Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group

Christian Chiarcos* — **Sebastian Hellmann**** — **Sebastian Nordhoff*****

* *Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, chiarcos@daad-alumni.de*

** *Department of Computer Science, University of Leipzig, Johannisgasse 26, 04103 Leipzig, hellmann@informatik.uni-leipzig.de*

*** *Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, sebastian_nordhoff@eva.mpg.de*

ABSTRACT. *The Open Linguistics Working Group (OWLG) is an initiative of experts from different fields concerned with linguistic data, including academic linguistics (e.g. typology, corpus linguistics), applied linguistics (e.g. computational linguistics, lexicography and language documentation), and NLP (e.g. from the Semantic Web community). The primary goals of the working group are 1) promoting the idea of open linguistic resources, 2) developing means for their representation, and 3) encouraging the exchange of ideas across different disciplines.*
To a certain extent, the activities of the Open Linguistics Working Group converge towards the creation of a Linguistic Linked Open Data cloud, which is a topic addressed from different angles by several members of the Working Group. In this article, some of these currently on-going activities are presented and described.

RÉSUMÉ. *Le groupe OWLG est une initiative d'experts provenant de différents domaines linguistiques, comprenant la linguistique académique (typologie, corpus), la linguistique appliquée (linguistique computationelle, lexicographie, documentation de langues) et le traitement automatique des langues (p.ex. Web Sémantique). Les objectifs principaux de ce groupe sont 1) la promotion de l'idée de ressources ouvertes et accessibles, 2) le développement de moyens pour représenter lesdites ressources et 3) la stimulation d'échanges entre les diverses disciplines et subdisciplines.*
Les activités de l'OWLG sont pour la plupart liées à la création d'un nuage de données linguistiques Linked Open Data. Les membres du groupe abordent ce thème sous des aspects différents, dont nous présenterons quelques-uns ci-dessous.

KEYWORDS: *Open Knowledge Foundation (OKFN), Open Linguistics Working Group (OWLG); Linked Data, OWL, RDF, open data; describing languages and language resources (Glottolog/Langdoc), modeling language resources and annotations (POWLA, OLiA), lexical-semantic resources (DBpedia), annotations for the Semantic Web (NIF).*

MOTS-CLÉS : *Open Knowledge Foundation (OKFN), Open Linguistics Working Group (OWLG); Linked Data, OWL, RDF, open data; linguistique descriptive (Glottolog/Langdoc), modélisation de ressources et annotations linguistiques (POWLA, OLiA), ressources lexico-sémantiques (DBpedia), annotations pour le Web Sémantique (NIF).*

1. Introduction

The Open Linguistics Working Group (OWLG) [1] of the Open Knowledge Foundation (OKFN) [2] is 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 information technology (e.g. Natural Language Processing, Semantic Web). The primary goals of the working group are to promote the idea of open linguistic resources, to develop means for their representation, and to encourage the exchange of ideas across different disciplines.

Within the OWLG, a general consensus has been established that Semantic Web formalisms provide crucial advantages for the publication of linguistic resources. As shown in this article, all major types of data and metadata relevant to linguistic data collections (lexical-semantic resources, annotated corpora, metadata repositories and typological databases) can be represented by means of RDF and OWL, they are thus structurally interoperable (using RDF as representation formalism), and conceptually interoperable (with metadata and annotations modeled in RDF, different resources can be directly linked to a single repository). The OWLG encourages the use of open licenses: 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. RDF is usually not the most appropriate format for every individual domain taken on its own; for linking data from different domains, however, it is the only viable option at present.

1.1. *Technical and terminological background*

Before coming to the description of the OWLG and its activities, we give a brief introduction of the technologies and terminological conventions applied throughout this article, in particular the notions of **RDF**, **OWL/DL**, and the concept of **Linked Data**.

The Resource Description Framework (RDF, Lassila and Swick, 1999) was originally invented to provide formal means to describe any resource, both offline (e.g. books in a library), and online (e.g. PDF documents in an electronic archive). The data structures provided by RDF were, however, so general that its use has extended far beyond its original application scenario. RDF is based on the notion of triples, consisting of a **predicate** that links a **subject** to an **object**. In other words, RDF formalizes relations between resources as edges in a directed labelled graph: subjects are identified using globally unique URIs and can point to (via the predicate) another URI in the object part. 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 on this basis, a rich ecosystem of format extensions and technologies has evolved, including APIs, RDF databases (triple stores), the query language SPARQL, etc. Infrastructures for linguistic resources can benefit from these achievements and the relatively large and active community maintaining and improving technologies and representation formalisms.

For the formalization of knowledge bases, several RDF extensions have been provided, for example the **Simple Knowledge Organization System** (SKOS, Miles and Bechhofer, 2009), which is naturally applicable to lexical-semantic resources, e.g. thesauri. A thorough logical modeling can be achieved by formalizing linguistic resources as ontologies, using the **Web Ontology Language** (OWL, McGuinness and Van Harmelen, 2004), another RDF extension. OWL comes in several dialects (profiles), the most important being OWL/DL and its sublanguages (e.g. OWL/Lite, OWL/EL) that have been designed to balance expressiveness and reasoning complexity (McGuinness and Van Harmelen, 2004; W3C OWL Working Group, 2009). OWL/DL is based on Description Logics (DL, Baader *et al.*, 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. OWL/DL can thus be employed to specify formal data models for linguistic resources. Primary data structures of OWL Ontologies are **concepts**

1. `http://linguistics.okfn.org`
2. `http://okfn.org`

(classes of objects), **individuals** (instances of concepts), and **properties** (relations between individuals). Ontologies further support **class operators** (e.g. `intersection`, `join`, `complementOf`, `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. In this article, we employ OWL/DL whenever we refer to the logical modeling (axioms) of a domain and just RDF (or OWL) for conceptual modeling (terminology).

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** [3] 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 LOD cloud. At the moment, the LOD cloud already covers a number of lexical-semantic resources, including WordNet, YAGO, OpenCyc, and the DBpedia. Other types of linguistic resources (linguistic corpora, typological data collections, linguistic terminology repositories) are not present in the LOD cloud at all (see Section 7.1). The ultimate goal of the OWLG can be seen in the development of a LOD (sub-)cloud of linguistic resources, the **Linguistic Linked Open Data (LLOD) cloud**, where linguistic resources (lexical-semantic resources, corpora, metadata repositories) are not only provided in an interoperable way (using RDF), but also freely accessible (under an open license) and linked with each other (so that applications can combine information from different knowledge sources). In this article, we describe ongoing activities in the OWLG that will eventually lead to the creation of such a LLOD cloud.

1.2. *Overview*

Section 2 presents the Open Linguistics Working Group, its goals, addressed problems, recent activities and on-going developments; Sections 3 to 6 introduce representative resources covered by the LLOD cloud; and Section 7 describes the interlinking of language resources within the LLOD cloud and applications for this data structure.

As for linguistic resource types addressed, Section 3 describes DBpedia, a lexical-semantic resource and one of the major free data sets in the Web of Data; Section 4 deals with the modeling of linguistic corpora by means of POWLA, an OWL/DL-based formalism to represent any linguistic corpus with text-based annotations in RDF. Sections 5 and 6 deal with metadata repositories and linguistic databases: Section 5 presents OLiA, Ontologies of Linguistic Annotations that provide linguistic reference categories for linguistic analysis and annotation; Section 6 describes the Glottolog/Langdoc project, an attempt to chart the documentary status of all the world's languages, it focuses on the development of a catalog of language resources and a taxonomy of language identifiers.

Section 7 motivates the advantages of open licenses and RDF and analyses the claim that both serve as a key enabler for collaboration to integrate data in a decentralized network. The Section shows how distributed efforts augment each other in the LLOD cloud, how they can be linked with each other and illustrates potential application scenarios; Section 7.1 sketches possibilities to interlink the resources described before and highlights the key advantages of RDF and the Linked Data paradigm, i.e. cross-domain

3. `http://lod-cloud.net`

interoperability and the possibility to create ties between related, but distributed resources; Section 7.3 demonstrates how LLOD resources can be utilized to annotate the largest existing corpus — the Web. The recently created NLP Interchange Format employs URI Fragment identifiers and RDF to represent Natural Language Processing (NLP) analyses of web documents (web annotations) as links to LLOD resources, and thus to integrate this data into the existing Linked Data infrastructure again.

2. The Open Linguistics Working Group

2.1. *The Open Knowledge Foundation*

The Open Knowledge Foundation (OKFN) is a non-profit 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." [4]

The OKFN provides a catalog system for open datasets, **CKAN**. [5] 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 18 OKFN **working groups** covering fields as different as government data, economics, archeology, open text books or cultural heritage. [6] 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 are summarized in this section: Section 2.2 specifies the goals of the working group; Section 2.3 identifies four major problems and challenges of the work with linguistic data; Section 2.4 gives an overview of recent activities and the current status of the group.

2.2. *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 fulfil:

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.

4. `http://www.opendefinition.org`

5. `http://ckan.org/`

6. For a complete overview see `http://okfn.org/wg`.

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, [7] the ACL Special Interest Group for Annotation (SIGANN), [8] and large multi-national initiatives such as the ISO initiative on Language Resources Management (ISO TC37/SC4), [9] the American initiative on Sustainable Interoperability of Language Technology (SILT), [10] or European projects such as the initiative on Common Language Resources and Technology Infrastructure (CLARIN), [11] the Fostering Language Resources Network (FLaReNet), [12] and the Multilingual Europe Technology Alliance (META). [13]

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 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.

2.3. *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: [14] (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. [15]

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.

---

7. `http://cyberling.org/`
8. `http://www.cs.vassar.edu/sigann/`
9. `http://www.tc37sc4.org`
10. `http://www.anc.org/SILT`
11. `http://www.clarin.eu`
12. `http://www.flarenet.eu`
13. `http://www.meta-net.eu`
14. 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`).
15. 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`).

2.4. *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.

At the moment, the Working Group assembles 67 people from 29 different organizations and 10 countries. Our group is relatively small, but continuously growing and sufficiently heterogeneous. It includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology, just to name a few, so, the ground for fruitful interdisciplinary discussions has been laid out.

The Working Group maintains a home page, [16] a mailing list [17], a wiki, [18] and a blog. [19] 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 LLOD cloud, which is the main topic of this article. [20]

The following sections summarize recent activities of different community members in this direction. Following a discussion of representative resources, we describe their interlinking and possible applications.

3. Working with lexico-semantic resources: DBpedia

One class of linguistic resources stands out with respect to its importance to the Semantic Web community, the class of lexical-semantic resources (LSRs). Accordingly, providing LSRs in RDF and as Linked Data is nowadays an established practice that has also been adopted in the linguistic community, see, for example, Martin *et al.* (2009) for an OWL/DL lexicon of Old French. Here, we illustrate this class of resources with a particularly prominent example, DBpedia, a general-purpose knowledge base which has evolved into the nucleus for the Web of Data.

