Linked Data in Linguistics 2014. Introduction and Overview

Christian Chiarcos, John McCrae, Elena Montiel, Antonio Branco, Nicoletta Calzolari, Petya Osenova, E
Goethe-Universität Frankfurt am Main, Universität Bielefeld, Affiliation3
Germany, Germany, Address3
chiarcos@uni-frankfurt.de, jmcrae@cit-ec.uni-bielefeld.de, author2@zzz.edu, author3@hhh.com, cimiano

Abstract

The Linked Data in Linguistics (LDL) workshop series brings together researchers from various fields of linguistics, natural language processing, and information technology to present and discuss principles, case studies, and best practices for representing, publishing and linking linguistic data collections. A major outcome of our work is the Linguistic Linked Open Data (LLOD) cloud, an LOD (sub-)cloud of linguistic resources, which covers various linguistic data bases, lexicons, corpora, terminology and metadata repositories. As a general introduction into the topic, we describe the concept of Linked Data, its application in linguistics and the development of the Linguistic Linked Open Data (LLOD) cloud since LDL-2013. We present the contributions of LDL-2014, the associated data challenge and its results and present the newly compiled LLOD cloud diagram.
The third instantiation of this series, collocated with the 9th Language Resources and Evaluation Conference (LREC-2014), May 27th, 2014, in Reykjavik, Iceland, is specifically dedicated to the study of ..., although contributions with respect to any application of Linked Data to linguistically and/or NLP-relevant resources are welcome.

Keywords: Linked Data in Linguistics (LDL), Linguistic Linked Open Data (LLOD) cloud

1. Background and Motivation

After half a century of computational linguistics (Dostert, 1955), quantitative typology (Greenberg, 1960), empirical, corpus-based study of language (Francis and Kucera, 1964), and computational lexicography (Morris, 1969), researchers in computational linguistics, natural language processing (NLP) or information technology, as well as in Digital Humanities, are confronted with an immense wealth of linguistic resources, that are not only growing in number, but also in their heterogeneity. Accordingly, the limited interoperability between linguistic resources has been recognized as a major obstacle for data use and re-use within and across discipline boundaries, and represents one of the prime motivations for adopting Linked Data to our field.
Interoperability involves two aspects (Ide and Pustejovsky, 2010):

How to access (read) a resource? (Structural interoperability)
Resources use comparable formalisms to represent and to access data (formats, protocols, query languages, etc.), so that they can be accessed in a uniform way and that their information can be integrated with each other.

How to interpret (understand) information from a resource? (Conceptual interoperability)
Resources share a common vocabulary, so that linguistic information from one resource can be resolved against information from another resource, e.g., grammatical descriptions can be linked to a terminology repository.

With the rise of the Semantic Web, new representation formalisms and novel technologies have become available, and different communities are becoming increasingly aware of the potential of these developments with respect to the challenges posited by the heterogeneity and multitude of linguistic resources available today. Many of these approaches follow the **Linked (Open) Data paradigm** (Berners-Lee, 2006), and this line of research, and its application to resources relevant for linguistics and/or NLP represent the focus of our work.

1.1. Linked Data

The Linked Open Data paradigm postulates four rules for the publication and representation of Web resources: (1) Referred entities should be designated by using URIs, (2) these URIs should be resolvable over HTTP, (3) data should be represented by means of W3C standards (such as RDF), (4) and a resource should include links to other resources. These rules facilitate information integration, and thus, interoperability, in that they require that entities can be addressed in a globally unambiguous way (1), that they can be accessed (2) and interpreted (3), and that entities that are associated on a conceptual level are also physically associated with each other (4).
In the definition of Linked Data, the **Resource Description Framework (RDF)** receives special attention. RDF was designed to provide metadata about resources that are available either offline (e.g., books in a library) or online (e.g., eBooks in a store). RDF provides a generic data model based on labeled directed graphs, which can be serialized in different formats. Information is expressed in terms of *triples* - consisting of a *property* (relation, i.e., a labeled edge) that connects a *subject* (a resource, i.e., a labeled node) with its *object* (another resource, or a literal, e.g., a string). RDF resources (nodes)[1] are represented by *Uniform Resource Identifiers (URIs)*. They are

[1]The term 'resource' is ambiguous: *Linguistic* resources are structured collections of data which can be represented, for example, in RDF. In RDF, however, 'resource' is the conventional name of a node in the graph, because, historically, these nodes were meant to represent objects that are described by metadata. We use the terms 'node' or 'concept' whenever *RDF* resources are meant in ambiguous cases.

thus globally unambiguous in the web of data. This allows resources hosted at different locations to refer to each other, and thereby to create a network of data collections whose elements are densely interwoven.

