty    = 0<pchar>
segment       = *pchar
query         = *( pchar / "/" / "?" )
fragment      = *( pchar / "/" / "?" )
pchar         = unreserved / pct-encoded / sub-delims / ":" / "@"
unreserved    = ALPHA / DIGIT / "-" / "." / "_" / "~"
pct-encoded   = "%" HEXDIG HEXDIG
sub-delims    = "!" / "$" / "&" / "'" / "(" / ")"
                / "*" / "+" / "," / ";" / "="
```

The idea behind NIF is to allow NLP tools to exchange annotations about text in RDF. Hence, the main prerequisite is that text becomes referenceable by URIs, so that they can be used as resources in RDF statements. To achieve this goal, we introduced the basic concepts for URIs and their syntax above. We developed an alternative ABNF for URIs, that allows us to generate them programatically for NIF based on the necessities of the client.

For NIF, we define the ABNF as below with a non-zero identifier and the exception that "path-empty" is not allowed. These ABNF rules imply that (1) they are a subset of the original ABNF ruling out some cases and (2) the main purpose of rewriting the ABNF is not functional, but rather used to have a simpler terminology (i.e. the split into prefix and identifier) for further explanations.

11 https://tools.ietf.org/html/rfc5234

```
URI        = prefix identifier
prefix     = scheme ":" hier-part [ "?" query ] [ "#" fragment ]
identifier = 1*pchar
```

In case the identifier is "char=0,7", the following examples are all valid NIF URIs:

```
http://example.com/char=0,7
http://example.com/ex/char=0,7
http://example.com/ex?char=0,7
http://example.com/ex#char=0,7
ftp://example.com/ex?p1=example&p2=example&char=0,7
ftp://example.com/ex?p1=example&p2=example#char=0,7
```

The reason for this different view is because NIF implementations gain two important features: On the one hand, NIF Web Services (NIF-WS) provide a convention to dynamically publish Linked Data for small pieces of text submitted via the GET parameter. On the other hand, NIF implementations can leave the choice of minting URIs to the client in a transparent way, thus facilitating a programatic approach for generating RDF from text.

We will give three examples here for easier understanding. Parameters and behaviour is explained in the respective online specification[12].

Example 1: NIF-WS publishes Linked Data:

```
curl -H "Accept: text/turtle" "http://nlp2rdf.lod2.eu/nif-ws.php?
  informat=text&input=My+favourite+actress+is+Natalie+Portman.#
      char=0,40"

<http://nlp2rdf.lod2.eu/nif-ws.php?
      informat=text&input=My+favourite+actress+is+Natalie+Portman
          .#char=0,40>
  rdf:type nif:RFC5147String , nif:Context ;
  nif:beginIndex "0" ;
  nif:endIndex "40" ;
  nif:isString "My favourite actress is Natalie Portman." .
```

Example 2: NIF-WS returns the text.

```
curl -H "Accept: text/plain" "http://nlp2rdf.lod2.eu/nif-ws.php?
  informat=text&input=My+favourite+actress+is+Natalie+Portman.#
      char=0,40"

My favourite actress is Natalie Portman.
```

Example 3: NIF-WS allows client to mint URIs via the prefix parameter

```
curl -H "Accept: text/turtle" "http://nlp2rdf.lod2.eu/nif-ws.php?
```

12 http://persistence.uni-leipzig.org/nlp2rdf/specification/api.html

```
  informat=text&input=My+favourite+actress+is+Natalie+Portman.
  &prefix=http://mydoc.de/doc5/"

<http://mydoc.de/doc5/char=0,40>
  rdf:type nif:RFC5147String , nif:Context ;
  nif:beginIndex "0" ;
  nif:endIndex "40" ;
  nif:isString "My favourite actress is Natalie Portman." .
```

Depending on the scenario, we can choose the prefix in the following manner:

WEB ANNOTATION. If we want to annotate a (web) resource directly, it is possible to use the existing document URL as the basis for the prefix and add a hash ('#'). The recommended prefix for the 26610 characters of `http://www.w3.org/DesignIssues/LinkedData.html` is: `http://www.w3.org/DesignIssues/LinkedData.html#`
This works best for plain text files either on the web or on the local file system (`file://`). For demonstration purposes, we minted a URI that contains a plain text extraction (19764 characters) created with 'lynx –dump', which we will use as the prefix for most of our examples: `http://persistence.uni-leipzig.org/nlp2rdf/examples/doc/LinkedData.txt#` and `http://persistence.uni-leipzig.org/nlp2rdf/examples/doc/LinkedData.txt#char=333,345` NIF can be used as a true stand-off format linking to external text.

WEB SERVICE. If the text is, however, sent around between web services or stored in a triple store, the prefix can be an arbitrarily generated URN[13]. Communication between the NLP tools in NIF is done via RDF and therefore mandates the inclusion of the text in the RDF during the POST or GET request. The main purpose here is to exchange annotations between client and server and the used URIs do not require to resolve to an information resource. NIF requires each web service to have a parameter "prefix" that empowers any client to modify the prefix of the created NIF output. The prefix parameter can be tested at `http://nlp2rdf.lod2.eu/demo.php`.

ANNOTATIONS AS LINKED DATA. For static hosting of annotations as Linked Data (e.g. for a corpus), the `/` and query component separator is advantageous. Often the basic unit of a corpus are the individual sentences and it makes sense to create individual prefixes on a per sentence basis.

13 `http://tools.ietf.org/html/rfc1737`

8.4.1 *Context-Hash-based URIs*

As an alternative to the offset-based scheme, context-hash-based URIs are designed to remain more robust regarding document changes. Context-hash-based URIs are constructed from five parts separated by an underscore '_':

1. a *scheme identifier*, in this case the string 'hash',
2. the *context length* (number of characters to the left and right used in the message for the hash-digest),
3. the *overall length* of the addressed string,
4. the *message digest*, a 32-character hexadecimal MD5 hash created from the string and the context. The message M consists of a certain number C of characters (see 2. context length above) to the left of the string, a bracket '(', the string itself, another bracket ')' and C characters to the right of the string: 'leftContext(String)rightContext'. If there are not enough characters to left or right, C is adjusted and decreased on the corresponding side (see the 'Hurra!' example below).
5. the *string itself*, the first 20 (or less, if the string is shorter) characters of the addressed string, urlencoded.

The additional brackets '(' and ')' around the string were introduced to make the identifier more uniquely distinguishable. If there is a sentence 'Hurra! Hurra!' and the context size is too large, e.g. 10, then the first and the second 'Hurra!' would have the same hash. By adding brackets, however, the hash is easily distinguishable: `md5("(Hurra! Hurra!)")` != `md5("(Hurra!) Hurra!")` != `md5("Hurra! (Hurra!)")`.

Note that context-hash-based URIs are unique identifiers of a specific string only if the context size is chosen sufficiently large. If, for example, a complete sentence is repeated in the document, parts of the preceding and/or subsequent sentences are to be included to make the reference to the string unique. However, in many NLP applications, a unique reference to a specific string is not necessary, but rather, all word forms within the same minimal context (e.g., one preceding and one following word) are required to be analysed in the same way. Then, a context-hash-based URI refers uniquely to words in the same context, not one specific string. Using a small context, one can refer to a whole class of words rather than just an individual one. For example, by using the string ' the ' (with one preceding and following white space as context) we obtain the digest: md5(' (the) '). The resulting URI is `http://www.w3.org/DesignIssues/LinkedData.html#hash_1_5_8dc0d6c8afa469c52ac4981011b3f582_%20the%20` and would denote all occurrences of 'the' in the given reference context, surrounded by a single white space on both sides.

Trivially, every string is uniquely addressable if the context-length is large enough. The algorithm for finding the addressed strings in a given text is simple: 1. URL decode the fifth part (the string itself) and

search for all occurrences and get the start indices. 2. From all found start indices generate the hash by calculating the end index (start index + overall length), adding brackets and including the context (if start index − context length < 0, then left context starts at index 0, right context starts at end index and does not go beyond end of text). The following algorithm computes the minimal context-length (MinCl) on a fixed document with a given set of annotations, so that each URI only denotes one substring.

```
1: procedure MinCl(annotations, cl)
2:     uris ← {}
3:     for all annotations a do
4:         uri ← makeUri(a)
5:         if uris contains uri then
6:             return MinCl (annotations, cl +1 )
7:         else
8:             uris ← uris ∪ uri
9:         end if
10:        return cl
11:    end for
12: end procedure
```

9 EVALUATION AND RELATED WORK

9.1 QUESTIONNAIRE AND DEVELOPERS STUDY FOR NIF 1.0

Hellmann, Lehmann, and Auer (2012); Hellmann, Lehmann, et al. (2013)

Hellmann, Lehmann, et al. (2013)

The NLP2RDF Project[1] provides reference implementations[2] and demo showcases to create a community around NIF and support its adoption.

NLP tools can be integrated using NIF, if an adapter is created, that is able to parse a *NIF Model* into the internal data structure and also to output the NIF as a serialization. The effort for this integration is usually very low; just a parser and a serializer have to be written. An NLP pipeline can then be formed by either passing the NIF RDF Model from tool to tool (sequential execution) or passing the text to each tool and then merge the NIF output to a large model (parallel execution). After the release of NIF version 1.0 in November 2011[3] a total of **32** implementations for different NLP tools and converters were created (8 by the authors, including Wiki-link corpus, 13 by people participating in our survey and 11 more, we have heard of). In 2011, we performed a first round of the NIF developer study by assigning the task of developing NIF 1.0 wrappers for 6 popular NLP tools to 6 postgraduate students at our institute. Wrappers were developed for UIMA, GATE-ANNIE, Mallet, MontyLingua, OpenNLP and DBpedia Spotlight (first six lines of Table 7 on page 107). The remaining entries were created in 2012 and 2013 by adopters of NIF 1.0, some even already implementing a draft version of 2.0. Table 7 on page 107 summarizes the results of our NIF developer study.

The first columns contain the self-assessment of the developers regarding their experience in Semantic Web, NLP, Web Services and application development frameworks on a scale from 1 (no experience) to 5 (very experienced). The middle columns summarize the required development effort in hours including learning the NLP tool, learning NIF and performing the complete wrapper implementation. The development effort in hours (ranging between 3 and 40 hours) as well as the number of code lines (ranging between 110 and 445) suggest, that the implementation of NIF wrappers is easy and fast for an average developer. The next section displays the NIF assessment by the developers regarding their experience during the development with respect to the adequacy of the general NIF framework, the coverage of the provided ontologies and the required extensibility. All devel-

1 http://nlp2rdf.org
2 https://github.com/NLP2RDF
3 http://nlp2rdf.org/nif-1-0/

opers were able to map the internal data structure to the NIF URIs to serialize RDF output (Adequacy). Although NIF did not provide a NLP Domain Ontology for Mallet the developer was able to create a compatible OWL Ontology to represent Topic Models. Both UIMA, GATE and Stanbol are extensible frameworks and NIF was currently not able to provide NLP domain ontologies for all possible domains, but only for the used plugins in this study. After inspecting the software the developers agreed however that NIF is general enough and adequate to provide a generic RDF output based on NIF using literal objects for annotations. In case of the UIMA Clerezza consumer an RDF serializer already exists and we have compared potential output in Section 11.1.

Finally, the last section contains an assessment of the NIF approach by the developers regarding the perceived scalability, interoperability, quality of the documentation, the usefulness of the reference implementation, the learning curve / entrance barrier and the performance overhead on a scale from 1 (low) to 5 (very high). The results[4] suggest, that NIF lives up to its promise of ease-of-use and increased interoperability and is generally perceived positive by developers.

9.2 QUALITATIVE COMPARISON WITH OTHER FRAMEWORKS AND FORMATS

Hellmann, Lehmann, et al. (2013)

In Ide and Suderman (2012), the *Graph Annotation Framework (GrAF)* was used to bridge the models of UIMA and GATE. GrAF is the XML serialization of the *Linguistic Annotation Framework* (LAF) and has recently been standardized by ISO. GrAF is meant to serve as a pivot format for conversion of different annotation formats and is able to allow a structural mapping between annotation structures. GrAF is similar to the Open Annotation effort, offering a very expressive and flexible framework, which is focused on XML (although not exclusively). *Extremely Annotational RDF Markup* (Peroni & Vitali, 2009, EARMARK) is a stand-off format to annotate text with markup (XML, XHTML) and represent the markup in RDF including overlapping annotations. The main method to address content is via ranges that are similar to the NIF URI scheme. *TELIX* (Rubiera, Polo, Berrueta, & Ghali, 2012) extends SKOS-XL[5] and suggests RDFa as annotation format. We were unable to investigate *TELIX* in detail, because neither an implementation nor proper documentation was provided. In Section 10.1, we have argued already that RDFa is not a suitable format for NLP annotations in general. The usage of SKOS-XL by TELIX only covers a very small part of NLP annotations, i.e. lexical entities. With the early *Tipster* and the more modern *UIMA* (Ferrucci & Lally, 2004),

4 more data at http://svn.aksw.org/papers/2013/ISWC_NIF/public/devstudy.pdf

5 http://www.w3.org/TR/skos-reference/skos-xl.html

GATE (Cunningham et al., 2002), *Ellogon*, *Heart-of-Gold* and *OpenNLP*[6] a number of comprehensive NLP frameworks already exist. NIF, however, focuses on interchange, interoperability as well as decentralization and is complementary to existing frameworks. Ultimately, NIF rather aims at establishing an ecosystem of interoperable NLP tools and services (including the ones mentioned above) instead of creating yet another monolithic (Java-)framework. A similar approach to NIF is the Weblicht framework based on the Text Corpus Format (Heid, Schmid, Eckart, & Hinrichs, 2010, TCF). TCF is based on XML and grounds all annotations on the token layer and not on the offsets. Given existing and established standoff XML formats such as GATE XML or GrAF, the question arises whether development of another XML format provides any improvements. By being directly based on RDF, Linked Data and ontologies, NIF also comprises crucial features such as *annotation type inheritance* and *alternative annotations*, which are cumbersome to implement or not available in other NLP frameworks (Schierle, 2011). With its focus on conceptual and access interoperability NIF also facilitates *language resource* and *access structure* interchangeability, which is hard to realize with existing frameworks. NIF does not aim at replacing NLP frameworks, which are tailored for high-performance throughput of terabytes of text; it rather aims to ease access to the growing availability of heterogeneous NLP web services as, for example, already provided by *Zemanta* and *Open Calais*.

9.3 URI STABILITY EVALUATION

Hellmann, Lehmann, and Auer (2012)

As the context-hash-based URI scheme differs significantly in terms of uniqueness and stability from the offset-based scheme, we evaluate both schemes with real revision histories from Wikipedia articles. Although Wikipedia pages are edited quite frequently (≈ 202,000 edits per day[7]), the senses of each page tend to remain relatively stable after a certain number of revisions (Hepp et al., 2007).

We downloaded a Wikipedia dump with the full edit revision history[8]. From this dump, we randomly selected 100 articles which had more than 500 edits total. We retrieved the last 100 revisions of these 100 articles and removed the wiki markup[9]. Then we split the resulting plain text into tokens at word level. We used a deterministic algorithm (mostly based on regular expressions) for the markup removal and the tokenisation to avoid any side effects. The text for each revision contained 57062.4 characters on average, which we split into 7410.7 tokens on average (around 7.7 chars/token). About 47.64 characters were added between each revision. For each token and each

6 http://opennlp.apache.org
7 http://www.wikistatistics.net/wiki/en/edits/365
8 http://dumps.wikimedia.org/enwiki/20111007/
9 Code from http://www.mediawiki.org/wiki/Extension:ActiveAbstract

revision, we generated one URI for the offset scheme and six URIs for the context-based scheme with context length 1, 5, 10, 20, 40 and 80. Cf. Section 9.4 for details why other approaches were not included in this evaluation. For every same URI that was generated within one revision *i* for two different tokens (a violation of the uniqueness property), the uniqueness ratio (*URatio*) decreases: $\frac{|\mathrm{UniqueURIs_i}|}{|\mathrm{Tokens_i}|}$. The stability was calculated by the intersection of UniqueURIs of two revisions (*i* and *i+1*) over the number of tokens of the second revision: $\frac{|\mathrm{UniqueURIs_i} \cap \mathrm{UniqueURIs_{i+1}}|}{|\mathrm{Tokens_i}+1|}$. Thus non-unique URIs were penalized for the calculation of stability (without this penalty the percentage was always about 99%). We did the same measurement between the first and the last revision (columns *1...100* and *Stab 1...100*) of each article. The results are summarized in Table 5.

While a high context length (cl=80) provides more stability between revisions (99.98%), $\mathrm{cl} = 10$ yields 87.12% of the URIs valid over 100 edits. The offset-based URIs have a probability of 54% to become invalid between revisions. This corresponds roughly to the theoretically probability for a random insertion to break a URI: $\frac{a-1}{n+1} + \frac{n-a+2}{2n+2} = \frac{a+n}{2n+2}$ (n = text length, a = annotation length). For context-hash URIs: $\frac{a+2cl-1}{n+1}$.

tok ≈ 7410.7	Unique	URatio	Stability	1...100	Stab 1...100
context 1	2830.2	0.3988	0.3946	2647.3	0.3680
context 5	7060.0	0.9548	0.9454	6417.7	0.8551
context 10	7311.4	0.9871	0.9771	6548.8	**0.8712**
context 20	7380.6	0.9963	0.9854	6429.1	0.8553
context 40	7402.2	0.9990	**0.9866**	6146.8	0.8183
context 80	7408.8	0.9998	0.9847	5678.6	0.7568
offset	7410.7	**1.00**	0.5425	104.4	0.0164

Table 5: Evaluation of URI stability with different context length versus the offset scheme. The second last column measures how many annotations remain valid over 100 edits on Wikipedia.

9.4 RELATED URI SCHEMES

Hellmann, Lehmann, and Auer (2012)

As the suitability of the string identifiers highly depends on the specific task, we present in the following a list of criteria, which allow to evaluate and design suitable identifiers:

Uniqueness. The URIs must uniquely identify the substring. **Validity.** The URI scheme must produce valid URIs for arbitrary substrings. Valid URIs must not contain invalid characters and must be limited in length, since most browsers limit the size of the URIs, they can handle[10]. **XML Compatibility.** The identifier part for the generated

10 MS Internet Explorer has a maximum URL length of 2,083 characters. http://support.microsoft.com/kb/q208427/

	Uniq	Val	XML	Stab	Addr	Self	Impl	Exp	Example
Context-Hash(NIF)	+	+	+	+	+	+	o	o	#hash_10_12_6ofo...
Offset(NIF)	++	++	+	- - -	++	+	++	o	#offset_717_729
Offset plain	++	++	-	- - -	++	-	++	o	#717-729
Yee (Context)	+	- -	+	+	- -	- -	- -	o	#:words:The-(Semantic We...
RFC 5147	++	++	+	- - -	++	++	+	+	#char=717,729
LiveURL (Content)	- -	+	-	+	+	-	++	o	#8Semantic12+0x206A73ED
LiveURL (Position)	+	+	-	- -	+	-	-	o	not available for text
Wilde et al. (Regex)	o	++	+	+	+	+	- -	++	#matching=`Semantic\sWeb`

Table 6: Comparison of URI schemes (first two are used in NIF)

URIs should be usable as an XML tag name (for RDF/XML serialisation). For example, XML tag elements can not begin with a number, thus prohibiting tags such as `<717-729>`. **Stability.** The URI should only become invalid if the referenced string is changed significantly, thus rightfully rendering the annotations void. It should not become invalid through unrelated changes. **Addressability.** The URIs can efficiently find the annotated substring within the text, i.e. calculate the start and end index (ideally rule based). **Self-Description.** Some URI schemes require certain parameters to find the appropriate substring in the document. The URIs should contain encoded information that can be used to identify the scheme itself and that can be used to reproduce the *configuration* of the scheme. As correct implementations are necessary to allow the creation of tool chains, it is beneficial, if the scheme has a low complexity to avoid **implementation** errors. **Expressivity.** This criteria measure how expressive the function is that references the strings (e.g. regex is more expressive than just start/end index).

Table 6 shows a comparison of various URI schemes. **LiveURLs** (Kannan & Hussain, 2006)[11] is realized as a Firefox plugin and offers two different ways to produce string identifiers: a context-based and a position based. The user can select a text in the browser and then the plugin creates the URL pointing to the corresponding fragment. This URL can be shared and the referenced string is highlighted. As the identifier starts with a number, it can create a conflict with XML serialisation. Furthermore, the identifier does not contain enough information to uniquely distinguish duplicates, i.e. it would match several occurrences. The position based method uses a combination of the parent node's id and index in the DOM tree alongside an identifier for the child position. The position based method is content-specific and works only on XHTML. Analogous to all position based methods, the scheme is highly susceptible to change. Wilde and Duerst (2008) filed an RFC in April 2008[12] proposing a parameter-like syntax using fragments that refer to statistics about the characters in the

11 http://liveurls.mozdev.org
12 http://tools.ietf.org/html/rfc5147

string (e.g. offsets, line, length), e.g. `ftp://example.com/text.txt#line=10,20;length=9876,UTF-8`. The basic properties of this scheme are a super set to the offset-based NIF scheme and the `owl:sameAs` relation holds: `:offset_717_729 owl:sameAs :char=717,729 .` The *line* parameter will be considered for further benchmarks, but lacks the necessary granularity. The spec of the RFC restricts this scheme to the "plain text" media type, which excludes XML and HTML. Furthermore the scheme contains many optional parameters for integrity checking. When used as RDF subjects, it is tedious to resolve such optional parts, as `#line=10,20` is neither syntactically the same URI as `#line=10,20;length=9876`, nor can we automatically infer an `owl:sameAs` relation. Yee (1998) proposed *Text-Search Fragment Identifiers*, which pinpoint the wanted substring with a fragment that includes the string and its context. Before the creation of the fragment identifier, however, the original HTML source is manipulated and all HTML tags are removed and special characters are normalized. The resulting URL for our example is: `#:words:The-(Semantic Web)-isnt-just-about-putting`. The problem is that the proposed normalization (i.e. remove HTML and tokenise context) can not be standardized easily as it relies on difficult to normalize NLP methods. Therefore, there is no guarantee to reproduce the manipulation bi-directionally (e.g. to find the annotated substring). Longer selected substrings lead to longer, invalid URIs. Wilde and Baschnagel (2005) propose to use regular expression patterns following the parameter "matching" as fragment identifiers, i.e. `matching=Semantic\sWeb` would match all nine occurrences of "Semantic Web" at once. Although being powerful, it is not straight-forward to implement an algorithm that produces regular expressions addressing the correct strings in a text and thus results in high implementation complexity and unpredictability regarding uniqueness. Considering the possibility to include the context in an URI, this scheme is a superset of the previous approach by Yee.

Tool	Developer	Type	SW	NLP	Web Services	Frameworks	Effort (h)	Tool	NIF	Implementation	LoC	Lang	Adequacy	Coverage	NIF Extension	Scalability	Interoperability	Documentation	Reference Impl.	Entrance barr.	Perf. overhead
UIMA	MB	w	3	2	3	4	35	20	5	10	271	Java	✓	no (✓ for POS)	n.a.	2	4	4	5	3	2
GATE	DC	w	4	1	4	4	20	3	5	14	445	Java	✓	no (✓ for POS)	n.a.	4	5	4	5	3	2
Mallet	MA	w	1	4	2	3	40	4	8	28	400	Java	✓	no (NIF 1.0)	✓	3	4	3	5	4	3
MontyLingua	MN	w	4	1	4	2	25	4	3	18	252	Python	✓	✓	n.a.	4	4	5	-	3	3
Spotlight	RS	w	3	3	5	1	20	4	4	12	110	Node-JS	✓	no (NIF 1.0)	✓	4	5	4	5	4	3
OpenNLP	MB	w	3	2	3	4	3/8	1	0*	2	267	Java	✓	no (NIF 1.0)	✓	2	4	4	5	3	2
OpenCalais	AL	w	4	4	3	4	32	6	6	20	201	PHP	✓	no (NIF 1.0)	✓	3	3	4	5	4	3
Zemanta	MV	w	3	4	4	4	24	1	3	20	235	Python	✓	✓	n.a.	3	4	3	5	4	3
SemanticQuran	MS	w	4	3	2	2	25	1	6	18	500	Java	✓	✓	n.a.	5	5	4	5	4	2
ITS2NIF	FS	w	3	3	3	3	20	7	7	6	72	XSLT	✓	✓	n.a.	3	3	3	3	1	3
THD	MD	w	4	2	5	3	20	7	8	5	300	Java	✓	no	✓	3	4	2	2	3	3
STANBOL	RW	w/i	5	4	4	4	28	?	8	20	400	Java	✓	no	~	?	?	3	?	2	2
Spotlight	MN	i	4	2	4	3	24	8	1	15	212	Scala	✓	✓	n.a.	4	4	3	4	3	2
Coat	SL	i	2	1	2	4	165	10	5	150	-	Java	✓	✓	n.a.	3	-	3	-	3	-
DrugExractor	AK	w	4	5	4	4	16	1	5	10	30	Java	~	no	✓	3	3	4	-	1	-

Table 7: Results of the NIF developer case study.

Part IV

THE NLP INTERCHANGE FORMAT IN USE

10 USE CASES AND APPLICATIONS FOR NIF

10.1 INTERNATIONALIZATION TAG SET 2.0

Heim et al. (2009); Hellmann, Lehmann, and Auer (2012); Hellmann, Lehmann, et al. (2013); Hellmann et al. (2010); Klebeck et al. (2011); Rizzo et al. (2012), ITS 2.0 W3C standard - http://www.w3.org/TR/its20/

The *Internationalization Tag Set* (ITS) Version 2.0 is a W3C working draft, which is in the final phase of becoming a W3C recommendation. Among other things, ITS standardizes HTML and XML attributes which can be leveraged by the localization industry (especially language service providers) to annotate HTML and XML nodes with processing information for their data value chain. In the standard, ITS defines 19 *data categories*[1], which provide a shared conceptualization by the W3C working group and its community of stakeholders. An example of three attributes in an HTML document is given here:

```
<html><body><h2 translate="yes">Welcome to <span
   its-ta-ident-ref="http://dbpedia.org/resource/Dublin" its-
       within-text="yes"
   translate="no">Dublin</span> in
   <b translate="no" its-within-text="yes">Ireland</b>!</h2></
       body></html>
```

As an outreach activity, the working group evaluated *RDFa*[2] to create a bridge to the RDF world, but concluded that the format was not suitable to serve as a best practice for RDF conversion. The main problem was that the defined ITS attributes annotate the text within the HTML nodes, but RDFa only has the capability to annotate resources with the text in the node as an object. RDFa lacks subject URIs, which refer to the text within the tags. Although it is theoretically possible to extract provenance information (i.e. offsets and position in the text), the RDFa standard does not include this use case and current RDFa parsers (with the exception of *viejs.org*) do not implement such an extraction.

In a joint effort, the ITS 2.0 RDF ontology[3] was developed using NIF, which was included within the proposed standard alongside an algorithm for a round-trip conversion of ITS attributes to NIF[4] (simple granularity). Provenance can be kept with an XPointer/XPath fragment identifier.

```
@base <http://example.com/exampledoc.html#> .
<char=0,> a nif:Context , nif:RFC5147String ;
```

1 http://www.w3.org/TR/its20/#datacategory-description
2 http://www.w3.org/TR/rdfa-syntax/
3 http://www.w3.org/2005/11/its/rdf#
4 ttp://www.w3.org/TR/its20/#conversion-to-nif

```
<char=11,17>
    nif:anchorOf      "Dublin" ;
    itsrdf:translate     "no";
    itsrdf:taIdentRef    dbpedia:Dublin ;
    # needed provenance for round-tripping
    prov:wasDerivedFrom  <xpath(/html/body[1]/h2[1]/span[1]/text
        ()[1])> ;
    nif:referenceContext <char=0,> .
```

NIF successfully creates a bridge between ITS and RDF and a round-trip conversion was recently implemented as a proof-of-concept. Therefore, NIF can be expected to receive a wide adoption by machine translation and industrial language service providers. Additionally, the ITS Ontology provides well modeled and accepted properties, which can in turn be used to provide best practices for NLP annotations.

10.1.1 *ITS2NIF and NIF2ITS conversion*

The ITS 2.0 standard describes technology that aims to enhance "the foundation to integrate automated processing of human language into core Web technologies. ITS 2.0 bears many commonalities with is predecessor, ITS 1.0 but provides additional concepts that are designed to foster the automated creation and processing of multilingual Web content. ITS 2.0 focuses on HTML, XML-based formats in general, and can leverage processing based on the XML Localization Interchange File Format (XLIFF), as well as the Natural Language Processing Interchange Format (NIF)."[5]

10.1.1.1 *Conversion to NIF*

ITS 2.0 W3C standard - Section F
http://www.w3.org/TR/its20/#conversion-to-nif

This section provides an informative algorithm to convert XML or HTML documents (or their DOM representations) that contain ITS metadata to the RDF format based on NIF. The conversion results in RDF triples.

The algorithm creates URIs that in the query part contain the characters "[" and "]", as part of XPath expressions. In the conversion output (see an example[6]), The URIs are escaped as "%5B" and "%5D". For readability the URIs shown in this section do not escape these characters.

The algorithm is intended to extract the text from the XML/HTML/DOM for an NLP tool. It can produce a lot of "phantom" predicates from excessive whitespace, which 1) increases the size of the intermediate mapping and 2) extracts this whitespace as text, and therefore might decrease NLP performance. It is strongly recommended to normalize whitespace in the input XML/HTML/DOM in order to mini-

5 cited from http://www.w3.org/TR/its20/#abstract
6 http://www.w3.org/TR/its20/examples/nif/EX-nif-conversion-output.ttl

mize such phantom predicates. A normalized example is given below. The whitespace normalization algorithm itself is format dependent (for example, it differs for HTML compared to general XML).

The output of the algorithm shown below uses the ITS RDF ontology and its namespace[7].
Like the algorithm, this ontology is not a normative part of the ITS 2.0 specification and is being discussed in the ITS Interest Group[8].

The following example is an HTML document[9] with whitespace character normalization as preparation for the conversion to NIF. Note that text nodes in the `head` element are not taken into account.

```
<!DOCTYPE html><html xmlns="http://www.w3.org/1999/xhtml">
<head><meta http-equiv="Content-Type" content="text/html;charset=
    utf-8" >
<title>NIF conversion example</title></head>
<body><h2 translate="yes">Welcome to <span
    its-ta-ident-ref="http://dbpedia.org/resource/Dublin" its-
        within-text="yes"
    translate="no">Dublin</span> in <b translate="no" its-within-
        text="yes">Ireland</b>!</h2></body></html>
```

The conversion algorithm to generate NIF consists of seven steps:

- STEP 1: Get an ordered list of all text nodes of the document.
- STEP 2: Generate an XPath expression for each non-empty text node of all leaf elements and memorize them.
- STEP 3: Get the text for each text node and make a tuple with the corresponding XPath expression (X,T). Since the text nodes have a certain order we now have a list of ordered tuples ((x0,t0), (x1,t1), . . . , (xn,tn)).
- STEP 4 (optional): Serialize as XML or as RDF. The list with the XPath-to-text mapping can also be kept in memory. Part of a serialization example is given below. The upper part is in RDF Turtle Syntax while the lower part is in XML (the `mappings` element).

```
# Turtle example:
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=b0,e0>
 nif:wasConvertedFrom  <http://example.com/myitsservice?informat=
     html&intype=url&input=http://example.com/doc.html&xpath=x0>
      .
```

7 http://www.w3.org/2005/11/its/rdf-content/its-rdf.html
8 http://www.w3.org/International/its/wiki/ITS-RDF_mapping
9 examples/html5/EX-HTML-whitespace-normalization.html

```
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=b1,e1>
 nif:wasConvertedFrom <http://example.com/myitsservice?informat=
     html&intype=url&input=http://example.com/doc.html&xpath=x1>
      .
# ...
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=bn,en>
 nif:wasConvertedFrom <http://example.com/myitsservice?informat=
     html&intype=url&input=http://example.com/doc.html&xpath=xn>
     .
<!-- XML Example -->
<mappings>
 <mapping x="xpath(x0)" b="b0" e="e0" />
 <mapping x="xpath(x1)" b="b1" e="e1" />
 <!-- ... -->
 <mapping x="xpath(xn)" b="bn" e="en" />
</mappings>
```

where

```
b0 = 0
e0 = b0 + (Number of characters of t0)
b1 = e0
e1 = b1 + (Number of characters of t1)
...
bn = e(n-1)
en = bn + (Number of characters of tn)
```

Example (continued)

```
# Turtle example:
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#> .
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
# "Welcome to "
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=0,11>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/text()
     [1]>.
# "Dublin"
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=11,17>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/span
     [1]/text()[1]>.
# " in "
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=17,21>
 nif:wasConvertedFrom
```

```
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/text()
     [2]> .
# "Ireland"
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/b[1]/
     text()[1]> .
# "!"
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=28,29>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/text()
     [3]> .
# "Welcome to Dublin Ireland!"
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=0,29>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/text()
     > .
<!-- XML Example -->
<mappings>
 <mapping x="xpath(/html/body[1]/h2[1]/text()[1])" b="0" e="11"
     />
 <mapping x="xpath(/html/body[1]/h2[1]/span[1]/text()[1])" b="11"
      e="17" />
 <mapping x="xpath(/html/body[1]/h2[1]/text()[2])" b="17" e="21"
     />
 <mapping x="xpath(/html/body[1]/h2[1]/b[1]/text()[1])" b="21" e=
     "28" />
 <mapping x="xpath(/html/body[1]/h2[1]/text()[3])" b="28" e="29"
     />
 <mapping x="xpath(/html/body[1]/h2[1])" b="0" e="29" />
</mappings>
```

- STEP 5: Create a context URI and attach the whole concatenated text `$(t0+t1+t2+...+tn)` of the document as reference.
- STEP 6: Attach any ITS metadata annotations from the XML/HTML/DOM input to the respective NIF URIs.
- STEP 7: Omit all URIs that do not carry annotations (to avoid bloating the data).

```
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/
    ontologies/nif-core#>
```

```
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=0,29>
    rdf:type             nif:Context ;
    rdf:type             nif:RFC5147String ;
# concatenate the whole text
    nif:isString         "$(t0+t1+t2+...+tn)" ;
    nif:beginIndex   "0" ;
    nif:endIndex     "29" ;
    itsrdf:translate     "yes";
    nif:sourceUrl      <http://example.com/doc.html> .
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=11,17>
    rdf:type              nif:RFC5147String ;
    nif:beginIndex   "11" ;
    nif:endIndex     "17" ;
    itsrdf:translate     "no";
    itsrdf:taIdentRef  <http://dbpedia.org/resource/Dublin> ;
    nif:referenceContext <http://example.com/myitsservice?
        informat=html&intype=url&input=http://example.com/doc.
        html&char=0,29> .
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
    rdf:type              nif:RFC5147String ;
    nif:beginIndex   "21" ;
    nif:endIndex     "28" ;
    itsrdf:translate     "no";
    nif:referenceContext <http://example.com/myitsservice?
        informat=html&intype=url&input=http://example.com/doc.
        html&char=0,29> .
```

A complete sample output in RDF/XML format after step 7, given the input document (HTML example above), is available at http://www.w3.org/TR/its20/examples/nif/EX-nif-conversion-output.ttl.

Note: The conversion to NIF is a possible basis for a natural language processing (NLP) application that creates, for example, named entity annotations. A non-normative algorithm to integrate these annotations into the original input document is given in section 10.1.1.2. Many decisions to be made in this algorithm depend on the particular NLP application being used.

Note: NIF allows an URL for a String resource to be referenced as URIs that are fragments of the original document in the form:
http://example.com/myitsservice?informat=html&intype=url&input=http://example.com/doc.html&char=0,11
or
http://example.com/myitsservice?informat=html&intype=url&input=http://example.com/doc.html&xpath=/html/body{[}1{]}/h2{[}1{]}/text(){[}1{]}
This offers a convenient mechanism for linking NIF resources in RDF

back to the original document. The NIF Web Service Access Specification[10] defines the parameters for NIF web services.

RDF treats URIs as opaque and does not impose any semantic constraints on the used fragment identifiers, thus enabling their usage in RDF in a consistent manner. However, fragment identifiers get interpreted according to the retrieved mime type, if a retrieval action occurs as is the case in Linked Data. The char fragment is defined currently only for text/plain while the xpath fragment is not defined for HTML. Therefore this URL recipe does fulfil the ITS requirements to support both XML and HTML and the aim of this mapping to produce resources adhering to the Linked Data principle of dereferenceablility. The future definition and registration of these fragment types, while a potentially attractive feature, is beyond the scope of this specification.

10.1.1.2 *Conversion NIF2ITS*

ITS 2.0 W3C standard - Section G - http://www.w3.org/TR/its20/#nif-backconversion

The following algorithm relies on the HTML example given in Section 10.1.1.1. It is assumed that the example has been converted to NIF, leading to the output[11] exemplified for the ITS2NIF conversion algorithm.

This example uses DBpedia Spotlight[12] as an example natural language processing (NLP) tool. In it, DBpedia Spotlight linked "Ireland" to DBpedia:

```
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
 rdf:type nif:RFC5147String;
 itsrdf:taIdentRef <http://dbpedia.org/resource/Ireland> .
<http://dbpedia.org/resource/Ireland>
 rdf:type <http:/nerd.eurecom.fr/ontology#Country> .
```

The conversion algorithm to generate ITS out of NIF consists of two steps:

- STEP 1: NIF Web services accept two different types of input. It is possible to either send the extracted text (the object of the `nif:isString` property) directly or NIF RDF to the NLP tool, i.e. the text is sent as a `nif:Context` node and included as `nif:isString`. Either way, the output of the Web service will be a NIF representation.

  Accepting text will be the minimal requirement of a NIF web service. Ideally, you would be able to send the `nif:Context` node with the isString as RDF directly, which has the advantage, that all other annotations can be used by the NLP tool:

10 http://persistence.uni-leipzig.org/nlp2rdf/specification/api.html
11 http://www.w3.org/TR/its20/examples/nif/EX-nif-conversion-output.ttl
12 https://github.com/dbpedia-spotlight/dbpedia-spotlight

```
<http://example.com/myitsservice?informat=html&intype=url&
    input=http://example.com/doc.html&char=0,29>
 rdf:type nif:RFC5147String ;
 rdf:type nif:Context ;
 nif:beginIndex "0" ;
 nif:endIndex "29" ;
 nif:isString  "Welcome to Dublin in Ireland!" .
```

- STEP 2: Use the mapping from ITS2NIF (available after step 7 of the ITS2NIF algorithm, section 10.1.1.1) to reintegrate annotations in the original ITS annotated document.

For step 2, three cases can occur.

CASE 1: The NLP annotation created in NIF matches the text node. Solution: Attach the annotation to the parent element of the text node.

```
# Based on:
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/b[1]/
     text()[1]> .
# and:
<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
 itsrdf:taIdentRef  <http://dbpedia.org/resource/Ireland> .
# we can attach the metadata to the parent node:
<b its-ta-ident-ref="http://dbpedia.org/resource/Ireland"
   translate="no">Ireland</b>
```

CASE 2: The NLP annotation created in NIF is a substring of the text node. Solution: Create a new element, e.g., for HTML "span". A different input example is given below as case 2 is not covered in the original example input.

```
# Input:

<html>
  <body>
     <h2>Welcome to Dublin in Ireland!</h2>
  </body>
</html>

# ITS2NIF

<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=0,29>
 nif:wasConvertedFrom
 <http://example.com/myitsservice?informat=html&intype=url&input=
     http://example.com/doc.html&xpath=/html/body[1]/h2[1]/text()
     [1]> .
```

```
# DBpedia Spotlight returns:

<http://example.com/myitsservice?informat=html&intype=url&input=
    http://example.com/doc.html&char=21,28>
    itsrdf:taIdentRef  <http://dbpedia.org/resource/Ireland> .

# NIF2ITS

<html>
  <body>
     <h2>Welcome to Dublin in <span
          its-ta-ident-ref="http://dbpedia.org/resource/Ireland">
              Ireland</span>!</h2>
  </body>
</html>
```

Case 3: The NLP annotation created in NIF starts in one region and ends in another. Solution: No straight mapping is possible; a mapping can be created if both regions have the same parent.

10.2 OLIA

Hellmann, Lehmann, et al. (2013)

The *Ontologies of Linguistic Annotation* (OLiA) Chiarcos (2012b)[13] provide stable identifiers for morpho-syntactical annotation tag sets, so that NLP applications can use these identifiers as an interface for interoperability. OLiA provides *Annotation Models (AMs)* for fine-grained identifiers of NLP tag sets, such as *Penn*[14]. The individuals of these annotation models are then linked via `rdf:type` to coarse-grained classes from a *Reference Model (RM)*, which provides the interface for applications. The coverage is immense: OLiA comprises over 110 OWL ontologies for over 34 tag sets in 69 different languages, the latest addition being the Korean *Sejong tagset*. The benefit for application developers is three-fold:

1. **Documentation.** OLiA allows tagging with URIs (e.g. `http://purl.org/olia/penn.owl#DT`) instead of just short cryptic strings such as "DT". Developers who are unfamiliar can open the URL in an ontology browser and read the included documentation collected from the literature.
2. **Flexible Granularity.** For a wide range of NLP tools who built upon POS tags, very coarse-grained tags are sufficient. For example for keyword extraction, entity recognition and lemmatization, it is often not necessary to distinguish between singular/plural or common/proper noun. OLiA maps all four tags to a common class `olia:Noun`. Such a mapping exists for almost

13 http://purl.org/olia
14 http://purl.org/olia/penn.owl

all tags and can be easily reused by developers for a wide range of tag sets.

3. **Language Independence.** AMs for different languages are mapped to the common RM providing an abstraction across languages.

NIF provides two properties: `nif:oliaLink` links a `nif:String` to an OLiA-AM. Although a reasoner could automatically deduce the abstract type of each OLiA individual from the RM, it was a requirement that the coarse-grained types should be linked redundantly to the strings as well in case reasoning services are not available or would cause high overhead. Therefore, an OWL annotation property `nif:oliaCategory` was created as illustrated in the following example.

