# The Open Linguistics Working Group

Chiarcos, Christian ; Hellmann, Sebastian ; Nordhoff, Sebastian ; Moran, Steven ; Littauer, Richard ; Eckle-Kohler, Judith ; Gurevych, Iryna ; Hartmann, Silvana ; Matuschek, Michael ; Meyer, Christian M

# The Open Linguistics Working Group

**Christian Chiarcos[*], Sebastian Hellmann[†], Sebastian Nordhoff[‡], Steven Moran[§], Richard Littauer[¶], Judith Eckle-Kohler[**], Iryna Gurevych[**,††], Silvana Hartmann[**], Michael Matuschek[**], Christian M. Meyer[**]**

[*]Information Science Institute, University of Southern California
chiarcos@daad-alumni.de

[†] Department of Computer Science, University of Leipzig
hellmann@informatik.uni-leipzig.de

[‡] Department of Linguistics, Max Planck Institute for Evolutionary Anthropology
sebastian_nordhoff@eva.mpg.de

[§] Research Unit "Quantitative Language Comparison", University of Munich
steve.moran@lmu.de

[¶] Computational Linguistics Department, Saarland University
littauer@coli.uni-saarland.de

[**] Ubiquitous Knowledge Processing Lab (UKP-TUDA), Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de

[††] Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information, Frankfurt/M.
http://www.ukp.tu-darmstadt.de

**Abstract**

This paper describes the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN). The OWLG is an initiative concerned with linguistic data by scholars from diverse fields, including linguistics, NLP, and information science. The primary goal of the working group is to promote the idea of open linguistic resources, to develop means for their representation and to encourage the exchange of ideas across different disciplines. This paper summarizes the progress of the working group, goals that have been identified, problems that we are going to address, and recent activities and ongoing developments.
Here, we put particular emphasis on the development of a Linked Open Data (sub-)cloud of linguistic resources that is currently being pursued by several OWLG members.

**Keywords:** open research, linguistics resources, open access, linked data, Linguistic Linked Open Data

## 1. Background

The Open Knowledge Foundation (OKFN)[1] is a community-based non-profit organization promoting open knowledge – data and content that is free to use, re-use and to be distributed without restriction. The OKFN defines standards, develops tools that allow people to create, find and share open material, and organizes working groups and events. For example, the **Open Definition** standard sets out principles to define "openness". The definition can be summed up in the statement that "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."[2]
One tool that the OKFN provides is **CKAN**,[3] a catalog system for open datasets. CKAN is an open-source data portal software developed to easily publish, find and reuse open content and data, particularly in ways that are also machine automatable. The OKFN also hosts various working groups addressing problems of open data in different domains. Currently, there are 18 OKFN **working groups** covering fields as diverse as government data, economics, archeology, text books and cultural heritage. The OKFN organizes various events, such as the Open Knowledge Conference (OKCon), and facilitates the exchange of ideas between different working groups. The vision of the OKFN is a world in which open knowledge is ubiquitous. In this paper, we apply the aims of open knowledge to data in the field of linguistics.

In 2010, the **Open Linguistics Working Group** (OWLG) was founded and it has grown steadily since. In this paper, we highlight our goals and current projects, and show how we hope to attain "openness" in linguistics through: 1) promoting the idea of open linguistic resources, 2) devel-

[1] http://okfn.org/
[2] http://opendefinition.org/
[3] http://ckan.org/

oping the means for the representation of open data, and 3) encouraging the exchange of ideas across different disciplines.

Publishing linguistic data under open licenses is an important issue in academic research, as well as in the development of applications. We see increasing support for this in the linguistics community (Pederson, 2008), and there are a growing number of resources published under open licenses, such as corpora (Meyers et al., 2007). There are many reasons for publishing resources under open licenses: for instance, freely available data can be more easily reused, double investments can be avoided, and results can be replicated. Also, other researchers can build on this data, and subsequently refer to the publications associated with it. Nevertheless, a number of ethical, legal and sociological problems are associated with open data, and technologies that establish interoperability (and thus, re-usability) of linguistic resources are still under development. The OWLG represents an open forum for interested researchers to address these and related problems.