Several data base implementations for RDF data are available, and these can be accessed using **SPARQL** (Prud'Hommeaux and Seaborne, 2008), a standardized query language for RDF data. SPARQL uses a triple notation similar to RDF, only that properties and RDF resources can be replaced by variables. SPARQL is inspired by SQL, variables can be introduced in a separate `SELECT` block, and constraints on these variables are expressed in a `WHERE` block in a triple notation. SPARQL does not only support running queries against individual RDF data bases that are accessible over HTTP (so-called 'SPARQL end points'), but also, it allows us to combine information from multiple repositories (federation). RDF can thus not only be used to *establish* a network, or cloud, of data collections, but also, to *query* this network directly.

RDF has been applied for various purposes beyond its original field of application. In particular, it evolved into a generic format for knowledge representation. It was readily adopted by disciplines as different as biomedicine and bibliography, and eventually it became one of the building stones of the **Semantic Web**. Due to its application across discipline boundaries, RDF is maintained by a large and active community of users and developers, and it comes with a rich infrastructure of APIs, tools, databases, query languages, and multiple sub-languages that have been developed to define data structures that are more specialized than the graphs represented by RDF. These sub-languages can be used to create *reserved vocabularies* and *structural constraints* for RDF data. For example, the Web Ontology Language (OWL) defines the datatypes necessary for the representation of ontologies as an extension of RDF, i.e., *classes* (concepts), *instances* (individuals) and *properties* (relations).

The concept of Linked Data is closely coupled with the idea of **openness** (otherwise, the linking is only partially reproducible), and in 2010, the original definition of Linked Open Data has been extended with a 5 star rating system for data on the Web.[2] The first star is achieved by publishing data on the Web (in any format) under an open license, and the second, third and fourth star require machine-readable data, a non-proprietary format, and using standards like RDF, respectively. The fifth star is achieved by linking the data to other people's data to provide context. If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects.

1.2. Linked Data for Linguistics and NLP

Publishing Linked Data allows resources to be globally and uniquely identified such that they can be retrieved through standard Web protocols. Moreover, resources can be easily linked to one another in a uniform fashion and thus become structurally interoperable. (Chiarcos et al., to appear) identified the five main benefits of Linked Data for Linguistics and NLP:

Conceptual Interoperability Semantic Web technologies allow to provide, to maintain and to share centralized, but freely accessible terminology repositories. Reference to such terminology repositories facilitates conceptual interoperability as different concepts used in the annotation are backed up by externally provided definitions, and these common definitions may be employed for comparison or information integration across heterogeneous resources.

Linking through URIs URIs provide globally unambiguous identifiers, and if resources are accessible over HTTP, it is possible to create resolvable references to URIs. Different resources developed by independent research groups can be connected into a cloud of resources.

Information Integration at Query Runtime (Federation) Along with HTTP-accessible repositories and resolvable URIs, it is possible to combine information from physically separated repositories in a single query at runtime: Resources can be uniquely identified and easily referenced from any other resource on the Web through URIs. Similar to hyperlinks in the HTML web, the web of data created by these links allows to navigate along these conenctions, and thereby to freely integrate information from different resources in the cloud.

Dynamic Import When linguistic resources are interlinked by references to resolvable URIs instead of system-defined IDs (or static copies of parts from another resource), we always provide access to the most recent version of a resource. For community-maintained terminology repositories like the ISO TC37/SC4 Data Category Registry (Wright, 2004; Windhouwer and Wright, 2012, ISOcat), for example, new categories, definitions or examples can be introduced occasionally, and this information is available immediately to anyone whose resources refer to ISOcat URIs. In order to preserve link consistency among Linguistic Linked Open Data resources, however, it is strongly advised to apply a proper versioning system such that backward-compatibility can be preserved: Adding concepts or examples is unproblematic, but when concepts are deleted, renamed or redefined, a new version should be provided.

