Linked Data in Linguistics 2014. Introduction and Overview

Christian Chiarcos[1], **John McCrae**[2], **Petya Osenova**[3], **Cristina Vertan**[4]
[1] Goethe-Universität Frankfurt am Main, Germany, chiarcos@uni-frankfurt.de
[2] Universität Bielefeld, Germany, jmcrae@cit-ec.uni-bielefeld.de
[3] University of Sofia, Bulgaria, petya@bultreebank.org
[4] Universität Hamburg, Germany,cristina.vertan@uni-hamburg.de

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 Multilingual Knowledge Resources and Natural Language Processing, although contributions with respect to any application of Linked Data to linguistically and/or NLP-relevant resources are welcome, as well.

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):

(a) How to access 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.

(b) How to interpret information from a resource? (Conceptual interoperability) Resources share a common vocabulary, so that 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 repre-

[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

sented by *Uniform Resource Identifiers (URIs).* They are 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 like 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 as triples in the `WHERE` block. SPARQL does not only support querying against individual RDF data bases that are accessible over HTTP ('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.

Beyond its original field of application, RDF 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. (2013a) identified five main benefits of Linked Data for Linguistics and NLP:

(1) 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.

(2) 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.

(3) 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 connections, and thereby to freely integrate information from different resources in the cloud.

(4) 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 or redefined, a new version should be provided.

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

(6) Distributed Development To these, Chiarcos et al. (2013b) add that the distributed approach of the Linked Data paradigm facilitates the distributed development of a web of resources and collaboration between researchers

are meant in ambiguous cases.

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

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

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. The **Open Linguistics Working Group (OWLG)**[3] is an interdisciplinary network open to any individual 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). In this context, 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 an emerging Linked Open Data (sub-)cloud of linguistic resources.

This Linguistic Linked Open Data (LLOD) cloud is a result of a coordinated effort of the OWLG, its members and collaborating initiatives, most noteably the W3C Ontology-Lexica Community Group (OntoLex, see below) specializes in lexical-semantic resources. As the OWLG organizes the LDL workshop series also as a vehicle to facilitate, to promote and to support this process, we would like to take the chance to unveil a revised cloud diagram on the occasion of LDL-2014.

2.1. The LLOD Cloud

In our current, informal understanding, **Linguistic Data** is pragmatically defined as any kind of resource considered relevant for linguistic research or Natural Language Processing tasks. Our assessment of relevance follows the classification of resources provided by data providers or the community, as reflected, for example, in tags assigned to resources at `datahub.io`, the meta data repository from which the LLOD cloud is currently being built. During diagram compilation, resources associated with the OWLG, or with tags like 'LLOD', 'linguistics', etc. are gathered, stored in a JSON document, categorized according to manually defined classification rules, and plotted and reformatted using a GraphML editor.[5]

Among these data sets, we encourage the use of **open** licenses and limit the diagram to such data sets. As defined by the Open Definition, "openness" refers to "[any] piece of content or data [that] is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and share-alike."[6]

Linguistic **Linked** Open Data, then, comprises resources that are provided under an open license and published in conformance with the Linked Data principles as stated above. Typically, these do not represent resources which are RDF-native, but resources that have been transformed into Linked Data.

This also has an impact on the types of linguistic resources considered here, in particular the concept of **corpora**: In empirical linguistics and NLP, *collections of primary data* represent the elementary foundation of research and development. Yet, while it is possible to represent primary data such as plain text in RDF, this is not necessarily the most efficient way of doing so – also given the fact that specialized XML-based standards such as the Text Encoding Iniative[7] are well-established and widely used. However, RDF provides highly flexible data structures that can be employed to represent linguistic annotations of arbitrary complexity. As understood *here*, a 'corpus' is thus always a linguistically analyzed resource: Along with classical representations where both annotations *and* primary data are modeled in RDF (e.g., in the seminal study of (Burchardt et al., 2008)), but also hybrid data sets where only annotations are provided as Linked Data, but the primary data is stored in a conventional format (e.g., (Cassidy, 2010)). At the moment, corpora in the LLOD cloud seem to be relatively rare (see 'CORPUS' resources in Fig. 1), but this only reflects the fact that several corpora had to be excluded from the diagram because they were not linked yet with other LLOD data sets such as lexical resources or repositories of annotation terminology.

