Towards Web-Scale Collaborative Knowledge Extraction

Sebastian Hellmann, Sören Auer

Abstract While the Web of Data, the Web of Documents and Natural Language Processing are well researched individual fields, approaches to combine all three are fragmented and not yet well aligned. This chapter analyzes current efforts in collaborative knowledge extraction to uncover connection points between the three fields. The special focus is on three prominent RDF data sets (DBpedia, LinkedGeoData and Wiktionary2RDF), which allow users to influence the knowledge extraction process by adding another crowd-sourced layer on top. The recently published NLP Interchange Format (NIF) provides a way to annotate textual resources on the Web through the assignment of URIs with fragment identifiers. We will show how this formalism can easily be extended to encompass new annotation layers and vocabularies.

1 Introduction

The vision of the Giant Global Graph[1] 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 now 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. [5] argues that

Sebastian Hellmann
AKSW, Universität Leipzig, e-mail: hellmann@informatik.uni-leipzig.de

Sören Auer
AKSW, Universität Leipzig, e-mail: auer@informatik.uni-leipzig.de

[1] http://dig.csail.mit.edu/breadcrumbs/node/215

[2] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

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 (cf. [1, 4, 5]) in Figure 1 and discuss these in more detail throughout this article.

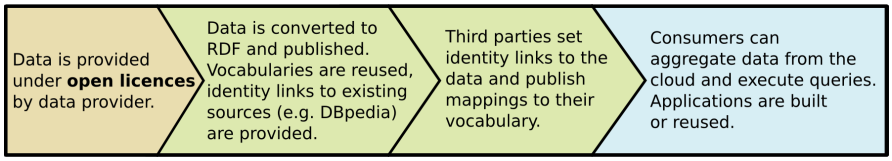

Fig. 1 Summary of the above-mentioned methodologies for publishing and exploiting Linked Data [10]. The data provider is only required to make data available under an open license (leftmost step). The remaining, data integration steps can be contributed by third parties and data consumers.

Natural Language Processing

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 *Zemanta*[3], *OpenCalais*[4], *Ontos*[5], *Enrycher*[6], *Extractiv*[7], *Alchemy API*[8]. 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 com-

[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/`

mon NLP tasks. Some first evidence in that direction is provided by tools such as *RDFaCE* [20], *Spotlight* and *Fox*,[9] 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 [1]. 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. 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: *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.

NLP Interchange Format

In order to simplify the combination of tools, improve their interoperability and facilitating the use of Linked Data, we developed the NLP Interchange Format (NIF). 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 ontology as REST services (access layer). Other than more centralized solutions such as *UIMA* and *GATE*, NIF 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. Ultimately, we envision an ecosystem of NLP tools and services to emerge using NIF for exchanging and integrating rich annotations.

The remainder of this article is structured as follows: In the next section, we will take up the cudgels on behalf of open licenses and RDF and give relevant background information and facts about the used technologies and the current state of the Web of Data. We will especially elaborate on the following aspects: The importance of open licenses and open access as an enabler for collaboration; the ability

[9] `http://aksw.org/Projects/FOX`

to interlink data on the Web as a key feature of RDF; a discussion about scalability and decentralization; as well as an introduction on how conceptual interoperability can be achieved by (1) re-using vocabularies and (2) agile ontology development (3) meetings to refine and adapt ontologies (4) tool support to enrich ontologies and match schemata. In Section 3, we will describe three data sets that were created by a knowledge extraction process and maintained collaboratively by a community of stakeholders. Especially, we will focus on DBpedia's[10] *Mappings Wiki*[11] (which governs the extraction from Wikipedia), the mapping approach of *LinkedGeoData*[12] (extracted from OpenStreetMaps) and the configurable extraction of RDF from Wiktionary. While Section 4 introduces key concepts of the NLP Interchange Format (NIF), Section 5 shows how to achieve interoperability between NIF and existing annotation ontologies which are modelling different layers of NLP annotations. Section 5, also shows how extensions of NIF have the potential to connect the Giant Global Graph (especially the resources introduced in Section 3), the Web of Documents and NLP tool output. The article concludes with a short discussion and an outlook on future work in Section 6.

2 Background

2.1 Open licenses, open access and collaboration

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[13]. DBpedia [23], for example, was initially converted to RDF solely from the openly licensed database dumps provided by Wikipedia. With 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) [28] and the RDF data used in the TIGER Corpus Navigator [13]. 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

[10] `http://dbpedia.org`

[11] `http://mappings.dbpedia.org/`

[12] `http://linkedgeodata.org`

[13] More data sets can be explored here: `http://thedatahub.org/tag/published-by-third-party`

unable to publish their annotations and resulting data. The open part of the American National Corpus (OANC[14]) on the other hand has been converted to RDF and was re-published successfully using the POWLA ontology [9]. 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.[15] Only around 18% of the data sets of the LOD cloud provide clear licensing information at all.[16] Of these 18% an even smaller amount is considered *open* in the sense of the open definition[17] 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. Very often self-made licenses allow scientific and non-commercial use, but do not specify how redistribution is handled.

2.2 RDF as a data model

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.

[14] http://www.anc.org/OANC/

[15] http://richard.cyganiak.de/2007/10/lod/#open

[16] http://www4.wiwiss.fu-berlin.de/lodcloud/state/#license

[17] http://opendefinition.org/

2.3 Performance and scalability

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 triples stores, inferencing by OWL reasoners and of course the combination thereof. Although the scalability is a constant focus of RDF data management research[18], 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 [2]. The NLP Interchange Format (cf. Section 4) 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*[19] or *Sindice*[20] 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.