Due to continuous reviewing by a large community of stakeholders, DBpedia has evolved into a paragon of best practices for Linked Data. We describe aspects that are relevant for the OWLG and the creation of a Linguistic Linked Open Data cloud, including the lexical data contained in DBpedia as well as the recent internationalization effort (including the creation of a French version) and DBpedia Spotlight, a multilingual entity linking software.

3.1. *DBpedia*

DBpedia (Lehmann *et al.*, 2009) 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 1). The extracted RDF knowledge from the

16. `http://linguistics.okfn.org`
17. `http://lists.okfn.org/mailman/listinfo/open-linguistics`
18. `http://wiki.okfn.org/Wg/linguistics`
19. `http://blog.okfn.org/category/working-groups/wg-linguistics`
20. Details on these can be found on the OWLG wiki, `http://wiki.okfn.org/Wg/linguistics`.

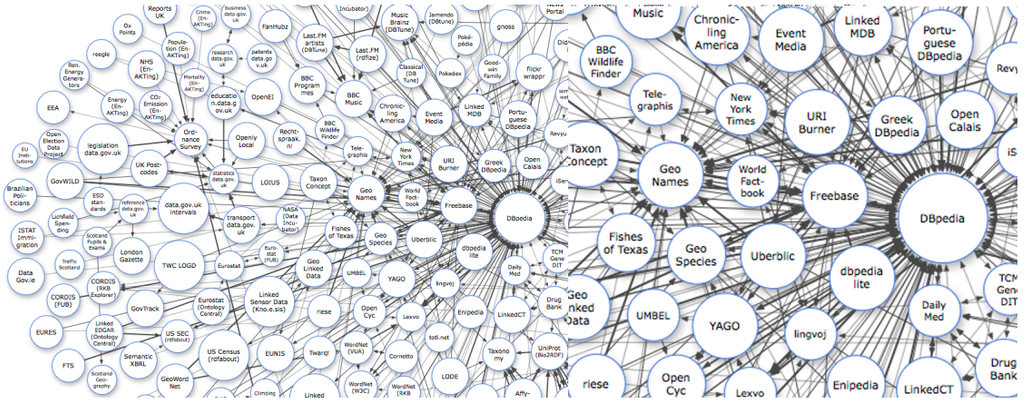

Figure 1. *Excerpt of the data sets interlinked with DBpedia. Source:* `http://lod-cloud.net` *(with kind permission of Anja Jentzsch and Richard Cyganiak).*

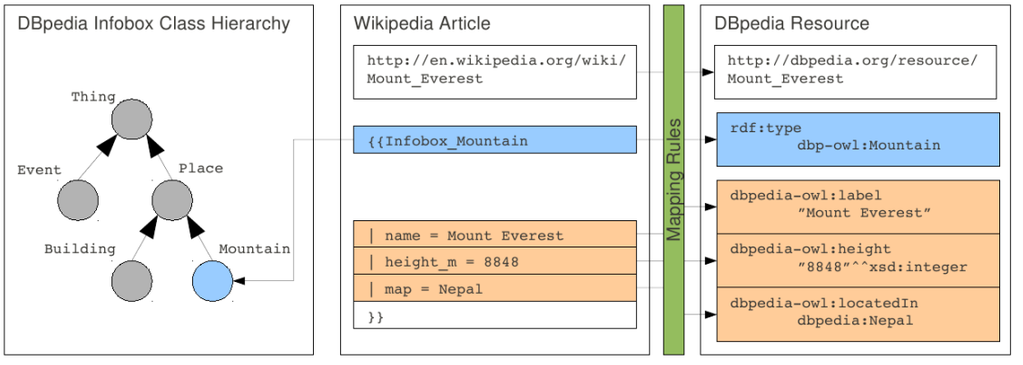

Figure 2. *Rule-based manipulation of extracted data in DBpedia Mappings Wiki (Hellmann* et al.*, 2012b).*

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 3.7 DBpedia contains more than 3.64 million things, of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases. The DBpedia data set features labels and abstracts for 3.64 million things in up to 97 different languages; 2,724,000 links to images and 6,300,000 links to external Web pages; 6,200,000 external links into other RDF datasets, and 740,000 Wikipedia categories. The dataset consists of 1 billion RDF triples out of which 385 million were extracted from the English edition of Wikipedia and roughly 665 million were extracted from other language editions and links to external datasets (Bizer, 2011a).

Currently, the DBpedia Ontology is maintained in a crowd-sourcing approach and thus freely editable on a *Mappings Wiki*:[21] each OWL class can be modeled on a Wiki page and the `subClassOf` axioms (shown on the left side of Figure 2) are created manually. The classification of articles according to the ontology classes is based on rules. In Figure 2, the article is classified as `dbp-owl:Mountain`, because it contains the Infobox "Infobox_Mountain" in its source.

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

3.2. *DBpedia as a sense repository and interlinking hub for linguistic resources*

DBpedia was created before the foundation of the OWLG and was motivated by the idea to query Wikipedia like a database. Not only do the OWLG community overlap with the DBpedia community, but also, 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 3.64 million entities (1.1 million in French) with highly stable identifier-to-sense assignment (Hepp *et al.*, 2007): 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 can be adapted for the creation of a LLOD cloud: 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 [22] 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.

3.3. *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 (Kontokostas *et al.*, 2011; Kontokostas *et al.*, 2012) 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 (Kontokostas *et al.*, 2011). Soon, the approach was generalized and applied to 15 other Wikipedia language editions (Bizer, 2011a), amongst them the localized French DBpedia. The French Wikipedia is currently the third largest Wikipedia [23] with about 1.1 million articles. Therefore it is also responsible for the third largest localized DBpedia with a total of 88.2 million RDF triples. The French community at the DBpedia Mappings Wiki has started to create mappings for infoboxes and achieve a coverage of about 38.81%. [24]

3.4. *DBpedia Spotlight*

The availability of large quantities of qualitative background knowledge provided by DBpedia and other lexical-semantic resources in the Web of Linked Data represents an important factor for improving the quality of NLP tools, especially with respect to tasks that involve (or can benefit from) Natural Language Understanding (Auer and Lehmann, 2010). 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 and maintenance, in particular for multi-domain NLP applications, was often impractical.

The band-width of applications of DBpedia data in NLP is thus immense, but here, we focus on a single example application, DBpedia Spotlight (Mendes *et al.*, 2011), 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.

22. For DBpedia Live see `http://live.dbpedia.org/`
23. Accessed on Oct 10th, 2011, `http://www.wikipedia.org/`
24. Accessed on Oct 10th, 2011, `http://mappings.dbpedia.org/server/statistics/fr/`

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. [25] 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.

4. Modeling linguistic corpora: POWLA

Besides lexical-semantic resources, the second major type of linguistic resources are **annotated corpora**. This section describes POWLA, as formalism to represent linguistic corpora by means of Semantic Web formalisms, in particular, OWL/DL. As compared to earlier approaches in this direction (Burchardt *et al.*, 2008; Hellmann *et al.*, 2010), POWLA is not tied to a specific selection of annotation layers, or a specific annotation scheme. Instead, it is designed to support any kind of text-oriented annotation.

The idea underlying POWLA is to represent linguistic annotations by means of RDF, to employ OWL/DL to define data types and consistency constraints for these RDF data, and to adopt these data types and constraints from an existing representation formalism applied for the loss-less representation of arbitrary kinds of text-oriented linguistic annotation within a generic exchange format. Here, we took the PAULA data model as our point of departure. PAULA XML is an XML standoff format developed at the Collaborative Research Center (SFB) 632 "Information Structure" (Dipper, 2005; Chiarcos *et al.*, 2008; Chiarcos *et al.*, 2011), it originates from early drafts of the Linguistic Annotation Framework (Ide and Romary, 2004), and it is thus closely related to the later ISO TC37/SC4 format GrAF (Ide and Suderman, 2007). With POWLA as an OWL/DL linearization of the PAULA data model, all annotations currently covered by PAULA (i.e. any text-oriented linguistic annotation) can be represented as part of the Linguistic Linked Open Data cloud.

When compared to current initiatives within the linguistics/NLP community such as the ISO TC37/SC4 (Ide and Suderman, 2007), which focus on complex standoff XML formats specifically designed for linguistic data, the POWLA approach offers several crucial advantages:

1) the increasing number of RDF databases provides us with convenient means for the management of linguistic data collections;

2) when an RDF representation of linguistic corpora is augmented with an OWL/DL specification of data types and constraints for these, existing reasoners are able to check the consistency of this representation;

3) such specifications and constraints are captured in OWL ontologies, which have a higher reusability than custom solutions;

4) resources can be freely interconnected with each other as well as with lexical-semantic resources available from the Linked Open Data cloud.

4.1. *PAULA data types*

PAULA implements the insight that any kind of linguistic annotation can be represented by means of **directed (acyclic) graphs** (Bird and Liberman, 2001), i.e. the basic triple structure underlying RDF: aside from the primary data (text), linguistic annotations consist of three principal components, i.e. segments (spans of text, e.g. a phrase, modeled as nodes), relations between segments (e.g. dominance relation between two phrases, modeled as edges) and annotations that describe different types of segments or relations (modeled as labels). As an illustrative example, Fig. 3 shows the first line of the Europarl corpus, v.3

25. `http://wiki.dbpedia.org/Lexicalizations`

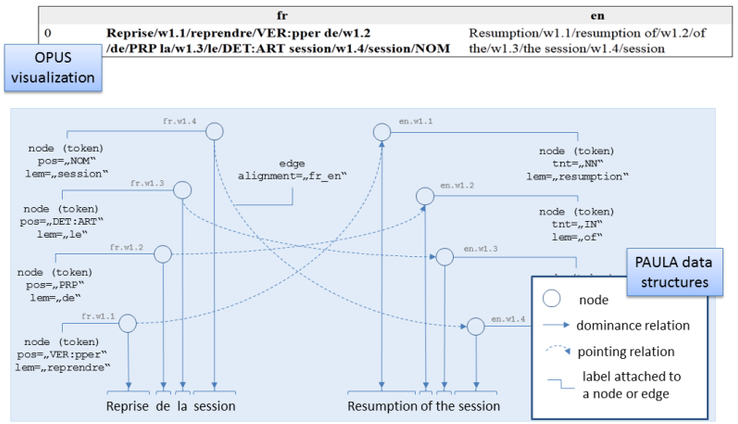

Figure 3. *Using PAULA data structures for a parallel corpus.*

(Koehn, 2005), with the alignment between French and English and annotations as provided as part of the Open Parallel Corpus OPUS (Tiedemann, 2009). [26]

PAULA data types relevant for linguistic annotations are the following:

node	(structural units of annotation)	
	token	character spans in the primary data
	markable	span of tokens (data structure of flat, layer-based annotations defined with respect to, e.g. a timeline)
	struct	hierarchical data structure (e.g. for a tree) establishes dominance relations between a struct (parent) and tokens, markables or other structs
edge	(relational unit of annotation, connecting nodes)	
	dominance relation	directed edge between a struct and its children, implying a hierarchical relationship
	pointing relation	general directed edge, non-hierarchical
label	(attached to nodes or edges)	
	features	linguistic annotations

A unique feature of PAULA is the differentiation of two types of edges with respect to their relationship to the primary data. Where **dominance relation**s are applied (e.g. for constituent syntax), the text covered by a child node is always covered by the parent node. **Pointing relations** do not impose such constraints, and can be used for all other types of relational annotations, where source and target may or may not overlap (e.g. dependency syntax, coreference, or alignment).