Beyond representing linguistic analyses for collections of examples, text fragments, or entire discourses, the Linked Data paradigm particularly facilitates the management of **information about language and language resources** ('METADATA' in Fig. 1). These include linguistic databases (collections of features and inventories of individual languages, e.g., from linguistic typology), repositories of linguistic terminology (e.g., grammatical categories or language identifiers), and metadata about language resources (incl. bibliographical data). While bibliographical data and terminology management represent classical Linked Data applications, our *databases* are a specifically linguistic resource: Databases of features of individual languages 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.

The third major group of resources in the diagram are **lexical-semantic resources** ('LEXICON', 1), i.e., resources focusing on the general meaning of words and the structure of semantic concepts. These represent by far the most established type of linguistic resources in the LD context: They have been of inherent interest to the Semantic Web

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

[4] `http://okfn.org/`

[5] The extraction scripts can be found under `https://github.com/jmccrae/llod-cloud.py`.

[6] `http://opendefinition.org`

[7] `http://www.tei-c.org`

community, and hence a long tradition in this regard, going back to earliest attempts to integrate WordNet into the SW world (Gangemi et al., 2003). In the diagram, we distinguish two types of lexical-semantic resources, i.e., *lexical resources* in a strict sense (which provide specifically linguistic information, e.g., grammatical features, as found, e.g., in a dictionary, or in a WordNet), and and *general knowledge bases* (such as classical thesauri or semantic repositories such as YAGO and DBpedia) whose origins lay outside of the stricter boundaries of linguistics or NLP. While the latter do not provide us with grammatical information, they formalize semantic knowledge, and in this respect, they are of immanent relevance for Natural Language Processing tasks such as Named Entity Recognition or Anaphora Resolution.

2.2. Recent Developments

Since the publication of the last LLOD cloud diagram at LDL-2013, Sep 2013 in Italy, Pisa, we have continued to gather and to convert data sets, to refine our classification of language resources and encouraged others to contribute, e.g., by organizing LDL-2014 and the associated data challenge (see below).

These efforts have met with success such that the number of candidate resources for the cloud has increased substantially, from 65 resources in September 2013 to 107 in April 2014. We thus enforced the constraints imposed on resources in the cloud diagram. As of April 2014, we limit datasets in the cloud diagram to those with links to other LLOD data sets. Applying these stricter filters, we arrive at 68 resources in the new diagram. For generating the diagram, we rely on the metadata as provided by Datahub.io, so only datasets are considered whose links with other LLOD data sets are explicitly documented there. During diagram generation, we test whether the URLs given for the data are responding. At the moment, we do not, however, validate the information provided there, but a stricter validation routine is envisioned.

Among others, novel data sets include resources prepared for LDL-2014 and the data challenge, but also resources that have not been covered by earlier diagram instantiations because they lacked the necessary tags to recognize them as being linguistically relevant. An example for the latter is the Greek WordNet (RDF edition released in early 2013),[8] but also several thesauri and multilingual vocabularies. This partially explains the growth of the cloud particular with respect to lexical resources.

At the same time, the growing number of linked lexical resources also reflects the activities of the W3C Ontology-Lexica Community Group (OntoLex). The OntoLex group is not only closely collaborating with the OWLG, but both also have a considerable overlap in terms of their members, and as for LDL-2013, several LDL-2014 organizers are active in both groups. While the OWLG is interested in open linguistic resources in general, the OntoLex group takes a specific focus on lexical resources, culminating in the proposal of a common model for machine-readable lexicons in RDF, the *lemon* model (McCrae et al., 2012). By now, already 41% of lexical resources (7 out of 17) in the diagram (lemonWordNet, PDEVlemon, Parole/Simple, lemonUby, lemonBabelNet, germlex, DBnary) employ *lemon* or *lemon*-derived vocabularies, so that we see a considerable degree of convergence in this field. The resulting degree of interoperability and visibility arising from the use of shared vocabularies is certainly one of the most concrete achievements of the community activities we aimed to initiate with forming the OWLG, preparing the LLOD diagram and conducting workshops at linguistic, NLP and IT conferences.