```
<char=342,345> a nif:String, nif:RFC5147String ;
    nif:oliaLink        penn:NNP  ;
    nif:oliaCategory    olia:Noun , olia:ProperNoun .
# deducable by a reasoner:
penn:NNP       a olia:Noun,  olia:ProperNoun .
```

The NLP2RDF project provides conversions of the OLiA OWL files to CSV and Java HashMaps for easier consumption.[15] Consequently, queries such as 'Return all strings that are annotated (i.e. typed) as `olia:PersonalPronoun` are possible, regardless of the underlying language or tag set.

All the ontologies are available under an open license.[16]

10.3 RDFACE

Hellmann, Lehmann, et al. (2013)

RDFaCE (RDFa Content Editor)[17] (Khalili et al., 2012) is a rich text editor that supports WYSIWYM (What-You-See-Is-What-You-Mean) authoring including various views of the semantically enriched textual content. One of the main features of RDFaCE is combining the results of different NLP APIs for automatic content annotation. The main challenge here is the heterogeneity of the existing NLP APIs in terms of API access, URI generation and output data structure. Different NLP APIs use different URL parameter identifiers such as "content", "text", "lookupText" etc. to indicate the input for the REST API. Furthermore, for identifying the discovered entities they use either their own URI schemes such as:

`http://d.opencalais.com/genericHasher-1/e7385008-0856-3afc-a40f-0000dcd27ded`

`http://api.evri.com/v1/organization/university-of-leipzig-0xbdb4d`

or external URIs such as:

`http://dbpedia.org/resource/University_of_Leipzig`

`http://mpii.de/yago/resource/University_of_Leipzig`

15 `http://olia.nlp2rdf.org/owl/{Penn.java|penn.owl.csv|penn-link.rdf.csv}`

16 http://sourceforge.net/projects/olia/

17 http://aksw.org/Projects/RDFaCE

Another important issue is that each API returns different properties with different identifiers and in a different structure.

To cope with these heterogeneity issues, RDFaCE uses a server-side proxy. At first, the proxy handled the access heterogeneity by hard coding the input parameters and connection requirements of each individual API. After implementing NIF, the integration process was simplified to a great extent by abstracting the diversity of different NLP APIs and introducing an interoperability layer. Adding new NLP APIs to RDFaCE became straightforward and additional efforts to handle heterogeneity between different data formats were removed.

10.4 TIGER CORPUS NAVIGATOR

Hellmann et al. (2010)

A large number of annotated corpora have become available over the past years. Still, the retrieval of dedicated linguistic knowledge for given applications or research questions out of these corpora remains a tedious process. An expert in linguistics might have a very precise idea of the concepts she would like to retrieve from a corpus. Yet, she faces a number of challenges when trying to retrieve corresponding examples out of a particular corpus:

ACCESS she needs a tool that is able to process the format of the corpus, that is easy to deploy, and that provides an intuitive user interface

DOCUMENTATION she needs to be familiar with the annotations and the query language

REPRESENTATION she needs a representation of the results so that these can be studied more closely or that they can be processed further with other NLP tools.

In this section, we describe a novel approach to this problem that starts from the premise that linguistic annotations can be represented by means of existing standards developed in the Semantic Web community: RDF and OWL[18] are well-suited for data integration, and they allow to represent different corpora and tagsets in a uniform way.

We present the Tiger Corpus Navigator, an Active Machine Learning tool that allows a user to extract formal definitions of extensionally defined concepts and the corresponding examples out of annotated corpora. Based on an initial seed of examples provided by the user, the Navigator learns a formal OWL Class Definition of the concept that the user is interested in. This definition is converted into

18 http://www.w3.org/TR/rdf-concepts, http://www.w3.org/TR/owl-ref

a SPARQL query[19] and passed to Virtuoso,[20] a triple store database with reasoning capabilities. The results are gathered and presented to the user to choose more examples, to refine the query, and to improve the formal definition. The data basis for the Navigator is an OWL/RDF representation of the Tiger corpus[21] and a set of ontologies that represent its linguistic annotations.

Our tool, available under http://tigernavigator.nlp2rdf.org, addresses and circumvents the barriers to the acquisition of knowledge out of corpora presented above:

(i) it does not need any deployment and provides a user interface in a familiar surrounding, the browser,

(ii) the concept descriptions acquired during the classifier refinement represent the (conceptual representation of the) annotations in the corpus in an explicit and readable way, and finally,

(iii) the Navigator uses OWL; the query results are thus represented in a readable, portable and sustainable way.

10.4.1 *Tools and Resources*

Several categories of tools and resources need to be integrated to enable the implementation of the goals presented above: We employ the **DL-Learner** (Lehmann, 2009) to learn class definitions for linguistic concepts; **NLP2RDF** (Hellmann, 2010) is applied for the conversion and ontological enrichment of corpus data; and the **OLiA ontologies** (Chiarcos, 2012b) provide linguistic knowledge about the annotations in the corpus.

10.4.1.1 *DL-Learner*

The DL-Learner extends Inductive Logic Programming to Descriptions Logics, OWL and the Semantic Web; it provides a OWL/DL-based machine learning tool to solve supervised learning tasks and support knowledge engineers in constructing knowledge. The induced classes are short and readable and can be stored in OWL and reused for classification. OWL/DL is based on Description Logics that can essentially be understood as fragments of first-order predicate logic with less expressive power, but usually decidable inference problems and a user-friendly variable free syntax. OWL Class definitions form a subsumption hierarchy that is traversed by DL-Learner starting from the top element (*owl:Thing*) with the help of a refinement operator and an algorithm that searches in the space of generated classes.

19 http://www.w3.org/TR/rdf-sparql-query
20 http://virtuoso.openlinksw.com
21 http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus

An example of such a refinement chain is (in Manchester OWL Syntax):

(Sentence) ⇝
(Sentence and hasToken some Thing) ⇝
(Sentence and hasToken some VVPP) ⇝
(Sentence and hasToken some VVPP and hasToken some (stts:AuxiliaryVerb and hasLemma value "werden"))

The last class can easily be paraphrased into: A sentence that has (at least) one Token, which is a past participle (*VVPP*), and another Token, which is an AuxiliaryVerb with the lemma *werden* (passive auxiliary, lit. 'to become'). Detailed information can be found in Lehmann (2009) and under `http://dl-learner.org`.

10.4.2 *NLP2RDF in 2010*

In 2010, NLP2RDF[22] was still a framework that integrated multiple NLP tools in order to assess the meaning of the annotated text by means of RDF/OWL descriptions: Natural language (a character sequence) is converted into a more expressive formalism – in this case OWL/DL – that grasps the underlying meaning and serves as input for (high-level) algorithms and applications.

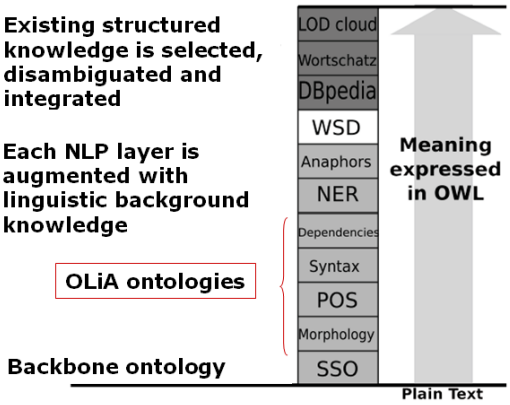


Figure 21: NLP2RDF stack

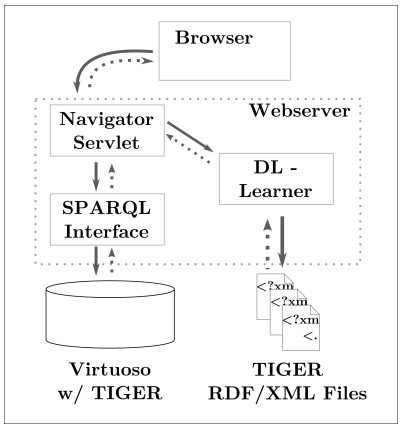


Figure 22: Architecture of the Tiger Corpus Navigator

In a first step, sentences are tokenized and aggregated in a *Structured Sentence ontology (SSO)*, which was a preliminary version of the NIF Core Ontology.. The SSO consists of a minimal vocabulary that denotes the basic structure of the sentence such as tokens and relative position of a token in a sentence.

As shown in Figure 21, the SSO serves as the backbone model, which is then augmented additional layers of annotations:

22 `http://nlp2rdf.org`, `http://code.google.com/p/nlp2rdf`

(1) features from NLP tools
in light grey: morphology, parts of speech (POS), syntactic structures and edge labels (syntax, dependencies), named entity recognition (NER), coreference (anaphors)

(2) rich linguistic ontologies for these features (Section 10.4.3)
combined in a *tagset-ontology pair* for every level mentioned in (item 1)

(3) background knowledge from the Web of Data
examples in dark grey: Linking Open Data (LOD) Cloud,[23] DBPedia,[24] and Wortschatz[25]

(4) additional knowledge
knowledge created by the Navigator (Section 10.4.1.1) or derived from the steps described above (e.g., in white: word sense disambiguation, WSD)

10.4.3 *Linguistic Ontologies*

The Ontologies of Linguistic Annotations (Chiarcos, 2012b, OLiA) represent an architecture of modular OWL/DL ontologies that formalize several intermediate steps of the mapping between concrete annotations, a Reference Model and external terminology repositories, such as GOLD[26] or the ISO TC37/SC4 Data Category Registry:[27]

- Multiple Annotation Models formalize annotation schemes and tag sets, e.g., STTS for the part of speech tags of the Tiger corpus.
- The Reference Model provides the integrating terminology for different annotation schemes (OLiA Annotation Models).
- For every Annotation Model, conceptual subsumption relationships between Annotation Model concepts and Reference Model concepts are specified in a Linking Model. Other Linking Models specify relationships between Reference Model concepts and external terminology repositories (Chiarcos, 2010).

For the tiger navigator, we focused on the STTS Annotation Model[28] that covers the morphosyntactic annotations in the Tiger corpus.

The usage of OLiA combined with NLP2RDF offers two major advantages: OLiA provides a growing collection of annotation models

23 http://richard.cyganiak.de/2007/10/lod
24 http://dbpedia.org
25 http://wortschatz.uni-leipzig.de
26 http://linguistics-ontology.org
27 http://www.isocat.org
28 available under http://nachhalt.sfb632.uni-potsdam.de/owl

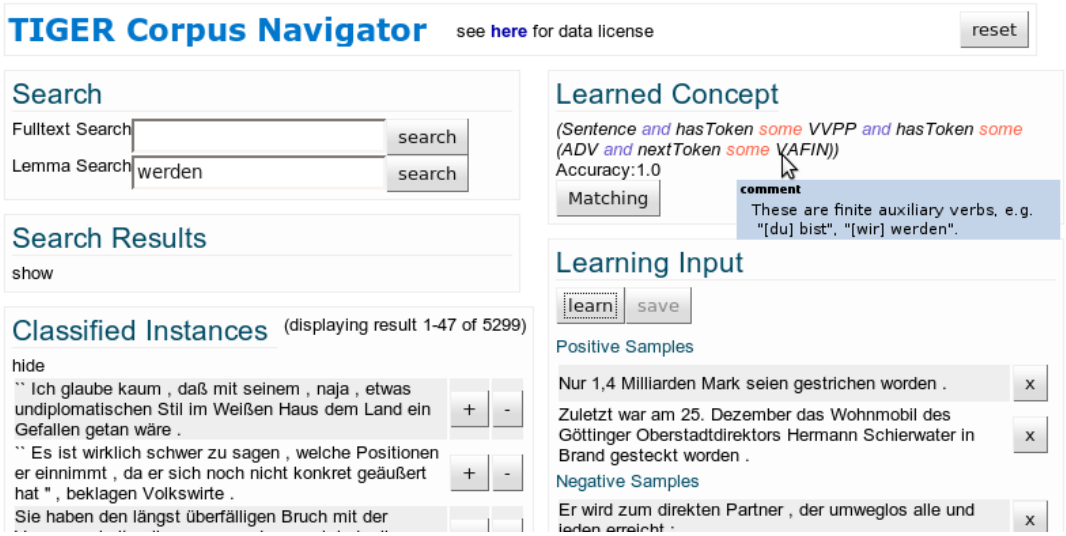


Figure 23: Screenshot of the Tiger Corpus Navigator

for more than 50 languages, that are interlinked with the OLiA Reference Model (and further to community-maintained repositories of linguistic terminology). The adaption of the Navigator to other corpora and other languages is thus easily possible. The interlinking further allows to reuse learned classes on other corpora and even to learn on a combination of different corpora.

10.4.4 *Implementation*

Figure 22 shows the architecture of the Tiger Corpus Navigator: The Virtuoso triple store contains the whole corpus in RDF and allows queries over the complete data for retrieval, the data used by DL-Learner consists of one file for the OWL schema and 50,474 RDF/XML files (one per sentence), which it loads on demand according to the given examples.

With the Navigator user interface (Figure 23), the user starts his research by searching for sentences with certain lemmas or words. The retrieved sentences are presented on the left side. They can be moved to the right panel and classified as positive or negative examples, i.e., as instances or counterinstances of the target concept. Upon pressing the *Learn* button, they are sent to the DL-Learner and the learned OWL Class Definition is displayed (right top). The *Matching* button triggers the retrieval of matching sentences. The user can choose more positive and negative examples from the classified instances and iterate the procedure until the learned definition has an acceptable quality.

To aid the user during this process, the accuracy of the definition on the training data is given below the definition. Additionally, the number of matching sentences is displayed (in this case 5,299, ≈10% of