2.4 Conceptual interoperability

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 [12], 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 [16]. In 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. Espe-

[18] `http://factforge.net` or `http://lod.openlinksw.com` provide SPARQL interfaces to query billions of aggregated facts.

[19] `http://swoogle.umbc.edu`

[20] `http://sindice.com`

cially 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 [2], 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*[21]. 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 [7] allow to extend ontologies based on the entity-level data .

3 Collaborative Knowledge Extraction

Knowledge Extraction is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources. The resulting knowledge needs to be in a machine-readable and machine-interpretable format and must represent knowledge in a manner that unambiguously defines its meaning and facilitates inferencing [31]. By this definition, almost all RDF/OWL knowledge bases that where created from "legacy" sources can be considered as being created by a knowledge extraction process. In this section, we will focus on three prominent knowledge bases that fall in this category: *DBpedia*, *LinkedGeoData* and *Wiktionary2RDF*. The crowd-sourcing process that yielded these knowledge bases stretched over different stages of their development process:
- All three knowledge bases originate from crowd-sourced wiki approaches, i.e. *Wikipedia*, *OpenStreetMaps* and *Wiktionary*.
- The knowledge extraction process itself is crowd-sourced: (1) DBpedia provides a mappings wiki, which allows to define extraction rules on Wikipedia's infoboxes; (2) LinkedGeoData provides a mapping XML file from terms occurring in OpenStreetMaps to RDF properties; (3) Wiktionary2RDF allows domain experts to create and maintain wrappers for language-specific Wiktionary editions

[21] `http://mappings.dbpedia.org`

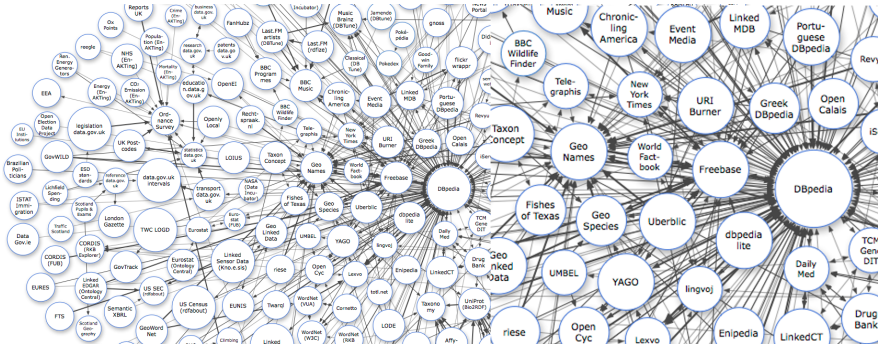

Fig. 2 Excerpt of the data sets interlinked with DBpedia. Source: `http://lod-cloud.net` (with kind permission of Anja Jentzsch and Richard Cyganiak).

- Each project has a mailing list and a bug tracker, where data consumers can report bugs and discuss modelling issues. Occasionally, patches are directly provided by the community.
- Third parties have provided link sets for inclusion into the data set itself (e.g. DBpedia contains links to Yago, WordNet, Umbel).
- Third parties publish links into one of the projects alongside their own data sets, as can be seen on the LOD cloud image.

Due to continuous reviewing by a large community of stakeholders, DBpedia has evolved into a paragon of best practices for linked data. The same accounts to a lesser extent for LinkedGeoData and Wiktionary2RDF as both projects are much younger.

3.1 DBpedia

DBpedia [23] is a community effort to extract structured information from Wikipedia and to make this information available on the Web. The main output of the DBpedia project is a data pool that (1) is widely used in academics as well as industrial environments, that (2) is curated by the community of Wikipedia and DBpedia editors, and that (3) has become a major crystallization point and a vital infrastructure for the Web of Data. DBpedia is one of the most prominent Linked Data examples and presently the largest hub in the Web of Linked Data (Figure 2). The extracted RDF knowledge from the English Wikipedia is published and interlinked according to the Linked Data principles and made available under the same license as Wikipedia (cc-by-sa).

In its current version 3.8 DBpedia contains more than 3.77 million things, of which 2.35 million are classified in a consistent ontology, including 764,000 persons, 573,000 places, 112,000 music albums, 72,000 films, 18,000 video games, 192,000 organizations, 202,000 species and 5,500 diseases. The DBpedia data set

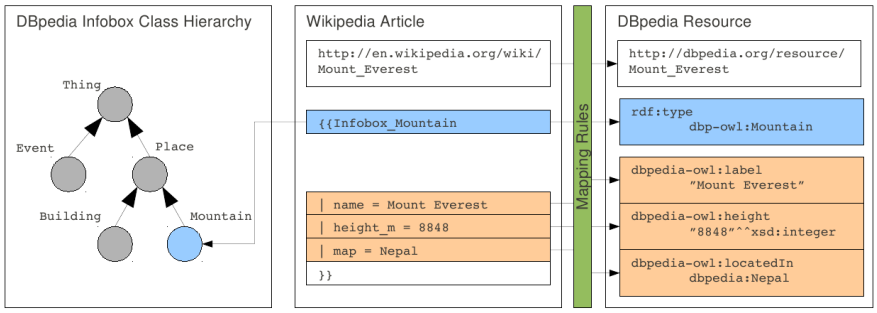

Fig. 3 Rule-based manipulation of extracted data in DBpedia Mappings Wiki [15].

features labels and abstracts in up to 111 different languages; 8.0 million links to images and 24.4 million links to external Web pages; 27.2 million data links into other RDF datasets, and 55.8 million Wikipedia categories. The dataset consists of 1.89 billion RDF triples out of which 400 million were extracted from the English edition of Wikipedia and 1.46 billion were extracted from other Wikipedia language editions and around 27 million links to external datasets [6].