2.3. Organizing LDL-2014

The LDL workshop series and LDL-2014 are organized by the Open Linguistics Working Group to bring together researchers from various fields of linguistics, NLP, and IT to present and discuss principles, case studies, and best practices for representing, publishing and linking linguistic data collections, and aims to facilitate the exchange of technologies, ideas and resources across discipline boundaries, that (to a certain extend) find a material manifestation in the emerging LLOD cloud.

LDL-2014, collocated with the 9th International Conference on Language Resources and Evaluation (LREC-2014), May 2014, Reykjavik, Iceland, is the third workshop on Linked Data in Linguistics following LDL-2012 (March 2012 in Frankfurt am Main, Germany), LDL-2013 (Sep 2013 in Pisa, Italy), as well as more specialized events such as the workshops on Multilingual Linked Open Data for Enterprises (MLODE-2012: Sep 2012 in Leipzig, Germany), and Natural Language Processing and Linked Open Data (NLP&LOD-2013: Sep 2013 in Hissar, Bulgaria), and the theme session on Linked Data in Linguistic Typology (at the 10th Biennial Conference of the Association for Linguistic Typology, ALT-2013, Aug 2013 in Leipzig, Germany), as well as presentations, panels and informal meetings at various conferences.

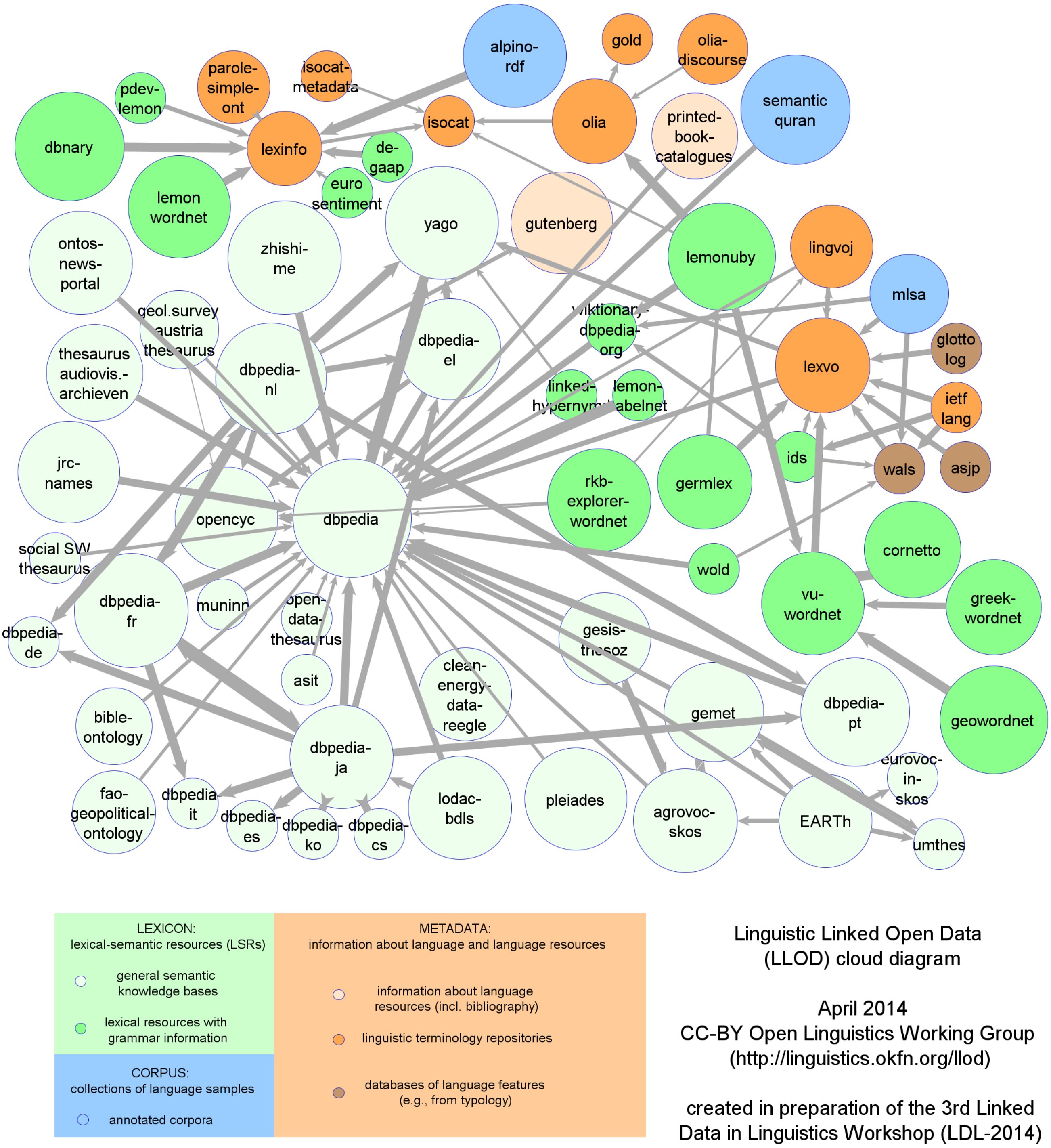
LDL-2014 is organized in the context of two closely related community efforts, the *Open Linguistics Working Group* (OWLG), and the *W3C Ontology-Lexica Community Group* (OntoLex), and supported by two recently started EU projects, *LIDER*, and *QTLeap*.