```
// root node
#root:[cat=/VP|S/] &
// root dominates "werden"
#root >* #werden:[lemma="werden"] &
// root dominates participle
#root >* #partizip:[pos=/V.PP/] &
// finite sentence
(#root >HD #werden |
// infinitive with zu
 (#root >HD #VZ:[cat="VZ"] & #VZ >HD #werden)) &
// root dominates participle directly
(#root >OC #partizip |
// participle is dominated by VP
 (#VP:[cat="VP"] >HD #partizip &
// root dominates VP directly
   (#root >OC #VP |
// or indirectly over a CVP
     (#root >OC #CVP:[cat="CVP"] & #CVP >CJ #VP))))
```

Figure 24: Rule for passive sentences in the Tiger Query Language (König & Lezius, 2003)

the corpus). Hovering over a named class in the concept description presents a tooltip explaining the meaning of the construct as specified by the OLiA Annotation Model. This allows to quickly gain insight into the annotations of the corpus and judge whether the learning result matches the needs of the user.

10.4.5 *Evaluation*

In this section, we evaluate recall and precision of automatically acquired concepts for passive identification in German. We describe two problems (with 4 experiments each), in which we vary several configuration options: training set size (how many examples a user needs to choose), learning time and usage of lemmas.

10.4.5.1 *Experimental Setup*

We consider the German *werden* passive that is formed by the auxiliary *werden* and a past participle Schoenthal (1975).
The task is to distinguish passive clauses from other auxiliary constructions, given only linguistic surface structure (SSO) and morphosyntactic annotations (POS). In the corpus, neither POS nor SSO alone are sufficient to distinguish passive from active clauses, so that information from both sources has to be combined. For our experiment, the DL-Learner was trained on POS and lemmas. Syntax annotation was used only to identify target classifications (with the query in Figure 24).
Three sets of sentences can be distinguished:

1. finite passive (finite auxiliary *werden*, 6,333 sentences, condition `#root >HD #werden`)

2. infinite passive with particle *zu* (lit. 'to') (37 sentences, condition `#root >HD #VZ`)

3. active (44,099 sentences that do not match the query)

From these sets we identified two learning problems to measure how well our approach can separate these sets from each another: (i) learn an OWL class that *covers* all finite passives (set 1) and the remainder (sets 2, 3), and (ii) distinguish between infinite passives (set 2) and the remainder. The second problem is especially difficult, as the number of correct sentences (37) is less than 0.07% of sentences in total.

For each problem, the data is split into training and test data (both positive and negative).[29]

As BASELINE, we randomly drew 5 positive (5p) and 5 negative (5n) sentences from the training data. In the experiments, we performed 4 iterations, starting with 5p+5n initial examples, and adding 5p+5n examples in every iteration. Precision and recall were measured on the intersection of retrieved sentences with the target classification.

We tested three configuration variations for the first problem: (1) we adapted the max. execution time to three times the number of examples (ADAPT, 30s, 60s, 90s, 120s),[30] (2) we reduced the number of initial examples to 2p+2n and added 2p+2n for each iteration (REDUCE, total 4,8,12,16), and (3) we deactivated the inclusion of *owl:hasValue* (lemmas) in the classes (NO_LEMMA).

As for the second problem, (1) we added 10 additional negative examples (ADD_10, total 20, 40, 60, 80), (2) we added 10n but adapted the runtime to 3 times the example size (ADD_10_X3, 60s, 120s, 180s, 240s), and (3) we used again the baseline (BASELINE) with no lemmas (NO_LEMMA).

For the first problem, we conducted a stratified leave-one-out 10-fold cross validation. As it was impossible to create 10 folds for the second set, we used a randomized 70%-30% split averaged over 10 runs (28 sentences for training, 11 for testing).

10.4.5.2 *Results*

Our results (summarized in Figure 25) show that the Tiger Corpus Navigator is capable of acquiring concepts that involve multiple knowledge sources, here, the SSO (lemma) and the OLiA ontologies (for POS) with a high recall and with reasonable speed.

The observed high recall is inherent in the learning algorithm: When exploring the search spaces, it automatically discards all classes that

29 Five overlapping sentences were removed.

30 DL-Learner provides an anytime algorithm called ROLearner2, it stops when finding a class with 100% accuracy or a given maximum execution time (default 30 sec) is reached and returns its (intermediate) results.

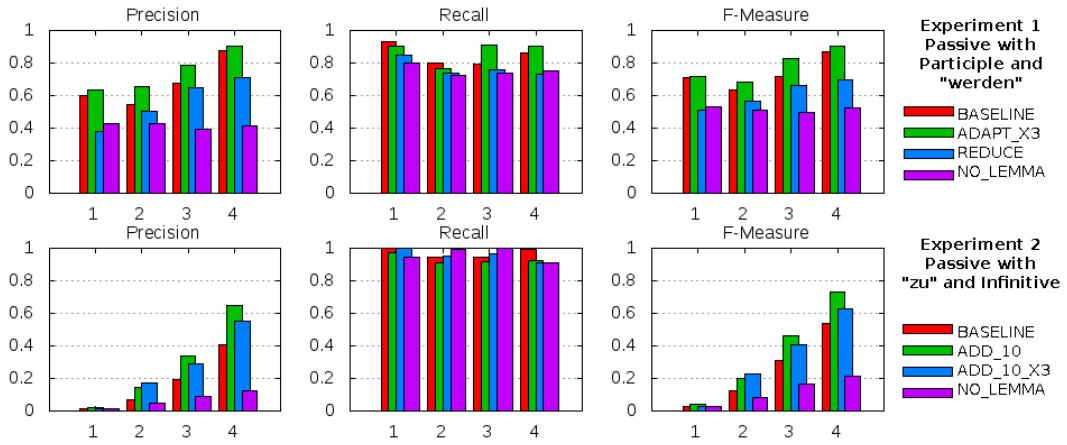


Figure 25: Evaluation results

do not cover all positive examples, so it produces very general results. High precision, however, can only be achieved after a certain number of iterations or by raising the *noise* parameter (zero in our experiments).

We found that our results are clearly dependent on lemmas, *owl:hasValue* inclusion yields better results. The selection of significant lemmas is done generically by DL-Learner according to a value frequency threshold, set equal to the number of positive examples. Users could also wish to manually configure this parameter or give certain lemmas in advance.

The size of the training set had a great influence on the performance with about 20% lower F-Measure in iteration 4 (REDUCE vs. ADD_10 to BASELINE). We observed marginal effects by increasing the maximum learning time with a slight F-Measure gain of 3.5% (ADAPT_X3 vs. BASELINE) and even a loss of more than 10% in the second experiment (ADD_10 vs. ADD_10_X3).

Although the second experiment amounts to a much lower F-Measure scores in iteration 4, the achieved results are interpretable: 40 % precision and 99 % recall mean that the retrieved set of sentences was reduced to about 100 sentences of which 40 would be correct. Such a small sample would be suitable for manual inspection and postprocessing.

Our implementation fulfills the speed requirements for a web scenario: For the first experiment, the average learning times for BASELINE were 1.8 sec, 22.6 sec, 31.9 sec and 29.5 sec, and for the second experiment 0.5 sec, 2.2 sec, 5.3 sec, 13.3 sec. The SPARQL queries needed 14.6 seconds on average and can be further improved by caching. The last example of the refinement chain in Section 10.4.1.1 was one of the highest scoring learned classes.

10.4.6 *Related Work and Outlook*

In the introduction, we identified three elementary functions a corpus tool has to fulfill, i.e., to **access**, to **document** and to **represent** linguistic annotations. We presented the Tiger Corpus Navigator, which provides *access* via a an intuitive user interface over the Web. The paradigm of navigating a corpus based on example sentences rids the necessity of being familiar with the *documentation* beforehand. Even more so, only the necessary information is presented unobstrusively on-the-fly. Learned classes *represent* the results in a formal, yet easily understandable way and the evaluation has shown that it is possible to extract the desired information without much time or effort.

10.4.6.1 *Access to Linguistic Annotations*

Linguistic corpora can be accessed by several corpus tools, e.g., GATE (Cunningham et al., 2002), TGrep2 (Rohde, 2005), TigerSearch (König & Lezius, 2003), the Stockholm TreeAligner (Marek, Lundborg, & Volk, 2008), or MMAX2 (Müller & Strube, 2006), just to name a few. Newer tools also provide web interfaces, such as the IMS Corpus Workbench (Christ, 1994), the Linguist's Search Engine (Resnik, Elkiss, Lau, & Taylor, 2005), or ANNIS (Chiarcos et al., 2008; Zeldes, Ritz, Lüdeling, & Chiarcos, 2009).

All these tools, however, have in common that they operate on a formal, complex query language that represents a considerable hurdle to their application by non-specialists.

The Tiger Corpus Navigator represents an innovative approach to access corpus data that may complement such traditional corpus interfaces. It provides access to the primary data of specific sentences on the basis of extensionally defined conceptual descriptions, it is thus even possible to search for concepts that are not directly annotated (as shown for the passive concept and the Tiger POS annotation).

10.4.6.2 *Document Linguistic Annotations*

In our approach, linguistic annotations are explicitly documented by their linking to repositories of linguistic terminology. These repositories contain descriptions, definitions and examples that are represented to the user as tooltips (Figure 23). In this way, the OWL representation of linguistic corpora and their linking with existing terminology repositories serves a documentation function.

And more than this, the application of the Tiger Corpus Navigator does not require the users to be familiar with the documentation at all: The automatic acquisition of query concepts allows a relatively uninformed user to run queries against a database without the necessity to be aware of the underlying data format, its expressivity

and even the kind of annotations available. Thereby, our approach extends and generalizes approaches to access annotated corpora on the basis of abstract, ontology-based descriptions such as Chiarcos et al. (2008); Rehm, Eckart, and Chiarcos (2007). As opposed to these, however, the concepts are not pre-defined in our scenario, but acquired by the system itself. The Tiger Corpus Navigator thus allows for corpus querying independently from the theoretical assumptions underlying the actual annotations in the corpus.

10.4.6.3 *Represent Linguistic Annotations*

As for exchange and representation formats, the linguistic community still struggles to define its own standards; several concurrent proposals are currently in use, e.g., NITE XML (Carletta et al., 2003), UIMA XML (Goetz & Suhre, 2004), LAF/GrAF (Ide, 2007), or PAULA (Chiarcos et al., 2008). Here, standards from the Semantic Web community are applied, RDF and OWL, that are maintained by a large community and supported by a number of tools. So far, only few NLP tools working with OWL are available, e.g., (Aguado de Cea, Puch, & Ramos, 2008), but a number of linguistic resources has already been transformed to OWL/DL (Burchardt et al., 2008; Scheffczyk et al., 2006), or linked with ontologies (Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2006). Also, existing ontologies have been extended with concepts and properties for linguistic features (Buitelaar et al., 2006; Davis, Handschuh, Troussov, Judge, & Sogrin, 2008). The Navigator represents another step in this development of convergence of ontological and NLP resources.

10.4.6.4 *Application Scenarios*

The Tiger Corpus Navigator may not constitute a full-fledged substitute for existing query tools, as the subsequent refinement of the classifier by the user may turn out to be a time-consuming task. It does, however, represent a prototype implementation of a technology that may be integrated with "traditional" tools to browse, query or access/distribute corpora. If used as a corpus exploration interface of an archive of linguistic resources, for example, the Tiger Corpus Navigator reduces the initial bias to assess the suitability of a corpus with unknown annotations. Such an archive may host different resources that require specialized tools for visualization and querying (e.g., TGrep2 for constituent syntax, MMAX2 for coreference, etc.), so that the efforts required to evaluate the suitability of a resource are enormous (a user has to acquaint itself not only with the annotations and some "standard" query language, but also with several specialized tools and their task-specific query languages). Using the Navigator, a user develops a classifier for a concept of interest, and the correctness of the classifier and the concept description obtained

and the tooltips that contain their documentation allow her to assess the suitability of a corpus and its annotations for the task at hand immediately . If indeed a resource appears to be useful for a particular task, the user may decide to obtain the corpus and to process it further with the appropriate corpus tools.

10.4.6.5 *Future Work*

Future work includes the ability to save learned OWL classes. They can be collaboratively reused and extended by multiple users (Web2.0). Furthermore, they can be utilized to classify previously untagged text, converted by NLP2RDF in the same manner as here and thus extend the discovery of matching sentences beyond the initial corpora. With a corresponding parser-ontology pair it is even possible to replace the initial full text search by entering any example sentences.

It should be noted here that we aimed primarily for a proof-of-concept implementation. The Tiger Corpus Navigator does currently not come with an appropriate visualization, and it is restricted to sentence-level classification. Given sufficient interest from the community, the corresponding extensions may, however, be possible in subsequent research. Another topic for further research may be the combination of existing corpus management and corpus query tools with the Tiger Corpus Navigator, resp. the underlying technologies.

10.5 ONTOSFEEDER – A VERSATILE SEMANTIC CONTEXT PROVIDER FOR WEB CONTENT AUTHORING

Klebeck et al. (2011)

One of the routine tasks of a content author (e.g. a journalist) during the time of writing is researching for context information required for the intended article. Without proper tool support, the author has to resort to manual searching (e.g. Google) and skimming through available information sources. The availability of structured data on the Web of Data allows to automate these routine activities by identifying topics within the article with the aid of Natural Language Processing (NLP) and subsequently presenting relevant context information by retrieving descriptions from the Linked Open Data Web (LOD).

We present the *Ontos Feeder*[31] – a system serving as context information provider, that can be integrated into Content Management Systems in order to support authors by supplying additional information on the fly. Ontos Feeder uses the Ontos Web Service (OWS, see Section 28) to analyse the article text and retrieve *Ontos Entity Identifier* (OEI) URIs for relevant topics. These OEIs are interlinked with several data sources on the web and enriched with internal facts from the *Ontos Knowledge Base*. The Feeder is open-source and currently

31 http://www.ontos.com

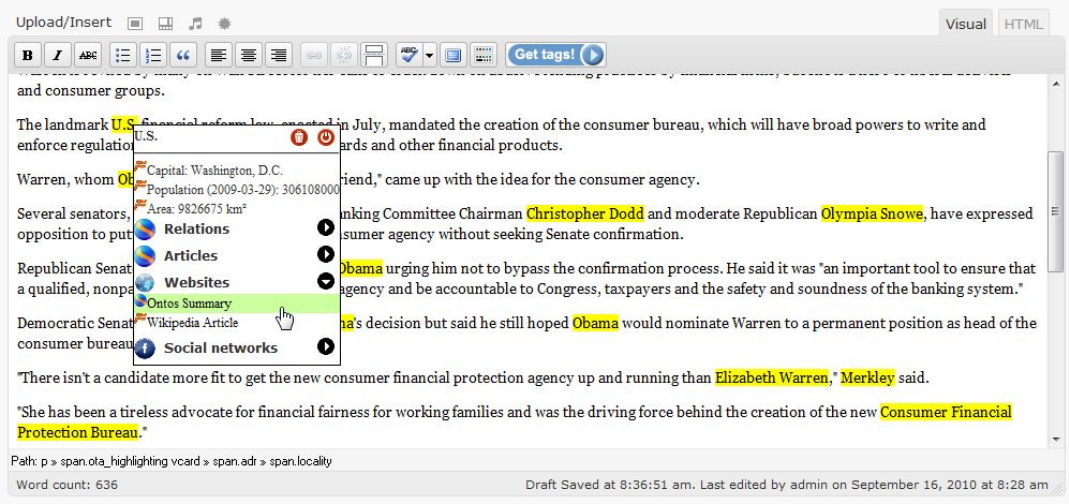


Figure 26: Entities are highlighted in the WYSIWYG editor of the CMS, Pop-ups allow to select further information.

available for the CMS (Drupal[32] and Wordpress[33]. Additionally, the Feeder can automatically annotate the article with Microformats and RDFa annotations. These are increasingly utilized by search engines such as Google or Bing [34].

10.5.1 *Feature Description and User Interface Walkthrough*

The content creation process begins with the writing of an article in a supported CMS system. Having written the content, the author clicks on the `get tags` button to send the text to the OWS. The OWS analyses the text and returns disambiguated URIs for the found entities. Then the Ontos Feeder annotates the returned entities in the original text within the CMS and highlights them in the WYSIWYG editor (see Figure 26). In a **context information area** of the CMS an overview of the found entities is given in the form of thumbnails (see Figure 27). Now the author has several choices:

- View additional information about the entities and navigate recursively.
- Adapt the filter (e.g. by type) in the config options and remove some of the entities.
- Revise the text and resend the text to the OWS.
- Accept all annotated entities and publish them along with the text.

32 http://sourceforge.net/projects/ontosfeeder
33 http://wordpress.org/extend/plugins/ontos-feeder/
34 http://ivan-herman.name/2009/12/12/rdfa-usage-spreading.../ and http://www.mahalo.com/rdfa

Figure 27: The context information area is displayed next to the WYSIWYG editor and allows to navigate recursively to relevant contextual information from the Data Web.

If an author requires additional information about a particular entity, pointing at each annotation or thumbnail results in showing an appropriate pop-up menu with further contextual information. Each entity type provides different context information depending on their source; some of them are gathered from LOD providers such as DBpedia or Freebase, and some are coming directly from the OWS itself. While the LOD providers are used to retrieve entity attributes like age, nationality or area, the OWS provides information from the Ontos Knowledge Base comprised of information about the relationships to other entities (persons or organisations), related news articles as well as a summarizing entity report. Clicking on the related persons or organisations link in the pop-up menu refreshes the context information area with the thumbnails of that entities, so that the author can navigate recursively through all the relationships.

The *Ontos Knowledge Base* contains aggregated information from various sources. The URIs assigned to the extracted entities by the Web Service are *Ontos Entity Identifiers* (OEI). OEIs are de-referencable identifiers and are connected via `owl:sameAs` links to other Linked Data entities. Therefore, additional information from other Linked Data providers such as *DBpedia* and *Freebase* is presented in the entity context as well.

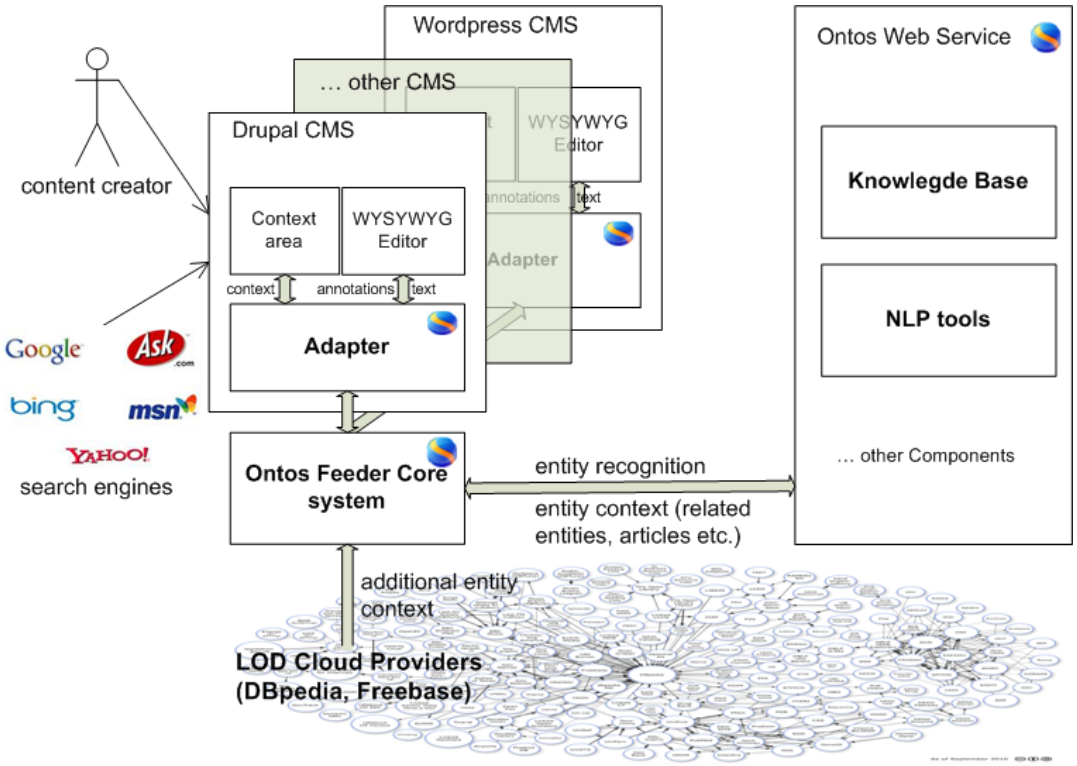


Figure 28: Ontos Feeder overall architecture

10.5.2 *Architecture*

While the server side consists of the OWS, the client side consists of the Core system and the CMS - adapters (see Figure 28). The core is CMS independent and can be embedded into a specific CMS by an appropriate adapter. Currently adapters for Drupal and WordPress are available.

ONTOS WEB SERVICE (OWS) The Core system sends queries to the OWS. The Ontos Knowledge Base contains aggregated and refined information from around 1 million documents mainly from English online news. The Ontos Semantic Engine (NLP) extracts entities and their relationships from the text and disambiguates entities based on significance (Efimenko, Minor, Starostin, Drobyazko, & Khoroshevsky, 2010). The significance is a complex measure based on the position of the occurrence in the text, the overall number of occurrences and the number of connected events and facts extracted from the text. The resulting information is returned to the Ontos Feeder.

ONTOS FEEDER The Ontos Feeder currently supports requesting information for persons, organisations, locations and products, but can generally be extended to handle any of the entity types supported by the OWS. The user can configure, which types of entities the OWS

should try to recognize in the provided text. The retrieval of each single piece of contextual information is encapsulated as a separate task by Ontos Feeder to increase the flexibility. The task engine supports task chaining, so if information could not be retrieved from a particular Linked Data source, it is requested from another one. The type of presented contextual information depends on the type of the recognized entity. The contextual information of a *Person* for example can consist of the age, nationality, status roles, connections to other persons and organisations, latest articles about this person, a Wikipedia article, a New York Times article, the personal homepage and a collection of different public social network profiles from Twitter or Facebook. Information about connections to other people and organisations, the status roles and the relevant articles are collected from the OWS. As every single information piece is requested by its own task, the variety of the presented contextual information can easily be adapted to personal needs.

10.5.3 *Embedding Metadata*

The OWS is able to annotate plain text as well as markup data such as HTML documents. The result is returned as a `stand-off annotation`, either in the form of start and end positions for text or an XPath expression for XML markup. A specialized annotation algorithm is used to: 1. highlight the annotations in the source HTML document in the editors. and 2. insert the annotations *inline* (as e.g. RDFa) into the HTML source of the article. Because all of the supported CMS WYSIWYG editors (currently FCKEditor and TinyMCE[35]) are capable of returning the current article as plain text, Ontos Feeder utilizes the Web Service in plain-text mode. As each of the editors have a different API, a special abstraction layer is put in front of the annotation algorithm to make it editor-independent. Furthermore, to make the annotation algorithm work faster for a plain-text document, all annotations are sorted in descended order and inserted bottom-up into the text. This avoids the recalculation of the annotation positions as compared to the top-down insertion. The annotation algorithm is capable of dealing with the entire supported semantic markup languages (RDFa and Microformats) and allows for annotation highlighting and on-the-fly binding of the contextual pop-up menu (see Figure 26).

10.5.4 *Related Work and Summary*

In recent years, several services have been published for suggesting annotations of tags to users. Among those services, OpenCalais and Zemanta are highly related to the Ontos Feeder as they also pro-

35 http://ckeditor.com/ and http://tinymce.moxiecode.com/

vide CMS integrations[36]. While Zemanta focuses on provide tag and link suggestions only, OpenCalais additionally extracts facts from the written text of the article. In contrast, the focus of the OWS is to provide disambiguated additional information, which is useful for the author. The data comes from the Ontos Knowledge Base and has been aggregated and fused from several sources. Also, the contribution of the Ontos Feeder, go well beyond the provision of a mere wrapper of a data service as it has a flexible, extensible architecture, is open-source and provides a context information area with recursive Linked Data navigation that aids the author. It transform standoff annotations into inline RDFa and thus allows for a more fine-grained annotation method. Future work will be devoted to the area of co-referencing (Glaser, Jaffri, & Millard, 2009) for example by using OKKAM. Furthermore, it is planned, that users are able to define own vocabularies for named entity recognition, thus personalizing the annotation process.

10.6 RELFINDER: REVEALING RELATIONSHIPS IN RDF KNOWLEDGE BASES

Heim et al. (2009)

The Semantic Web enables answers to new kinds of user questions. Unlike searching for keywords in Web pages (as e.g. in *Google*), information can be accessed according to its semantics. The information is stored in structured form in knowledge bases using formal languages such as *RDF*[37] or *OWL*[38] and consisting of statements about real world objects like 'Washington' or 'Barack Obama'. Each object has a unique identifier (*URI*) and is usually assigned to ontological classes, such as 'city' or 'person', and an arbitrary number of properties that define links between the objects (e.g., 'lives in'). Given this semantically annotated and linked data, new ways to reveal relationships within the contained information are possible.

A common visualization for linked data are graphs, such as the *Paged Graph Visualization* (Deligiannidis, Kochut, & Sheth, 2007). In order to find relationships in these visualizations, users normally apply one of the following two strategies: They either choose a starting point and incrementally explore the graph by following certain edges, or they start with a visualization of the entire graph and then filter out irrelevant data. Some more sophisticated solutions are based on the concept of faceted search. The tool *gFacet* (Heim, Ziegler, & Lohmann, 2008), for instance, groups object data into facets that are represented by nodes and can be used to filter a user-defined result set. However, all these approaches require the user to manually explore the visualization in order to find relationships between two ob-

36 http://drupal.org/project/[opencalais|zemanta]
37 http://www.w3.org/RDF/
38 http://www.w3.org/TR/owl-features/

jects of interest. This kind of trial-and-error search can be very time consuming, especially in large knowledge bases that contain many data links.

As a solution to this problem, we propose an approach that automatically reveals relationships between two known objects and displays them as a graph. The relationships are found by an algorithm that is based on a concept proposed in Lehmann, Schüppel, and Auer (2007) and that can be applied to large knowledge bases, such as *DBpedia* (Lehmann et al., 2009) or the whole *LOD-Cloud*[39]. Since the graph that visualizes the relationships can still become large, we added interactive features and filtering options to the user interface that enable a reduction of displayed nodes and facilitate understanding. We present an implementation of this approach – the *RelFinder* – and demonstrate its applicability by an example from the knowledge base *DBpedia* (Lehmann et al., 2009).

10.6.1 *Implementation*

The RelFinder is implemented in Adobe Flex[40] and runs in all Web browsers with an installed Flash Player[41]. In the following, we first explain its general functionality before describing the involved mechanisms in more detail.

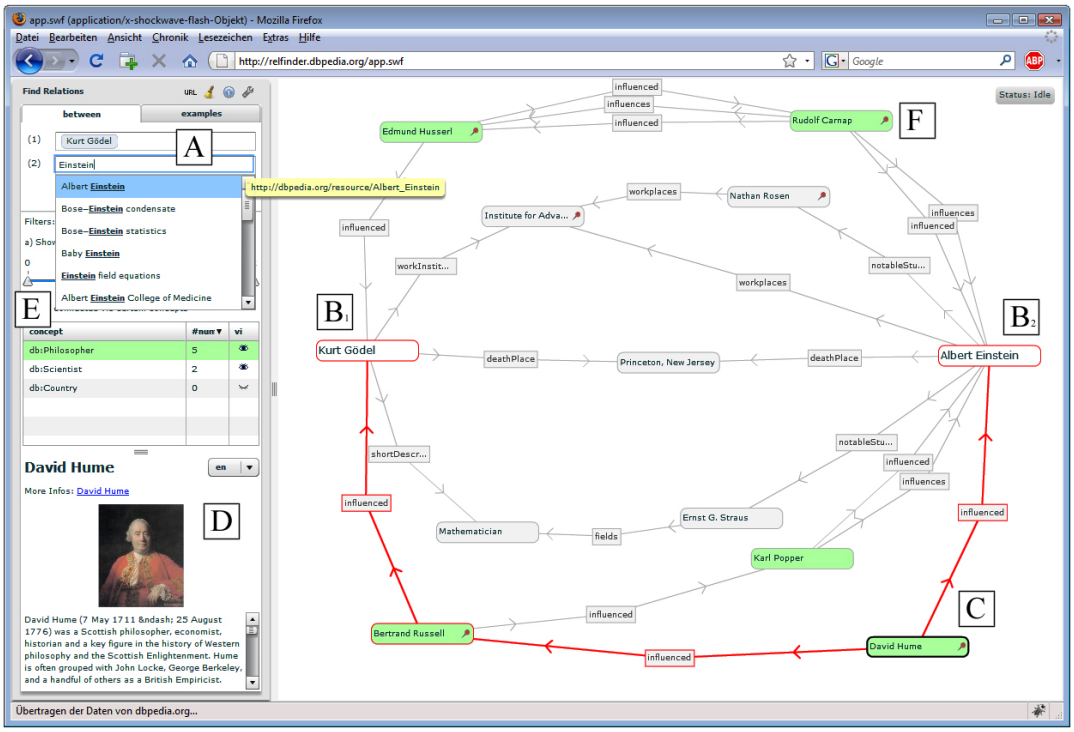


Figure 29: Revealing relationships between Kurt Gödel und Albert Einstein.

39 http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
40 http://www.adobe.com/products/flex
41 The current version of the RelFinder is accessible at http://relfinder.dbpedia.org

The search terms that are entered by the user in the two input fields in the upper left corner (Fig. 29, A) get mapped to unique objects of the knowledge base. These constitute the left and right starting nodes in the graph visualization (Fig. 29, B) that get then connected by relations and objects found in between them by the algorithm. If a certain node is selected all graph elements that connect this node with the starting nodes are highlighted forming one or more paths through the graph (Fig. 29, C). In addition, further information about the selected object is displayed in the sidebar (Fig. 29, D). Filters can be applied to increase or reduce the number of relationships that are shown in the graph and to focus on certain aspects of interest (Fig. 29, E).

10.6.2 *Disambiguation*

Ideally, the search terms that are entered by the user can be uniquely matched to objects of the knowledge base without any disambiguation. However, if multiple matches are possible (e.g., in case of homonyms, polysemes, or incomplete user input) the user is supported by a disambiguation feature. Generally, a list of objects with labels that enclose the search terms is already shown below the input box while the user enters the terms (Fig. 29, A). This disambiguation list results from a query against the SPARQL endpoint of the selected knowledge base. The following code shows the DBpedia optimized SPARQL query for the user input 'Einstein'[42]:

```
SELECT ?s ?l count(?s) as ?count WHERE {
?someobj ?p ?s .
?s <http://www.w3.org/2000/01/rdf-schema#label> ?l .
?l bif:contains '"Einstein"' .
FILTER (!regex(str(?s), '^http://dbpedia.org/resource/Category:')).
FILTER (!regex(str(?s), '^http://dbpedia.org/resource/List')).
FILTER (!regex(str(?s), '^http://sw.opencyc.org/')).
FILTER (lang(?l) = 'en').
FILTER (!isLiteral(?someobj)).
} ORDER BY DESC(?count) LIMIT 20
```

The disambiguation list is sorted by relevance using the 'count' value (or alternatively a string comparison if 'count' is not supported by the endpoint). 'count' is also used to decide if a user's search term can be automatically matched to an object of the knowledge base or if a manual disambiguation is necessary. An automatic match is performed in one of the following two cases: 1) if the user input and

42 A configuration file allows to freely define the queried endpoint, the element that is queried (typically 'rdfs:label'), and properties that should be ignored. It is also possible to deactivate specific syntax elements, such as 'count' or 'bif:contains', in case a SPARQL endpoint does not support these.

the label of the most relevant object are completely equal, or 2) if the user input is contained in the label of the most relevant object and this object has a much higher count value than the second relevant object of the disambiguation list (ten times higher by default). Thus, the automatic disambiguation is rather defensive in order to prevent false matches.

If the user does not select an entry from the disambiguation list and if no automatic match is possible, the entries from the disambiguation list are shown again in a pop-up dialog that explicitly asks the user to provide the intended meaning of the search term by selecting the corresponding object.

10.6.3 *Searching for Relationships*

A query building process composed of several SPARQL queries searches for relationships between the given objects of interest. Since the shortest connection is not known in advance, the process searches iteratively for connections with increasing length, starting from zero. As a constraint, the direction of the property relations within each connection chain is only allowed to change once. We defined this constraint due to performance reasons and because multiple changes in the direction of the edges are difficult to be followed and understood by the user. If our objects of interests are a and b this results in the following search patterns:

$$
\begin{aligned}
a &\rightarrow \cdots \rightarrow b \\
a &\leftarrow \cdots \leftarrow b \\
a \rightarrow \cdots \rightarrow c &\leftarrow \cdots \leftarrow b \\
a \leftarrow \cdots \leftarrow c &\rightarrow \cdots \rightarrow b
\end{aligned}
$$

Thus, we are looking either for one-way relationships (first two lines) or those with an object c in between such that there is a one-way relationship each from a and from b to c or from c to a and b (last two lines). Note that c is not known in advance but found within the searching process.

The algorithm has several parameters: 1) It can be configured to suppress circles in extracted relationships. With the help of SPARQL filters, any object is only allowed to occur once in each connecting relationship. 2) Objects and properties can be ignored using regular expression patterns on their labels or URIs, which is useful if someone is not interested in certain objects or properties. More importantly, also structural relations between objects can be omitted, such as whether two objects belong to the same class or to the same part of a class hierarchy (i.e., ignoring *rdf:type* and *rdfs:subClassOf* properties). We decided to remove these by default, since they normally yield a multitude of relationships of minor interest which can be better explored in more traditional ways such a hierarchy browsers. 3) A maximum

length of the returned relationships can be defined. 4) The SPARQL endpoint to use can be configured.

An exemplary SPARQL query that searches for relationships of the type *Kurt Gödel* ← of1 ← c → os1 → *Albert Einstein* is given here (filter omitted):

```
SELECT * WHERE {
  db:Kurt_Goedel ?pf1 ?of1 .
  ?of1 ?pf2 ?c .
  db:Albert_Einstein ?ps1 ?os1 .
  ?os1 ?ps2 ?c .
    FILTER ...
} LIMIT 20
```

10.6.4 *Graph Visualization*

The found relationships are added one by one to the graph, beginning with the shortest (i.e., direct relationships and relationships with only one object in between, if there are any). All objects are visualized as nodes connected by edges that are labeled and directed according to the property relation they represent (Fig. 29, B). Since the labels of the edges are crucial for understanding the relationships they serve as flexible articulations in the force-directed layout (Fruchterman & Reingold, 1991), what reduces overlaps but cannot completely avoid them.

10.6.4.1 *Interactive Features*

To further reduce overlaps in the graph, we implemented a pinning feature that enables users to manually drag single nodes away from agglomerations and forces them to stay at the position they got dropped (pinned nodes are indicated by needle symbols as can be seen in Fig. 29, F). Especially in situations where many nodes are connected by many edges and thus are likely to overlap in the automatic layout, manual adjustments in combination with our pinning feature are helpful to produce an understandable graph layout that facilitates visual tracking. As already mentioned, visual tracking is additionally supported by the possibility of highlighting all paths that connect a selected node with the starting nodes (Fig. 29, C).

If a certain node is selected in the graph, further information about the corresponding object is displayed in the sidebar (e.g., in case of DBpedia these are a title, short abstract, and image extracted from Wikipedia, Fig. 29, D). Moreover, the ontological class an object belongs to is highlighted in the list that is shown in the sidebar (Fig. 29, E). Vice versa, all corresponding nodes in the graph are highlighted if an ontological class is selected from the list.

10.6.4.2 *Filtering Options*

The shown relationships can be filtered in two ways: 1) According to their length (i.e., the number of objects in between) and 2) according to the ontological classes the objects belong to. For instance, relationships consisting of several objects could be regarded as too far-fetched or objects belonging to certain classes might not be of interest for a user's goals (Fig. 29, E) and are therefore removed from the graph.

Filtering helps to reduce the number of displayed relationships in the graph and can hence prevent the graph from getting overly cluttered. For each search process, the filters are automatically set to initial values that avoid an over-cluttered graph, if possible.

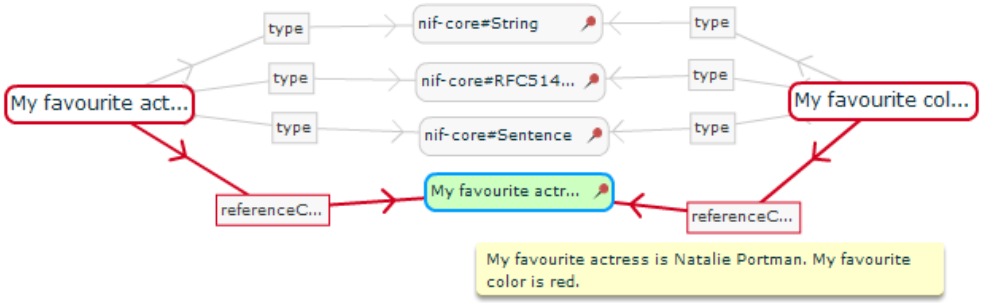


Figure 30: Two sentence in NIF from the same reference context.

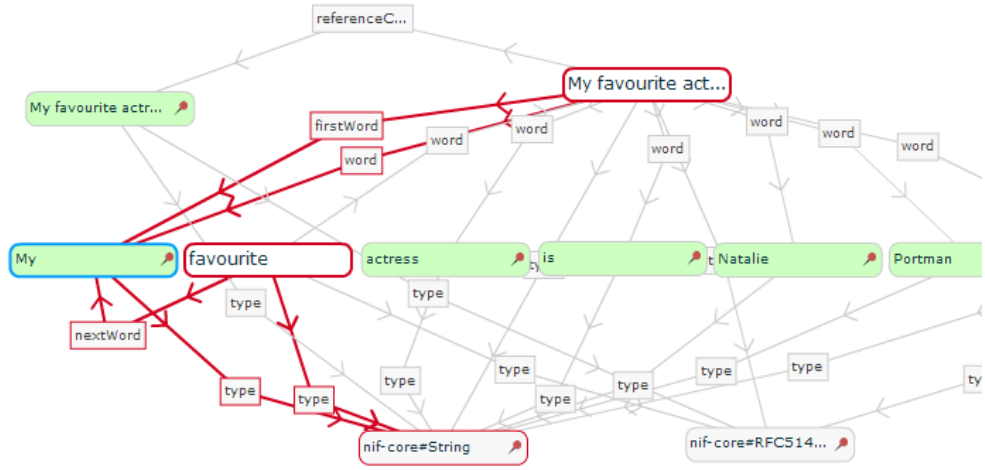


Figure 31: Relation between the first and the second word in our example sentence.

10.6.5 *Conclusion*

We introduced an approach in this section that uses properties in semantically annotated data to automatically find relationships between any pair of user-defined objects and visualizes them in a force-

Figure 32: Two adjectives from different sentences linked to the same OLiA identifier.

directed graph layout. The RelFinder can therefore save a lot of time which would otherwise be lost in searching these relationships manually. Since the amount of found relationships can be large, we additionally provide two types of filters that can be used to reduce the displayed relationships according to their length and the ontological classes they belong to and thus allow focusing on only a relevant part of the relationships. Together with a pinning feature that lets users rearrange the graph layout manually, we therefore provide a semi-automatic approach.

The basic mechanisms of the RelFinder work with every SPARQL endpoint and can therefore be applied to NIF knowledge bases with only little configuration effort. Figure 30, Figure 31 and Figure 31 show the RefFinder applied to our example sentence from Chapter 7.

11 PUBLICATION OF CORPORA USING NIF

Gerber et al. (2013); Hellmann, Lehmann, et al. (2013)

11.1 WIKILINKS CORPUS

Hellmann, Lehmann, et al. (2013)

To evaluate NIF against other formats for modeling NLP annotations as RDF, we converted the *Wikilinks Corpus* (Singh, Subramanya, Pereira, & McCallum, 2012) to linked data using NIF.

11.1.1 *Description of the corpus*

The Google Wikilinks Corpus[1] is a large-scale corpus, which collects found hyperlinks to Wikipedia from text fragments gathered from over 3 million web sites. Every item consist of the *website URI* of the crawled sites and a number of *mentions*, including the English Wikipedia link, the hyperlink anchor text, its byte offset and in most cases a *context string*, i.e. suffix and prefix around the anchor of variable length. With over 3 million items and 40 million mentions it surpasses most free corpora by far and serves as a very good testbed for measuring scalability of RDF as well as performance of NER Disambiguation tools in a noisy and multi-domain environment. An example[2] of the original format is displayed below (numbers are byte offset):

```
URL http://1967mercurycougar.blogspot.com/2009_10_01_archive.html
MENTION Lincoln Continental Mark IV 40110 http://en.wikipedia.org
    /wiki/Lincoln_Continental_Mark_IV
MENTION 1975 MGB roadster 41481 http://en.wikipedia.org/wiki/MG_
    MGB
MENTION Buick Riviera 43316 http://en.wikipedia.org/wiki/Buick_
    Riviera
MENTION Oldsmobile Toronado 43397 http://en.wikipedia.org/wiki/
    Oldsmobile_Toronado
TOKEN seen 58190
TOKEN crush 63118
TOKEN owners 69290
TOKEN desk 59772
TOKEN relocate 70683
TOKEN promote 35016
TOKEN between 70846
TOKEN re 52821
TOKEN getting 68968
```

1 we used the https://code.google.com/p/wiki-link/wiki/ExpandedDataset

2 http://googleresearch.blogspot.de/2013/03/learning-from-big-data-40-million.html

	NS	NSI	NSTAN	NSTANI	OA	UC
# triples	477 250 589	316 311 355	511 220 514	350 281 280	577 488 725	607 563 176
# generated URIs	76 850 241	42 880 316	110 820 166	76 850 241	169 849 625	189 342 046
# percentage	100%	66.28%	107.12%	73.40%	121.00%	127.30%
# percentage URIs	100%	55.79%	144.2%	100%	221.01%	246.38%

Table 8: Comparison of triple count and minted URIs. Percentage relative to NS. (NIF Simple (NS), NIF Simple Ideal (NSI), NIF Stanbol (NSTAN), NIF Stanbol Ideal (NSTANI), Open Annotation (OA), UIMA Clerezza (UC)

```
TOKEN felt 41508
```

11.1.2 *Quantitative Analysis with Google Wikilinks Corpus*

15% of the items did not contain any mention with context strings and where therefore omitted. Every mention was then converted into two resources, a `nif:Context` resource for each context string and the mention resource itself with `nif:beginIndex`, `nif:endIndex`, `itsrdf:taIdentRef` and `nif:referenceContext`. The created context resource was then linked via `nif:broaderContext` to a URI of the form:[3] `http://wiki-link.nlp2rdf.org/api.php?uri=$websiteURI#char=0,` The corpus resulted in 10,526,423 files hosted in an Apache2 file system[4] and a 5.6 GB turtle dump (37.8 GB uncompressed), while the size of the original format was 5.3 GB (18 GB uncompressed). Table 8 gives a comparison of created triples and URIs by different profiles as well as OA and UIMA Clerezza[5] . Because we only have text snippets for each mention, we were forced to create one context resource per mention. If the whole plain text of the website were available (as according to the creators is planned in the near future), NIF could further reduce the number of triples to 66.28% (NSI), by using the whole document text as context. This is not the underspecified variant, which would even cause another large reduction of triples.

11.2 RDFLIVENEWS

Gerber et al. (2013)

One further example were NIF is used to publish linked data is RDFLiveNews. Implementing the original vision behind the Semantic Web requires the provision of a Web of Data which delivers timely data at all times. The foundational example presented in Berners-Lee et al's seminal paper on the Semantic Web (Berners-Lee, Hendler, &

3 e.g. `http://wiki-link.nlp2rdf.org/api.php?uri=http://phish.net/song/on-green-dolphin-street/history#char=0,`

4 `http://wiki-link.nlp2rdf.org/`

5 NS was generated, all others calculated based on `http://persistence.uni-leipzig.org/nlp2rdf/doc/wikilink-stats.txt`

Lassila, 2001) describes a software agent who is tasked to find medical doctors with a rating of excellent or very good within 20 miles of a given location at a given point in time. This requires having timely information on which doctors can be found within 20 miles of a particular location at a given time as well as having explicit data on the rating of said medical doctors. Even stronger timeliness requirements apply in decision support, where software agents help humans to decide on critical issues such as whether to buy stock or not or even how to plan their drive through urban centers. Furthermore, knowledge bases in the Linked Open Data (LOD) cloud would be unable to answer queries such as "Give me all news of the last week from the New York Times pertaining to the director of a company". Although the current LOD cloud has tremendously grown over the last years (Auer, Lehmann, & Ngomo, 2011), it delivers mostly encyclopedic information (such as albums, places, kings, etc.) and fails to provide up-to-date information that would allow addressing the information needs described in the examples above.

The idea which underlies our work is thus to alleviate this current drawback of the Web of Data by developing an approach that allows extracting RDF from unstructured (i.e., textual) data streams in a fashion similar to the live versions of the DBpedia[6] and LinkedGeoData[7] datasets. The main difference is yet that instead of relying exclusively on structured data like LinkedGeoData or on semi-structured data like DBpedia, we rely mostly on unstructured, textual data to generate RDF. By these means, we are able to unlock some of the potential of the document Web, of which up to 85% is unstructured (Gaag, Kohn, & Lindemann, 2009). To achieve this goal, our approach, dubbed RdfLiveNews, assumes that it is given unstructured data streams as input. These are deduplicated and then used as basis to extract patterns for relations between known resources. The patterns are then clustered to labeled relations which are finally used as basis for generating RDF triples. We evaluate our approach against a sample of the RDF triples we extracted from RSS feeds and show that we achieve a very high precision.

The remainder of this work is structured as follows: We first give an overview of our approach and give detailed insights in the different steps from unstructured data streams to RDF. Then, we evaluate our approach in several settings. We then contrast our approach with the state of the art and finally conclude.

11.2.1 *Overview*

We implemented the general architecture of our approach dubbed RDFLiveNews according to the pipeline depicted in Figure 33. First,

6 http://live.dbpedia.org/sparql
7 http://live.linkedgeodata.org/sparql

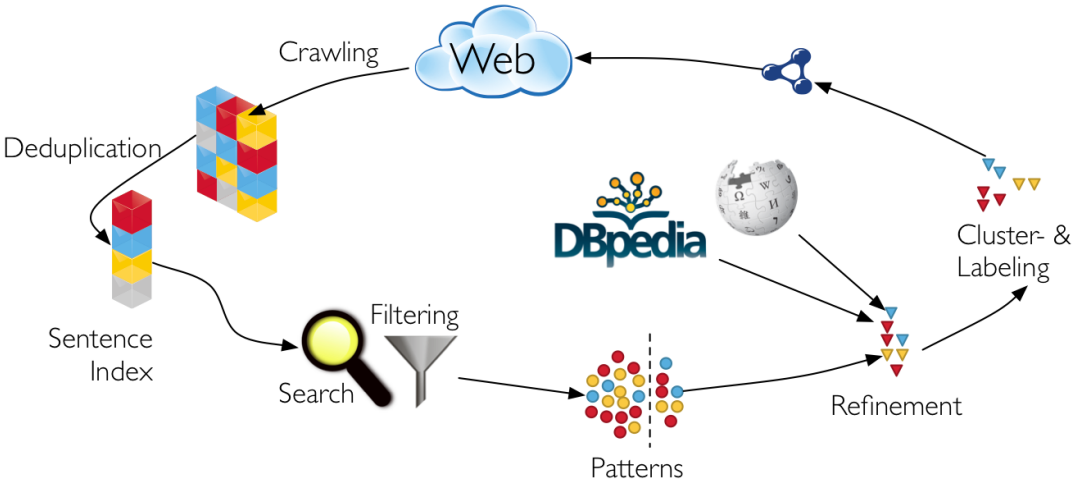


Figure 33: Overview of the generic time slice-based stream processing (Gerber et al., 2013).

we gather textual data from data streams by using RSS feeds of news articles. Our approach can yet be employed on any unstructured data published by a stream. Since input streams from the Web can be highly redundant (i.e., convey the same information), we then deduplicate the set of streams gathered by our approach. Subsequently, we apply a pattern search to find lexical patterns for relations expressed in the text. After a refinement step with background knowledge, we finally cluster the extracted patterns according to their semantic similarity and transform this information into RDF using NIF.

11.2.2 *Mapping to RDF and Publication on the Web of Data*

To close the circle of the round-trip pipeline of RDFLiveNews, the following prerequisite steps are required to re-publish the extraction results in a sensible way:

1. The facts and properties contained in the internal data structure of our tool have to be mapped to OWL.
2. Besides the extracted factual information several other aspects and meta data are interesting as well, such as extraction and publication data and provenance links to the text the facts were extracted from.
3. URIs need to be minted to provide the extracted triples as linked data.

Mapping to OWL. Each cluster $c_i \in \mathcal{C}$ represents an `owl:ObjectProperty` $prop_{c_i}$. The `rdfs:domain` and `rdfs:range` of $prop_{c_i}$ is determined by a majority voting algorithm with respect to δ and ρ of all $p_r \in \mathcal{C}$.

The `skos:prefLabel`[8] of $prop_{c_i}$ is the label determined by the cluster labeling step and all other NLRs of the patterns in c_i get associated with $prop_{c_i}$ as `skos:altLabels`. For each subject-object pair in $S_{\subseteq}{}'$ we produce a triple by using $prop_{c_i}$ as predicate and by assigning learned entity types from DBpedia or `owl:Thing`.

Provenance tracking with NIF. Besides converting the extracted facts from the text, we are using the current draft of the NLP Interchange Format (NIF) Core ontology[9] to serialize the following information in RDF: the sentence the triple was extracted from, the extraction date of the triple, the link to the source URL of the data stream item and the publication date of the item on the stream. Furthermore, NIF allows us to link each element of the extracted triple to its origin in the text for further reference and querying.

NIF is an RDF/OWL based format to achieve interoperability between language tools, annotation and resources. NIF offers several URI schemes to create URIs for strings, which can then be used as subjects for annotation. We employ the NIF URI scheme, which is grounded on URI fragment identifiers for text (RFC 5147[10]). NIF was previously used by NERD (Rizzo et al., 2012) to link entities to text. For our use case, we extended NIF in two ways: (1) we added the ability to represent extracted triples via the ITS 2.0 / RDF Ontology[11]. `itsrdf:taPropRef` is an `owl:AnnotationProperty` that links the NIF String URI to the `owl:ObjectProperty` by RDFLiveNews. The three links from the NIF String URIs (str_1, str_2, str_3) to the extracted triple (s, p, o) itself make it well traceable and queryable: $str_1 \mapsto s$, $str_2 \mapsto p$, $str_3 \mapsto o$, $s \mapsto p \mapsto o$. An example of NIF RDF serialization is shown in Listing 2. (2) Although (Rizzo et al., 2012) already suggested the minting of new URIs, a concrete method for doing so was not yet researched. In RDFLiveNewswe use the source URL of the data stream item to re-publish the facts for individual sentences as linked data. We strip the scheme component (http://) of the source URL and percent encode the ultimate part of the path and the query component[12] and add the md5 encoded sentence to produce the following URI:

```
http://rdflivenews.aksw.org/extraction/ + example.com:8042/over/ +
    urlencode(there?name=ferret) + / + md5('sentence')
```

Listing 2: Example RDF extraction of RDFLiveNews

```
@base <http://rdflivenews.aksw.org/extraction/www.necn.com
    /07/04/12/Scientists-discover-new-subatomic-partic/landing.
    html%3FblockID%3D735470%26feedID%3D4213/8
    a1e5928f6815c99b9d2ce613cf24198#>.
## prefixes: please use http://prefix.cc, e.g. http://prefix.cc/
    rlno
## extracted property + result of linking
```

8 http://www.w3.org/2004/02/skos/
9 http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
10 http://tools.ietf.org/html/rfc5147
11 http://www.w3.org/2005/11/its/rdf#
12 http://tools.ietf.org/html/rfc3986#section-3