As the mapping of morphosyntactic annotations to PAULA (and further to POWLA) has been described before (Chiarcos, 2012a), we focus here on the modeling of alignment relations as required for our French-English example in Fig. 3: word-level alignment is a directed relation between a source language word and a (set of) target language word(s); in PAULA, this can be appropriately represented by means of a special type of pointing relation. A key advantage of this modeling as compared to the original formalization in the Europarl corpus is that both annotations and alignment can be expressed using the same formalism, whereas most tools for annotated parallel corpora keep both levels of representation apart. In this sense, alignment and grammatical annotations are interoperable with each other in PAULA. Bringing this approach to RDF then extends this notion of interoperability even beyond the realm of annotated corpora to lexical-semantic resources, and metadata repositories.

26. `http://opus.lingfil.uu.se`

4.2. *The POWLA ontology*

The POWLA ontology implements the PAULA data model in OWL/DL. The top-level concept of `POWLAElement`, with subconcepts `Document`, `Layer`, `Relation` and `Node`. Here, we concentrate on the latter two, `Document` and `Layer` are more important for corpus organization.

A `POWLAElement` is anything that can carry a label (property `hasLabel`). For `Node` and `Relation`, `hasLabel` contains string values of linguistic annotation (subproperty `hasAnnotation`). For every attribute (e.g., `pos` for part-of-speech annotation), a corresponding subproperty of `hasAnnotation` (e.g., `has_pos`) is created.

A `Node` is a `POWLAElement` that covers a stretch of primary data. It can carry `hasChild` properties that link it with another `Node`, and it can be source or target of a `Relation`. A `Relation` is a `POWLAElement` that is used for edges that carry annotations. The properties `hasSource` and `hasTarget` assign a `Relation` source and target `Node`. Dominance relations are relations whose source and target are also connected by a `hasChild` property. Pointing relations are relations where source and target are not connected by `hasChild`. It is thus not necessary to distinguish pointing relations and dominance relations as separate concepts in the POWLA ontology.

Two basic subclasses of `Node` are distinguished: a `Terminal` is a `Node` which does not have a `hasChild` property. It corresponds to a "token" in PAULA. A `Nonterminal` is a `Node` which has at least one `hasChild` property. The differentiation between PAULA struct and markable can be inferred and is therefore not explicitly represented in the ontology: a struct is a `Nonterminal` that has another `Nonterminal` as its child, or that is connected to at least one of its children by means of a (dominance) `Relation`, any other `Nonterminal` corresponds to a PAULA markable.

Both `Terminals` and `Nonterminals` are characterized by a string value (property `hasString`), and a particular position (properties `hasStart` and `hasEnd`) with respect to the primary data. `Terminals` are further connected with each other by means of `nextTerminal` properties.

4.3. *Modelling linguistic annotations in POWLA*

With the data types defined within the POWLA ontology, linguistic annotations can now be represented in OWL/RDF, see, for example, the following listing for the French word *Reprise* (`fr.w1.1`), its English translation *Resumption* (`en.w1.1`) and their alignment (`fr.w1.1_en.w1.1`) from Fig. 3:

```
<powla:Terminal rdf:ID="fr.w1.1">
  <powla:has_pos>VER:pper</powla:has_cat>
  <powla:has_lem>reprendre</powla:has_lem>
  <powla:hasString>Reprise</powla:hasString>
  ...
</powla:Terminal>
```

```
<powla:Terminal rdf:ID="en.w1.1">
     <powla:has_tnt>NN</powla:has_tnt>
     <powla:has_lem>resumption</powla:has_lem>
     <powla:hasString>Resumption</powla:hasString>
     ...
  </powla:Terminal>
```

```
<powla:Relation rdf:ID="fr.w1.1_en.w1.1">
   <powla:hasSource rdf:resource="...#fr.w1.1"/>
   <powla:hasTarget rdf:resource="...#en.w1.1"/>
   <powla:has_alignment>fr_en</powla:has_alignment>
</powla:Relation>
```

The properties `has_pos`, `has_lem`, `has_tnt` are subproperties of `hasAnnotation` that have been created to reflect the `pos`, `lem` and `tnt` attributes of the nodes in Fig. 3. The alignment relation `fr.w1.1_en.w1.1` preserves its PAULA attribute `has_alignment` as another subproperty of `hasAnnotatation`, it is marked as a pointing relation by the absence of a `hasChild` property connecting its source and target node.

Although illustrated here for morphosyntactic annotations and alignment only, the conversion of other annotation layers from PAULA to POWLA is similarly straight-forward, cf. Chiarcos (2012a) for an exam-

ple with syntax annotations. Thus, all PAULA data types can be represented in OWL/DL, through PAULA, various corpora in different formats can be converted to OWL/RDF, and subsequently be linked with each other (e.g. through alignment) or other resources from the LLOD.

4.4. *Application*

A key advantage of the OWL/RDF formalization is that it represents a standardized representation formalism for different corpora, an issue further explored in Section 7. Datatypes in OWL/DL assure that the validity of corpora can be automatically checked (according to consistency constraints posited by the POWLA ontology). POWLA represents a possible solution to the **structural interoperability** challenge for linguistic corpora (Ide and Pustejovsky, 2010). Unlike state-of-the-art formalisms developed in this direction (e.g. GrAF, Ide and Suderman, 2007 and PAULA), it does not involve a special-purpose XML standoff format, but it builds on established standards with broad technical support from an active and comparably large community. Standard formats specifically designed for linguistic annotations as developed in the context of the ISO TC37/SC4 (e.g. GrAF), are, however, still under development.

Also, RDF allows us to store and to query linguistic corpora with off-the-shelf databases. While PAULA data requires a conversion to the table format of a relational database for storing and querying (Zeldes *et al.*, 2009), POWLA data can be directly processed with an RDF triple store and queried with SPARQL. For the example of alignment relations shown above, a SPARQL query for English translations of French past participle verbs could be as follows:

```
PREFIX powla:<http://purl.org/powla/powla.owl#>      #
PREFIX ep:<http://opus.lingfil.uu.se/cwb/Europarl#>
SELECT ?fr ?en
WHERE {
  ?fr a powla:Node.                                  #
  ?en a powla:Node.                                  #
  ?fr powla:has_pos "VER:pper".
  ?alignment a powla:Relation.                       #
  ?alignment powla:has_alignment="fr_en".            # ?fr ->[has_alignment="fr_en"] ?en
  ?alignment powla:hasSource ?fr.                    #
  ?alignment powla:hasTarget ?en.                    #
}
```

Using the data structures defined by POWLA, SPARQL macros can be defined that provide shorthands for frequent combination of attributes. In the listing, the expression to the right can replace the lines marked with a # using a query preprocessor. [27] Like PAULA, POWLA can thus represent the basis to develop an infrastructure capable to store, to process and to query any kind of text-oriented annotation. Unlike PAULA, however, POWLA is not based on a domain-specific XML standoff format, but on RDF and it can be stored and queried by means of RDF databases without further conversion.

Moreover, as an RDF-based formalism, POWLA does not only provide structural interoperability among linguistic annotations and corpora, but also interoperability with other types of linguistic resources: within the LLOD, also lexical-semantic resources and linguistic knowledge bases can be stored in RDF databases and queried with SPARQL, and if linked with annotated corpora, they can be used, for example, to provide formal semantics for annotations and metadata of linguistic corpora. The following section illustrates such an application with a terminology repository for linguistic categories.

27. POWLA sample data and SPARQL macros for all operators of AQL, a query language for multi-layer corpora (Chiarcos *et al.*, 2008), can be found under `http://purl.org/powla`.

5. Representing linguistic annotations: OLiA

The Ontologies of Linguistic Annotation (OLiA) are a repository of annotation terminology for various linguistic phenomena on a great band-width of languages. In combination with RDF-based formats like POWLA (Section 4) and NIF (Section 7.3), or with lexical-semantic resources like Lemon (McCrae *et al.*, 2011), the OLiA ontologies allow to represent linguistic annotations in corpora, grammatical specifications in dictionaries, and their respective meaning within the Linguistic Linked Open Data cloud in an interoperable way.

5.1. *Modular specifications for reference terminology and annotation terminology*

It is generally agreed that **repositories of linguistic annotation terminology** represent a key element in the establishment of conceptual interoperability for NLP tools and linguistic resources, yet multiple – and partially divergent – terminology repositories have been developed by different communities, including the General Ontology of Linguistic Description (Farrar and Langendoen, 2003b, GOLD) and the ISO TC37/SC4 Data Category Registry (Kemps-Snijders *et al.*, 2009, ISOcat).

The Ontologies of Linguistic Annotations – briefly, OLiA ontologies (Chiarcos, 2008) – represent a modular architecture of OWL/DL ontologies that formalize the mapping between annotations and multiple existing terminology repositories (External Reference Models) by means of the OLiA Reference Model that mediates between both.

In the OLiA architecture, four different types of ontologies are distinguished:

– the OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.

– multiple OLiA ANNOTATION MODELs formalize annotation schemes and tagsets. Annotation Models are based on the original documentation of an annotation scheme, they provide an interpretation-independent representation.

– for every Annotation Model, a LINKING MODEL defines `subClassOf` relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model.

– existing terminology repositories can be integrated as EXTERNAL REFERENCE MODELs, if they are represented in OWL/DL. Then, Linking Models specify `subClassOf` relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g. `olia:Determiner`) and grammatical features (e.g. `olia:Accusative`), as well as properties that define relations between these (e.g. `olia:hasCase`). Far from being yet another annotation terminology ontology, the OLiA Reference Model does not stipulate its own view on the linguistic world, but rather, it is a derivative of EAGLES (Leech and Wilson, 1996), MULTEXT/East (Erjavec, 2004), and GOLD (Farrar and Langendoen, 2003b) that was introduced as a technical means to allow to interpret linguistic annotations with respect to these terminological repositories and extended with respect to the annotation schemes linked with it. These extensions are also further communicated to the communities behind GOLD and ISOcat. The Reference Model specifies for example that a past participle is a participle that is morphologically marked for past tense:

$$\texttt{PastParticiple} \equiv \texttt{Participle and hasTense some Past}$$

Annotation Models differ conceptually from the Reference Model in that they include not only concepts and properties, but also individuals: individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model. As an example, consider the tag `VER:pper` from the tagset of the French Treetagger (Stein, 2003) and the corresponding individual `french-tt:VER_pper` in the Annotation Model `http://purl.org/olia/french-tt.owl`:

```
VER_pper system:hasTag 'VER:pper'
  VER_pper a VerbPastParticiple
```

Linking Models then import an Annotation Model and the Reference Model and specify `subClassOf` ($\sqsubseteq$) relations between their concepts:

```
french-tt:VerbPastParticiple ⊑ olia:PastParticiple
```

The Linking with External Reference Models like ISOcat is analogous: `olia:Participle` $\sqsubseteq$ `isocat:DC-1341`, and `olia:Past` $\sqsubseteq$ `isocat:DC-1347`.[28] In consequence, is is true that `french-tt:VER_pper a isocat:DC-1341`.