## 2. Goals

As a result of discussions with interested linguists, NLP engineers and information technology experts, we have identified seven goals for the OWLG. By pursuing these goals, we can help these communities, each of which has their own way of using, accessing, and sharing linguistic data:

1. Promote the idea and principles behind "openness" of content and data, as defined by the Open Definition, in linguistics and in relation to language data.

2. Act as a central point of reference and support for those interested in open linguistic data.

3. Facilitate communication between researchers from different sub-communities within the large field of linguistics that use, distribute, or maintain open linguistic data.

4. Serve as a mediator between providers and users of technical infrastructure.

5. Build and maintain an index of open linguistic data sources and tools that link existing resources. This involves registering resources in CKAN, as well as collecting candidates for the creation of a Linked Open Data cloud of linguistic resources (Sect. 4.).

6. Assemble best-practice guidelines and use cases concerning creating, using and distributing data.

7. Gather information on legal issues surrounding linguistic data to the community. This is a recurring problem in various domains, touching laws on copyright (as in corpus linguistics, or in compiled works such as databases or bibliographies) and privacy (endangered language archives).

In many aspects, the aims of the OWLG are not unique. Indeed, there are numerous initiatives with similar motivation and overlapping aims, e.g., Cyberling,[4], the ACL Special Interest Group for Annotation (SIGANN),[5] or the W3C Ontology-Lexica Community Group (OntoLex).[6] While it supports these sub-community-specific efforts, the OWLG is not restricted to either academic linguistics, or Natural Language Processing, or computational lexicography communities *alone*, but emphasizes its interdisciplinary character.

Unlike large multi-national initiatives such as the ISO Technical Committee on Terminology and other language and content resources (ISO TC37),[7] the American initiative on Sustainable Interoperability of Language Technology (SILT),[8] and European projects such as the initiative on Common Language Resources and Technology Infrastructure (CLARIN),[9] the Fostering Language Resources Network (FLaReNet),[10] and the Multilingual Europe Technology Alliance (META),[11] that also pursue partially similar goals, the OWLG is a network of individual researchers rather than institutions, and it facilitates the exchange of ideas rather than deliverables. The OWLG may thus complement the institutional approach on interdisciplinary collaboration. Cooperations with other initiatives in the field are highly welcome – in fact, several OWLG members are engaged in these.

One central aspect in our work is that we focus on *open* linguistic resources and on the problems and benefits associated with using, maintaining, and distributing open linguistic resources. The OWLG provides a platform for sharing experiences and technology across discipline boundaries, as researchers work with field-specific technologies, but face similar issues. For instance, heterogeneous data, interoperability or the question of exit strategies arise in lexicography, corpus research, and linguistic typology alike.

Shared technology can facilitate collaboration and reusability across discipline borders, but the OWLG does not act as a creator of software, nor does it collect and preserve linguistic resources. It is instead a forum that connects researchers addressing such problems, and through the exchange of experience and data between its members, it contributes to the development of interoperable infrastructures. Our community aims at sharing resources through technological infrastructures and assisting in addressing more general questions that arise out of open data.

One set of technologies that is particularly appealing to several OWLG members is represented by the Linked (Open) Data paradigm (Berners-Lee, 2006; Heath and Bizer, 2011):

[4] http://cyberling.org
[5] http://www.cs.vassar.edu/sigann
[6] http://www.w3.org/community/ontolex
[7] http://www.iso.org/tc37
[8] http://www.anc.org/SILT
[9] http://www.clarin.eu
[10] http://www.flarenet.eu
[11] http://www.meta-net.eu

1. Referred entities should be designated unambiguously by URIs,
2. these URIs should be resolvable over HTTP,
3. data should be represented by means of established standards (e.g., RDF),
4. and a resource should include links to other resources.

Historically, Linked Data is coupled with specific formats like RDF, but the OWLG does not prescribe the use of any specific format. Yet, cooperations between individual OWLG members have been initiated that may eventually lead to the conversion of further data sets to RDF.

Linked Data is closely associated with the idea of openness (otherwise, links to other resources can only be resolved under certain circumstances), and in 2010, the definition of Linked Data has been extended with a 5 star rating system for data on the web.[12] The first star is achieved by publishing data on the web (in any format) under an open license. Publishing, or working towards publishing linguistic resources with this condition represents the minimal requirement for interested researchers to participate in the OWLG.