Currently, the DBpedia Ontology is maintained in a crowd-sourcing approach and thus freely editable on a *Mappings Wiki*:[22] each OWL class can be modeled on a Wiki page and the `subClassOf` axioms (shown on the left side of Figure 3) are created manually. The classification of articles according to the ontology classes is based on rules. In Figure 3, the article is classified as `dbp-owl:Mountain`, because it contains the Infobox "Infobox_Mountain" in its source.

3.1.1 Internationalization of DBpedia

While early versions of the DBpedia Information Extraction Framework (DIEF) used only the English Wikipedia as their sole source, its focus later shifted integrate information from many different Wikipedia editions. During the fusion process, however, language-specific information was lost or ignored. The aim of the current research in internationalization [21, 22] is to establish best practices (complemented by software) that allow the DBpedia community to easily generate, maintain and properly interlink language-specific DBpedia editions. In a first step, we realized a language-specific DBpedia version using the Greek Wikipedia [21]. Soon, the approach was generalized and applied to 15 other Wikipedia language editions [6]

[22] `http://mappings.dbpedia.org`

3.1.2 DBpedia as a sense repository and interlinking hub for common entities

DBpedia data can be directly exploited for NLP and linguistic applications, e.g. NLP processing pipelines and the linking of linguistic concepts to their encyclopedic counterparts. Most importantly, DBpedia provides background knowledge for around 3.77 million entities with highly stable identifier-to-sense assignment [17]: Once an entity or a piece of text is correctly linked to its DBpedia identifier, it can be expected that this assignment remains correct over time. DBpedia provides a number of relevant features and incentives which are highly beneficial for NLP processes: 1. the senses are curated in a crowd-sourced community process and remain stable; 2. Wikipedia is available in multiple languages; 3. data in Wikipedia and DBpedia[23] remains up-to-date and users can influence the knowledge extraction process in the Mappings Wiki; 4. the open licensing model allows all contributors to freely exploit their work.

Note that most of the above-mentioned properties are inherited from Wikipedia. The additional benefit added by DBpedia is the standardization and re-usability of the data for NLP developers. Especially, the community around DBpedia Spotlight has specialized in providing datasets refined from DBpedia that are directly tailored towards NLP processes [26].

3.1.3 DBpedia Spotlight

The band-width of applications of DBpedia data in NLP research is immense, but here, we focus on a single example application, DBpedia Spotlight by [25], a tool for annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. DBpedia Spotlight performs named-entity extraction, including entity detection and Name Resolution. Several strategies are used to generate candidate sets and automatically select a resource based on the context of the input text.

The most basic candidate generation strategy in DBpedia Spotlight is based on a dictionary of known DBpedia resource names extracted from page titles, redirects and disambiguation pages. These names are shared in the DBpedia Lexicalization dataset.[24] The graph of labels, redirects and disambiguations in DBpedia is used to extract a lexicon that associates multiple surface forms to a resource and interconnects multiple resources to an ambiguous name. One recent development is the internationalization of DBpedia Spotlight, and the development of entity disambiguation services for German and Korean has begun. Other languages will follow soon including the evaluation of the performance of the algorithms in other languages.

[23] For DBpedia Live see `http://live.dbpedia.org/`

[24] `http://wiki.dbpedia.org/Lexicalizations`

3.2 LinkedGeoData

With the *OpenStreetMap* (OSM)[25] project, a rich source of spatial data is freely available. It is currently used primarily for rendering various map visualizations, but has the potential to evolve into a crystallization point for spatial Web data integration (e.g. as gazetteer for NLP applications focusing on recognition of spatial entities). The goal of the *LinkedGeoData* (LGD) [30] project is to lift OSM's data into the Semantic Web infrastructure. This simplifies real-life information integration and aggregation tasks that require comprehensive background knowledge related to spatial features. Such tasks might include, for example, to locally depict the offerings of the bakery shop next door, to map distributed branches of a company, or to integrate information about historical sights along a bicycle track.

The majority of LGD data, which comprises 15 billion spatial facts, is obtained by converting data from the popular OpenStreetMap community project to RDF and deriving a lightweight ontology from it. Furthermore, interlinking is performed with *DBpedia*, *GeoNames* and other datasets as well as the integration of icons and multilingual class labels from various sources. As a side effect, LGD is striving for the establishment of an OWL vocabulary with the purpose of simplifying exchange and reuse of geographic data. Besides coarse-grained spatial entities such as countries, cities and roads LGD also contains millions of buildings, parking lots, hamlets, restaurants, schools, fountains or recycling trash bins. Since the initial LGD release in [3], a substantial effort was invested in maintaining and improving LinkedGeoData, which includes improvements of the project infrastructure, the generated ontology, and data quality in general. To date, the LinkedGeoData project comprises in particular:

- A flexible system for mapping OpenStreetMap data to RDF including support for nice URIs (camel case), typed literals, language tags, and a mapping of the OSM data to classes and properties.
- Support for ways: Ways are OpenStreetMap entities used for modelling things such as streets but also areas. The geometry of a way (a line or a polygon) is stored in a literal of the corresponding RDF resource, which makes it easy to e.g. display such a resource on a map. Furthermore, all nodes referenced by a way are available both via the Linked Data interface and the SPARQL endpoints.
- A *REST interface* with integrated search functions as well as a publicly accessible *live SPARQL endpoint* that is being interactively updated with the minutely changesets that OpenStreetMap publishes.
- A simple *republication method* of the corresponding RDF changesets so that LinkedGeoData data consumers can replicate the LinkedGeoData store.
- Direct *interlinking* with *DBpedia*, *GeoNames* and the *UN FAO* data. Integration of appropriate *icons and multilingual labels* for LinkedGeoData ontology elements from external sources.