Within an application, the French tag (individual) `VER_pper` can then be circumscribed by means of the concepts it is associated with. For instance, an expression like `olia:Participle and olia:hasTense some olia:Present`. This description is concept-based and thus independent from any particular tagset, and, applied to another Annotation Model, it would retrieve another set of individuals that represent the same meaning with different annotations, e.g. the tag `VBN` from the Penn Treebank tagset (Marcus *et al.*, 1994) that was used for the annotation of the English part of the Europarl corpus (see Section 4).[29]

Using the SPARQL macros described before, we can now formulate tagset-independent corpus queries, e.g. for an alignment between French and English participles in the Europarl corpus:

```
PREFIX ep:<http://opus.lingfil.uu.se/cwb/Europarl#>
SELECT ?fr ?en
WHERE {
  ?fr a olia:Participle.
  ?fr ->[has_alignment="fr_en"] ?en.
  ?en a olia:Participle.
```

5.2. *Current status of the OLiA ontologies*

The OLiA ontologies are available from `http://purl.org/olia`. They will be officially released under a Creative Commons Attribution license in mid-2012.

The OLiA ontologies cover different grammatical phenomena, including inflectional morphology, word classes, phrase and edge labels of different syntax annotations, as well as prototypes for discourse annotations (coreference, discourse relations, discourse structure and information structure). Annotations for lexical semantics are only covered by the OLiA ontologies to the extent that they are encoded in syntactic and morphosyntactic annotation schemes (e.g. as grammatical roles). For lexical semantic annotations in general, a number of reference resources are already available, including RDF versions of WordNet[30], FrameNet,[31] and Wikipedia (i.e. DBpedia).

The OLiA Reference Model comprises 14 `MorphologicalCategorys` (morphemes), 263 `MorphosyntacticCategorys` (word classes/part-of-speech tags), 83 `SyntacticCategorys` (phrase labels), and 326 different values for 16 `MorphosyntacticFeatures`, 4 `MorphologicalFeatures`, 4 `SyntacticFeatures` and 4 `SemanticFeatures`.

As for morphological, morphosyntactic and syntactic annotations, the OLiA ontologies include 32 Annotation Models for about 70 different languages, including several multi-lingual annotation schemes,

28. `olia:PastParticiple` does not seem to have an exact counterpart in the morphosyntactic profile of ISOcat, data category `isocat:DC-1596`, labeled `pastParticipleAdjective`, is defined as "Adjective based on a past participle" which may excludes potential non-adjectival uses of participles.
29. `http://purl.org/olia/penn.owl`
30. `http://thedatahub.org/dataset/w3c-wordnet`
31. `http://wiki.loa-cnr.it/index.php?title=LoaWiki:OFN`

Table 1. *Languages of the Francophonie covered by OLiA Annotation Models.*

Annotation Model	Languages	Phenomena	Associated resources
`http://purl.org/olia/french.owl`	French	inflectional morphology, parts-of-speech, constituent syntax	French Treebank (Abeillé *et al.*, 2000)
`http://purl.org/olia/french-tt.owl`	French	parts-of-speech	French TreeTagger (Stein, 2003)
`http://purl.org/olia/connexor.owl`	French	inflectional morphology, parts-of-speech, dependency syntax	Connexor parser (Tapanainen and Järvinen, 1997)
`http://purl.org/olia/eagles.owl`	French, Greek	inflectional morphology, parts-of-speech	various resources (EAGLES annotation standard) (Leech and Wilson, 1996)
`http://purl.org/olia/mte`	Bulgarian, Macedonian, Romanian	inflectional morphology, parts-of-speech	corpora, lexica (Erjavec, 2004)
`http://purl.org/olia/sfb632.owl`	Québécois, Greek, African languages	linguistic glosses	questionnaire for information structure (QUIS) data (Skopeteas *et al.*, 2006)

e.g. EAGLES (Chiarcos, 2008) for 11 Western European languages, and Multext-East (Chiarcos and Erjavec, 2011) for 15 (mostly) Eastern European languages. As for non-(Indo-)European languages, the OLiA ontologies include morphosyntactic annotation schemes for languages of the Indian subcontinent, for Arabic, Basque, Chinese, Estonian, Finnish, Hausa, Hungarian, and Turkish. Other languages, including languages of Africa, the Americas, the Pacific and Australia are covered by Annotation Models developed for glosses as produced in typology and language documentation. The OLiA ontologies also cover historical language stages, including Old High German, Old Norse and Old/Classical Tibetan. The current OLiA ontologies for the morphosyntactic annotation of languages within the Francophonie are summarized in Table 1.

As mentioned above, application of modular OWL/DL ontologies allows to link annotations with terminological repositories: annotation schemes and reference terminology are formalized as OWL/DL ontologies, and the linking is specified by `subClassOf` descriptions. This mechanism has also been applied to link the OLiA Reference Model with existing terminology repositories, including GOLD (Chiarcos, 2008), the OntoTag ontologies (Buyko *et al.*, 2008) and ISOcat (Chiarcos, 2010a). Thereby, the OLiA Reference Model provides a stable intermediate representation between existing terminology repositories and ontological models of annotation schemes. This allows any concept that can be expressed in terms of the OLiA Reference Model also to be interpreted in the context of ISOcat or GOLD. Using the OLiA Reference Model, it is thus possible to develop applications that are interoperable in terms of GOLD *and* ISOcat even though both are still under development and both differ in their conceptualizations.

Within the LLOD, the OLiA ontologies can thus be used to describe linguistic categories for any kind of linguistic resource, including corpora (as in the example above) and lexical-semantic resources (McCrae *et al.*, 2011) and thus contribute to their **conceptual interoperability**. Another possible application is in the development of **tag-set independent NLP architectures** (Buyko *et al.*, 2008; Rehm *et al.*, 2007; Chiarcos, 2010b), also see Section 7.3.

6. Providing information about languages and language resources: Glottolog/Langdoc

Like annotations, other forms of information about languages and language resources can be represented by LLOD resources as well. Unlike OLiA, resources of this type are not necessarily designed to serve an interoperability-enhancing purpose – although they can be applied as such within the LLOD –, but as a self-contained database. To exemplify this kind of resource, we describe the Glottolog/Langdoc project, that applies RDF in accordance with its original function, i.e., to describe resources, e.g. books in a library, here, to collect and to formalize **information about languages and language resources**.

By doing so, Glottolog/Langdoc covers the band-width of languages in the world as far as possible, i.e. with a certain emphasis – albeit not a strict focus – on less-resourced languages. Section 6.1 gives an overview of the bibliographical part of the project (Langdoc), Section 6.2 introduces the notion of **languoid**, a data structure for the modeling of genealogical relationships between language families, languages and dialects (Glottolog), and Section 6.3 summarizes the resource types provided for the Linguistic Linked Open Data cloud.

6.1. *Langdoc: charting the documentary status of all the world's languages*

The computational treatment and modeling of language resources has so far mainly concentrated on major languages with a research tradition in NLP and some commercial viability. The Wiki of the Association for Computational Linguistics, [32] for example, lists 64 languages, but most of these are European (64%, 41/64) or official languages in nation states (77%, 49/64), there are only six languages, or nine percent, that are neither. These are Greenlandic, Iñupiaq, Kurdish, Navajo, Punjabi, and Sanskrit.

This corresponds to a general tendency: outside industrialized countries, languages resources become very scarce. This is true in particular for languages that are spoken by ethnical minorities (i.e. the vast majorities of languages in the world). Treebanks or annotated corpora seem like fanciful ideas when the total of resources for a language amounts to a description of its verbs and a treatise of its phonology from a local university, which is the case, for instance, for the Niger-Congo language Aduge.

Before one can start thinking about developing a WordNet or similar larger resources for these languages, one must take stock of the resources which exist, however arcane they might be. This is the aim of Langdoc under the umbrella of the Glottolog/Langdoc [33] project (Hammarström and Nordhoff, 2011). Building upon bibliographical work by dedicated scholars, [34] Langdoc lists 166,459 resources providing information about the world's linguistic diversity. [35] The resources are tagged for resource type (grammar, word list, text collection, etc.), macroarea (geographic region), and language. Table 2 gives an overview of the resources covered so far, classified by macroarea and document type.

32. `http://www.aclweb.org/aclwiki/index.php?title=Category:Resources_by_language`
33. `http://glottolog.org`.
34. ASJP Automated Similarity Judgment Program bibliography `http://lingweb.eva.mpg.de/asjp/index.php/ASJP`; Alain Fabre's *Diccionario etnolingüÃstico y guÃa bibliogrÃ¡fica de los pueblos indÃgenas sudamericanos* `http://www.tut.fi/~fabre/BookIntervetVersio`; The bibliography of the Papua New Guinea branch of SIL `http://www.sil.org/pacific/png/`; Randy LaPolla's Tibeto-Burman bibliography `http://victoria.linguistlist.org/~lapolla/bib/index.htm`; The bibliography of the South Asian Linguistics Archive `http://www.sealang.net/library/`; Frank Seifart's bibliography `www.eva.mpg.de/lingua/staff/seifart.html`; The World Atlas of Language Structures `www.wals.info`; Harald Hammarström's bibliography `http://haraldhammarstrom.ruhosting.nl/`; The catalogue of the Max Planck Institute for Evolutionary Anthropology in Leipzig, `www.eva.mpg.de/library`; The SIL bibliography `www.ethnologue.com/bibliography.asp`; The Web-version of EBALL, by Jouni Maho and Guillaume SÃ©gerer `http://sumale.vjf.cnrs.fr/Biblio/`; Jouni Maho's bibliography of Africa `http://goto.glocalnet.net/maho/eball.html`; Tom Güldemann's bibliography of Africa `http://www2.hu-berlin.de/asaf/Afrika/Mitarbeiter/Gueldemann.html`; Chintang-Puma Documentation Project `http://www.uni-leipzig.de/~ff/cpdp/`
35. Note that we only provide the reference, but no copy of the work itself. We link to WorldCat, GoogleBooks and Open Library to help users retrieve a copy.

Table 2. *Langdoc language resources according to geographic region and document type.*

Area	Refs
Africa	74,787
South America	32,897
Eurasia	16,879
Pacific	15,424
Australia	7,557
North America	3,815
Middle America	1,897

document type	refs
comparative treatise	13,827
grammar sketch	13,810
ethnographic treatise	13,504
grammar	10,209
overview	8,273
dictionary	7,408
wordlist	5,552

document type	refs
phonology	1,942
bibliography	1,464
specific feature	1,362
text	1,039
sociolinguistics	943
dialectology	797
new testament	143

Langdoc has two main goals:

1) for every language, provide a reference of the most extensive piece of documentation.