The **Open Linguistics Working Group** was founded in October 2010, and since its formation, it 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. 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 130 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

[8] `http://datahub.io/de/dataset/greek-wordnet`, cf. `http://okfn.gr/2013/01/983/`.



Linguistic Linked Open Data (LLOD) cloud diagram

April 2014
CC-BY Open Linguistics Working Group
(http://linguistics.okfn.org/llod)

created in preparation of the 3rd Linked Data in Linguistics Workshop (LDL-2014)

Figure 1: Linguistic Linked Open Data cloud as of April 2014.

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 by the OWLG as sketched above. 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 and Third Workshop on Linked Data in Linguistics continued this series of workshops. In order to further contribute to the integration of the field, their organizers involved members of both the OWLG and the W3C Ontology-Lexica Community Group.

The **Ontology-Lexica Community (OntoLex) Group**[9]

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

was founded in September 2011 as a W3C Community and 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 on realizing a proposal for a standard ontology lexicon model, currently discussed under the the designation *lemon*. By now, the core specification of the model is almost complete, the group started to develop additional modules for specific tasks and use cases, and some of these are presented at LDL-2014.

As mentioned above, LDL-2014 is supported by two recently started EU Projects. The project **Linked Data as an Enabler of Cross-Media and Multilingual Content Analytics for Enterprises Across Europe** (LIDER) aims to provide an ecosystem for the establishment of linguistic linked open data, as well as media resources metadata, for a free and open exploitation of such resources in multilingual, cross-media content analytics across Europe. The project **Quality Translation with Deep Language Engineering Approaches** (QTLeap) explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets (including Linked Open Data) and by recent advances in deep language processing.

To accomodate the importance of multilinguality and semantically-oriented NLP that we encounter in the community as well as these initiatives, LDL-2014 takes a focus on Multilingual Knowledge Resources and Natural Language Processing, although contributions on Linked Data emphasising other aspects of linguistics or NLP were explicitly encouraged.

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

For the 3rd edition of the workshop on Linked Data in Linguistics, we invited contributions discussing the application of the Linked Open Data paradigm to linguistic data in various fields of linguistics, natural language processing, knowledge management and information technology in order to to present and discuss *principles*, *case studies*, and *best practices* for representing, publishing and linking mono- and multilingual linguistic and knowledge data collections, including corpora, grammars, dictionaries, wordnets, translation memories, domain specific ontologies etc. In this regard, the Linked Data paradigm 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. The adaptation of some processes and best practices to **multilingual linguistic resources and knowledge bases** acquires special relevance in this context. Some processes may need to be modified to accommodate the publication of resources that contain information in several languages. Also the linking process between linguistic resources in different languages poses important research questions, as well as the development and application of freely available knowledge bases and crowdsourcing to compensate the lack of publicly accessible language resources for various languages.