Ecosystem RDF as a data exchange framework is maintained by an interdisciplinary, large and active community, and it comes with a developed infrastructure that provides APIs, database implementations, technical support and validators for various RDF-based languages, e.g., reasoners for OWL. For developers of linguistic resources, this ecosystem can provide technological support or off-the-shelf implementations for common problems, e.g., the development

[2] `http://www.w3.org/DesignIssues/LinkedData.html`, paragraph 'Is your Linked Open Data 5 Star?'

of a database that is capable of support flexible, graph-based data structures as necessary for multi-layer corpora (Ide and Suderman, 2007).

To these, we may add that the distributed approach of the Linked Data paradigm facilitates the distributed development of a web of resources and collaboration between researchers that provide and use this data and that employ a shared set of technologies. One consequence is the emergence of interdisciplinary efforts to create large and interconnected sets of resources in linguistics and beyond. The LDL workshop series provides a forum to discuss and to facilitate such on-going developments, in particular, the emerging Linguistic Linked Open Data cloud.

2. Linguistic Linked Open Data: Building the Cloud

Recent years have seen not only a number of approaches to provide linguistic data as Linked Data, but also the emergence of larger initiatives that aim at interconnecting these resources. Among these, the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN) has spearheaded the creation of new data and the republishing of existing linguistic resources as part of the emerging Linguistic Linked Open Data (LLOD) cloud. As the OWLG organizes the LDL workshop series as a vehicle to facilitate this process, we would like to take the chance to unveil the revised cloud diagram on the occasion of LDL-2014.

2.1. The LLOD Cloud

Aside from benefits arising from the actual *linking* of linguistic resources, various linguistic resources from very different fields have been provided in RDF and related standards in the last decade. In particular, this is the case for **lexical resources** (Fig. 1, LEXICON), e.g., WordNet (Gangemi et al., 2003), which represent a cornerstone of the Semantic Web and which are firmly integrated in the Linked Open Data (LOD) cloud. In a broader sense, also general knowledge bases from the LOD such as the DBpedia have been rendered as lexical resources, because of their immanent relevance for Natural Language Processing tasks such as Named Entity Recognition or Anaphora Resolution. Other types of linguistically relevant resources with less importance to AI and Knowledge Representation, however, are not a traditional part of the LOD cloud and motivate the creation of a sub-cloud dedicated to linguistic resources.

As such, the Linked Data paradigm also facilitates the management of information about language (Fig. 1, LANGUAGE_DESCRIPTION), i.e., linguistic terminology and linguistic databases. **Terminology repositories** serve an important role to establish conceptual interoperability between language resources. If resource-specific annotations or abbreviations are expanded into references to repositories of linguistic terminology and/or metadata categories, linguistic annotations, grammatical features and metadata specifications become more easily comparable. Important repositories developed by different communities include GOLD (Farrar and Langendoen, 2003) and ISOcat (Wright, 2004; Windhouwer and Wright, 2012), yet, only recently these terminology repositories were put in relation with each other using Linked Data principles and with linguistic resources, e.g., within the OLiA architecture (Chiarcos, 2012b). **Linguistic databases** are a particularly heterogeneous group of linguistic resources; they contain complex and manifold types of information, e.g., feature structures that represent typologically relevant phenomena, along with examples for their illustration and annotations (glosses) and translations applied to these examples (structurally comparable to corpus data), or word lists (structurally comparable to lexical-semantic resources). RDF as a generic representation formalism is thus particularly appealing for this class of resources.

Finally, for **linguistic corpora** (Fig. 1, CORPORA), the potential of the Linked Data paradigm for modeling, processing and querying of corpora is immense, and RDF conversions of semantically annotated corpora have been proposed early (Burchardt et al., 2008). RDF provides a graph-based data model as required for the interoperable representation of arbitrary kinds of annotation (Bird and Liberman, 2001; Ide and Suderman, 2007), and this flexibility makes it a promising candidate for a general means of representation for corpora with complex and heterogeneous annotations. RDF does not only establish interoperability between annotations within a corpus, but also between corpora and other linguistic resources (Chiarcos, 2012a). In comparison to other types of linguistic resources, corpora are currently underrepresented in the LLOD cloud, but the development of schemes for corpora and/or NLP annotations represents an active line of research (Chiarcos, 2012c; Hellmann et al., 2012) also addressed in the workshop.

2.2. Recent Developments

At the moment, we have 107 data sets, we decided to restrict the cloud to resources whose links are specified in Datahub.io.

hence linked datasets only

The first draft of the LLOD diagram has been sketched ...