2) beyond that, provide as many references as possible, including grey literature like manuscripts, Ph.D. theses, and M.A. theses

Langdoc data are searchable by standard bibliographical data such as author, year, title, etc. Every reference has its own URI with XHTML and RDF representations. The bibliographical data employ standard ontologies such as DCMI and BIBO. The novel feature of Langdoc is the possibility for genealogical searches in a stepless manner. This is accomplished by using a set-theoretic approach: French is a subset of Romance, and a subset of Indo-European. This means that a reference associated with French is associated with Romance (and Indo-European) at the same time. A researcher interested in languages of the Pacific Ocean could search at any level of the deeply nested tree of Austronesian languages (Fig. 4). Queries like "Give me any grammar of an Oceanic language" or "Give me any dictionary of a Polynesian language" become possible. The genealogical data just mentioned lead us to the second part of the Glottolog/Langdoc project to be discussed here: Glottolog.

6.2. *Glottolog: an empirical approach to definitions of languages*

Linguists and laymen often debate whether a given linguistic variety is a language or not (e.g. are Serbian and Croatian the same, are Hindi and Urdu the same, how many varieties of Quechua or Kurdish are there, etc.). While there is little hope to solve these social conflicts with structural linguistic means, there is nevertheless a way to move the debate from an essentialist issue to a labeling issue, making use of the set-theoretic Linked Data approach of the Glottolog/Langdoc project. As explained above, linguistic varieties are linked to references. This can be used to provide an extensional definition of linguistic varieties that we refer to as **languoid**: languoid X is defined as the set of all observations found in the references associated with it (Nordhoff and Hammarström, 2011). French for instance would be defined as the set containing the *Petit Robert*, the *Grand Larousse*, the *Bescherelle*, etc.

This extensional definition has a number of consequences:

1) spurious languoids disappear. There are a number of ISO 639-3 codes which SIL (the ISO 639-3 registrar) assigned to "languages", but it is not clear what they refer to. Consultations with experts of the relevant areas have proven fruitless. Examples for such dubious cases are Cumeral [cum], Omejes [ome], Ponares [pod], and Tomedes [toe], all supposedly spoken in Colombia. At the time of writing, it is unclear whether these languages exist at all. We have SIL's word for it, but no way of tracing the chain of scientific argumentation. Changing the procedure from fiat definitions by a registrar to a document-centric empirical definition means that such spurious cases become undefined, which corresponds to our intuitions.

2) family relations can be modeled in a set-theoretic fashion: let AE be the set of references associated with American English and CE the set of references associated with Commonwealth English. The languoid "Modern English" can then be defined as the union of those two sets (and possibly others, like Indian English, New Zealand English, etc.). This procedure is recursive, and Indo-European as a language family

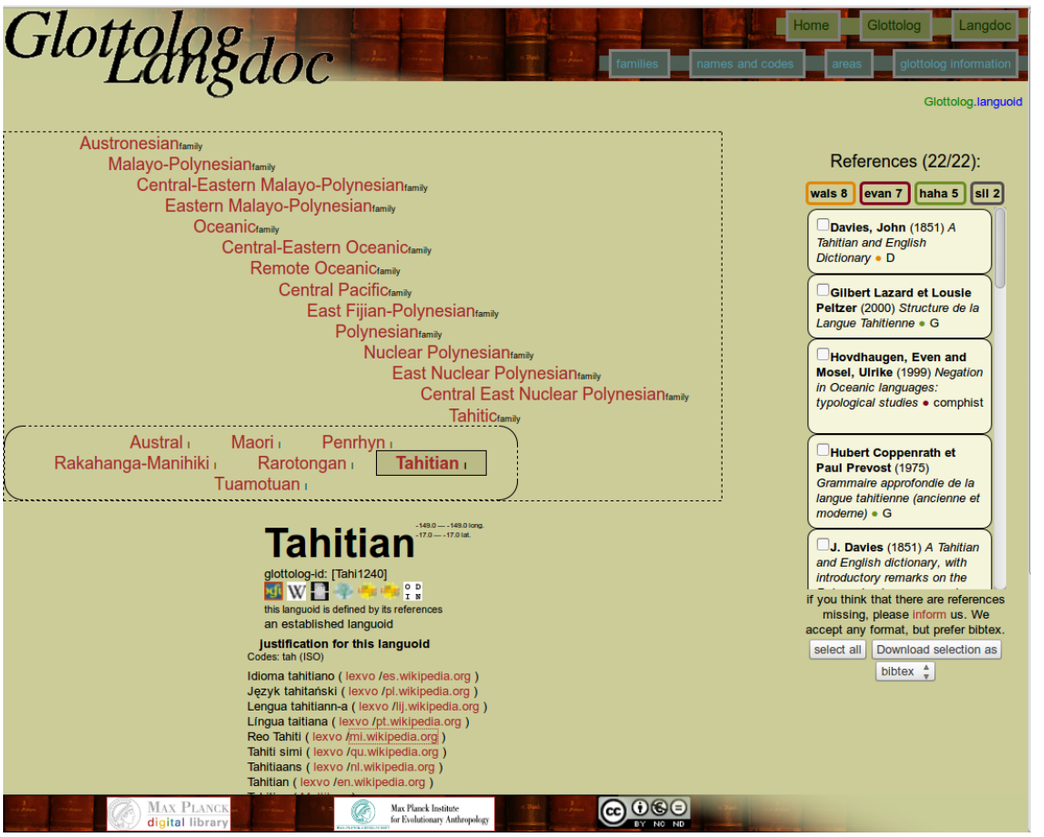

Figure 4. *XHTML representation of the languoid "Tahitian", left: genealogy, right: references. Bottom: names, codes, and geographic location.*

can be defined as the union of all sets of references associated with its daughters.

3) the question of whether Serbian or Croatian are distinct languages boils down to the (uncontroversial) observation that there are documents describing a languoid "Croatian" while others describe a languoid "Serbian". These two languoids are distinct, but have a common mother "Serbo-Croatian", which is associated with all references associated with its daughters, as well as a couple of additional resources. This does not solve the question which node is a "language", a "family" or a "dialect", but it provides unique IDs to all of the nodes, which structurally-oriented linguists can use for their scientific purposes without having to delve into the issues in the realm of sociolinguistics.

4) the provision of URIs for every languoid means that conflicting opinions can be modeled. The Glottolog/Langdoc project does for instance not believe that there is sufficient evidence for a node "Altaic". But the project provides URIs for Turkic, Mongolic, and Tungusic, and third-party projects can reuse the Glottolog data to integrate them into their divergent classification as children of their node "Altaic".

This leads to the modeling employed by Glottolog. As stated above, we employ a set-theoretic approach. Every languoid is seen as a set. Subset and superset relations can model genealogical relationships. In this particular case, Glottolog employs `skos:narrower` and `skos:broader` to model the relation between a larger languoid like Romance and a smaller languoid like French. Note that languages are seen as concepts here, and not as individuals, similar to biological taxonomies. A particular lion is an instance of *panthera leo*, and at the same time an instance of felidae, carnivora, mammals, and animals. The variety described in a particular document, e.g. the *Grand Larousse* is an instance of the languoid *French*, the

languoid *Romance*, and the languoid *Indo-European*, all at the same time. Documents can be associated with a languoid of a particular level directly or indirectly via one of its children. In the former case, we use `dcmi:bibliographicCitation`, in the latter case, we use `glottolog:fullEmpiricalGrounding`, which recursively collects all `dcmi:bibliographicalCitations` of dominated nodes (Nordhoff and Hammarström, 2011).

6.3. *Glottolog/Langdoc and Linked Data*

Glottolog/Langdoc provides two types of resources as Linked Open Data: languoids and resources.

Languoid is a cover term for dialect, language, and language family (Good and Hendryx-Parker, 2006). Every languoid has its own URI and is annotated for ancestors, siblings, children, names, codes, geographic location and references. Links are provided to Multitree, [36] LL-Map (Xie *et al.*, 2009), LinguistList, [37] Ethnologue (Lewis, 2009), ODIN, [38] WALS, [39] OLAC, [40] lexvo, [41] and Wikipedia. Languoids are modeled using SKOS and RDFS and linked with ontologies like GOLD, [42] lexvo, [43], and wgs84. [44]

Language resources are available in XHTML and RDF. Resources make use of Dublin Core (Weibel *et al.*, 1998) and are annotated for the languoids they are applied to. Additionally, resources are linked to WorldCat, [45] GoogleBooks, [46] and Open Library. [47]

The database currently covers 166,459 resources and 94,008 languoids (Tables 2 and 3). This information is essential for language documentation and typological research, and it was originally intended for these. Within the LLOD cloud, however, this data can serve different purposes: on the one hand, Langdoc can inform researchers and engineers from other disciplines about language resources for a particular languoid, and these resources may play a role in, for example, the development of NLP tools for this particular language (see Section 7.1 for an example). On the other hand, Glottolog provides a fine-grained and literature-based classification of languages that can be used to define the linguistic content of a resource with great level of detail (in particular if compared with the ISO 693 standard that is represented in the LLOD through lexvo). Furthermore, integrating Glottolog/Langdoc with the LLOD may help to improve the exchange of information between typology/language documentation and computational linguistics: recent years have seen an increased interest in the computational linguistics community to develop NLP resources for less-resourced languages, to explore statistical approaches on annotation projection and annotation-sparse NLP algorithms, but little of this is known in more theoretically communities. Integrating Glottolog/Langdoc in the LLOD cloud along with NLP resources could encourage NLP researchers to register their resources in Langdoc, and thereby to overcome the gap between theoretically-oriented research and statistical NLP.