Further, LDL-2014 provides a forum for researchers on natural language processing and semantic web technologies to present case studies and best practices on the exploitation of linguistic resources exposed on the Web for **Natural Language Processing** applications, or other content-centered applications such as content analytics, knowledge extraction, etc. The availability of massive linked open knowledge resources raises the question how such data can be suitably employed to facilitate different NLP tasks and research questions. Following the tradition of earlier LDL workshops, we encouraged contributions to the Linguistic Linked Open Data (LLOD) cloud and research on this basis. In particular, this pertains to contributions that demonstrate an added value resulting from the combination of linked datasets and ontologies as a source for semantic information with linguistic resources published according to as linked data principles. Another important question to be addressed in the workshop is how Natural Language Processing techniques can be employed to further facilitate the growth and enrichment of linguistic resources on the Web.

The call for papers emphasized the following topics:

1. **Use cases** for creating or publishing linked linguistic data collections

2. **Modelling** linguistic data and metadata with OWL and/or RDF

3. **Ontologies** for linguistic data and metadata collections as well as for cross-lingual retrieval

4. Description of **data sets** following Linked Data principles

5. **Applications of such data**, other ontologies or linked data from any subdiscipline of linguistics

6. **NLP&LLOD**: Application and applicability of (Linguistic) Linked Open Data in NLP / NLP contributions to (Linguistic) Linked Open Data

7. Challenges of **multilinguality** and **collaboratively constructed open resources** for knowledge extraction, machine translation and other NLP tasks.

8. **Legal and social aspects** of (L)LOD

9. **Best practices** for the publication and linking of multilingual knowledge resources

Along with regular workshop submissions, we invited contributions to the associated data challenge (see below) for data sets together with data set descriptions. In total, we received 19 submissions in response to our calls, including 5 data set descriptions for the associated data challenge. Regular submissions were reviewed by at least 3 members of the program committee. On this basis, we accepted 6 submissions as full papers and 4 as short papers.

The 10 accepted papers address a wide range of problems in the area of NLP and (Linguistic) Linked Open Data, pertaining to modeling, representation, analysis and publishing of various data or metadata.

Taken together, the contributions cover a vast and heterogeneous field, they involve different types of linguistic resources, such as machine-readable lexicons, etymological and diachronic databases, web, movies, and grammar terminology, but also address issues of localization and multilinguality. Our tentative classification, that we apply both to the proceedings and the remainder of this section, is a compromise between a classification on grounds of resource types and prospective applications:

A particularly popular branch of research is concerned with **modeling lexical-semantic resources** using RDF-based vocabularies and lexicon-to-ontology mappings, most noteably *lemon*. This group of submissions partially overlaps with a surprisingly large number of papers concerned with the modeling of multilingual resources in more academic fields of linguistics, namely **cross-linguistic studies** in linguistic typology and comparative linguistics. A third group of papers involves different conceptions of **metadata**, i.e., terminology for linguistic categories and language resources, but also annotations to multimedial content. Finally, we sketch the contributions to the data set challenge, all of which were concerned with lexical-semantic resources.

3.1. Modelling Lexical-Semantic Resources with *lemon*

In their paper **Attaching translations to proper lexical senses in DBnary**, Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian and Didier Schwab present the current status of the DBnary project: DBnary aims at extracting linked open data from Wiktionaries in various languages, for which the authors present a similarity technique for disambiguation of linked translations.

John Philip McCrae, Christiane Fellbaum and Philipp Cimiano describe their approach on **Publishing and linking WordNet using *lemon* and RDF** where they propose a strategy for publishing the Princeton WordNet as linked data through an open model. The advantage of this approach is that it provides linking also to the resources which have been already integrated into WordNet.

