THESIS SUMMARY

Title: *Integrating Natural Language Processing (NLP) and Language Resources Using Linked Data*
Author: *Sebastian Hellmann*
Bib. Data: *2013, XX, 195p.*
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 thesis are summarized here:

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.