In summary, the OWLG is more of a network than an organization. It follows a grass-roots approach. Because of this, it does not need to depend on centralized funding. Our technical infrastructure is provided by the OKFN, and our work is indirectly supported through the institutions of different members, e.g., the European LOD2 project and the Max-Planck Institute for Evolutionary Anthropology Leipzig who sponsored the OWLG-organized Workshop on Linked Data in Linguistics (LDL-2012), held in March 2012 in Frankfurt/M., Germany.[13]

## 3. Recent Activities and On-going Developments

Among the broad range of problems associated with linguistic resources, we identified four major classes of problems and challenges that the OWLG can address:

**identifying open linguistic resources** So far, there is no established common point of reference for existing open linguistic resources. Furthermore, there are multiple metadata collections. The OWLG is working to extend CKAN for open resources from linguistics in order to bridge this gap. Although there are other metadata repositories (e.g., those maintained by META-NET,[14] FLARENET,[15] or CLARIN[16]) available, the CKAN repository is qualitatively different in two respects: (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 OKF working groups, as well as interested individuals outside these working groups. Example resources of potential relevance to linguists include collections of open textbooks,[17] the complete works of Shakespeare,[18] or the Open Richly Annotated Cuneiform Corpus (ORACC).[19]

**technical problems** Often, researchers have questions regarding the choice of tools, representation formats and metadata standards for different types of linguistic annotation. These questions are being addressed by the OWLG: proposals for the interoperable representation of linguistic resources and NLP analyses by means of W3C standards such as RDF are actively being explored.

**legal questions** There is great uncertainty with respect to legal questions regarding the creation and distribution of linguistic data. The OWLG represents a platform to discuss such problems and experiences and to develop recommendations, e.g., the publication of linguistic resources under open licenses.

**spread the word** Finally, there is an agitation challenge for open data in linguistics, i.e., how we should convince our collaborators to release their data under open licenses (and what may be potential obstacles). Towards this end, we have conducted workshops at the Open Knowledge Conference (OKCon-2011, June 2011, Berlin, Germany),[20] and at the 34th Annual Meeting of the German Linguistics Society (DGfS-2012, March 2012, Frankfurt/M., Germany),[21] and published a contributed volume about Linked Data in Linguistics (Chiarcos et al., 2012). The OWLG is present at venues like LREC to actively promote the goals of openness, interoperability and reusability in linguistics.

The Working Group has reached a critical step in its formation process: With a defined set of (preliminary) goals and principles, we can now concentrate on the tasks at hand, and focus on collecting resources and attracting interested people in order to address the challenges identified above.

As of March, 16th, 2012, the Working Group assembles 90 people from 20 different countries.[22] 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: the ground for fruitful interdisciplinary discussions has been laid out.

[12] http://www.w3.org/DesignIssues/LinkedData.html
[13] http://ldl2012.lod2.eu
[14] http://www.meta-net.eu
[15] http://www.flarenet.eu/?q=Documentation_about_Individual_Resources
[16] http://catalog.clarin.eu/ds/vlo
[17] http://wiki.okfn.org/Wg/opentextbooks
[18] http://openshakespeare.org
[19] http://oracc.museum.upenn.edu
[20] http://okcon.org/2011
[21] http://ldl2012.lod2.eu
[22] Austria, Australia, Belgium, Benin, Canada, France, Germany, Greece, Hungary, India, Ireland, Japan, the Netherlands, Portugal, Serbia, Slovenia, Spain, Turkey, UK, US.

The OWLG maintains a home page,[23] a mailing list,[24] a wiki,[25] and a (guest) blog.[26] Recent activities include the collection information about legal issues,[27] discussing the creation of a workflow repository,[28] and initial steps towards the formation of a Linked Open Data (sub-)cloud of linguistic resources, the Linguistic Linked Open Data (LLOD) cloud. In this paper, we focus on the latter aspect, because it involves a large number of OWLG members: We found that independent research activities of many community members include the development of formalisms to represent linguistic corpora in OWL and RDF, the conversion of lexical-semantic resources to RDF, the creation of metadata collections about linguistic data collections and publications, and the development of terminology repositories.