[25] `http://openstreetmap.org`

- The spatial-semantic user interface *LinkedGeoData browser* as well as the *Vicibit* application to facilitate the integration of LGD facet views in external web pages.

In essence, the transformation and publication of the OpenStreetMap data according to the Linked Data principles in LinkedGeoData adds a new dimension to the Data Web: spatial data can be retrieved and interlinked on an unprecedented level of granularity. For NLP applications, the LinkedGeoData resource opens possibilities previously hardly thinkable. For example, entity references in text such as 'the bakery on Broad Street' can possibly be resolved by using the vast knowledge comprised in LGD's 15 billion spatial facts.

3.3 Wiktionary2RDF

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[26]), containing information about hundreds of spoken and even ancient languages. For example, the English *Wiktionary* contains nearly 3 million words[27]. 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. [27] 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 corpora (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 (ELE) 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

[26] `http://s23.org/wikistats/wiktionaries_html.php`

[27] See `http://en.wiktionary.org/wiki/semantic` for a simple example page

showed, that only a fraction of the available data is extracted 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 straight-forward approach (implementing software adapters for scraping), Wiktionary2RDF employs 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 "entry layout explained" (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.[28] 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.

In order to conceive a flexible, effective and efficient solution, we survey in this section the challenges associated with Wiki syntax, *Wiktionary* and large-scale extraction.

3.3.1 Processing Wiki Syntax

Pages in *Wiktionary* are formatted using the *wikitext* markup language[29]. 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[30] instead, is convenient as we could reuse the DBpedia extraction framework[31] 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.

[28] `http://wiktionary.dbpedia.org/`

[29] `http://www.mediawiki.org/wiki/Markup_spec`

[30] `http://dumps.wikimedia.org/backup-index.html`

[31] `http://wiki.dbpedia.org/Documentation`

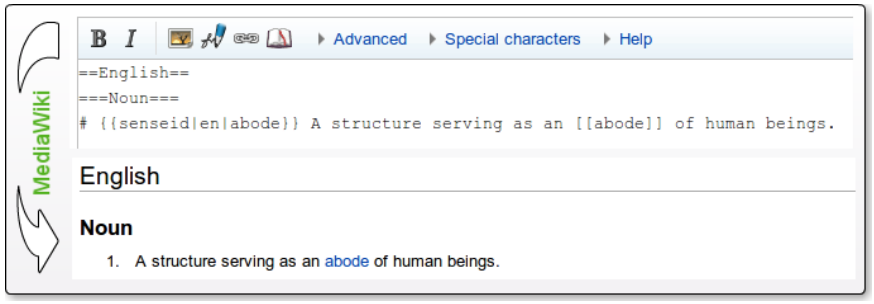

Fig. 4 An excerpt of the *Wiktionary* page *house* with the rendered HTML.

An excerpt of the *Wiktionary* page *house* and the resulting rendered page are shown in Figure 4.

The markup == is used to denote headings, # denotes a numbered list ($\star$ 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 4, the `senseid` template[32] 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.

3.3.2 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

[32] `http://en.wiktionary.org/wiki/Template:senseid`

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[33]).

- **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*[34] as well as the *GNU Free Documentation License (GFDL)*.[35] All the data extracted by our approach falls under the same licenses.
- **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)[36]. 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.[37]

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

---

[33] http://stats.wikimedia.org/wiktionary/EN/TablesWikipediaEN.htm

[34] http://en.wiktionary.org/wiki/Wiktionary:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

[35] http://en.wiktionary.org/wiki/Wiktionary:GNU_Free_Documentation_License

[36] http://meta.wikimedia.org/wiki/Template:Wikimedia_Growth

[37] http://stats.wikimedia.org/wiktionary/EN/TablesWikipediaEN.htm

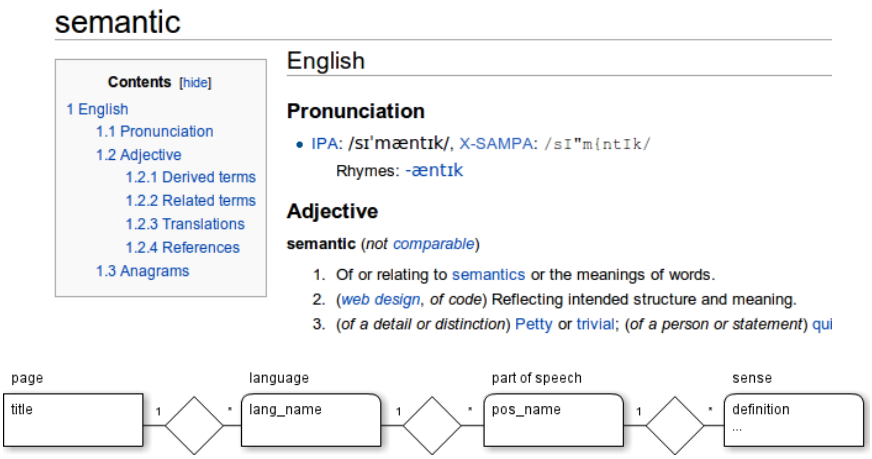

Fig. 5 Example page `http://en.wiktionary.org/wiki/semantic` and underlying schema (only valid for the English *Wiktionary*, other WLE might look very different.)

for this tree is restricted by the ELE guidelines[38]. The entities illustrated in Figure 5 of the ER diagram will be called *block* from now on. The schema can differ between WLEs and normally evolves over time.