The paper **Releasing genre keywords of Russian movie descriptions as Linked Open Data: An experience report** by Andrey Kutuzov and Maxim Ionov describes efforts on publishing genre-classified movie keywords as LOD using the *lemon* model. The resource is also linked to Russian component of the Wiktionary RDF dump created by the DBpedia team.[10]

3.2. Cross-linguistic Studies: Applications in Comparative Linguistics and Typology

Although most of the following papers also involve lexical resources, they are special in their domain of application, i.e., the study of cross-linguistic and/or diachronic relationships in linguistics.

In **Linking etymological databases. A case study in Germanic**, Christian Chiarcos and Maria Sukhareva describe the modeling of etymological dictionaries of various Germanic languages in a machine-readable way as Linguistic Linked Open Data. The authors adopted *lemon*, and identified several problematic aspects in its application to this kind of data. The work is challenging, since it handles different language stages, but the current model represents a solid basis to discuss possible adjustments of both *lemon* and the authors' approach in order to develop a *lemon*-conformant representation that meets the requirements of diachronic data.

More focusing on semantic shift than etymological (phonological) continuity, but operating in a similar setting, Fahad Khan, Federico Boschetti and Francesca Frontini describe an approach on **Using *lemon* to model lexical semantic shift in diachronic lexical resources**. They propose *lemonDIA*, an ontology-based extension of the *lemon* model for representing lexical semantic change in temporal context that formalizes notions of perdurance and temporal anchoring of lexical senses.

Coming from the slightly different angle of cross-linguistic language comparison in linguistic typology, the paper **Typology with graphs and matrices** by Steven Moran and Michael Cysouw describes how to extract information from LLOD representations of different typological data sets, and how to transform and operate with the extracted information in order to determine associations between syntactic and phonological features.

Robert Forkel introduces **The Cross-Linguistic Linked Data project**, an ongoing initiative and its infrastructure aiming towards establishing a platform for interoperability among various language resources assembled in typological research. The important role of Linguistic Linked Open Data has long been recognized as publishing strategy for typological datasets (Chiarcos et al., 2012), but here, a unified publication platform is described which may have a considerable effect on the typological publicating practice.

3.3. Metadata

As used here, metadata refers to information provided *about* another resource, including language resources, linguistic terminology and multimedia contents.

From CLARIN Component Metadata to Linked Open Data by Matej Durco and Menzo Windhouwer describes the conversion from CMDI resource descriptions to LOD. As a result, the RDF metadata can be accessed with standard query languages using SPARQL endpoints.

In **Towards a Linked Open Data Rrepresentation of a grammar terms index**, Daniel Jettka, Karim Kuropka, Cristina Vertan and Heike Zinsmeister introduce onoing work on creating a Linked Open Data representation of German grammatical terminology, an effort which nicely complements established efforts to create repositories for

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

linguistic terminology used in language documentation, NLP and the development of machine-readable lexicons. Given the great amount of language-specific terminology, the proposed strategy is also applicable to other languages and their linking may eventually improve the multilingual coverage of linguistic terminology repositories.

A different kind of metadata is subject to **A brief survey of multimedia annotation localization on the web of Linked Data** by Gary Lefman, David Lewis and Felix Sasaki. The authors focus on the localization of multimedia ontologies and Linked Data frameworks for Flickr data. In this respect, Linguistic Linked Open Data may serve as a mediator between multimedia annotation in social media and the Web of Linked Data.

3.4. Data Challenge

The workshop was associated with an open challenge for the creation of datasets for linguistics according to linked data principles. Unlike the preceding Monnet challenge[11] that was organized by the W3C OntoLex community at MLODE-2012, the LDL-2014 was not restricted to the application of the *lemon* format. Nevertheless, all submissions were, indeed, lexical-semantic resources.

This challenge required submissions of new or substantially updated linked datasets and was evaluated by reviewers on technical grounds. The following criteria were applied:

1. *Availability*, i.e. (a) whether the resource uses Linked Data and RDF, (b) whether it is hosted on a publicly accessible server and is available both during the period of the evaluation and beyond, and (c) whether it uses an open license.