The development of a collection of open, freely accessible linguistic resources that are represented in interoperable standards represents a concrete goal for the working group and may be seen as a long-term vision of the OWLG. At the time of writing, this collection comprises 103 resources, including lexicons, word lists, corpora and collections of linguistic meta data. The license status of these resources varies. Some are free, others are partially free (i.e., annotations free, but text under copyright), and a few have been included that are available under restrictive licenses, but representative of a particular type of resource.

One major goal in the recent past has been the creation of a Linguistic Linked Data (LLOD) cloud from this compilation. We selected 28 of these resources to investigate the possibility of establishing cross-links between them. A draft for the LLOD cloud diagram, inspired by the Linked Open Data diagram by Cyganiak and Jentzsch[29] is shown in Fig. 1. A subset of these resources will be discussed in the next section.

## 4. Towards a LLOD Cloud

If published as Linked Data, linguistic resources represented in RDF/OWL can be linked with other resources already available in the Linked Open Data cloud (LOD) if they share certain URIs. The OWLG aims to nurture the growth of a sub-cloud within the LOD, a Linguistic Linked Open Data (LLOD) cloud, where diverse data sets can be made discoverable through interoperable technological infrastructure. This allows for data federation and querying across distributed resources. Currently, there are several types of data sets spanning semantic knowledge bases, lexical-semantic resources, annotated corpora, linguistic databases, and metadata and terminological ontologies. Some of these resources are described in detail in this section.

### 4.1. DBpedia: A General-Purpose Knowledge Base for the Semantic Web

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web (Lehmann et al., 2009). 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. The extracted RDF knowledge from the English Wikipedia is published and interlinked according to the Linked Data principles and made available under the same license as Wikipedia (CC-BY-SA). In its current version (3.7), 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. DBPedia is a general purpose knowledge base for the Semantic Web. It provides much information that can be linked to linguistics data in the LLOD.

### 4.2. DBpedia+Wiktionary: Linking DBpedia to a Lexical-Semantic Resource

A recent effort by the Agile Knowledge Engineering and Semantic Web (AKSW)[30] research group in Leipzig is dedicated to the development of an DBpedia-based open-source framework to extract semantic lexical resources (an ontology about language use) from Wiktionary.[31] Wiktionary is a wiki-based open content dictionary, a collaborative project for creating a free lexical database in multiple languages.[32] The extracted data currently includes language, part of speech, senses, definitions, synonyms, taxonomies and translations for each lexical entry. At the moment, we focus on improving flexibility (to adapt to the loose schema) and configurability (to adapt to differences between different languages). The configuration uses an XML encoding language-mappings and templates containing placeholders and thus enables the addition of languages without altering the source code. The extracted data can (due to its semantic richness) be automatically transformed