36. `http://multitree.linguistlist.org`
37. `http://linguistlist.org`
38. `http://www.csufresno.edu/odin`
39. `http://wals.info`
40. `http://www.language-archives.org`
41. `http://lexvo.org`
42. `http://linguistics-ontology.org`
43. `http://lexvo.org`
44. `http://www.w3.org/2003/01/geo/wgs84_pos`
45. `http://www.worldcat.org`
46. `http://books.google.com`
47. `http://openlibrary.org`

Table 3. *Documentation status of the languages in Langdoc (excluding resource-heavy languages like English, French or German).*

Language	Refs	Language	Refs
Swahili	1,916	Nyanja	504
Hausa	1,609	Arabic, Tunisian	498
Nama	1,288	Tachelhit	490
Zulu	1,060	Wolof	487
Arabic, South Levantine	1,033	Tibetan	483
Yoruba	925	Sotho, Northern	467
Kabyle	897	Aymara, Central	462
Thai	745	Aymara, Southern	454
Pulaar	743	Vietnamese	439
Xhosa	739	Paraguayan Guaraní	436
Akan	729	Singa	405
Éwé	713	16 languages	300-399
Tswana	703	31 languages	200-299
Mapudungun	610	159 languages	100-199
Shona	597	389 languages	50-99
Somali	591	647 languages	25-49
Amharic	554	611 languages	15-24
Igbo	550	533 languages	10-14
Sotho, Southern	539	1,033 languages	5-9
Arabic, Algerian	526	351 languages	4
Oromo, Borana-Arsi-Guji	516	436 languages	3
Turkish	511	612 languages	2
Tarifit	505	1,045 languages	1

7. Integrating and using language resources: LLOD

Tim Berners-Lee coined the idea of the Giant Global Graph [48], which connects all data and allows discovery of new relations between the data. This idea has been pursued in the Linked Open Data community, where the Linked Open Data cloud now numbers 295 repositories and 31,634,213,770 RDF triples. [49]

Although it is difficult to objectively identify reasons for the success of the LOD cloud, advocates generally argue that open licences as well as open access are a key enabler for the growth of such a network as they provide a strong incentive for collaboration and contribution by third parties. Bizer (2011b) 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 summarized several methodologies (Auer and Lehmann, 2010; Berners-Lee, 2006; Bizer, 2011b) in Figure 5 and will relate to the steps throughout this Section. Before we go into detail on the creation of a Linguistic LOD cloud, however, we will elaborate on three aspects (Open licences, RDF as a data format and scalability) regarding the evolution of the original LOD cloud.

Open licences, open access and collaboration: DBpedia, FlickrWrappr, 2000 U.S. Census, LinkedGeoData, LinkedMDB are prominent examples of LOD data sets, where the conversion, interlinking, and the hosting of the links and the converted RDF data has been completely provided by third-party stakeholders with almost no development cost for the original data providers [50]. DBpedia, for example, was initially converted to RDF by a university from the open data dumps provided by Wikipedia. A company

48. Accessed on Jan, 20th, 2012, `http://dig.csail.mit.edu/breadcrumbs/node/215`, November 2007.
49. Accessed on Jan, 20th, 2012, `http://www4.wiwiss.fu-berlin.de/lodcloud/state/`, September 2011.
50. More data sets can be explored here: `http://thedatahub.org/tag/published-by-third-party`

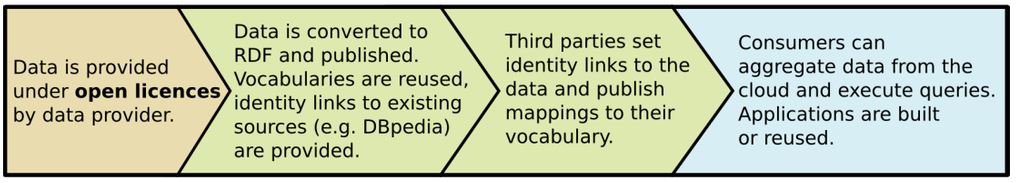

Figure 5. *Summary of several methodologies for publishing and exploiting Linked Data. The data provider is only required to make data available under an open licence (left-most step). The remaining steps for data integration can be contributed by third parties and data consumers.*

then provided the hosting 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 by members of the OWLG, could not be re-published due to licensing issues. In particular, 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). Very often (as is the case in the previous two examples), the reason for closed licences is the strict copyright of the primary data (such as newspaper texts) and researchers are unable to publish their resulting data. The open part of the American National Corpus (OANC [51]) on the other hand has been converted to RDF and was re-published successfully using POWLA (Chiarcos, 2012b). In this manner, the work contributed to OANC was directly reusable for other scientists and likewise the same accounts for the RDF conversion as it will also be public.

Note that the *Open* in Linked Open Data refers mainly to *open access*, i.e. retrievable by HTTP. [52] Only around 18% of the data sets of the LOD cloud provide clear licensing information at all. [53] Of these 18% an even smaller amount is considered *open* after the definition of the OKFN.

RDF as a data model: RDF as a data model has distinctive features when compared to its alternatives. [54] 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, which 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) cannot be supplemented by transformations such as XSLT, because the linking and mappings are implicit. All three aspects are important to enable ad-hoc collaboration by interest groups such as the OWLG. The resulting technology mix provided by RDF allows any collaborator to join their data into the decentralized data network over HTTP with immediate benefits for the original collaborator and others. Although workarounds might be constructed (possibly leading to something similar to RDF), the creation of a Web of Data with alternative technologies to RDF can currently be considered infeasible due to the lack of alternatives.

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 storage systems. This expressivity poses a performance challenge to query answering in RDF triples stores and inferencing in OWL reasoners, and of course also to the combination thereof. Although there are efforts to load as many triples as possible within one store, [55] the strength of RDF is its flexibility and suitability for data

51. `http://www.anc.org/OANC/`
52. Accessed on Jan, 20th, 2012 `http://richard.cyganiak.de/2007/10/lod/\#open`
53. Accessed on Jan, 20th, 2012 `http://www4.wiwiss.fu-berlin.de/lodcloud/state/\#license`
54. We deliberatively omitted Topic Maps, which offer similar features, but are not widely supported.
55. `http://factforge.net` or `http://lod.openlinksw.com` provide SPARQL interfaces to query billions of aggregated facts.

integration and not superior performance for specific use cases. Therefore many RDF libraries are often integrated with indexing systems (e.g. Jena and Lucene [56]), which provide the required performance. Furthermore, many RDF systems are designed to be deployed in parallel to existing high-performance systems and not as a replacement. An overview over systems that provide Linked Data and SPARQL on top of relational database systems can be found in Auer *et al.* (2009). The NLP Interchange Format (cf. Section 7.3) allows to express the output of highly optimized NLP systems (e.g. UIMA [57]) as RDF/OWL.

7.1. *Current state of the LLOD*

Since its foundation in late 2010, the OWLG has made progress in the creation and identification of resources, and in providing resources as RDF. The primary resource types identified in this context are lexical-semantic resources (e.g. DBpedia, Section 3), linguistic corpora (e.g. the POWLA formalization of the French and English Europarl, Section 4), repositories of linguistic terminology (e.g. OLiA, Section 5) and typological databases and metadata repositories (e.g. Glottlog/Langdoc, Section 6).

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 assembled in Chiarcos *et al.* (2012), 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 density of the graph, as well. Figure 6 shows a current sketch of the LLOD cloud. [58]

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 *et al.*, 2010). Like these, other corpora with comparable annotations can be represented in RDF/OWL using the POWLA scheme.

Using tools like DBpedia Spotlight (Section 3.4), 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.) Other general knowledge bases that are available in the LOD have been included in the diagram besides DBpedia: YAGO, [59] OpenCyc, [60] the Open Data Thesaurus, [61] different versions of the English WordNet [62] and the Dutch WordNet Cornetto. [63] Lemon is a formalism to publish lexical resources as Linked Data and has been applied to WordNet, Wiktionary and other resources (McCrae *et al.*, 2011), and it builds on earlier models such as LingInfo (Buitelaar *et al.*, 2006) and LexOnto (Cimiano *et al.*, 2007). At the moment, other groups are actively working on further Wiktionary instantiations that may be integrated into the LLOD, e.g. Meyer and Gurevych (2010). RDF versions of FrameNet are also developed, but have not yet been publicly released (Scheffczyk *et al.*, 2006; Picca *et al.*, 2008).

56. `http://jena.sourceforge.net/ARQ/lucene-arq.html`
57. Apache UIMA project `http://uima.apache.org`
58. It should be noted that the LLOD cloud is still work in progress. The resources in Figure 6 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`).
59. `http://www.mpi-inf.mpg.de/yago-naga/yago`
60. `http://www.opencyc.org`
61. `http://thedatahub.org/dataset/open-data-thesaurus`
62. `http://ckan.net/dataset/w3c-wordnet`, `http://ckan.net/dataset/vu-wordnet`, `http://ckan.net/dataset/rkb-explorer-wordnet`
63. `http://ckan.net/dataset/cornetto`

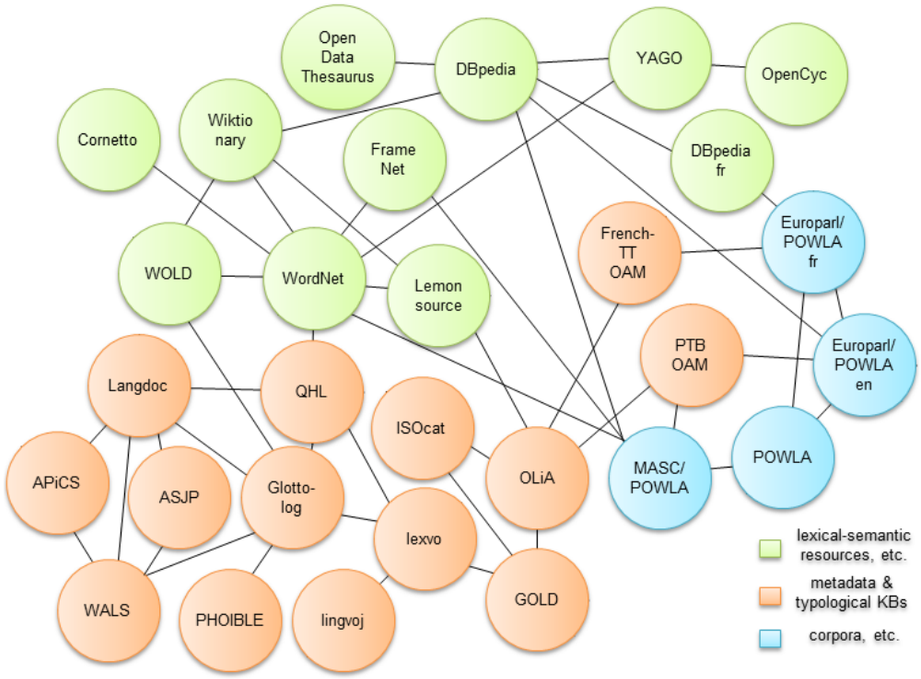

Figure 6. *Current draft for the Linguistics Linked Open Data (LLOD) cloud. Source:* `http://linguistics.okfn.org/resources/llod`

In the group of lexical-semantic resources, the World Loanword Database (WOLD) [64] 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 and Cysouw, 2012). Besides Glottolog and Langdoc, other typological databases in the diagram include the World Atlas of Language Structures (WALS); [65] the Atlas of Pidgin and Creole Language Structures (Michaelis *et al.*, in preparation, APiCS); the Phonetics Information Base and Lexicon (PHOIBLE), [66] containing phoneme inventories from over 1,000 languages (Moran, 2012); and the Automated Similarity Judgment Program (Brown *et al.*, 2008, ASJP), [67] 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 [68] 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.

64. `http://wold.livingsources.org`
65. `http://www.wals.info`
66. `http://phoible.org`
67. `http://cldbs.eva.mpg.de/asjp`
68. `http://ckan.net/dataset/lexvo`, `http://ckan.net/dataset/lingvoj`

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 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 5 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. [69]

7.2. *Querying linked resources in the LLOD*

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, we have given an example for the querying of multiple interlinked resources in Section 4, 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 5 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.