3.3.3 Wiki-scale Data Extraction

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

3.3.4 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

[38] For English see `http://en.wiktionary.org/wiki/Wiktionary:ELE`

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 1 Statistical comparison of extractions for different languages. XML lines measures the number of lines of the XML configuration files

more than 80 million facts. The data is available as N-Triples dumps[39], Linked Data[40], via the *Virtuoso Faceted Browser*[41] or a SPARQL endpoint[42]. Table 1 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 one of 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.

Community Process. For each of the languages, a configuration XML file was created, which describes how the Wiktionary2RDF framework should transform the Wiki syntax into triples. Existing configuration files are public and can be altered by everybody without touching the source code of the project and patches can be submitted back into the project. Additionally, they serve as templates to aid creation of config files for more languages by a community. We can identify three sources for low data quality during the extraction process: 1. An error or missing feature in the extraction algorithm of the software framework 2. An erroneous or incomplete configuration file 3. a *Wiktionary* page that does not adhere to the ELE guidelines. While the Wiktionary2RDF project requires a developer for the first point, two and three can be fixed by domain experts and *Wiktionary* users. Providing a live extraction, similar to DBpedia also has the potential to become a great supportive resource to help editors of *Wiktionary* in spotting inconsistencies.

4 The NLP Interchange Format

The motivation behind NIF is to allow NLP tools to exchange annotations about documents in RDF. Hence, the main prerequisite is that parts of the documents (i.e. strings) are referenceable by URIs, so that they can be used as subjects in RDF statements. We call an algorithm to create such identifiers *URI Scheme*: For a given text t (a sequence of characters) of length $|t|$ (number of characters), we are looking for a *URI Scheme* to create a URI, that can serve as a *unique* identifier for a substring

[39] `http://downloads.dbpedia.org/wiktionary`

[40] for example `http://wiktionary.dbpedia.org/resource/dog`

[41] `http://wiktionary.dbpedia.org/fct`

[42] `http://wiktionary.dbpedia.org/sparql`

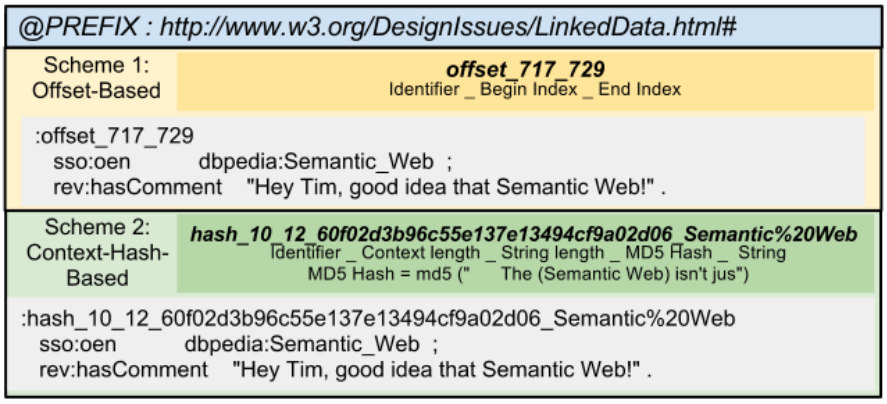

Fig. 6 NIF URI schemes: Offset (top) and context-hashes (bottom) are used to create identifiers for strings [14]

s of t (i.e. $|s| \leq |t|$). Such a substring can (1) consist of adjacent characters only and it is therefore a unique character sequence within the text, if we account for parameters such as context and position or (2) derived by a function which points to several substrings as defined in (1).

NIF provides two URI schemes, which can be used to represent strings as RDF resources. In this section, we focus on the first scheme using offsets. In the top part of Figure 6, two triples are given that use the following URI as subject:

`http://www.w3.org/DesignIssues/LinkedData.html#offset_717_729`

According to the above definition, the URI points to a substring of a given text t, which starts at index 717 until the index 729.

For the URI creation scheme, there are three basic requirements – *uniqueness*, *ease of implementation* and *URI stability* during document changes. Since these three conflicting requirements can not be easily addressed by a single URI creation scheme, NIF defines two URI schemes, which can be chosen depending on which requirement is more important in a certain usage scenario. Naturally further schemes for more specific use cases can be developed easily. After discussing some guidelines on the selection of URI namespaces, we explain in this section how stable URIs can be minted for parts of documents by using *offset-based* and *context-hash* based schemes (see Figure 6 for examples).