[23] http://linguistics.okfn.org

[24] http://lists.okfn.org/mailman/listinfo/open-linguistics

[25] http://wiki.okfn.org/Wg/linguistics

[26] http://blog.okfn.org/category/working-groups/wg-linguistics

[27] http://wiki.okfn.org/Working_Groups/linguistics/legal_issues

[28] http://wiki.okfn.org/Wg/linguistics/workflows

[29] http://lod-cloud.net

[30] http://aksw.org/

[31] http://wiktionary.dbpedia.org/

[32] http://www.wiktionary.org

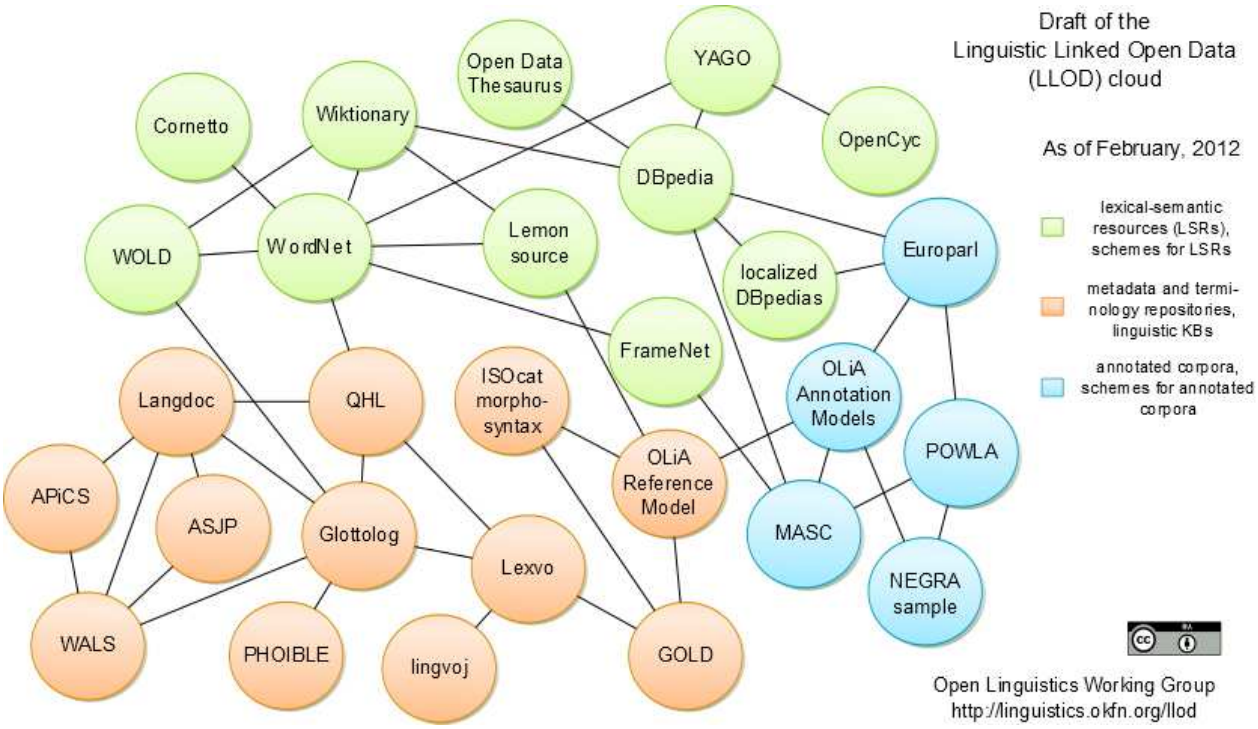


Figure 1: Draft version of the Linguistic Linked Open Data cloud.

into the Lemon model (McCrae et al., 2011) or simpler domain specific formats. The dataset contains information about 3.4 million word forms. A live data set that is constantly synchronized with the wiki is planned to be released soon.

### 4.3. Uby: A Network of Lexical-Semantic Resources

Uby (Gurevych et al., to appear) is a large integrated lexical resource developed at the Ubiquitous Knowledge Processing Lab, TU Darmstadt. It currently contains interoperable versions of 8 open resources in two languages: English WordNet, Wiktionary, Wikipedia, FrameNet and VerbNet, German Wikipedia, Wiktionary, and multilingual OmegaWiki. A subset of these resources is linked at the word sense level and these sense alignments are open as well. There are monolingual sense alignments between VerbNet–FrameNet[33] and VerbNet–WordNet[34] as well as between WordNet–Wikipedia (Niemann and Gurevych, 2011) and WordNet–Wiktionary (Meyer and Gurevych, 2011). In addition, Uby provides cross-lingual sense alignments between WordNet and the German OmegaWiki (Gurevych et al., to appear), also including the interlanguage links already given in Wikipedia and OmegaWiki. Uby is released along with a Java-API[35] and conversion tools licensed under the open Apache license. Uby is based on Uby-LMF, an instantiation of the ISO-LMF meta-model (Francopoulo et al., 2009). The Lexical Markup Framework (LMF) establishes structural and semantic interoperability between linguistic resources. Uby-LMF is currently serialized in XML, and this does not require the use of globally unique identifiers (URIs). It is therefore not part of the cloud diagram. However, XML is just one way of expressing an LMF model. An extension of LMF to include URIs (Francopoulo et al., 2007), and full-fledged RDF linearizations of LMF have been suggested, e.g., in the context of the Lexicon Model for Ontologies (Lemon) as described by McCrae et al. (2011).

### 4.4. MASC in POWLA: An Open Corpus as Linked Data

The Manually Annotated Sub-Corpus (MASC) is a corpus of 500K tokens of contemporary American English text drawn from the Open American National Corpus (Ide et al., 2008).[36] The MASC project is committed to a fully open model of distribution, without restriction, for all data and annotations produced or contributed.