```
rlno:directorOf a owl:ObjectProperty ;
  skos:prefLabel "director of" , skos:altLabel ", director of " ;
  owl:equivalentProperty dbp:director .
## extracted facts:
rlnr:Rolf_Heuer a dbo:Person ;
  rdfs:label "Rolf Heuer"@en ;
  rlno:directorOf dbpedia:CERN .
dbpedia:CERN a owl:Thing ;
  rdfs:label "CERN"@en .
## provenance tracking with NIF:
<char=0,10>   itsrdf:taClassRef dbo:Person ;
  itsrdf:taIdentRef rlnr:Rolf_Heuer .
<char=14,18>  itsrdf:taIdentRef dbpedia:CERN .
<char=11,24>  nif:anchorOf      ", director of"^^xsd:string ;
  itsrdf:taPropRef rlno:directorOf .
## detailed NIF output with context, indices and anchorOf
<char=0,> a nif:String, nif:Context, nif:RFC5147String ;
  nif:isString "Rolf Heuer , director of CERN , said the newly
      discovered particle is a boson , but he stopped just shy of
       claiming outright that it is the Higgs boson itself - an
      extremely fine distinction." ;
  nif:sourceUrl <http://www.necn.com/07/04/12/Scientists-discover
      -new-subatomic-partic/landing.html?blockID=735470&feedID
      =4213>;
## extraction date:
  dcterms:created "2013-05-09T18:27:08+02:00"^^xsd:dateTime .
## publishing date:
<http://www.necn.com/07/04/12/Scientists-discover-new-subatomic-
    partic/landing.html?blockID=735470&feedID=4213>
  dcterms:created "2012-08-15T14:48:47+02:00"^^xsd:dateTime .
<char=0,10> a nif:String, nif:RFC5147String .
  nif:referenceContext <char=0,>; nif:anchorOf "Rolf Heuer" ;
  nif:beginIndex "0"^^xsd:long ; nif:endIndex "10"^^xsd:long ;
```

Republication of RDF. The extracted triples are hosted on: `http://rdflivenews.aksw.org`. The data for individual sentences is crawlable via the file system of the Apache2 web server. We assume that source URLs only occur once in a stream when the document is published and the files will not be overwritten. Furthermore, the extracted properties and entities are available as linked data at *http://rdflivenews.aksw.org/{ontology | resource}/$name* and they can be queried via SPARQL at `http://rdflivenews.aksw.org/sparql`.

Part V

CONCLUSIONS

12

LESSONS LEARNED, CONCLUSIONS AND FUTURE WORK

12.1 LESSONS LEARNED FOR NIF

Hellmann, Lehmann, et al. (2013)

Hellmann, Lehmann, et al. (2013)

Our evaluation of NIF since the publication of NIF 1.0 in the developers study has been accompanied by extensive feedback from the individual developers and it was possible to increase ontological coverage of NLP annotations in version 2.0, especially with the ITS 2.0 / RDF Ontology, NERD (Rizzo et al., 2012), FISE and many more ontologies that were available. Topics that dominated discussions were scalability, reusability, open licenses and persistence of identifiers. Consensus among developers was that RDF can hardly be used efficiently for NLP in the internal structure of a framework, but is valuable for exchange and integration. The implementation by Apache Stanbol offered a promising perspective on this issue as they increased scalability by transforming the identifiers used in OLiA into efficient Java code structures (enums). Hard-compiling ontological identifiers into the type systems of Gate and UIMA seems like a promising endeavour to unite the Semantic Web benefits with the scalability requirements of NLP. A major problem in the area remains the URI persistence. Since 2011 almost all of the mentioned ontologies either changed their namespace and hosting (OLiA and NIF itself) or might still need to change (Lemon, FISE), which renders most of the hard-coded implementations useless.

12.2 CONCLUSIONS

During his keynote at the Language Resource and Evaluation Conference in 2012, Sören Auer stressed the decentralized, collaborative, interlinked and interoperable nature of the Web of Data. The keynote provides strong evidence that *Semantic Web technologies such as Linked Data are on its way to become main stream for the representation of language resources*. Chapter 3 and Chapter 6 summarize the results of the Linked Data in Linguistics (LDL) Workshop in 2012 and the NLP & DBpedia Workshop in 2013 and give a preview of the MLOD special issue. In total, five proceedings – three published at CEUR (OKCon 2011, WoLE 2012, NLP & DBpedia 2013), one Springer book (Linked Data in Linguistics, LDL 2012) and one journal special issue (Multilingual Linked Open Data, MLOD to appear) – have been (co-)edited to create incentives for scientists to convert and publish Linked Data and thus *to contribute open and/or linguistic data to the LOD cloud*. Based

on the disseminated call for papers, *152 authors contributed one or more accepted submissions* to our venues and 120 reviewers were involved in peer-reviewing.

Chapter 4 and Chapter 5 contain this thesis' contribution to the DBpedia Project in order to further increase the size and inter-linkage of the LOD Cloud with lexical-semantic resources. Our contribution comprises extracted data from Wiktionary (an online, collaborative dictionary similar to Wikipedia) in more than four languages (now six) as well as language-specific versions of DBpedia, including a quality assessment of inter-language links between Wikipedia editions and internationalized content negotiation rules for Linked Data. In particular the work described in Chapter 4 created the foundation for a DBpedia Internationalisation Committee with *members from over 15 different languages with the common goal to push DBpedia as a free and open multilingual language resource*.

The NIF core specification was presented in Chapter 7 and describes which URI schemes and RDF vocabularies must be used for (parts of) natural language texts and annotations in order to create *an RDF/OWL-based interoperability layer with NIF built upon Unicode Code Points in Normal Form C*. In Chapter 8, classes and properties of the *NIF Core Ontology* were described to formally define the relations between text, substrings and their URI schemes. Chapter 9 contains the evaluation of NIF.

In the questionnaire, we asked questions to 13 developers using NIF with the result that: UIMA, GATE and Stanbol are extensible NLP frameworks and NIF was not yet able to provide off-the-shelf NLP domain ontologies for all possible domains, but only for the plugins used in this study. After inspecting the software, the developers agreed however that NIF is general enough and adequate to provide a generic RDF output based on NIF using literal objects for annotations. All developers were able to map the internal data structure to NIF URIs to serialize RDF output (Adequacy). The development effort in hours (ranging between 3 and 40 hours) as well as the number of code lines (ranging between 110 and 445) suggest, that the implementation of NIF wrappers is easy and fast for an average developer. Furthermore the evaluation contains a comparison to other formats and an evaluation of the available URI schemes for web annotation.

In order to collect input from the wide group of stakeholders, a total of 16 presentations were given with extensive discussions and feedback, which has lead to a constant improvement of NIF from 2010 until 2013. After the release of NIF (Version 1.0) in November 2011, a total of *32 employments and implementations for different NLP tools and converters were reported* (8 by the (co-)authors, including Wiki-link corpus (Section 11.1), 13 by people participating in our survey and 11 more, of which we have heard). Several roll-out meetings and tutori-

als were held (e.g. in Leipzig and Prague in 2013) and are planned (e.g. at LREC 2014).

One major contribution in Chapter 10 is the usage of NIF as the recommended RDF mapping in the Internationalization Tag Set 2.0 W3C standard (Section 10.1) and the conversion algorithms from ITS to NIF will integrate software components of AR and back (Section 10.1.1). One outcome of the discussions in the standardization meetings and telephone conferences for ITS 2.0 resulted in the conclusion *no alternative RDF format or vocabulary other than NIF* with the required features to fulfill the working group charter. Five further uses of NIF are described for the Ontology of Linguistic Annotations (OLiA), the RDFaCE tool, the Tiger Corpus Navigator, the OntosFeeder and visualisations of NIF using the RelFinder tool. The 8 mentioned instances provide an implemented proof-of-concept of the features of NIF.

Chapter 11 described the conversion and hosting of the huge Google Wikilinks corpus with 40 million annotations for 3 million web sites. The resulting RDF dump contains 477 million triples in a 5.6 GB compressed dump file in turtle syntax. Section 11.2 describes how NIF can be used to publish extracted facts from news feeds in the RDFLiveNews tool as Linked Data.

12.3 FUTURE WORK

Hellmann, Lehmann, et al. (2013)

The NIF/NLP2RDF project can be seen as an umbrella project creating bridges between different communities to achieve interoperability in the NLP domain via ontologies. The currently active and fruitful collaborations such as Stanbol, Spotlight, Open Annotation, ITS, OLiA, NERD are yet mostly centered on stakeholders from the Semantic Web. With the soon-to-start LIDER EU project[1], NLP2RDF will outreach to core NLP projects such as CLARIN, ELRA and LanguageGrid.[2] Identifying incentives relevant for stakeholders outside the Semantic Web community remains an open challenge as in this initial phase NIF focused primarily on middleware interfaces and not directly on end user problems. We will investigate existing (and hopefully directly reusable) approaches on Semantic Web workflows such as SADI, Taverna and WSMO-Lite.[3] A NIF workflow, however, can obviously not provide any better performance (F-measure, efficiency) than a properly configured UIMA or GATE pipeline with the same components. NIF rather targets and benefits developers in terms of entry barrier, data integration, reusability of tools, conceptualisation and off-the-shelf solutions. Early adoption of open-source as

1 http://lider-project.eu/

2 http://www.clarin.eu/node/3637, http://elra.info, http://langrid.org

3 http://sadiframework.org, http://www.taverna.org.uk, http://www.w3.org/Submission/WSMO-Lite

well as industry projects is manifesting, but an exhaustive overview and a machine-readable collection of available implementations and deployments is yet missing.

One particular aspect worth mentioning is the increasing number of NIF-formated corpora for Named Entity Recognition (NER) that have come into existence after the publication of the main NIF paper *Integrating NLP using Linked Data* at ISWC 2013. These include the corpora converted by Steinmetz, Knuth, and Sack (2013) for the NLP & DBpedia workshop and an OpenNLP-based CoNLL-format converter by Brümmer. Furthermore, we are aware of three LREC 2014 submissions that leverage NIF: *NIF4OGGD - NLP Interchange Format for Open German Governmental Data*, N^3 – *A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format* and *Global Intelligent Content: Active Curation of Language Resources using Linked Data* as well as an early implementation of a GATE-based NER/NEL evaluation framework by Dojchinovski and Kliegr. Further funding for the maintenance, interlinking and publication of Linguistic Linked Data as well as support and improvements of NIF is available via the expiring LOD2 EU project, as well as the CSA EU project called LIDER (`http://lider-project.eu/`), which started in November 2013. Based on the evidence of successful adoption presented in this thesis, we can expect a decent to high chance of reaching critical mass of Linked Data technology as well as the NIF standard in the field of Natural Language Processing and Language Resources.

BIBLIOGRAPHY

Aguado de Cea, G., Puch, J., & Ramos, J. (2008, May). Tagging Spanish texts: The problem of "se". In *Proc. sixth international conference on language resources and evaluation (lrec 2008).* Marrakech, Morocco.
130

Aprosio, A. P., Giuliano, C., & Alberto Lavelli, L. (2013, October). Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
63, 67, 70

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., & Aumueller, D. (2009). Triplify: Light-weight linked data publication from relational databases. In J. Quemada, G. León, Y. S. Maarek, & W. Nejdl (Eds.), *Proceedings of the 18th international conference on world wide web, WWW 2009, madrid, spain, april 20-24, 2009* (pp. 621–630). ACM. Retrieved from http://doi.acm.org/10.1145/1526709.1526793 doi: doi:10.1145/1526709.1526793
xii, 22, 23

Auer, S., & Hellmann, S. (2012, may). The web of data: Decentralized, collaborative, interlinked and interoperable. In N. Calzolari et al. (Eds.), *Proceedings of the eight international conference on language resources and evaluation (lrec'12).* Istanbul, Turkey: European Language Resources Association (ELRA). (companion publication for the Keynote at LREC 2012)
xii, 3, 9, 195

Auer, S., & Lehmann, J. (2010). Making the web a data washing machine - creating knowledge out of interlinked data. *Semantic Web Journal*. Retrieved from http://www.jens-lehmann.org/files/2010/washing_machine_swj.pdf
3, 4, 61

Auer, S., Lehmann, J., & Hellmann, S. (2009). LinkedGeoData - adding a spatial dimension to the web of data. In *Proc. of 8th international semantic web conference (iswc).* Retrieved from http://www.informatik.uni-leipzig.de/~auer/publication/linkedgeodata.pdf doi: doi:10.1007/978-3-642-04930-9_46
xii

Auer, S., Lehmann, J., & Ngomo, A.-C. N. (2011). Introduction to linked data and its lifecycle on the web. In *Reasoning web* (p. 1-75).

145

Auger, A., & Barrière, C. (2010). *Probing Semantic Relations: Exploration and identification in specialized texts* (Benjamins ed.). John Benjamins. Retrieved from http://benjamins.com/#catalog/books/bct.23/main

67

Baader, F., Horrocks, I., & Sattler, U. (2005). Description logics as ontology languages for the Semantic Web. In D. Hutter & W. Stephan (Eds.), *Mechanizing mathematical reasoning* (p. 228-248). Springer Berlin / Heidelberg.

20

Berners-Lee, T. (2006). *Design issues: Linked data.* http://www.w3.org/DesignIssues/LinkedData.html.

3, 7, 20

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

144

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, *33*(1-2), 23–60.

19

Bizer, C. (2011). *Evolving the web into a global data space.* http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-GlobalDataSpace-Talk-BNCOD2011.pdf. (Keynote at 28th British National Conference on Databases (BNCOD2011))

3

Blume, M., Flynn, S., & Lust, B. (2012). Creating Linked Data for the interdisciplinary international collaborative study of language acquisition and use: Achievements and challenges of a new virtual linguistics lab. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 85-96). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2

30

Bolikowski, Ł. (2009). Scale-free topology of the interlanguage links in wikipedia. Retrieved from http://arxiv.org/abs/0904.0564

42

Bouda, P., & Cysouw, M. (2012). Treating dictionaries as a Linked-Data corpus. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 15-23). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (com-

panion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
28, 29, 35

Bratsas, C., Alexiou, S., Kontokostas, D., Parapontis, I., Antoniou, I., & Metakides, G. (2011). Greek open data in the age of linked data: A demonstration of lod internationalization. In *Proceedings of the 3rd international conference on web science (acm websci'11)*. Koblenz, Germany: ACM. Retrieved from http://www.websci11.org/fileadmin/websci/Posters/102_paper.pdf
39

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F. (1997). Extensible markup language (xml). *World Wide Web Journal*, *2*(4), 27–66.
18

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai., V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF - Language Typology and Universals*, *61*(4), 286-308.
35, 37

Bühmann, L., & Lehmann, J. (2012). Universal OWL axiom enrichment for large knowledge bases. In *Proceedings of ekaw 2012* (pp. 57–71). Springer. Retrieved from http://jens-lehmann.org/files/2012/ekaw_enrichment.pdf
23

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., ... Cimiano, P. (2006). LingInfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the 5th international conference on language resources and evaluation (lrec 2006)*. Genoa, Italy.
35, 130

Burchardt, A., Padó, S., Spohr, D., Frank, A., & Heid, U. (2008). Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In *Proc. 3rd international joint conf on nlp (ijcnlp 2008)*. Hyderabad.
30, 130

Cabrio, E., Cojan, J., Villata, S., & Gandon, F. (2013, October). Argumentation-based Inconsistencies Detection for Question-Answering over DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
68, 69

Carl, M., & Hœg Müller, H. (2012). Integrating treebank annotation and user activity data in translation research. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 77-84). Heidelberg:

Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
30

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., & Voormann, H. (2003). The NITE XML Toolkit: flexible annotation for multimodal language data. *Behavior Research Methods, Instruments, and Computers*, *35*(3), 353–363.
130

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr. E. R. H., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In M. Fox & D. Poole (Eds.), *Aaai*. AAAI Press.
64

Cassidy, S. (2010, Jan). An RDF realisation of LAF in the DADA annotation server. In *Proceedings of the 5th joint iso-acl/sigsem workshop on interoperable semantic annotation (isa-5)*. Hong Kong.
30

Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *Aaai spring symposium: Computational approaches to analyzing weblogs*.
49

Chiarcos, C. (2010, May). Grounding an ontology of linguistic annotations in the Data Category Registry. In *Lrec-2010 workshop on language resource and language technology standards (lt&lts)*. Valetta, Malta.
124

Chiarcos, C. (2012a). Interoperability of corpora and annotations. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 161-179). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
28, 30, 32, 36

Chiarcos, C. (2012b, May). Ontologies of linguistic annotation: Survey and perspectives. In *Proceedings of the eight international conference on language resources and evaluation (lrec'12)*. Istanbul, Turkey.
36, 83, 119, 122, 124

Chiarcos, C. (2012c). Powla: Modeling linguistic corpora in owl/dl. In E. Simperl, P. Cimiano, A. Polleres, Ã. Corcho, & V. Presutti (Eds.), *Eswc* (Vol. 7295, p. 225-239). Springer.

6

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., & Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *TAL (Traitement automatique des langues)*, *49*(2).
129, 130

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a linguistic linked open data cloud : The open linguistics working group. *TAL*, *52*(3), 245 - 275. Retrieved from http://www.atala.org/Towards-a-Linguistic-Linked-Open
xii, 3, 5, 15, 21, 22, 33, 36, 195

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2012a). Introduction and overview. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 1-12). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
15, 18, 27, 29, 30, 31, 32

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2012b). The Open Linguistics Working Group of the Open Knowledge Foundation. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 153-160). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
xiii, 15, 32

Chiarcos, C., Hellmann, S., Nordhoff, S., Moran, S., Littauer, R., Eckle-Kohler, J., ... Meyer, C. M. (2012, May). The open linguistics working group. In *Proceedings of the 8th international conference on language resources and evaluation* (pp. 3603–3610). Retrieved from http://www.christian-meyer.org/research/publications/lrec2012owlg/
xiii, 27, 62

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). *Linked Data in Linguistics. Representing Language Data and Metadata*. Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
xi, 3, 6, 7, 8, 27

Chiarcos, C., Ritz, J., & Stede, M. (2012). By all these lovely tokens... merging conflicting tokenizations. *Language Resources and Evaluation*, *46*(1), 53-74.
92

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Papers in computational lexicography (complex '94)* (pp. 22–32).
129

Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., & Clark, T. (2011). An open annotation ontology for science on web 3.0. *Biomedical Semantics*, *2*, S4+. Retrieved from http://dx.doi.org/10.1186/2041-1480-2-S2-S4 doi: 10.1186/2041-1480-2-S2-S4
94

Cimiano, P., Haase, P., Herold, M., Mantel, M., & Buitelaar, P. (2007). Lexonto: A model for ontology lexicons for ontology-based nlp. In *Proceedings of the ontolexo7 workshop held in conjunction with iswc* (Vol. 7).
35

Cristea, D. (2009). Textual entailment. *Computational Linguistics*(June), 1140–1143. Retrieved from http://portal.acm.org/citation.cfm?doid=1654536.1654556 doi: 10.3115/1654536.1654556
69

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. 40th anniversary meeting of the association for computational linguistics* (pp. 168–175).
8, 103, 129

Davis, B., Handschuh, S., Troussov, A., Judge, J., & Sogrin, M. (2008, May). Linguistically light lexical extensions for ontologies. In *Proc. sixth international conference on language resources and evaluation (lrec 2008)*. Marrakech, Morocco.
130

Declerck, T., Lendvai, P., Mörth, K., Budin, G., & Váradi, T. (2012). Towards Linked Language Data for Digital Humanities. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 109-116). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
28, 31, 48

Deligiannidis, L., Kochut, K., & Sheth, A. (2007). RDF data exploration and visualization. In *Proceedings of the acm first workshop on cyberinfrastructure 2007* (p. 39-46). ACM Press.

136

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proc. berliner xml tage 2005 (bxml 2005)* (pp. 39–50). Berlin, Germany.
19

Dojchinovski, M., & Kliegr, T. (2013, October). Datasets and GATE Evaluation Framework for Benchmarking Wikipedia-Based NER Systems. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
66, 68, 69

Dostert, L. (1955). The georgetown-ibm experiment. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages* (pp. 124–135). New York: John Wiley & Sons.
6

Dutta, A., Meilicke, C., Niepert, M., & Ponzetto, S. (2013, October). Integrating Open and Closed Information Extraction: Challenges and First Steps. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
64

Eckart, K., Riester, A., & Schweitzer, K. (2012). A discourse information radio news database for linguistic analysis. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 65-75). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
30

Efimenko, I., Minor, S., Starostin, A., Drobyazko, G., & Khoroshevsky, V. (2010). Semantic web. In G. Wu (Ed.), (p. 39-62). InTech.
134

Elbedweihy, K., Wrigley, S., & Ciravegna, F. (2013, October). Using BabelNet in Bridging the Gap Between Natural Language Queries and Linked Data Concepts. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
65, 66, 68

Erdmann, M., K.Nakayama, Hara, T., , & Nishio, S. (2008). Extraction of bilingual terminology from a multilingual web-based encyclopedia. *Journal of Information Processing*, *16*, 68-79.