Since LDL-2013, the categories have been increasingly clarified

Novel data sets include

2.3. Behind LDL-2014

The LLOD cloud is a result of a coordinated effort of the **Open Linguistics Working Group (OWLG)**,[3] a network open to anyone interested in linguistic resources and/or the publication of these under an open license. The OWLG is a working group of the Open Knowledge Foundation (OKFN),[4] a community-based non-profit organization promoting open knowledge (i.e., data and content that is free to use, re-use and to be distributed without restriction).

Since its formation in 2010, the Open Linguistics Working Group has grown steadily. One of our primary goals is to attain openness in linguistics through:

1. Promoting the idea of open linguistic resources,
2. Developing the means for the representation of open data, and

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

[4] http://okfn.org/

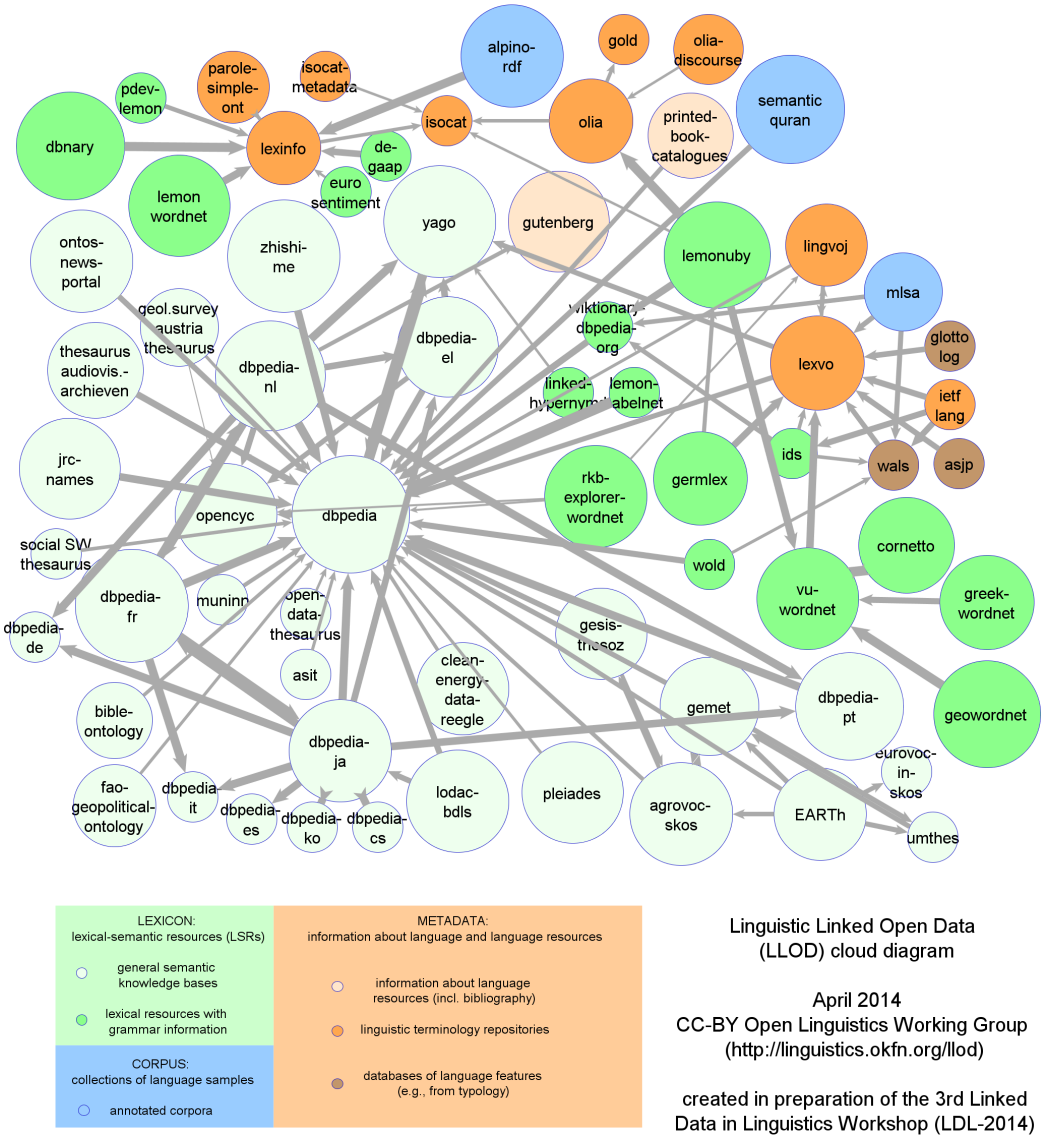

Figure 1: Linguistic Linked Open Data cloud as of September 2013.