4.1 Namespace Prefixes

A NIF URI is constructed from a namespace prefix and the actual *identifier* (e.g. “offset_717_729“). Depending on the selected context, different prefixes can be chosen. For practical reasons, it is recommended that the following guidelines should be met for NIF URIs: If we want to annotate a (web) resources, the whole content of

the document is considered as `str:Context`, as explained in the next section, and it is straightforward to use the existing document URL as the basis for the prefix. The prefix should then either end with slash ('/') or hash ('#')[43].

Recommended prefixes for `http://www.w3.org/DesignIssues/LinkedData.html` are:
– `http://www.w3.org/DesignIssues/LinkedData.html/`
– `http://www.w3.org/DesignIssues/LinkedData.html#`

4.2 Offset-based URIs

The offset-based URI scheme focuses on ease of implementation and is compatible with the position and range definition of RFC 5147 by [32] (esp. Section 2.1.1) and builds upon it in terms of encoding and counting character positions (See [14] for a discussion). Offset-based URIs are constructed of three parts separated by an underscore '_': (1) a *scheme identifier*, in this case the string 'offset', (2) *start index*, (3) the *end index*. The indexes are counting the gaps between the characters starting from 0 as specified in RFC 5147 with the exception that the encoding is defined to be Unicode Normal Form C (NFC)[44] and counting is fixed on Unicode Code Units[45]. This scheme is easy and efficient to implement and the addressed string can be referenced unambiguously. Due to its dependency on start and end indexes, however, a substantial disadvantage of offset-based URIs is the *instability* with regard to changes in the document. In case of a document change (i.e. insertion or deletion of characters), all offset-based URIs after the position the change occurred become invalid. The context-hash-based scheme is explained in more detail by [14].

4.3 Usage of Identifiers in the String Ontology

We are able to fix the referent of NIF URIs in the following manner: To avoid ambiguity, NIF requires that the whole string of the document has to be included in the RDF output as an `rdf:Literal` to serve as the reference point, which we will call *inside context* formalized using an OWL class called `str:Context`[46]. By typing NIF URIs as `str:Context` we are referring to the content only, i.e. an arbitrary grouping of characters forming a unit. The term *document* would be inappropriate to capture the real intention of this concept as `str:Context` could also be applied to a *paragraph* or a *sentence* and is **absolutely independent** upon the *wider context* in which the string is actually used such as a Web document reachable via HTTP.

[43] Note that with '/' the identifier is sent to the server during a request (e.g. Linked Data), while everything after '#' can only be processed by the client.

[44] `http://www.unicode.org/reports/tr15/#Norm_Forms`

[45] `http://unicode.org/faq/char_combmark.html#7`

[46] for the resolution of prefixes, we refer the reader to `http://prefix.cc`

We will distinguish between the notion of outside and inside context of a piece of text. The *inside context* is easy to explain and formalize, as it is the text itself and therefore it provides a *reference context* for each substring contained in the text (i.e. the characters before or after the substring). The *outside context* is more vague and is given by an outside observer, who might arbitrarily interpret the text as a "book chapter" or a "book section".

The class `str:Context` now provides a clear reference point for all other relative URIs used in this context and blocks the addition of information from a larger (outside) context. `str:Context` is therefore disjoint with `foaf:Document`, because labeling a context resource as a document is an information, which is not contained within the context (i.e. the text) itself. It is legal, however, to say that the string of the context occurs in (`str:occursIn`) a `foaf:Document`. Additionally, `str:Context` is a subclass of `str:String` and therefore its instances denote textual strings as well.

```
@prefix : <http://www.w3.org/DesignIssues/LinkedData.html#> .
@prefix str: <http://nlp2rdf.lod2.eu/schema/string/> .
:offset_0_26546
    rdf:type str:Context ;
# the exact retrieval method is left underspecified
    str:occursIn <http://www.w3.org/DesignIssues/LinkedData.html> ;
# [...] are all 26547 characters as rdf:Literal
    str:isString "[...]" .
:offset_717_729
    rdf:type str:String ;
    str:referenceContext :offset_0_26546 .
```

As mentioned in Section 4, NIF URIs are grounded on Unicode Characters using Unicode Normalization Form C counted in Code Units. For all resources of type `str:String`, the universe of discourse will then be the **words over the alphabet of Unicode characters** (sometimes called Σ^*). According to the "*RDF Semantics W3C Recommendation*", such an interpretation is considered a "semantic extension"[47] of RDF, because "extra semantic conditions" are "imposed on the meanings of terms"[48]. This "semantic extension" allows – per definitionem – for an unambiguous interpretation of NIF by machines. In particular, the `str:isString` term points to the string that fixes the referent of the context. The meaning of a `str:Context` NIF URI is then exactly the string contained in the object of `str:isString`. Note that Notation 3 even permits literals as subjects of statements, a feature, which might even be adopted to RDF[49].

[47] `http://www.w3.org/TR/2004/REC-rdf-mt-20040210/#urisandlit`

[48] `http://www.w3.org/TR/2004/REC-rdf-mt-20040210/#intro`

[49] `http://lists.w3.org/Archives/Public/www-rdf-comments/2002JanMar/0127.html`

5 Interoperability Between Different Layers of Annotations

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*.

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 namespace. 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 conventions are followed for uniform processing, while on the other hand developers can still use their own annotations using 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. In the following, we will present several vocabulary modules, including design choices, so they can serve as templates for adaption and further extensions.