41

Farrar, S., & Langendoen, D. (2003, July). Markup and the GOLD ontology. In *Emeld workshop on digitizing and annotating text and*

field recordings. Michigan State University.
37

Fellbaum, C. (Ed.). (1998). *WordNet: an electronic lexical database*. MIT Press.
65

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3/4), 327-348. doi: 10.1017/S1351324904003523
8, 102

Francis, W. N., & Kucera, H. (1964). *Brown Corpus manual. Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers* (Tech. Rep.). Providence, Rhode Island: Brown University. http://khnt.aksis.uib.no/icame/manuals/brown. (revised edition 1979)
7

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., ... others (2006, May). Lexical markup framework (lmf). In *Proceedings of the 5th international conference on language resources and evaluation (lrec 2006).* Genoa, Italy.
36, 48

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed placement. In *Softw. pract. exper., 21(11)* (p. 1129-1164).
140

Gaag, A., Kohn, A., & Lindemann, U. (2009). Function-based solution retrieval and semantic search in mechanical engineering. In *Idec '09* (pp. 147–158).
145

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, *26*(5-6), 415-439.
63

Garlik, S. H., Seaborne, A., & Prud'hommeaux, E. (2013). *SPARQL 1.1 Query Language.* http://www.w3.org/TR/sparql11-query/. Retrieved from http://www.w3.org/TR/sparql11-query/
27, 43

Gerber, D., Ngonga Ngomo, A.-C., Hellmann, S., Soru, T., Bühmann, L., & Usbeck, R. (2013). Real-time rdf extraction from unstructured data streams. In *Proceedings of iswc.*
xii, 143, 144, 146, 196

Glaser, H., Jaffri, A., & Millard, I. (2009, April). Managing co-reference on the semantic web. In *Www2009 workshop: Linked data on the web (ldow2009).* Retrieved from http://eprints.ecs.soton.ac.uk/17587/
136

Goetz, T., & Suhre, O. (2004). Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal*, *43*(3), 476–489.
130

Goldfarb, C. F., & Rubinsky, Y. (Eds.). (1990). *The SGML handbook*. New York: Oxford University Press.
18

Grishman, R. (1997). Information Extraction: Techniques and Challenges. *New York*, *i*(4), 10–27. Retrieved from http://www.springerlink.com/index/k454643746325537.pdf doi: 10.1007/3-540-63438-X_2
66

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199-220.
22

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., & Wirth, C. (2012). Uby - a large-scale unified lexical-semantic resource based on lmf. In *Eacl 2012*.
48

Héder, Mihály and Mendes, Pablo N. (2012). Round-trip semantics with sztakipedia and dbpedia spotlight. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, & S. Staab (Eds.), *Www (companion volume)* (p. 357-360). ACM.
63

Heid, U., Schmid, H., Eckart, K., & Hinrichs, E. W. (2010). A corpus representation format for linguistic web services: The d-spin text corpus format and its relationship with iso standards. In *Lrec*.
103

Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., & Stegemann, T. (2009). RelFinder: Revealing relationships in RDF knowledge bases. In *Proceedings of the 3rd international conference on semantic and media technologies (samt)* (Vol. 5887, p. 182-187). Springer. Retrieved from http://jens-lehmann.org/files/2009/relfinder_samt.pdf
xii, 111, 136

Heim, P., Ziegler, J., & Lohmann, S. (2008). gFacet: A browser for the web of data. In *Proceedings of the international workshop on interacting with multimedia content in the social semantic web (imc-ssw'08)* (p. 49-58). CEUR-WS.
136

Hellmann, S. (2010). The semantic gap of formalized meaning. In L. Aroyo et al. (Eds.), *Eswc (2)* (Vol. 6089, p. 462-466). Springer. Retrieved from http://svn.aksw.org/papers/2010/ESWC_PHDSym_NLP2RDF/PhD_ESWC_hellmann_public.pdf
xii, 30, 122

Hellmann, S., & Auer, S. (2013). Towards web-scale collaborative knowledge extraction. In I. Gurevych & J. Kim (Eds.), *The people's web meets nlp* (p. 287-313). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-35085-6_11 doi: 10.1007/978-3-642-35085-6_11
xiii, 3, 22, 82

Hellmann, S., Brekle, J., & Auer, S. (2012). Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In *Jist*. Retrieved from http://svn.aksw.org/papers/2012/JIST_Wiktionary/public.pdf
xii, 27, 47

Hellmann, S., Filipowska, A., Barriere, C., Mendes, P., & Kontokostas, D. (Eds.). (2013b). *Proceedings of the NLP and DBpedia Workshop in conjunction with the 12th International Semantic Web Conference (ISWC 2013) Sydney, Australia, October, 2013*. (Vol. 1064). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1064
xi, 27, 63, 71

Hellmann, S., Filipowska, A., Barriere, C., Mendes, P. N., & Kontokostas, D. (2013a, October). NLP & DBpedia - An Upward Knowledge Acquisition Spiral. In *Proceedings of 1st international workshop on nlp and dbpedia, october 21-25, sydney, australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
27, 63

Hellmann, S., Frischmuth, P., Auer, S., & Dietrich, D. (Eds.). (2011). *Proceedings of the 6th open knowledge conference, okcon 2011, berlin, germany, june/july, 2011* (Vol. 739). CEUR-WS.org.
xi

Hellmann, S., Lehmann, J., & Auer, S. (2009). Learning of OWL class descriptions on very large knowledge bases. *International Journal on Semantic Web and Information Systems*, *5*(2), 25–48. Retrieved from http://jens-lehmann.org/files/2009_dllearner_sparql.pdf doi: doi:10.4018/jswis.2009040102
xii

Hellmann, S., Lehmann, J., & Auer, S. (2011). Learning of owl class expressions on very large knowledge bases and its applications. In I. Semantic Services & W. A. E. Concepts (Eds.), *Learning of owl class expressions on very large knowledge bases and its applications* (p. 104-130). IGI Global. doi: doi:10.4018/978-1-60960-593-3
xiii

Hellmann, S., Lehmann, J., & Auer, S. (2012). Linked-data aware uri schemes for referencing text fragments. In *Ekaw 2012*. Springer. Retrieved from http://svn.aksw.org/papers/2012/NIF/EKAW_short_paper/public.pdf doi: doi:10.1007/978-3-642-16438-5_10
xii, 75, 95, 101, 103, 104, 111

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. (2013). Integrating nlp using linked data. In *12th international semantic web conference, 21-25 october 2013, sydney, australia.* Retrieved from http://svn.aksw.org/papers/2013/ISWC_NIF/public.pdf
xii, 3, 8, 10, 75, 89, 91, 93, 101, 102, 111, 119, 120, 143, 151, 153

Hellmann, S., Lehmann, J., Auer, S., & Nitzschke, M. (2012). Nif combinator: Combining nlp tool output. In *Ekaw* (p. 446-449). Retrieved from http://jens-lehmann.org/files/2012/ekaw_nif_combinator.pdf
xii, 91, 196

Hellmann, S., Lehmann, J., Unbehauen, J., Stadler, C., Lam, T. N., & Strohmaier, M. (2012). Navigation-induced knowledge engineering by example. In *Jist.* Retrieved from http://svn.aksw.org/papers/2012/JIST_NKE/public.pdf
xii

Hellmann, S., Moran, S., Brümmer, M., & McCrae, J. (Eds.). (to appear). *Multilingual linked open data (mlod) 2012 data post proceedings* (Vol. Special Issue on Dataset Descriptions). (to appear)
xi, 27

Hellmann, S., Stadler, C., & Lehmann, J. (2012). The German DBpedia: A sense repository for linking entities. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 181-189). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
xiii, 28, 33

Hellmann, S., Stadler, C., Lehmann, J., & Auer, S. (2009). DBpedia live extraction. In *Proc. of 8th international conference on ontologies, databases, and applications of semantics (odbase)* (Vol. 5871, pp. 1209–1223). Retrieved from http://svn.aksw.org/papers/2009/ODBASE_LiveExtraction/dbpedia_live_extraction_public.pdf doi: doi:10.1007/978-3-642-05151-7_33
xii, 40, 55

Hellmann, S., Unbehauen, J., Chiarcos, C., & Ngonga Ngomo, A.-C. (2010). The tiger corpus navigator. In *Proceedings of the ninth international workshop on treebanks and linguistic theories (tlt9).* Retrieved from http://dspace.utlib.ee/dspace/bitstream/10062/15937/1/tlt9_submission_33.pdf
xii, 6, 111, 121

Hepp, M., Siorpaes, K., & Bachlechner, D. (2007). Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management. *IEEE Internet Computing, 11*(5), 54-

65. Retrieved from http://dblp.uni-trier.de/rec/bibtex/journals/internet/HeppSB07
22, 103

Herold, A., Lemnitzer, L., & Geyken, A. (2012). Integrating lexical resources through an aligned lemma list. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 35-44). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
29

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text mining wissensrohstoff text konzepte, algorithmen, ergebnisse*. Herdecke [u.a.]: W3L-Verl.
89

Holtman, K., & Mutz, A. (1998, March). *Transparent Content Negotiation in HTTP* (No. 2295). RFC 2295 (Experimental). IETF. Retrieved from http://www.ietf.org/rfc/rfc2295.txt
40

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006, June). Ontonotes: the 90% solution. In *Conference of the north american chapter of the association for computational linguistics on human language technology (hlt-naacl 2006)* (pp. 57–60). New York.
130

Ide, N. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *Proceedings of the first international language resources and evaluation conference (lrec 1998)* (pp. 463–70).
18

Ide, N. (2007, June). GrAF: A graph-based format for linguistic annotations. In *Acl-2007 linguistic annotation workshop.* Prague, Czech Republic.
130

Ide, N., Fellbaum, C., Baker, C., & Passonneau, R. (2010). The Manually Annotated Sub-Corpus: A community resource for and by the people. In *Proceedings of the acl 2010 conference short papers* (pp. 68–73).
34

Ide, N., & Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. second international conference on global interoperability for language resources (icgl 2010).* Hong Kong, China.
7, 8, 9

Ide, N., & Suderman, K. (2007). GrAF: A graph-based format for linguistic annotations. In *Proc. linguistic annotation workshop (law 2007)* (pp. 1–8). Prague, Czech Republic.
8, 18, 19, 36

Ide, N., & Suderman, K. (2012). Bridging the Gaps: Interoperability for Language Engineering Architectures using GrAF. *LRE Journal*, *46*(1), 75-89.
102

K. Moerth and T. Declerck and P. Lendvai and T. Váradi. (2011). Accessing multilingual data on the web for the semantic annotation of cultural heritage texts. In *2nd workshop on the multilingual semantic web, iswc.*
48

Kannan, N., & Hussain, T. (2006). Live urls: breathing life into urls. In *Proceedings of the 15th international conference on world wide web* (pp. 879–880). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1135777.1135924 doi: http://doi.acm.org/10.1145/1135777.1135924
105

Khalili, A., Auer, S., & Hladky, D. (2012). The rdfa content editor - from wysiwyg to wysiwym. In *Proceedings of compsac 2012 - trustworthy software systems for the digital society, july 16-20, 2012, izmir, turkey.* (Best paper award)
4, 120

Kim, E., Weidl, M., Choi, K.-S., & Auer, S. (2010). Towards a korean dbpedia and an approach for complementing the korean wikipedia based on dbpedia. In *Proceedings of the 5th open knowledge conference (okcon 2010)* (Vol. 575, pp. 12–21). London, UK: CEUR-WS. Retrieved from http://ceur-ws.org/Vol-575/paper3.pdf
41

Klebeck, A., Hellmann, S., Ehrlich, C., & Auer, S. (2011). Ontosfeeder – a versatile semantic context provider for web content authoring. In G. Antoniou et al. (Eds.), *The semanic web: Research and applications* (Vol. 6644, p. 456-460). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-21064-8_34
xii, 111, 131

Kobilarov, G., Bizer, C., Auer, S., & Lehmann, J. (2009, April). DBpedia - a linked data hub and data source for web applications and enterprises. In *Proceedings of developers track of 18th international world wide web conference (www 2009), april 20th-24th, madrid, spain.* Retrieved from http://www2009.eprints.org/228/
39

König, E., & Lezius, W. (2003). *The TIGER language - a description language for syntax graphs, Formal definition.* (Tech. Rep.). IMS.

126, 129, 196

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., & Metakides, G. (2012). Internationalization of linked data: The case of the greek dbpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, *15*(0), 51 - 61. Retrieved from http://www.sciencedirect.com/science/article/pii/S1570826812000030 doi: 10.1016/j.websem.2012.01.001

xii, 27, 39, 40, 41, 46, 54

Krizhanovsky, A. A. (2010). Transformation of wiktionary entry structure into tables and relations in a relational database schema. *CoRR*. (http://arxiv.org/abs/1011.1368)

49

Lassila, O., & Swick, R. R. (1999). *Resource Description Framework (RDF) model and syntax specification* (Tech. Rep.). World Wide Web Consortium. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.

19

Lehmann, J. (2009). DL-Learner: learning concepts in description logics. *Journal of Machine Learning Research*.

122, 123

Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, *7*(3), 154–165. Retrieved from http://jens-lehmann.org/files/2009/dbpedia_jws.pdf doi: doi:10.1016/j.websem.2009.07.002

xii, 5, 27, 39, 40, 42, 137

Lehmann, J., Schüppel, J., & Auer, S. (2007, September). Discovering unknown connections - the DBpedia relationship finder. In *Proceedings of 1st conference on social semantic web. leipzig (cssw'07), 24.-28. september* (Vol. P-113 of GI-Edition). Bonner Köllen Verlag. Retrieved from http://www.informatik.uni-leipzig.de/~auer/publication/relfinder.pdf

137

Lewis, D., O'Connor, A., Molines, S., Finn, L., Jones, D., Curran, S., & Lawless, S. (2012). Linking localisation and language resources. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 45-54). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2

29

Marek, T., Lundborg, J., & Volk, M. (2008, October). Extending the TIGER query language with universal quantification. In *Proceed-*

ing of KONVENS-2008 (p. 3-14). Berlin.
129

McCrae, J., Montiel-Ponsoda, E., & Cimiano, P. (2012). Integrating WordNet and Wiktionary with *lemon*. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 25-34). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
29, 50, 57

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with Lemon. *The Semantic Web: Research and Applications*, 245–259.
35, 50, 58, 86

McGuinness, D., & Van Harmelen, F. (2004). *OWL Web Ontology Language overview. w3c recommendation* (Tech. Rep.). World Wide Web Consortium. http://www.w3.org/TR/owl-features.
20

Mendes, P. N., Jakob, M., & Bizer, C. (2012, May). Dbpedia for nlp: A multilingual cross-domain knowledge base. In *Proc. of the eight international conference on language resources and evaluation (lrec'12)*. Istanbul, Turkey.
27

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proc. 7th international conference on semantic systems (i-semantics)*.
4, 34

Meyer, C., & Gurevych, I. (2010). Worth its weight in gold or yet another resource – A comparative study of Wiktionary, OpenThesaurus and GermaNet. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 38–49). Springer.
35, 48

Meyer, C. M., & Gurevych, I. (2010). How web communities analyze human language: Word senses in wiktionary. In *Second web science conference*.
48

Meyer, C. M., & Gurevych, I. (2011). OntoWiktionary - Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In M. Pazienza & A. Stellato (Eds.), *Semi-automatic ontology development: Processes and resources*. IGI Global.
47, 48, 60

Michaelis, S., Maurer, P., Haspelmath, M., & Huber, M. (Eds.). (in preparation). *Atlas of pidgin and creole language structures*. Oxford: Oxford University Press.

35

Miles, A., & Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System reference. W3C Recommendation* (Tech. Rep.). World Wide Web Consortium. http://www.w3.org/TR/skos-reference.

19

Moran, S. (2012). Using Linked Data to create a typological knowledge base. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 129-138). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2

32, 35

Morris, W. (Ed.). (1969). *The American heritage dictionary of the English language*. New York: Houghton Mifflin.

7

Morsey, M., Lehmann, J., Auer, S., & Ngonga Ngomo, A.-C. (2011). DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data. In *Iswc 2011*. Retrieved from http://jens-lehmann.org/files/2011/dbpsb.pdf

69

Morsey, M., Lehmann, J., Auer, S., Stadler, C., & Hellmann, S. (2012). DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, *46*, 27. Retrieved from http://svn.aksw.org/papers/2011/DBpedia_Live/public.pdf

xii

Müller, C., & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In *Corpus technology and language pedagogy* (pp. 197–214). Frankfurt am Main: Peter Lang.

129

Navigli, R., Jurgens, D., & Vannella, D. (2013). SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. In *Proceedings of the 7th international workshop on semantic evaluation semeval 2013 in conjunction with the second joint conference on lexical and computational semantics sem 2013*.

69

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, *193*, 217-250.

65

Nebhi, K. (2013, October). A Rule-Based Relation Extraction System using DBpedia and Syntactic Parsing. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Pro-

ceedings.
67, 69, 70

Ngonga Ngomo, A.-C., & Auer, S. (2011). Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of ijcai.*
62

Ngonga Ngomo, A.-C., Heino, N., Lyko, K., Speck, R., & Kaltenböck, M. (2011). Scms - semantifying content management systems. In *Iswc 2011.*
4

Nordhoff, S. (2012). Linked Data for linguistic diversity research: Glottolog/Langdoc and ASJP. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 191-200). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
8, 28, 33

Nuzzolese, A., Gangemi, A., & Presutti, V. (Submitted, June). Gathering lexical linked data and knowledge patterns from FrameNet. In *Proceedings of the sixth international conference on knowledge capture (k-cap 2011).* Banff, Alberta, Canada: ACM.
61

Okumura, Manabu Fukusima, T., & Nanba, H. (2003). Text Summarization Challenge 2 - Text summarization evaluation at NTCIR Workshop 3. In *Proceedings of the hlt-naacl 03 text summarization workshop* (pp. 49–56). Retrieved from http://www.aclweb.org/anthology/W03-0507.pdf
69

Pareja-Lora, A. (2012). OntoLingAnnot's ontologies: Facilitating interoperable linguistic annotations (up to the pragmatic level). In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 117-127). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
31

Paulheim, H. (2013, October). DBpediaNYD, A Silver Standard Benchmark Dataset for Semantic Relatedness in DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.

70

Paulheim, H., & Ponzetto, S. P. (2013, October). Extending DBpedia with Wikipedia List Pages. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.

64

Peroni, S., & Vitali, F. (2009). Annotations with earmark for arbitrary, overlapping and out-of order markup. In U. M. Borghoff & B. Chidlovskii (Eds.), *Acm symposium on document engineering* (p. 171-180). ACM.

102

Picca, D., Gliozzo, A., & Gangemi, A. (2008). LMM: An OWL-DL metamodel to represent heterogeneous lexical knowledge. *Proceedings of LREC, Marrakech, Morocco*.

35

Poornima, S., & Good, J. (2010, July). Modeling and encoding traditional wordlists for machine applications. In *Proc. 2010 workshop on nlp and linguistics: Finding the common ground* (pp. 1–9). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W10-2101

37

Quasthoff, M., Hellmann, S., & Höffner, K. (2009). *Standardized multilingual language resources for the web of data: http://corpora.uni-leipzig.de/rdf.* Retrieved from http://triplify.org/files/challenge_2009/languageresources.pdf (3rd prize at the LOD Triplification Challenge, Graz, 2009)

xiii, 6

Quilitz, B., & Leser, U. (2008). Querying Distributed RDF Data Sources with SPARQL. In *Eswc*.

27

Rehm, G., Eckart, R., & Chiarcos, C. (2007). An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. ranlp 2007*. Borovets, Bulgaria.

130

Resnik, P., Elkiss, A., Lau, E., & Taylor, H. (2005, February). The web in theoretical linguistics research: Two case studies using the Linguist's Search Engine. In *31st meeting of the berkeley linguistics society* (p. 265-276).

129

Rizzo, G., Mendes, P., Charton, E., Hellmann, S., & Kalyanpur, A. (Eds.). (2012). *Proceedings of the Web of Linked Entities Workshop in conjunction with the 11th International Semantic Web Conference (ISWC 2012) Boston, USA, November 11, 2012.* (Vol. 906). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-906/

xi

Rizzo, G., Troncy, R., Hellmann, S., & Brümmer, M. (2012, 04). NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France.* Lyon, FRANCE. Retrieved from http://nerd.eurecom.fr/ui/paper/Rizzo_Troncy_Hellmann_Bruemmer-ldow2012.pdf
xii, 9, 66, 70, 84, 85, 86, 111, 147, 151

Rohde, D. (2005, May). *TGrep2 user manual, version 1.15* (Tech. Rep.). Cambridge, MA: MIT. Retrieved from http://tedlab.mit.edu/~dr/Tgrep2
129

Rubiera, E., Polo, L., Berrueta, D., & Ghali, A. E. (2012). Telix: An rdf-based model for linguistic annotation. In *Eswc.*
102

Sajous, F., Navarro, E., Gaume, B., Prévot, L., & Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In *Advances in natural language processing* (Vol. 6233, pp. 332–344).
50

Schalley, A. C. (2012). TYTO – A collaborative research tool for linked linguistic data. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 139-149). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
28, 32

Scheffczyk, J., Pease, A., & Ellsworth, M. (2006, November). Linking FrameNet to the Suggested Upper Merged Ontology. In *Proceedings of the fourth international conference on formal ontology in information systems (fois 2006)* (p. 289-300). Baltimore, Maryland, USA.
35, 130

Schierle, M. (2011). *Language engineering for information extraction* (PhD thesis). Universität Leipzig.
103

Schoenthal, G. (1975). *Das Passiv in der deutschen Standardsprache*. M?nchen: Hueber.
126

Sekine, S., & Nobata, C. (2004). Definition, dictionaries and tagger for extended named entity hierarchy. In A. Zampolli & M. T. Lino (Eds.), *Proceedings of the language resources and evaluation conference lrec* (pp. 1977–

1980). European Language Resources Association. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.918&rep=rep1&type=pdf doi: 10.1.1.3.91
66

Singh, S., Subramanya, A., Pereira, F., & McCallum, A. (2012). *Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia* (Tech. Rep. No. UM-CS-2012-015).
143

Sparck Jones, K. (2000). Further reflections on TREC. *Information Processing & Management*, *36*(1), 37–85. Retrieved from http://dx.doi.org/10.1016/S0306-4573(99)00044-8
67

Steinmetz, N., Knuth, M., & Sack, H. (2013, October). Statistical Analyses of Named Entity Disambiguation Benchmarks. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
66, 69, 154

Stubbs, M. (2007). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89–105 385). Rodopi.
65

Text Encoding Initiative. (1990, Nov). *TEI P1 guidelines for the encoding and interchange of machine readable texts* (Tech. Rep.). Text Encoding Initiative. http://www.tei-c.org/Vault/Vault-GL.html. (Draft Version 1.1 1)
7

Thompson, S. (1955-58). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends.* Bloomington: Indiana University Press.
31

Tobies, S. (2001). Complexity results and practical algorithms for logics in knowledge representation. *arXiv preprint cs/0106031*.
90