MASC is designed as a balanced selection of written and spoken text from several genres. As an open corpus, it has become increasingly popular in different projects. Therefore, it comprises various layers of annotations, including parts-of-speech, nominal and verbal chunks, constituent syntax, annotations of WordNet senses, frame-semantic annotations, document structure, illocutionary structure and other layers of annotation. As a multi-layer corpus, MASC is distributed in the Graph Annotation Format (GrAF) (Ide and Suderman, 2007), an XML standoff format with all annotations of a document grouped together in a set of XML files pointing to the same piece of primary data. XML standoff formats can be difficult to process, and specifically

[33] http://verbs.colorado.edu/semlink/

[34] http://verbs.colorado.edu/~mpalmer/projects/verbnet

[35] http://www.ukp.tu-darmstadt.de/data/uby/

[36] http://www.anc.org/MASC

for querying, it has been suggested to convert MASC to RDF.

POWLA constitutes such a formalism to represent linguistic corpora by means of semantic web formalisms, in particular, OWL/DL (Chiarcos, 2012; Chiarcos, this vola). 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 PAULA, an existing representation formalism applied for the loss-less representation of arbitrary kinds of text-oriented linguistic annotation within a generic exchange format (Chiarcos et al., 2008). With POWLA, 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. A converter from GrAF to POWLA, applied to data from the MASC, can be found under `http://purl.org/powla`.

### 4.5. MASC+WordNet: Linking Corpora to Lexical-Semantic Resources

Lexical-semantic resources represent first-class citizens of the Semantic Web world, and aside from newly created resources mentioned above, a number of resources are already available in the Linked Open Data cloud, including several instantiations of WordNet.[37] With corpora being represented in RDF, existing annotations for WordNet senses can thus be transformed into links between these resources.

The MASC corpus includes WordNet sense annotations (Baker and Fellbaum, 2009). Within GrAF, such WordNet annotations can only be represented as string values and processed as such, corpus-oriented representation formalisms would not allow to formulate queries that *access* the WordNet specifications for both senses (nor any other information from outside the corpus). Given a mapping between WordNet sense keys and the URIs of an RDF instantiation of WordNet, these links can be generated automatically.

### 4.6. OLiA: Data Categories for Linguistic Annotation

The Ontologies of Linguistic Annotation (OLiA) represent a repository of annotation terminology for various linguistic phenomena currently applied to about 70 languages (Chiarcos, 2008; Chiarcos, this volb). The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al., 2006), and their primary fields of application include the formalization of annotation schemes and concept-based querying over heterogeneously annotated corpora (Rehm et al., 2008), although recently, a broader application, especially in Natural Language Processing, has been suggested (Chiarcos, this volb).

In a Linked Data context, the OLiA ontologies act as a central reference hub for linguistic annotations, in that they provide formal definitions of annotation schemes as OWL/DL ontologies. Further, OLiA establishes interoperability between different annotation schemes by linking them to an overarching 'Reference Model'. Through the OLiA Reference Model, interoperability with community-maintained data category registries can be achieved, because it is linked to the General Ontology of Linguistic Description (Farrar and Langendoen, 2003, GOLD) and to an OWL/DL representation of the morphosyntactic profile of the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al., 2008, ISOcat).

### 4.7. Glottolog/Langdoc: A Global Database of Language Identifiers and Language Resources

Glottolog/Langdoc is as knowledge base for bibliographical references for and genealogical relationships between languages. It provides access to 180k references to descriptive literature treating (mostly) lesser-known languages which are interlinked with a very detailed language classification (Nordhoff and Hammarström, 2011).[38] Due to restrictions inherited from the original bibliographies the data are free for non-commerical use only (CC-BY-NC). The references are collated from 20 different bibliographies. For standard bibliographical data such as author and title, Glottolog/Langdoc uses Dublin Core Metadata Initiative (Weibel et al., 1998) metadata and the Bibliographic Ontology (BIBO).[39] Additional information includes document type, language, and geographical region. The bibliographical part Langdoc is complemented by the genealogical database Glottolog which lists names, codes, location, and family relations for 21288 "languoids" (languages, dialects, families), as well as a justification for why a particular languoid was included. Links to the related projects like Open Language Archives Community (OLAC)[40], Ethnologue (Lewis, 2009), Multitree[41] etc., are provided wherever possible. For Glottolog, a special purpose ontology was developed that can represent language classifications in a very granular fashion. The representation of both bibliographical and genealogical information allows to formulate combined queries such as, "Give me a list of all dictionaries of Afro-Asiatic languages from Africa written after 1975". All languoids and all references are available as XHTML and RDF and have their own URIs, allowing easy integration with other LLOD resources.