2. *Quality*, i.e. (a) whether the resource represents useful linguistically or NLP-relevant information, (b) whether it reuses relevant standards and models, and (c) wheter it contains complex, non-trivial information (e.g., multiple levels of annotation, manually validated analyses).

3. *Linking*, i.e., (a) wheter the resource contains links to external resources, and (b) whether it reuses existing properties and categories.

4. *Impact/usefulness* of the resource, i.e., (a) whether it is relevant and likely to be reused by many researchers in NLP and beyond, and (b) whether it uses linked data to improve the quality of and access to the resource.

5. *Originality*, i.e., (a) whether the data set represents a type of resource or a community currently underrepresented in (L)LOD cloud activities, or (b) whether the approach facilitates novel and unforeseen applications or use cases (as described by the authors) enabled through Linked Data technology.

This year there were five accepted submissions to the challenge. Every challenge committee member provided a ranking of these resources, and the average rank was taken as decisive criterion. In this process, we chose two joint winners and one highly commended paper.

The winners were **DBnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations** by Gilles Sérraset and Andon Tchechmedjiev and **Linked-data based domain-specific sentiment lexicons** by Gabriela Vulcu, Raul Lario Monje, Mario Munoz, Paul Buitelaar and Carlos A. Iglesias, describing the EuroSentiment lexicon. An outstanding characteristic of the DBnary data is its high degree of maturity (quality, usefulness, linking, availability). The EuroSentiment dataset is specifically praised for its originality and quality, as it represents the *only* manually corrected sentiment lexicon currently available as Linguistic Linked Open Data.

Sérraset and Tchechmedjiev describe the extraction of multilingual data from Wiktionary based on 12 language editions of Wiktionary, and as such represents a large and important lexical resource that should have application in many linguistic areas. **Vulcu et al.** describe the creation of a lexicon for the EuroSentiment project, which tackles the important field of sentiment analysis through the use of sophisticated linguistic processing. The resource described extends the *lemon* model with the MARL vocabulary to provide a lexicon that is unique in the field of sentiment analysis due to its linguistic sophistication.

Beyond this, we highly commend the work presented in **A multilingual semantic network as linked data: Lemon-BabelNet** by Maud Ehrmann, Francesco Cecconi, Daniele Vannelle, John P. McCrae, Philipp Cimiano and Roberto Navigli, which describes the expression of BabelNet using the *lemon* vocabulary. BabelNet is one of the largest lexical resources created to date and its linked data version at over 1 billion triples will be one of the largest resources in the LLOD cloud. As such, the clear usefulness of the resource as a target for linking and also the use of the widely-used *lemon* model make this conversion a highly valuable resource for the community as noted by the reviewers.

Finally, we will note that our two runner-up participants **PDEV-LEMON: A linked data implementation of the pattern dictionary of English verbs based on the *lemon* model** by Ismail El Maarouf, Jane Bradbury and Patrick Hanks, and **Linked Hypernyms Dataset - Generation Framework and Use Cases** by Tomáš Kliegr, Vaclav Zeman and Milan Dojchinovski were also well received as resources that continue to grow the linguistic linked open data cloud and are likely to find applications for a number of works in linguistics and natural language processing.

3.5. Invited Talks

In addition to regular papers and dataset descriptions, LDL-2014 features two invited speakers, Piek Vossen, VU Amsterdam, and Gerard de Melo, Tsinghua University.

Piek Th.J.M. Vossen is a Professor of computational lexicology at the Vrije Universiteit Amsterdam, The Netherlands. He graduated from the University of Amsterdam in Dutch and general linguistics, where he obtained a PhD in computational lexicology in 1995, and is probably most well-known for being founder and president of the Global WordNet Association.

In his talk, he will describe and elaborate on the application of **The Collaborative Inter-Lingual-Index for harmonizing WordNets**. The Inter-Lingual-Index, originally

[11] http://sabre2012.infai.org/mlode/monnet-challenge

developed in the context of EuroWordNet, provides a set of common reference points through which WordNets can be linked with each other across different languages and thereby establishes a semantic layer for the interpretation of text in a multilingual setting. Although devised before the advent of modern Linked Data technology, the applications developed on this basis are inspiring for applications of Linguistic Linked Open Data and we are therefore very happy to welcome Piek for discussions and exchange of ideas.