5.1 OLiA

The *Ontologies of Linguistic Annotation* (OLiA) [8][50] 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[51]. 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

[50] `http://purl.org/olia`

[51] `http://purl.org/olia/penn.owl`

the following example, we show how *Penn Tag Set*[52] identifiers are combined with NIF:

```
@prefix sso: <http://nlp2rdf.lod2.eu/schema/sso/> .
# POS tags produced by Stanford Parser online demo
# http://nlp.stanford.edu:8080/parser/index.jsp
:offset_713_716
    str:anchorOf "The" ;
    str:referenceContext :offset_0_26546 ;
    sso:oliaIndividual <http://purl.org/olia/penn.owl#DT> ;
    sso:oliaCategory <http://purl.org/olia/olia.owl#Determiner>  .
:offset_717_725
    str:anchorOf "Semantic" ;
    str:referenceContext :offset_0_26546 ;
    sso:oliaIndividual <http://purl.org/olia/penn.owl#NNP> ;
    sso:oliaCategory <http://purl.org/olia/olia.owl#ProperNoun>  .
```

`oliaIndividual` and `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 the modelling of linguistic features, but to provide a mapping for data integration. Thus 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[53].

5.2 ITS 2.0 and NERD

At the time of writing the *MultilingualWeb-LT Working Group*[54] is working on a new specification for the *Internationalization Tag Set* (ITS) Version 2.0[55], which will allow to include coarse-grained NLP annotation into XML and HTML via custom attributes. Because attributes can only occur once per element, a corresponding NIF vocabulary module would require to reflect that in its design. Complementary to the ITS standardization effort, the *Named Entity Recognition and Disambiguation* (NERD) project [29] has created mappings between different existing entity type hierarchies to normalize named entity recognition tags. In this case, a vocabulary module can be composed of (1) DBpedia identifiers, (2) the functional OWL property `disambigIdentRef` to connect NIF with DBpedia (3) and additional type attachment to the included DBpedia identifier (`nerd:Organisation` in this case):

[52] `http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html`

[53] `http://sourceforge.net/projects/olia/`

[54] `http://www.w3.org/International/multilingualweb/lt/`

[55] `http://www.w3.org/TR/2012/WD-its20-20120829/`

```
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
:offset_23107_23110
    str:anchorOf "W3C" ;
    itsrdf:disambigIdentRef <http://dbpedia.org/resource/World_Wide_Web_Consortium> ;
    str:referenceContext :offset_0_26546 .
<http://dbpedia.org/resource/World_Wide_Web_Consortium>
    rdf:type  <http:/nerd.eurecom.fr/ontology#Organisation> .
```

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

```
:offset_23107_23110
    itsrdf:disambigIdentRef <http://dbpedia.org/resource/World_Wide_Web_Consortium> ;
    itsrdf:disambigIdentRef <http://rdf.freebase.com/ns/m.082bb> .
# entails that
<http://dbpedia.org/resource/dbpedia:World_Wide_Web_Consortium>
    owl:sameAs  <http://rdf.freebase.com/ns/m.082bb> .
```

5.3 *lemon and Wiktionary2RDF*

URIs of RDF datasets using lemon [24] can be attached to NIF URIs employing two properties, which link to lexical entries and senses contained in a lemon lexicon.

```
@prefix wiktionary: <http://wiktionary.dbpedia.org/resource/> .
:offset_717_725
    str:anchorOf "Semantic" ;
    str:referenceContext :offset_0_26546 ;
    sso:lexicalEntry wiktionary:semantic ;
    sso:lexicalSense wiktionary:semantic-English-Adjective-1en .
```

5.4 *Apache Stanbol*

Apache Stanbol[56] 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, it is necessary to keep provenance, confidence of annotations as well as full information about alternative annotations (often ranked by confidence) and not only the best estimate. In this case the vocabulary module uses an extra RDF node with a uniform resource name (urn)[57].

[56] http://stanbol.apache.org

[57] http://tools.ietf.org/html/rfc1737

```
@prefix fise: <http://fise.iks-project.eu/ontology/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
:offset_23107_23110
    str:anchorOf "W3C" ;
    str:referenceContext :offset_0_26546 ;
    sso:annotation <urn:enhancement-3f794cd6-11d1-3cae-f514-154d4e6a3b59>  .
<urn:enhancement-3f794cd6-11d1-3cae-f514-154d4e6a3b59>
    fise:confidence 0.9464704504529554 ;
    fise:entity-label "W3C"@en ;
    fise:entity-reference <http://dbpedia.org/resource/World_Wide_Web_Consortium> ;
    fise:entity-type <http:/nerd.eurecom.fr/ontology#Organisation> ;
    fise:entity-type dbo:Organisation, owl:Thing,
                     <http://schema.org/Organization> ;
    dcterms:created "2012-07-25T09:02:38.703Z"^^xsd:dateTime ;
    dcterms:creator "stanbol.enhancer.NamedEntityTaggingEngine"^^xsd:string ;
    dcterms:relation <urn:enhancement-c5377650-41af-7ea2-8ac8-44356007821a> ;
    rdf:type fise:Enhancement ;
    rdf:type fise:EntityAnnotation .
```

6 Discussion and Outlook

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 [18]. 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 [19] and the ISOcat data registry [33], 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 [10]. An overview can be found in [11].

One 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 [1]. 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 potentially be boosted when using background knowledge from LinkedGeoData, Wiktionary2RDF, DBpedia, Geonames or other LOD sources as crowd-sourced and community-reviewed and timely-updated gazetteers. 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 and maintenance in particular for multi-domain NLP applications was often impractical.