7.2.1. *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 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. [70] 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 {
```

69. 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.

70. 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 and Good, 2010) whose lexicons are linked to the GOLD ontology (Farrar and Langendoen, 2003a). The queries necessary for this purpose would be, however, almost identical.

```
  ?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.

7.2.2. *Enriching lexical-semantic resources with linguistic information (DBpedia ($\mapsto$ POWLA) $\mapsto$ 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 scms:  <http://ns.aksw.org/scms/>.
CONSTRUCT { ?semClass <#realizedAs> ?syntClass }
WHERE {
  ?x a powla:Node.
  ?x scms:means ?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.

7.3. *Conducting Web annotations*

So far, we have focused on the description of *resource* modeling and querying of the (Linguistic) Linked Open Data cloud. Here, we address the *application aspect* of these resources.

For example, with the DBpedia serving as an entity repository it is possible to link the Web of Documents with the Web of Data via DBpedia identifiers. This function is provided by DBpedia Spotlight (Section 3.4). This section describes the NLP Interchange Format (NIF), and using its URI Recipes, annotations like those of DBpedia Spotlight can be used to represent the linking transparently for applications. Using the OLiA ontologies to represent linguistic annotations, NIF allows to represent the output of classical NLP tools (tagger, parser, etc.), thus mingling documents, annotations and data in a uniform way.

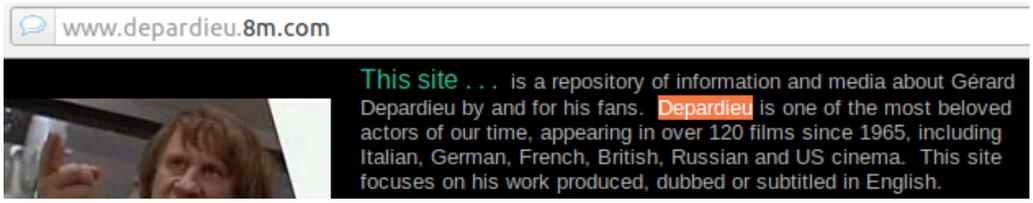

Figure 7. *The second occurrence of* `Depardieu` *is highlighted and is linked in the example with the French DBpedia resource about Gérard Depardieu. Source: http://www.depardieu.8m.com/.*

7.3.1. *NLP Interchange Format*

The NLP Interchange Format (NIF) (Hellmann *et al.*, 2012a) [71] brings together most of the resources described so far. It is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. The core of NIF consists of a vocabulary, which can represent Strings as RDF resources. A special URI Design is used to pinpoint annotations to a part of a document. These URIs can then be used to attach arbitrary annotations to the respective character sequence. Based on these URIs, annotations can be interchanged between different NLP tools and applications. NIF consists of 3 components:

Structural interoperability URI recipes are used to anchor annotations in documents with the help of fragment identifiers (Section 7.3.2). The URI recipes are complemented by two ontologies (String Ontology and Structured Sentence Ontology), which are used to describe the basic types of these URIs (String, Document, Word, Sentence) as well as the relations between them (sub/super string, next/previous word).

Conceptual interoperability Best practices for annotating these URIs are given to provide interoperability. OLiA, as presented in Section 5, is used for the grammatical features, the SCMS Vocabulary (Ngonga Ngomo *et al.*, 2011), [72] DBpedia and NERD (Rizzo *et al.*, 2012) [73] are used for sense tagging. Furthermore NIF can be used with Lemon (McCrae *et al.*, 2011) and other data in RDF.

Access interoperability An interface description for NIF Components and Web Services allows NLP tools to interact on a programmatic level.

7.3.2. *Anchoring Web annotations*

One basic use case of NIF is to allow NLP tools to exchange annotations about (Web) documents in RDF. The first prerequisite is that Strings are referenced by URIs, so they can be used as a subject in RDF triples. A quite simple example is depicted in Figure 7, which can be addressed with two possible URI recipes according to the NIF 1.0 specification:

1) a URI scheme is used with offsets, which is easy to compute and handle programmatically. Note that the HTML document is treated as a string or a character sequence. The # is used in this case to address a fragment of the whole document, hence the naming "fragment identifier":

`http://www.depardieu.8m.com/#offset_22295_22304_Depardieu`

2) a URI scheme based on the context and md5 hashes, which is more stable w.r.t. to offset changes. Here the context is 6 characters before and after the occurrence and the actual substring has to be enclosed in brackets to produce the message for the md5 digest: "`nbsp; (Depardieu) is on`":

`http://www.depardieu.8m.com/#hash_6_9_e7146a74239c3878aedf0c45c6276618_Depardieu`

A NIF 1.0 model, which links the second occurrence of *Depardieu* to the French DBpedia, has to contain the following RDF triples:

71. Specification 1.0: `http://nlp2rdf.org/nif-1.0`
72. `http://scms.eu`
73. `http://nerd.eurecom.fr/ontology/`

```
@prefix :       <http://www.depardieu.8m.com/#>
@prefix fr:     <http://fr.dbpedia.org/resource>
@prefix str:    <http://nlp2rdf.lod2.eu/schema/string/> .
@prefix scms:   <http://ns.aksw.org/scms/> .

:offset_22295_22304_Depardieu                :offset_0_54093_%3C!doctype%20html%20publi
    scms:means  fr:Gerard_Depardieu ;            rdf:type    str:OffsetBasedString ;
    rdf:type    str:OffsetBasedString .          rdf:type    str:Document ;
                                                 rdf:type    str:Document ;
                                                 str:subString :offset_22295_22304_Depardieu;
                                                 str:sourceUrl <http://www.depardieu.8m.com/> .
```

Similarly, the output of NLP tools can be represented, e.g. by associating *Depardieu* with its language (e.g. a Glottolog languoid), with a syntactic parse tree (as specified in POWLA), or with morphosyntactic annotations (as provided by OLiA). Future research has to show whether NIF can additionally serve as a Meaning Representation Language (MRL) (Hellmann, 2010).

8. Summary

In this article, we have introduced the Open Linguistics Working Group (OWLG), an initiative of experts from different fields concerned with linguistic data, including academic linguistics, applied linguistics and information technology. The primary goals of the working group are to promote the idea of open linguistic resources, to develop means for their representation, and to encourage the exchange of ideas across different disciplines.

Although it is difficult to measure the benefit of open licenses and free language resources, closed and unclearly licensed data are a major obstacle for data conversion and republishing processes, and prevent collaborative evolution of data sets. This obstacle became particularly obvious when data providers (linguists and domain experts), re-publishers (focused on representation and integration of data) and data consumers (NLP engineers) met in the context of the OWLG and discussed these issues.

The activities of the OWLG converge towards the creation of a Linguistic Linked Open Data (LLOD) cloud. This article described formalisms and methodologies relevant for the major types of resources within this LLOD cloud, illustrated here by representative examples, including lexical-semantic resources (DBpedia), linguistic corpora (POWLA), and data collections about linguistic terminology (OLiA), as well as languages and language resources (Glottolog/Langdoc). We described how RDF can be employed to achieve interoperability between these and other resources, and the possibility to integrate information from different sources on the basis of the concept of Linked Data. RDF is a formalism with a sufficient degree of generality, with broad technological support, and maintained by an active community, so that it is at the moment the most promising formalism to achieve interoperability and information integration from an unrestricted set of linguistic resources. Moreover, necessary preconditions for concrete applications built on this basis were described, in particular, the NLP Interchange Format (NIF) that enables the creation of NLP pipelines that directly assess resources from the LLOD cloud.

Acknowledgements

We would like to thank the members of the Open Linguistics Working Group, in particular Jonas Brekle, Philipp Cimiano, Judith Eckle-Kohler, Richard Littauer, Michael Matuschek, John McCrae and Steve Moran. Further, we would like to thank Pablo Mendes and Claus Stadler for their contributions to this work in regard to DBpedia, three anonymous reviewers and the editorial board of the TAL special issue for comments and feedback.

The research of Christian Chiarcos was partially funded by the German Research Foundation (DFG) through the Collaborative Research Center (SFB) 632 “Information Structure” at the University of Potsdam. The work of Sebastian Hellmann was carried out in the context of the LOD2 project, co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement No. 257943). Sebastian Nordhoff’s research was conducted at the Max Planck Institute for Evolutionary Anthropology, Leipzig.

We would also like to note that this article summarizes a number of earlier publications of the authors on the topic, in parts, it builds on descriptions about the progress of the Open Linguistics Working Group and the development of a Linguistic Linked Open Data cloud published in the accompanying volume for the Workshop on Linked Data in Linguistics (LDL-2012, Chiarcos *et al.*, 2012) that were substantially extended, revised and augmented with novel material for the current publication.

9. References

Abeillé A., Clement L., Kinyon A., "Building a treebank for French", *In Proc. LREC 2000*, 2000.

Auer S., Dietzold S., Lehmann J., Hellmann S., Aumueller D., "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*, ACM, p. 621-630, 2009.

Auer S., Lehmann J., "Making the Web a Data Washing Machine - Creating Knowledge out of Interlinked Data", *Semantic Web Journal*, 2010.

Baader F., Horrocks I., Sattler U., "Description logics as ontology languages for the Semantic Web", *in* D. Hutter, W. Stephan (eds), *Mechanizing Mathematical Reasoning*, Springer Berlin / Heidelberg, p. 228-248, 2005.

Berners-Lee T., "Design Issues: Linked Data", , `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

Bird S., Liberman M., "A formal framework for linguistic annotation", *Speech Communication*, vol. 33, n° 1-2, p. 23-60, 2001.

Bizer C., "DBpedia 3.7 released, including 15 localized Editions", , `http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/`, 2011a.

Bizer C., "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`, 2011b. Keynote at 28th British National Conference on Databases (BNCOD2011).

Bouda P., Cysouw M., "Treating Dictionaries as a Linked-Data Corpus", *in* Chiarcos *et al.* (2012), p. 15-23, 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.

Brown C. H., Holman E. W., Wichmann S., Velupillai. V., "Automated classification of the world's languages: A description of the method and preliminary results", *STUF - Language Typology and Universals*, vol. 61, n° 4, p. 286-308, 2008.

Buitelaar P., Declerck T., Frank A., Racioppa S., Kiesel M., Sintek M., Engel R., Romanelli M., Sonntag D., Loos B., Micelli V., Porzel R., Cimiano P., "LingInfo: Design and applications of a model for the integration of linguistic information in ontologies", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2006.

Burchardt A., Padó S., Spohr D., Frank A., Heid U., "Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control", *Proc. 3rd International Joint Conf on NLP (IJCNLP 2008)*, Hyderabad, 2008.

Buyko E., Chiarcos C., Pareja-Lora A., "Ontology-Based Interface Specifications for a NLP Pipeline Architecture", *Proc. LREC 2008*, Marrakech, Morocco, 2008.