### 4.8. PHOIBLE: A Typological Data Base of Phoneme Inventories

PHOIBLE (PHOnetics Information Base and LExicon) is a repository of cross-linguistic phonological segment inventory data that encompasses several legacy segment inventory databases and contains additional linguistic (e.g., distinctive features, genealogical information) and non-linguistic information (e.g., population figures, geographic data) about a large number of languages.[42] As part of

[37] `http://wordnet.rkbexplorer.com`, `http://www.w3.org/TR/wordnet-rdf` (WordNet 2.0), `http://semanticweb.cs.vu.nl/lod/wn30` (WordNet 3.0).

[38] `http://glottolog.livingsources.org`

[39] `http://bibliontology.com/`

[40] `http://language-archives.org`

[41] `http://multitree.org/`

[42] `http://phoible.org`

the LLOD, PHOIBLE will be published online in RDF. PHOIBLE uses a Linked Data graph to model segment inventories and their distinctive features and it can be used to investigate descriptive universals of phonological inventories (Moran, 2012), such as those stated in (Hyman, 2008). Additional linguistic information about languages (e.g. genealogical classification) and non-linguistic information (e.g. geographic information, population figures) is linked via other resources by ISO 639-3 codes, so that data can be queried and extracted for various statistical analyses, e.g. (Moran et al., 2012).

### 4.9. Further Typological Data Sets

The World Loanword Database (WOLD)[43] (Haspelmath and Tadmor, 2009) and the World Atlas of Language Structures (WALS) (Haspelmath et al., 2008) are typological data collections being integrated into the LLOD. WOLD is a lexical-semantic resource that provides mini-dictionaries for 41 languages, WALS is a linguistic database of typological features. WALS explicitly excludes Pidgin and Creole Languages. The Atlas of Pidgin and Creole Language Structures (APiCS) will remedy this shortcoming (Michaelis et al., in preparation), and these data will also be available as RDF. The Automated Similarity Judgment Program ASJP (Brown et al., 2008) provides information about the basic phonological shape of 40 words for over 5000 languages. ASJP data will be made available later this year. The project quantitative modeling of historical-comparative linguistics[44] (abbreviated QHL here) provides the content of dictionaries of South American languages as Linked Data. Due to copyright issues, the publication of the entire QHL dataset is difficult, but partial publication should be possible.

### 4.10. Other Resources

Aside from the resources mentioned here, the diagram includes further ontologies, lexical-semantic resources, corpora, linguistic databases and terminology repositories. For reasons of space, not all of them can be discussed here with detail, but the official SVG version of the LLOD diagram under `http://linguistics.okfn.org/llod` includes hyperlinks pointing to these resources.

### 4.11. Contributing to the LLOD Cloud

Based on the the principles postulated by Cyganiak and Jentzsch for the Linked Open Data cloud,[45] we apply the following criteria for a new linguistic resource to be included in the LLOD cloud diagram: (1) The data is resolvable through HTTP, (2) it is provided as RDF, (3) it contains links to another dataset in the diagram, and (4) the entire dataset must be available. In order to add a new dataset, a contributor would have to create a web or wiki page and announce the resource on the OWLG mailing list. The diagram itself is maintained in a repository and can be edited collaboratively.

At the moment, the LLOD cloud diagram has *draft* status. This means that resources and their linkings do not yet have to be provided (even though many of them are available already), but that their publication under LLOD conditions is promised by the data providers. The shift from draft to official status will require that all resources shown in the diagram are published under LLOD conditions and is expected to take place within the next two years.