3. Encouraging the exchange of ideas across different disciplines.

The OWLG represents an open forum for interested individuals to address these and related issues. At the time of writing, the group consists of about 100 people from 20 different 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; the ground for fruitful interdisciplinary discussions has been laid out. One concrete result emerging out of collaborations between a large number of OWLG members is the LLOD cloud as already sketched above.

The emergence of the LLOD cloud out of a set of isolated resources was accompanied and facilitated by a series of **workshops and publications** organized under the umbrella of the OWLG, including the Open Linguistics track at the Open Knowledge Conference (OKCon-2010, July 2010, Berlin, Germany), the First Workshop on Linked Data in Linguistics (LDL-2012, March 2012, Frankfurt am Main, Germany), the Workshop on Multilingual Linked Open Data for Enterprises (MLODE-2012, September 2012, Leipzig, Germany), the Linked Data for Linguistic Typology track at ALT-2012 (September 2013, Leipzig, Germany). Plans to create a LLOD cloud were first publicly announced at LDL-2012, and subsequently, a first instance of the LLOD materialized as a result of the MLODE-2012 workshop, its accompanying hackathon and the data postproceedings that will appear as a special issue of the Semantic Web Journal (SWJ). The Second Workshop on Linked Data in Linguistics (LDL-2013) continues this series of workshops. In order to further contribute to the integration of the field, it is organized as a joint event of the OWLG and the W3C Ontology-Lexica Community Group. The **Ontology-Lexica Community (OntoLex) Group**[5] was founded in September 2011 as a W3C Community and

[5] http://www.w3.org/community/ontolex

Business Group. It aims to produce specifications for a lexicon-ontology model that can be used to provide rich linguistic grounding for domain ontologies. Rich linguistic grounding include the representation of morphological, syntactic properties of lexical entries as well as the syntax-semantics interface, i.e., the meaning of these lexical entries with respect to the ontology in question. An important issue herein will be to clarify how extant lexical and language resources can be leveraged and reused for this purpose. As a byproduct of this work on specifying a lexicon-ontology model, it is hoped that such a model can become the basis for a web of lexical linked data: a network of lexical and terminological resources that are linked according to the Linked Data Principles forming a large network of lexico-syntactic knowledge.
The OntoLex W3C Community Group has been working for more than a year on realizing a proposal for a standard ontology lexicon model, currently discussed under the the designation *lemon*. As the core specification of the model is almost complete, the group started to develop of additional modules for specific tasks and use cases, and some of these are presented at LDL-2013.

3. LDL-2014: The 3rd Workshop on Linked Data in Linguistics

The goal of the 2nd Workshop on Linked Data in Linguistics (LDL-2013) has been to bring together researchers from various fields of linguistics, NLP, and information technology to present and discuss principles, case studies, and best practices for representing, publishing and linking linguistic data collections, including corpora, dictionaries, lexical networks, translation memories, thesauri, etc., infrastructures developed on that basis, their use of existing standards, and the publication and distribution policies that were adopted.
For the 2nd edition of the workshop on Linked Data in Linguistics, we invited contributions discussing the application of the Linked Open Data paradigm to linguistic data as it might provide an important step towards making linguistic data: i) easily and uniformly queryable, ii) interoperable and iii) sharable over the Web using open standards such as the HTTP protocol and the RDF data model. Recent research in this direction has lead to the emergence of a Linked Open Data cloud of linguistic resources, the Linguistic Linked Open Data (LLOD) cloud, where Linked Data principles have been applied to language resources, allowing them to be published and linked in a principled way. Although not restricted to lexical resources, these play a particularly prominent role in this context. The topics of interest mentioned in the call for papers were the following ones:

1. Use cases for creation, maintenance and publication of linguistic data collections that are linked with other resources
2. Modelling linguistic data and metadata with OWL and/or RDF
3. Ontologies for linguistic data and metadata collections
4. Applications of such data, other ontologies or linked data from any subdiscipline of linguistics
5. Descriptions of data sets, ideally following Linked Data principles
6. Legal and social aspects of Linguistic Linked Open Data

CHECK ACCEPTANCE RATE

LDL-2014 is collocated with the 9th International Conferences for Language Resources and Evaluation (LREC-2014), and at this edition, we put a particular focus on multilingual knowledge resources and ...