Chiarcos C., "An ontology of linguistic annotations", *LDV Forum*, vol. 23, n° 1, p. 1-16, 2008.

Chiarcos C., "Grounding an Ontology of Linguistic Annotations in the Data Category Registry", *LREC 2010 Workshop on Language Resource and Language Technology Standards (LT&LTS)*, Valetta, Malta, p. 37-40, May, 2010a.

Chiarcos C., "Towards Robust Multi-Tool Tagging. An OWL/DL-Based Approach", *ACL 2010*, Uppsala, Sweden, p. 659-670, July, 2010b.

Chiarcos C., "Interoperability of Corpora and Annotations", *in* C. Chiarcos, S. Nordhoff, S. Hellmann (eds), *Linked Data in Linguistics*, Springer, 2012a. p. 161-179.

Chiarcos C., "POWLA: Modeling linguistic corpora in OWL/DL", *Proceedings of 9th Extended Semantic Web Conference (ESWC2012)*, 2012b.

Chiarcos C., Dipper S., Götze M., Leser U., Lüdeling A., Ritz J., Stede M., "A Flexible Framework for Integrating Annotations from Different Tools and Tagsets", *TAL (Traitement automatique des langues)*, 2008.

Chiarcos C., Erjavec T., "OWL/DL formalization of the MULTEXT-East morphosyntactic specifications", *5th Linguistic Annotation Workshop*, Portland, p. 11-20, 2011.

Chiarcos C., Nordhoff S., Hellmann S. (eds), *Linked Data in Linguistics. Representing Language Data and Metadata*, 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.

Chiarcos C., Ritz J., Stede M., "By all these lovely tokens ... Merging Conflicting Tokenizations", *Journal of Language Resources and Evaluation (LREJ)*, 2011. to appear.

Cimiano P., Haase P., Herold M., Mantel M., Buitelaar P., "LexOnto: A model for ontology lexicons for ontology-based NLP", *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC*, vol. 7, 2007.

Dipper S., "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation", *Proc. Berliner XML Tage 2005 (BXML 2005)*, Berlin, Germany, p. 39-50, 2005.

Erjavec T., "MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora", *Fourth International Conference on Language Resources and Evaluation, (LREC 2004)*, Lisboa, Portugal, p. 1535-1538, May, 2004.

Farrar S., Langendoen D., "Markup and the GOLD ontology", *EMELD Workshop on Digitizing and Annotating Text and Field Recordings*, Michigan State University, July, 2003a.

Farrar S., Langendoen D. T., "A Linguistic Ontology for the Semantic Web", *GLOT International*, vol. 7, p. 97-100, 2003b.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. *et al.*, "Lexical markup framework (LMF)", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May, 2006.

Good J., Hendryx-Parker C., "Modeling Contested Categorization in Linguistic Databases", *Proceedings of the EMELD Workshop on Digital Language Documentation*, East Lansing, Michigan, 2006.

Hammarström H., Nordhoff S., "LangDoc: Bibliographic Infrastructure for Linguistic Typology", *Oslo Studies in Language*, vol. 3, n° 2, p. 31-43, 2011. Language Variation Infrastructure.

Hellmann S., "The Semantic Gap of Formalized Meaning.", *in* L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, T. Tudorache (eds), *ESWC (2)*, vol. 6089 of *Lecture Notes in Computer Science*, Springer, p. 462-466, 2010.

Hellmann S., Lehmann J., Auer S., "Linked-Data Aware URI Schemes for Referencing Text Fragments", *EKAW 2012*, Lecture Notes in Artificial Intelligence (LNAI), Springer, 2012a.

Hellmann S., Stadler C., Lehmann J., "The German DBpedia: A Sense Repository for Linking Entities", *in* Chiarcos *et al.* (2012), p. 181-190, 2012b. 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.

Hellmann S., Unbehauen J., Chiarcos C., Ngonga Ngomo A.-C., "The TIGER Corpus Navigator", *9th International Workshop on Treebanks and Linguistic Theories (TLT-9)*, Tartu, Estonia, p. 91-102, 2010.

Hepp M., Siorpaes K., Bachlechner D., "Harvesting Wiki Consensus: Using Wikipedia Entries as Vocabulary for Knowledge Management", *IEEE Internet Computing*, vol. 11, n° 5, p. 54-65, 2007.

Ide N., Fellbaum C., Baker C., Passonneau R., "The Manually Annotated Sub-Corpus: A community resource for and by the people", *Proceedings of the ACL 2010 Conference Short Papers*, Association for Computational Linguistics, p. 68-73, 2010.

Ide N., Pustejovsky J., "What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability", *Proc. Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China, 2010.

Ide N., Romary L., "International standard for a linguistic annotation framework", *Natural language engineering*, vol. 10, n° 3-4, p. 211-225, 2004.

Ide N., Suderman K., "GrAF: A Graph-based Format for Linguistic Annotations", *Proc. Linguistic Annotation Workshop (LAW 2007)*, Prague, Czech Republic, p. 1-8, 2007.

Kemps-Snijders M., Windhouwer M., Wittenburg P., Wright S., "ISOcat: Remodelling metadata for language resources", *International Journal of Metadata, Semantics and Ontologies*, vol. 4, n° 4, p. 261-276, 2009.

Koehn P., "Europarl: A parallel corpus for statistical machine translation", *MT summit*, vol. 5, 2005.

Kontokostas D., Bratsas C., Auer S., Hellmann S., Antoniou I., Metakides G., "Towards Linked Data Internationalization - Realizing the Greek DBpedia", *Proceedings of the ACM WebSci'11*, 2011.

Kontokostas D., Bratsas C., Auer S., Hellmann S., Antoniou I., Metakides G., "Internationalization of Linked Data: The case of the Greek DBpedia edition", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. , n° 0, p. -, 2012.

Lassila O., Swick R. R., Resource Description Framework (RDF) Model and Syntax Specification, Technical report, World Wide Web Consortium, 1999.

Leech G., Wilson A., "EAGLES Recommendations for the Morphosyntactic Annotation of Corpora", , `http://www.ilc.cnr.it/EAGLES/annotate/annotate.html`, 1996.

Lehmann J., Bizer C., Kobilarov G., Auer S., Becker C., Cyganiak R., Hellmann S., "DBpedia - A Crystallization Point for the Web of Data", *Journal of Web Semantics*, vol. 7, n° 3, p. 154-165, 2009.

Lewis M. P. (ed.), *Ethnologue: Languages of the World*, 16 edn, SIL, Dallas, 2009.

Marcus M., Santorini B., Marcinkiewicz M., "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics*, vol. 19, n° 2, p. 313-330, 1994.

Martin F., Spohr D., Stein A., "Representing a resource of formal lexicalsemantic descriptions in the Web Ontology Language", *Journal for Language Technology and Computational Linguistics*, vol. 21, p. 1-22, 2009.

McCrae J., Spohr D., Cimiano P., "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon", *The Semantic Web: Research and Applications*, vol. , p. 245-259, 2011.

McGuinness D., Van Harmelen F., OWL Web Ontology Language overview. W3C recommendation, Technical report, World Wide Web Consortium, 2004.

Mendes P. N., Jakob M., García-Silva A., Bizer C., "DBpedia Spotlight: Shedding Light on the Web of Documents", *Proc. 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

Meyer C., Gurevych I., "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*, Springer, p. 38-49, 2010.

Michaelis S., Maurer P., Haspelmath M., Huber M. (eds), *Atlas of Pidgin and Creole Language Structures*, Oxford University Press, Oxford, in preparation.

Miles A., Bechhofer S., SKOS Simple Knowledge Organization System reference. W3C Recommendation, Technical report, World Wide Web Consortium, 2009.

Moran S., "Using Linked Data to create a typological knowledge base", *in* Chiarcos *et al.* (2012), p. 129-138, 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.

Ngonga Ngomo A.-C., Heino N., Lyko K., Speck R., Kaltenböck M., "Semantifying Content Management Systems", *International Semantic Web Conference (ISWC)*, 2011.

Nordhoff S., Hammarström H., "Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources", *Proceedings of the First International Workshop on Linked Science 2011*, vol. 783 of *CEUR Workshop Proceedings*, 2011.

Picca D., Gliozzo A., Gangemi A., "LMM: An OWL-DL metamodel to represent heterogeneous lexical knowledge", *Proceedings of LREC, Marrakech, Morocco*, 2008.

Poornima S., Good J., "Modeling and Encoding Traditional Wordlists for Machine Applications", *Proc. 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, Association for Computational Linguistics, Uppsala, Sweden, p. 1-9, July, 2010.

Quasthoff M., Hellmann S., Höffner K., "Standardized Multilingual Language Resources for the Web of Data: http://corpora.uni-leipzig.de/rdf", *3rd prize at the LOD Triplification Challenge, Graz, 2009*, 2009.

Rehm G., Eckart R., Chiarcos C., "An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora", *Proc. RANLP 2007*, Borovets, Bulgaria, 2007.

Rizzo G., Troncy R., Hellmann S., Bruemmer M., "NERD meets NIF: Lifting NLP extraction results to the linked data cloud", *LDOW, 5th Workshop on Linked Data on the Web, April 16, 2012, Lyon, France*, 04, 2012.

Scheffczyk J., Pease A., Ellsworth M., "Linking FrameNet to the Suggested Upper Merged Ontology", *Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS 2006)*, Baltimore, Maryland, USA, p. 289-300, November, 2006.

Skopeteas S., Fiedler I., Hellmuth S., Schwarz A., Stoel R., Fanselow G., Krifka M., "Questionnaire on Information Structure: Reference Manual", *Interdisciplinary Studies on Information Structure (ISIS),*, 2006.

Stein A., "French TreeTagger Part-of-Speech Tags", , `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`, 2003.

Tapanainen P., Järvinen T., "A non-projective dependency parser", *Proc. fifth conference on Applied natural language processing*, Association for Computational Linguistics, p. 64-71, 1997.

Tiedemann J., "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces", *in* N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds), *Recent Advances in Natural Language Processing (Volume V)*, John Benjamins, Amsterdam/Philadelphia, p. 237-248, 2009.

W3C OWL Working Group, OWL 2 Web Ontology Language. Document Overview. W3C Recommendation, Technical report, World Wide Web Consortium, 2009.

Weibel S., Kunze J., Lagoze C., , Wolf M., "RFC 2413 - Dublin Core Metadata for Resource Discovery", , `http://www.isi.edu/in-notes/rfc2413.txt`, September, 1998.

Xie Y., Aristar-Dry H., Aristar A., Lockwood H., Thompson J., Parker D., Cool B., "Language and Location: Map Annotation Project - A GIS-based infrastructure for linguistics information management", *Computer Science and Information Technology, 2009. IMCSIT '09. International Multiconference on*, p. 305 -311, oct., 2009.

Zeldes A., Ritz J., Lüdeling A., Chiarcos C., "ANNIS: A search tool for multi-layer annotated corpora", *Proc. Corpus Linguistics*, Liverpool, UK, p. 20-23, July, 2009.