Turney, P. D., & Littman, M. L. (2005). Corpus-based Learning of Analogies and Semantic Relations. *Machine Learning*, *60*(1-3), 1–3. Retrieved from http://cogprints.org/4518/
67

Unbehauen, J., Hellmann, S., Auer, S., & Stadler, C. (2012). Knowledge extraction from structured sources. In S. Ceri & M. Brambilla (Eds.), *Search computing - broadening web search* (Vol. 7538, p. 34-52). Springer. Retrieved from http://svn.aksw.org/papers/2012/SearchComputing_KnowledgeExtraction/public.pdf
xiii

Unger, C., Mccrae, J., Walter, S., Winter, S., & Cimiano, P. (2013, October). A lemon lexicon for DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
65, 68

Uszkoreit, H., & Xu, F. (2013, October). From Strings to Things SAR-Graphs: A New Type of Resource for Connecting Knowledge and Language. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia* (Vol. 1064). Sydney, Australia: CEUR Workshop Proceedings.
65, 67

van Erp, M. (2012). Reusing linguistic resources: Tasks and goals for a Linked Data approach. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 57-64). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
30

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Discovering and maintaining links on the web of data. In A. Bernstein et al. (Eds.), *International semantic web conference* (Vol. 5823, p. 650-665). Springer.
62

W3C OWL Working Group. (2009). *OWL 2 Web Ontology Language. document overview. W3C Recommendation* (Tech. Rep.). World Wide Web Consortium. http://www.w3.org/TR/owl2-overview.
20

Weale, T., Brew, C., & Fosler-Lussier, E. (2009). Using the wiktionary graph structure for synonym detection. In *Proc. of the workshop on the people's web meets nlp, acl-ijcnlp*.
48

Wilcock, G. (2007, June). An OWL ontology for HPSG. In *Proc. 45th annual meeting of the association for computational linguistics* (pp. 169–172). Prague, Czech Republic.
30

Wilde, E., & Baschnagel, M. (2005). Fragment identifiers for plain text files. In *Proceedings of the sixteenth acm conference on hypertext and hypermedia* (pp. 211–213). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1083356.1083398 doi: http://doi.acm.org/10.1145/1083356.1083398
106

Wilde, E., & Duerst, M. (2008). *URI Fragment Identifiers for the text/-plain Media Type.* http://tools.ietf.org/html/rfc5147. ([Online; accessed 13-April-2011])
105

Windhouwer, M., & Wright, S. E. (2012). Linking to linguistic data categories in ISOcat. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), (p. 99-107). Heidelberg: Springer. Retrieved from http://www.springer.com/computer/ai/book/978-3-642-28248-5 (companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany) doi: 10.1007/978-3-642-28249-2
8, 27, 31

Yee, K. (1998). *Text-Search Fragment Identifiers.* http://zesty.ca/crit/draft-yee-url-textsearch-00.txt. ([Online; accessed 13-April-2011])
106

Zeldes, A., Ritz, J., Lüdeling, A., & Chiarcos, C. (2009, July). ANNIS: A search tool for multi-layer annotated corpora. In *Proc. corpus linguistics* (pp. 20–23). Liverpool, UK.
129

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Lrec.*
48

Zesch, T., Müller, C., & Gurevych, I. (2008). Using wiktionary for computing semantic relatedness. In *Aaai.*
48

SELBSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 8.1.2015

Sebastian Hellmann

Curriculum Vitae

Sebastian Hellmann

Education and Employment

since 2013 **Head of the NLP2RDF research group as sub group of AKSW**, *Universität Leipzig*.

since 2012 **WP leader of the LIDER EU project**, *AKSW, Universität Leipzig*.

since 2012 **WP leader of the LOD2 EU project**, *AKSW, Universität Leipzig*.

since 2009 **Researcher at the AKSW research group**, *Universität Leipzig*.

2002-2009 **Graduation in Computer Science**, *Universität Leipzig*.
Graduation in Computer Science with specialistation in Computational Linguistics at the University of Leipzig, Germany, title of master's thesis (Diplomarbeit): "Comparison of Concept Learning Algorithms With Emphasis on Ontology Engineering for the Semantic Web".

01-07/2006 **Erasmus in Portugal**, *Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa*.

2003–2005 **Student assistant**, *Institut für Linguistik, Universität Leipzig*.
for Prof. Steube and Prof. Bickel

2001–2002 **Civil Service**, *Deutsches Rotes Kreuz*, Fulda.

1991–2001 **Abitur**, *Freiherr-vom-Stein Gymnasium*, Fulda.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

1998–1999 **Student exchange scholarship**, *Harrison High School, Michigan*.
US High School Diploma, 1 year stay in USA, scholarship within the framework of the "Parlamentarischen Patenschafts Programm (PPP)" of the German Congress.

Organisation and Chairing of Scientific Events

Oct 2013 **NLP & DBpedia 2013**, *http://nlp-dbpedia2013.blogs.aksw.org/*, Sydney, Australia.
Proceedings of the NLP and DBpedia Workshop in conjunction with the 12th International Semantic Web Conference (ISWC 2013) Sydney, Australia, October, 2013. Sebastian Hellmann, Agata Filipowska, Caroline Barriere, Pablo Mendes, and Dimitris Kontokostas (Eds.). Volume 1064 of CEUR Workshop Proceedings,http://ceur-ws.org/Vol-1064 – 11 accepted papers by 31 authors, 29 PC members

Sep 2013 **LSWT 2013**, *http://aksw.org/Events/2013/LeipzigerSemanticWebTag.html*, 5. Leipziger Semantic Web Tag – Von Big Data zu Smart Data, Leipzig, Germany.

Sep 2013 **Tutorial Content Analysis and the Semantic Web**, *http://nlp2rdf.org/leipzig-24-9-2013*, Tutorial co-located with the LSWT 2013, Leipzig, Germany.

Nov 2012 **WoLE 2012**, *http://wole2012.eurecom.fr/*, 1st International Workshop on Web of Linked Entities 2013, Boston, USA.
Proceedings of the Web of Linked Entities Workshop in conjunction with the 11th International Semantic Web Conference (ISWC 2012) Boston, USA, November 11, 2012. Giuseppe Rizzo, Pablo Mendes, Eric Charton, Sebastian Hellmann, and Aditya Kalyanpur (Eds.). Volume 906 of CEUR Workshop Proceedings, http://ceur-ws.org/Vol-906 – 11 accepted papers by 38 authors, 37 PC members

Sep 2012 **MLODE 2012**, *http://sabre2012.infai.org/mlode*, Multilingual Linked Open Data for Enterprises 2012, Leipzig, Germany.
Multilingual Linked Open Data (MLOD) 2012 Data Post Proceedings (to appear). Sebastian Hellmann, Steven Moran, Martin Brümmer, and John McCrae (Eds.), Special Issue on Dataset Descriptions, Semantic Web Journal

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

Sep 2012 **Linked Data Cup 2012**, *http://i-challenge.blogs.aksw.org/*, Challenge, co-located with the I-SEMANTICS 2012, Graz, Austria.

Mar 2012 **LDL 2012**, *http://ldl2012.lod2.eu/*, Workshop at the 34th Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft (DGfS 2012), Frankfurt a. M., Germany.
Linked Data in Linguistics. Representing Language Data and Metadata, Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann (Eds.) Berlin; Heidelberg: Springer, (2012) companion volume of the Workshop on Linked Data in Linguistics 2012 LDL-2012, held in conjunction with the 34th Annual Meeting of the German Linguistic Society DGfS, March 2012, Frankfurt/M., Germany. `http://www.springer.com/computer/ai/book/978-3-642-28248-5` – 13 accepted papers by 34 authors, 24 PC members

Jun 2011 **OKCon 2011**, *http://2011.okcon.org*, Open Knowledge Conference 2011, Berlin, Germany.
Proceedings of the 6th Open Knowledge Conference, OKCon 2011, Berlin, Germany, June/July, 2011, Sebastian Hellmann, Philipp Frischmuth, Sören Auer, and Daniel Dietrich (Eds.). Volume 739 of CEUR Workshop Proceedings, `http://ceur-ws.org/Vol-739` – 12 accepted papers by 29 authors, 30 PC members

Jun 2011 **Open Linguistics Workshop**, *http://2011.okcon.org/2011/programme/open-linguistics-workshop*, in conjunction with the Open Knowledge Conference 2011, Berlin, Germany.

Community Service (Selection)

2009-2013 **PC Membership and Reviewing**.
International Semantic Web Conference (ISWC), Extended Semantic Web Conference (ESWC), Language Resource and Evaluation Conference (LREC), I-SEMANTICS, ISCS, KEOD, Composable Web, CLEF, AIMAS, Language Resources and Evaluation (Journal), Journal of Artificial Intelligence Research, Semantic Web Journal, Journal of Web Semantics

Presentations and Talks

May 27, 2010, 45 min. **The Semantic Gap of Formalized Meaning**, –, Research visit at UKP Darmstadt lead by Iryna Gurevych, Darmstadt, Germany.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

May 31, 2010, 30 min. **The Semantic Gap of Formalized Meaning**, *http://www.eswc2010.org/program-menu/phd-symposium.html*, PhD Symposium at ESWC 2010, Heraklion, Greece.

Jun 30, 2011, 20 min. **NIF: NLP Interchange Format**, *http://2011.okcon.org/2011/programme/open-linguistics-workshop*, Open Linguistics Workshop, OKCon 2012, Berlin, Germany.
http://www.slideshare.net/kurzum/nif-nlp-interchange-format

Aug 16, 2011, 40 min. **DBpedia – Extraction of Knowledge from Wikipedia**, *http://semanticweb.kaist.ac.kr/workshop2011/schedule.html*, Korea – Germany Joint Workshop for LOD2 Development and Application, Daejeon, Korea.
http://semanticweb.kaist.ac.kr/workshop2011/presentation/3_dbpedia_sebastian.pdf

Sep 14, 2011, 90 min. **Text Mining and Annotation on the Data Web**, *http://lod2.eu/Article/ISSLOD2011.html*, Indian-summer school on Linked Data, Leipzig, Germany.
with Pablo Mendes

Sep 21, 2011, 10 min. **NLP Interchange Format (NIF)**, *http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-program*, A Local Focus for the Multilingual Web - W3C Workshop, Limerick, Ireland.
http://www.w3.org/International/multilingualweb/limerick/slides/hellmann.pdf

Oct 23, 2011, 60 min. **NLP Interchange Format (NIF) - Version 1.0**, *http://msw2.deri.ie*, 2nd Workshop on the Multilingual Semantic Web, Bonn, Germany.
Invited Talk – http://msw2.deri.ie/sites/default/files/presentations/1400.pdf

May 24, 2012, 45 min. **The Web of Data: Decentralized, collaborative, interlinked and interoperable**, *http://www.lrec-conf.org/lrec2012/?Keynote-Speeches-and-Invited-Talk*, LREC 2012 Keynote by Sören Auer, Istanbul, Turkey.
http://www.lrec-conf.org/proceedings/lrec2012/keynotes/LREC%202012.Keynote%20Speech%201.Soeren%20Auer.pdf – the keynote was given by Sören Auer

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

Jun 11, 2012, 10 min. **Linked Data in Linguistics for NLP and Web Annotation**, *http://www.multilingualweb.eu/en/documents/dublin-workshop/dublin-program*, The Multilingual Web – Linked Open Data and MultilingualWeb-LT Requirements, Dublin, Ireland.
http://www.w3.org/International/multilingualweb/dublin/slides/06-hellmann.pdf

Jun 14, 2012, 45 min. **A Transparent Formalization of Text for Machines**, *http://www.computing.dcu.ie/videos/sebastian-hellman-dclrs-seminar-14th-june-2012*, Research Visit at CNGL, DCLRS Seminar, Dublin City University, Dublin, Ireland.
http://de.slideshare.net/kurzum/thesis-presentation-11928355

Sep 02, 2012, 15 min. **4 Challenges for Semantic Web, NLP and Web Annotation**, *http://www.dagstuhl.de/12362*, The Multilingual Semantic Web, Dagstuhl, Germany.
http://de.slideshare.net/kurzum/hellmann-dagstuhl-seminar

Sep 18, 2012, 15 min. **Annotation in NLP and Linguistics**, *http://www.w3.org/community/openannotation/wiki/September_2012,_Chicago*, Open Annotation Meeting of W3C Community Group, Chicago, USA.

Oct 10, 2012, 15 min. **Linked-Data Aware URI Schemes for Referencing Text Fragment**, *http://ekaw2012.ekaw.org/*, EKAW 2012, Galway, Ireland.
http://www.youtube.com/watch?v=-M8Ozr1QWo0&list=PLA_PbqRGYBCFApiBWxVs_r3quMVjzhlxb&index=4

Mar 13, 2013, 15 min. **The LOD2 Stack and the NLP2RDF Project**, *http://www.multilingualweb.eu/documents/rome-workshop/rome-program*, Making the Multilingual Web Work, W3C Workshop, Rome, Italy.
http://www.w3.org/International/multilingualweb/rome/slides/23-hellman.pdf

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

Sep 24, 2013, 90 min. **NIF 2.0 Tutorial: Content Analysis and the Semantic Web**, *http://nlp2rdf.org/leipzig-24-9-2013*, LSWT Tutorial: Content Analysis and the Semantic Web, Leipzig, Germany.
http://de.slideshare.net/kurzum/nif-20-tutorial-content-analysis-and-the-semantic-web

Oct 06, 2013, 20 min. **Why Open Data Should Be Linked Open Data**, *http://www.meetup.com/OpenKnowledgeFoundation/Open-data-Prague/1017292/*, Keynote for the Prague Open Data Meetup #7: Linked Open Cities, Prague, Czech Republic.
http://de.slideshare.net/osf/sebastian-hellman-why-open-data-should-be-openlinkeddata

Oct 23, 2013, 15 min. **Integrating NLP using Linked Data**, *http://iswc2013.semanticweb.org/content/program-wednesday#nlp*, International Semantic Web Conference 2013, In Use Tack, Sydney, Australia.
http://de.slideshare.net/kurzum/integrating-nlp-using-linked-data

Publications

[1] Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. Triplify: Light-weight linked data publication from relational databases. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 621–630. ACM, 2009.

[2] Sören Auer and Sebastian Hellmann. The web of data: Decentralized, collaborative, interlinked and interoperable. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). (companion publication for the Keynote at LREC 2012).

[3] Sören Auer, Jens Lehmann, and Sebastian Hellmann. Linked-

GeoData - adding a spatial dimension to the web of data. In *Proc. of 8th International Semantic Web Conference (ISWC)*, 2009.

[4] Didier Cherix, Sebastian Hellmann, and Jens Lehmann. Improving the performance of a sparql component for semantic web applications. In *JIST*, 2012.

[5] Christian Chiarcos and Sebastian Hellmann. Working group for open data in linguistics: Status quo and perspectives. In *Proceedings of the Open Knowledge Conference in 2011*. Open Knowledge Foundation, June 2011.

[6] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a linguistic linked open data cloud : The open linguistics working group. *TAL*, 52(3):245 – 275, 2011.

[7] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. The open linguistics working group of the open knowledge foundation. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics*, pages 153–160. Springer Berlin Heidelberg, 2012.

[8] Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. The open linguistics working group. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3603–3610, May 2012.

[9] Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. *Linked Data in Linguistics. Representing Language Data and Metadata*. Berlin; Heidelberg: Springer, 2012. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

[10] Sebastian Dietzold, Sebastian Hellmann, and Martin Peklo. Using javascript rdfa widgets for model/view separation inside

read/write websites. In *Proceedings of the 4th Workshop on Scripting for the Semantic Web*, 2008.

[11] Daniel Gerber, Axel-Cyrille Ngonga Ngomo, Sebastian Hellmann, Tommaso Soru, Lorenz Bühmann, and Ricardo Usbeck. Real-time rdf extraction from unstructured data streams. In *Proceedings of ISWC*, 2013.

[12] Philipp Heim, Sebastian Hellmann, Jens Lehmann, Steffen Lohmann, and Timo Stegemann. RelFinder: Revealing relationships in RDF knowledge bases. In *Proceedings of the 3rd International Conference on Semantic and Media Technologies (SAMT)*, volume 5887 of *Lecture Notes in Computer Science*, pages 182–187. Springer, 2009.

[13] Sebastian Hellmann. Comparison of concept learning algorithms with emphasis on ontology engineering for the semantic web. Diploma thesis, University of Leipzig, 2008.

[14] Sebastian Hellmann. The semantic gap of formalized meaning. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *ESWC (2)*, volume 6089 of *Lecture Notes in Computer Science*, pages 462–466. Springer, 2010.

[15] Sebastian Hellmann and Sören Auer. Towards web-scale collaborative knowledge extraction. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 287–313. Springer Berlin Heidelberg, 2013.

[16] Sebastian Hellmann, Jonas Brekle, and Sören Auer. Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In *JIST*, 2012.

[17] Sebastian Hellmann, Agata Filipowska, Caroline Barriere, Pablo Mendes, and Dimitris Kontokostas, editors. *Proceedings of the NLP and DBpedia Workshop in conjunction with the 12th International Semantic Web Conference (ISWC 2013) Sydney,*

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

Australia, October, 2013., volume 1064 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[18] Sebastian Hellmann, Agata Filipowska, Caroline Barriere, Pablo N. Mendes, and Dimitris Kontokostas. NLP & DBpedia - An Upward Knowledge Acquisition Spiral. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.

[19] Sebastian Hellmann, Philipp Frischmuth, Sören Auer, and Daniel Dietrich, editors. *Proceedings of the 6th Open Knowledge Conference, OKCon 2011, Berlin, Germany, June/July, 2011*, volume 739 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.

[20] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Learning of OWL class descriptions on very large knowledge bases. In Christian Bizer and Anupam Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[21] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Learning of OWL class descriptions on very large knowledge bases. *International Journal on Semantic Web and Information Systems*, 5(2):25–48, 2009.

[22] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Learning of owl class expressions on very large knowledge bases and its applications. In Interoperability Semantic Services and Web Applications: Emerging Concepts, editors, *Learning of OWL Class Expressions on Very Large Knowledge Bases and its Applications*, chapter 5, pages 104–130. IGI Global, 2011.

[23] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Linked-data aware uri schemes for referencing text fragments. In *EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

[24] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*, 2013.

[25] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. Nif combinator: Combining nlp tool output. In *EKAW*, pages 446–449, 2012.

[26] Sebastian Hellmann, Jens Lehmann, Jörg Unbehauen, Claus Stadler, Thanh Nghia Lam, and Markus Strohmaier. Navigation-induced knowledge engineering by example. In *JIST*, 2012.

[27] Sebastian Hellmann, Steven Moran, Martin Brümmer, and John McCrae, editors. *Multilingual Linked Open Data (MLOD) 2012 Data Post Proceedings*, volume Special Issue on Dataset Descriptions, to appear. to appear.

[28] Sebastian Hellmann, Claus Stadler, and Jens Lehmann. The german dbpedia: A sense repository for linking entities. In Chiarcos et al. [9], pages XIV, 216 S. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

[29] Sebastian Hellmann, Claus Stadler, Jens Lehmann, and Sören Auer. DBpedia live extraction. In *Proc. of 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE)*, volume 5871 of *Lecture Notes in Computer Science*, pages 1209–1223, 2009.

[30] Sebastian Hellmann, Jörg Unbehauen, Christian Chiarcos, and Axel-Cyrille Ngonga Ngomo. The tiger corpus navigator. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, NEALT Proceeding Series, 2010.

[31] Sebastian Hellmann, Jörg Unbehauen, and Jens Lehmann. Hanne - a holistic application for navigational knowledge engineering. In *Posters and Demos of ISWC 2010*, 2010.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

[32] Alex Klebeck, Sebastian Hellmann, Christian Ehrlich, and Sören Auer. Ontosfeeder – a versatile semantic context provider for web content authoring. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 456–460. Springer Berlin / Heidelberg, 2011.

[33] Dimitris Kontokostas, Charalampos Bratsas, Sören Auer, Sebastian Hellmann, Ioannis Antoniou, and George Metakides. Towards linked data internationalization - realizing the greek dbpedia. In *Proceedings of the ACM WebSci'11*, 2011.

[34] Dimitris Kontokostas, Charalampos Bratsas, Sören Auer, Sebastian Hellmann, Ioannis Antoniou, and George Metakides. Internationalization of linked data: The case of the greek dbpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15(0):51 – 61, 2012.

[35] Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.

[36] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27, 2012.

[37] Matthias Quasthoff, Sebastian Hellmann, and Konrad Höffner. Standardized multilingual language resources for the web of data: http://corpora.uni-leipzig.de/rdf, 2009. 3rd prize at the LOD Triplification Challenge, Graz, 2009.

[38] Giuseppe Rizzo, Pablo Mendes, Eric Charton, Sebastian Hellmann, and Aditya Kalyanpur, editors. *Proceedings of the Web of Linked Entities Workshop in conjunction with the 11th International Semantic Web Conference (ISWC 2012) Boston, USA, November 11, 2012.*, volume 906 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

[39] Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Brümmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France*, Lyon, FRANCE, 04 2012.

[40] Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. N3 - a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. In *Submitted to The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland*, 2014.

[41] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau Jr, Sören Auer, Juan Sequeda, and Ahmed Ezzat. A survey of current approaches for mapping of relational databases to rdf, 01 2009.

[42] Saeedeh Shekarpour, Sören Auer, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Sebastian Hellmann, and Claus Stadler. Keyword-driven sparql query generation leveraging background knowledge. In *International Conference on Web Intelligence*, 2011.

[43] Saeedeh Shekarpour, Sören Auer, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Sebastian Hellmann, and Claus Stadler. Generating sparql queries using templates. In *WIAS journal,Vol. 11, No. 3, 2013.*, 2013.

[44] Mohamed A. Sherif, Sandro Coelho, Ricardo Usbeck, Sebastian Hellmann, Jens Lehmann, Martin Brümmer, and Andreas Both. Nif4oggd - nlp interchange format for open german governmental data. In *Submitted to The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland*, 2014.

[45] Claus Stadler, Michael Martin, Jens Lehmann, and Sebastian Hellmann. Update Strategies for DBpedia Live. In *6th Workshop on Scripting and Development for the Semantic Web Colocated with ESWC 2010 30th or 31st May, 2010 Crete, Greece*, 2010.

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

[46] Sebastian Tramp, Philipp Frischmuth, Timofey Ermilov, Saeedeh Shekarpour, and Sören Auer. An Architecture of a Distributed Semantic Social Network. *Semantic Web Journal*, Special Issue on The Personal and Social Semantic Web, 2012.

[47] Jörg Unbehauen, Sebastian Hellmann, Sören Auer, and Claus Stadler. Knowledge extraction from structured sources. In Stefano Ceri and Marco Brambilla, editors, *Search Computing - Broadening Web Search*, volume 7538 of *Lecture Notes in Computer Science*, pages 34–52. Springer, 2012.

[48] Jörg Unbehauen, Sebastian Hellmann, Michael Martin, Sebastian Dietzold, and Sören Auer. xoperator - chat with the semantic web, 2008. Poster @ the ISWC 2008.

[49] Wikipedia. Knowledge extraction — Wikipedia, the free encyclopedia, 2011. [Online; accessed 10-August-2011; The article was bootstrapped by the LOD2 Project].

✉ *hellmann@informatik.uni-leipzig.de*
http://bis.informatik.uni-leipzig.de/SebastianHellmann
Mar 14, 1981 in Göttingen

LIST OF TABLES

LIST OF FIGURES

LISTINGS