Acknowledgements

We thank the organizers of the 9th International Conference on Language Resource and Evaluation (LREC-2014) for hosting and supporting LDL-2014, and in particular for joining us in the promotion of Linked Open Data in our field by establishing it as a main topic of the main conference.
The LDL workshop series is organized by the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation. LDL-2014 was supported by the LIDER project ..., as well as the QTLeap project ...
We thank the OWLG and its members for active contributions to the LLOD cloud, to the workshop and beyond. In particular, we have to thank the contributors, the program committee and the organizers of the data challenge for their invaluable work and engagement.
The introduction into Linked Data given in Sect. 1. is a revised and updated version of earlier introductions into the field, most noteably ... LDL-2013.

4. References

Berners-Lee, T. (2006). Design issues: Linked data. URL `http://www.w3.org/DesignIssues/LinkedData.html` (July 31, 2012).

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60.

Burchardt, A., Padó, S., Spohr, D., Frank, A., and Heid, U. (2008). Formalising multi-layer corpora in OWL/DL – lexicon modelling, querying and consistency control. In *3rd International Joint Conference on NLP (IJCNLP 2008)*, Hyderabad, India.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (to appear). Towards open data for linguistics: Linguistic linked data. In Oltramari, A., Lu-Qin, Vossen, P., and Hovy, E., editors, *New Trends of Research in Ontologies and Lexical Resources*. Springer, Heidelberg.

Chiarcos, C. (2012a). Interoperability of corpora and annotations. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics*, pages 161–179. Springer, Heidelberg.

Chiarcos, C. (2012b). Ontologies of linguistic annotation: Survey and perspectives. In *8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 303–310, Istanbul, Turkey, May.

Chiarcos, C. (2012c). POWLA: Modeling linguistic corpora in OWL/DL. In *9th Extended Semantic Web Conference (ESWC-2012)*, pages 225–239, Heraklion, Crete, May.

Dostert, L. (1955). The Georgetown-IBM experiment. In Locke, W. and Booth, A., editors, *Machine Translation of Languages*, pages 124–135. John Wiley & Sons, New York.

Farrar, S. and Langendoen, T. (2003). Markup and the GOLD ontology. In *EMELD Workshop on Digitizing and Annotating Text and Field Recordings*. Michigan State University, July.

Francis, W. N. and Kucera, H. (1964). Brown Corpus manual. Technical report, Brown University, Providence, Rhode Island. revised edition 1979.

Gangemi, A., Navigli, R., and Velardi, P. (2003). The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In Meersman, R. and Tari, Z., editors, *Proceedings of On the Move to Meaningful Internet Systems (OTM2003)*, pages 820–838, Catania, Italy, November.

Greenberg, J. (1960). A quantitative approach to the morphological typology of languages. *International Journal of American Linguistics*, 26:178–194.

Hellmann, S., Lehmann, J., and Auer, S. (2012). Linked-data aware URI schemes for referencing text fragments. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW-2012)*, Galway, Ireland.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China.

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

Morris, W., editor. (1969). *The American Heritage Dictionary of the English Language*. Houghton Mifflin, New York.

Prud'Hommeaux, E. and Seaborne, A. (2008). SPARQL query language for RDF. *W3C working draft*, 4(January).

Windhouwer, M. and Wright, S. (2012). Linking to linguistic data categories in ISOcat. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics*, pages 99–107. Springer, Heidelberg.

Wright, S. (2004). A global data category registry for interoperable language resources. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, pages 123–126, Lisboa, Portugal, May.

LDL-2013 Organizing Committee

Christian Chiarcos (Goethe-Universität Frankfurt am Main, Germany)
John McCrae (Universität Bielefeld, Germany)
...

LDL-2014 Program Committee

...

Data Challenge Organizers

Philipp Cimiano (Universität Bielefeld, Germany)
Christian Chiarcos (Goethe-Universität Frankfurt am Main, Germany)
John McCrae (Universität Bielefeld, Germany)

Data Challenge Committee