In this article, we have:

- described challenges and benefits of RDF for NLP.
- investigated the collaborative nature of three large data sets, which were created by a knowledge extraction process from crowd-sourced community projects.
- provided the extension mechanism of the NLP Interchange Format as a proof of concept, that NLP tool output can be represented in RDF as well as connected with exisiting LOD data sets.

7 Acknowledgments

We would like to thank our colleagues from AKSW research group and the LOD2 project for their helpful comments during the development of NIF. Especially, we would like to thank Christian Chiarcos for his support while using OLiA and Jonas Brekle for his work on Wiktionary2RDF. This work was partially supported by a grant from the European Union's 7th Framework Programme provided for the project LOD2 (GA no. 257943).

References

[1] Auer S, Lehmann J (2010) Making the web a data washing machine - creating knowledge out of interlinked data. Semantic Web Journal

[2] Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D (2009) Triplify: Light-weight linked data publication from relational databases. In: Proc. of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, ACM, pp 621–630

[3] 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)

[4] Berners-Lee T (2006) Design issues: Linked data. http://www.w3.org/DesignIssues/LinkedData.html

[5] 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)

[6] Bizer C (2012) Dbpedia 3.8 released, including enlarged ontology and additional localized versions. URL http://tinyurl.com/dbpedia-3-8

[7] Bühmann L, Lehmann J (2012) Universal owl axiom enrichment for large knowledge bases. In: Proceedings of EKAW 2012, URL http://jens-lehmann.org/files/2012/ekaw_enrichment.pdf

[8] Chiarcos C (2012) Ontologies of linguistic annotation: Survey and perspectives. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey

[9] Chiarcos C (2012) Powla: Modeling linguistic corpora in owl/dl. In: Proceedings of 9th Extended Semantic Web Conference (ESWC2012)

[10] Chiarcos C, Hellmann S, Nordhoff S (2011) Towards a linguistic linked open data cloud : The open linguistics working group. TAL 52(3):245 – 275, URL http://www.atala.org/Towards-a-Linguistic-Linked-Open

[11] Chiarcos C, Nordhoff S, Hellmann S (eds) (2012) Linked Data in Linguistics. Representing Language Data and Metadata. Springer

[12] Gruber TR (1993) A translation approach to portable ontology specifications. Knowledge Acquisition 5(2):199–220

[13] Hellmann S, Unbehauen J, Chiarcos C, Ngonga Ngomo AC (2010) The TIGER Corpus Navigator. In: 9th International Workshop on Treebanks and Linguistic Theories (TLT-9), Tartu, Estonia, pp 91–102

[14] Hellmann S, Lehmann J, Auer S (2012) Linked-data aware uri schemes for referencing text fragments. In: EKAW 2012, Springer, Lecture Notes in Artificial Intelligence (LNAI)

[15] Hellmann S, Stadler C, Lehmann J (2012) The German DBpedia: A Sense Repository for Linking Entities. In: [11], pp 181–190

[16] Hepp M, Bachlechner D, Siorpaes K (2006) Harvesting wiki consensus - using wikipedia entries as ontology elements. In: Völkel M, Schaffert S (eds) Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics, co-located with the 3rd Annual European Semantic Web Conference (ESWC 2006), ESWC2006, Workshop on Semantic Wikis

[17] 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

[18] 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

[19] Ide N, Suderman K (2007) GrAF: A graph-based format for linguistic annotations. In: Proc. Linguistic Annotation Workshop (LAW 2007), Prague, Czech Republic, pp 1–8

[20] 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

[21] Kontokostas D, Bratsas C, Auer S, Hellmann S, Antoniou I, Metakides G (2011) Towards linked data internationalization - realizing the greek dbpedia. In: Proceedings of the ACM WebSci'11

[22] Kontokostas D, Bratsas C, Auer S, Hellmann S, Antoniou I, Metakides G (2012) Internationalization of Linked Data: The case of the Greek DBpedia edition. Journal of Web Semantics

[23] 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

[24] McCrae J, Cimiano P, Montiel-Ponsoda E (2012) Integrating WordNet and Wiktionary with lemon. In: Chiarcos C, Nordhoff S, Hellmann S (eds) Linked Data in Linguistics, Springer

[25] Mendes PN, 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)

[26] Mendes PN, Jakob M, Bizer C (2012) 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] Meyer CM, Gurevych I (2011) OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary. In: Pazienza M, Stellato A (eds) Semi-Automatic Ontology Development: Processes and Resources, IGI Global

[28] Quasthoff M, Hellmann S, Höffner K (2009) Standardized multilingual language resources for the web of data: http://corpora.uni-leipzig.de/rdf. In: 3rd prize at the LOD Triplification Challenge, Graz, URL `http://triplify.org/files/challenge_2009/languageresources.pdf`

[29] Rizzo G, Troncy R, Hellmann S, Brümmer M (2012) NERD meets NIF: Lifting NLP Extraction Results to the LinkedData Cloud. In: Proceedings of Linked Data on the Web Workshop (WWW)

[30] Stadler C, Lehmann J, Höffner K, Auer S (2011) Linkedgeodata: A core for a web of spatial open data. Semantic Web Journal

[31] Unbehauen J, Hellmann S, Auer S, Stadler C (2012) Knowledge Extraction from Structured Sources. In: Search Computing - Broadening Web Search, Springer, lNCS volume 7538

[32] 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]

[33] Windhouwer M, Wright SE (2012) Linking to linguistic data categories in isocat. In: [11]