[43] `http://wold.livingsources.org/`

[44] `http://web.me.com/cysouw/projects/quanthistling.html`

[45] `http://richard.cyganiak.de/2007/10/lod/#how-to-join`

## 5. Conclusion

The OWLG has now established an interdisciplinary community of researchers wishing to explore ways to share their data in interoperable ways. This community addresses scientific issues, as well as legal or technological questions and best practices. The setup phase of the OWLG is completed, and it is now focusing on addressing concrete problems. This paper has highlighted some of these joint activities which aim to create and to interlink open linguistic resources.

**Acknowledgements**

We would like to thank the OKFN for infrastructural support, and all members of the Open Linguistics Working Group, the participants of the OKCon-2011 Open Linguistics Workshop and the Workshop on Linked Data in Linguistics (LDL-2012) for discussions that have guided the development of the working group. We would also like to thank three anonymous reviewers for comments and feedback.

## 6. References

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop, held in conjunction with ACL-IJCNLP 2009*, pages 125–129, Suntec, Singapore, August.

Tim Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`.

Cecil H. Brown, Eric W. Holman, Sören Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF - Language Typology and Universals*, 61(4):286–308.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. *TAL (Traitement automatique des langues)*, 49(2).

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer, Heidelberg.

Christian Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Christian Chiarcos. 2012. Interoperability of corpora and annotations. In Christian Chiarcos, Sebastian Nordhoff,

and Sebastian Hellmann, editors, *Linked Data in Linguistics*. Springer. p. 161-179.

Christian Chiarcos. this vol.a. A generic formalism to represent linguistic corpora in RDF and OWL/DL.

Christian Chiarcos. this vol.b. Ontologies of Linguistic Annotation: Survey and perspectives.

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

Gil Francopoulo, Núria Bel, Monte Georg, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2007. Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. In *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*, Tübingen, Germany.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. to appear. Uby - A large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Apr.

Martin Haspelmath and Uri Tadmor, editors. 2009. *World Loanword Database*. Max Planck Digital Library.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. 2008. The world atlas of language structures online. Munich: Max Planck Digital Library. Available online at http://wals.info/.

Tom Heath and Christian Bizer. 2011. *Linked Data - Evolving the Web into a Global Data Space*. Morgan & Claypool, San Rafael.

Larry M. Hyman. 2008. Universals in phonology. *The Linguistic Review*, 25:83–137.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proc. Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, Charles J. Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proc. 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, May.

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. ISOcat: Corralling data categories in the wild. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Jens Lehmann, Christian Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia – A crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL, Dallas, 16 edition.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *The Semantic Web: Research and Applications*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer Berlin / Heidelberg.

Christian M. Meyer and Iryna Gurevych. 2011. What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892, Chiang Mai, Thailand.

Adam Meyers, Nancy Ide, Ludovic Denoyer, and Yusuke Shinyama. 2007. The shared corpora working group report. In *Proceedings of the First Linguistic Annotation Workshop (LAW-I), held in conjunction with ACL-2007*, pages 184–190, Prague, Czech Republic.

Susanne Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. in preparation. *Atlas of Pidgin and Creole Language Structures*. Oxford University Press, Oxford.

Steven Moran, Daniel McCloy, and Richard Wright. 2012. Revisiting the population vs phoneme-inventory correlation. In *LSA Annual Meeting Extended Abstracts 2012. eLanguage.*

Steven Moran. 2012. Using Linked Data to create a typological knowledge base. In Chiarcos et al. (Chiarcos et al., 2012), pages 129–138.

Elisabeth Niemann and Iryna Gurevych. 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 205–214, Oxford, UK.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science 2011*, volume 783 of *CEUR Workshop Proceedings*.

Ted Pederson. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Georg Rehm, Richard Eckart, Christian Chiarcos, and Johannes Dellert. 2008. Ontology-based XQuery'ing of XML-encoded language resources on multiple annotation layers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Thomas Schmidt, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. 2006. Avoiding data graveyards: From heterogeneous data collected in multiple research projects to sustainable linguistic resources. In *Proceedings of the E-MELD workshop on Digital Language Documentation*, East Lansing.

Stuart L. Weibel, John A. Kunze, Carl Lagoze, and Misha Wolf. 1998. RFC 2413 - Dublin Core metadata for resource discovery. `http://www.ietf.org/rfc/rfc2413.txt`, September.