Gerard de Melo is an Assistant Professor at Tsinghua University, where he is heading the Web Mining and Language Technology group. Previously, he was a post-doctoral researcher at the the ICSI AI group of the UC Berkeley, and a doctoral candidate at the Max Planck Institute for Informatics.

In his talk, Gerard de Melo will describe the transition **From Linked Data to Tightly Integrated Data**. He argues that the true potential of Linked Data can only be appreciated when extensive cross-linkage and integration leads to an even higher degree of interconnectedness. Gerard compares different approaches on integration into unified, coherent knowledge bases and develops ideas on how to address some remaining challenges that are currently impeding a more widespread adoption of Linked Data.

Acknowledgements

We thank the organizers of the 9th International Conference on Language Resource and Evaluation (LREC-2014) for hosting and supporting LDL-2014, in particular for joining us in promoting Linked Open Data in our field by establishing it as a main topic of the main conference.

The LDL workshop series and LDL-2014 are organized by the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation. LDL-2014 is supported by two EU projects, LIDER (Linked Data as an Enabler of Cross-Media and Multilingual Content Analytics for Enterprises Across Europe), as well as QTLeap (Quality Translation with Deep Language Engineering Approaches).

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 our contributors, our co-organizers, the committee and the organizers of the data challenge for their invaluable work and engagement.

4. References

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

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.

Cassidy, S. (2010). An RDF realisation of LAF in the DADA Annotation Server. In *5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5)*, Hong Kong, January.

Chiarcos, C., Nordhoff, S., and Hellmann, S., editors. (2012). *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013a). 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., Cimiano, P., Declerck, T., and McCrae, J. (2013b). Linguistic linked open data (llod). introduction and overview. In Chiarcos, C., Cimiano, P., Declerck, T., and McCrae, J., editors, *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i–xi, Pisa, Italy, Sep.

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.

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.

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.

McCrae, J., Montiel-Ponsoda, E., and Cimiano, P. (2012). Integrating WordNet and Wiktionary with lemon. In Chiarcos, C., Nordhoff, S., and Hellmann, S., editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 25–34, Heidelberg. Springer.

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-2014 Organizing Committee

Christian Chiarcos (Goethe-Universität Frankfurt am Main, Germany)
John McCrae (Universität Bielefeld, Germany)
Elena Montiel (Universidad Politécnica de Madrid, Spain)
Kiril Simov (Bulgarian Academy of Sciences, Sofia, Bulgaria)
Antonio Branco (University of Lisbon, Portugal)
Nicoletta Calzolari (ILC-CNR, Italy)
Petya Osenova (University of Sofia, Bulgaria)
Milena Slavcheva (JRC-Brussels, Belgium)
Cristina Vertan (University of Hamburg, Germany)

LDL-2014 Program Committee

Eneko Agirre (University of the Basque Country, Spain)
Guadalupe Aguado (Universidad Politécnica de Madrid, Spain)
Peter Bouda (Interdisciplinary Centre for Social and Language Documentation, Portugal)
Steve Cassidy (Macquarie University, Australia)
Damir Cavar (Eastern Michigan University)
Eric Charton (Ecole Polytechnique de Montréal, Canada)
Walter Daelemans (University of Antwerp, Belgium)
Ernesto William De Luca (University of Applied Sciences Potsdam, Germany)
Gerard de Melo (University of California at Berkeley)
Thierry Declerck (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany)
Dongpo Deng (Institute of Information Sciences, Academia Sinica, Taiwan)
Alexis Dimitriadis (Universiteit Utrecht, The Netherlands)
Jeff Good (University at Buffalo)
Asunción Gómez Pérez (Universidad Politécnica de Madrid, Spain)
Jorge Gracia (Universidad Politécnica de Madrid, Spain)
Walther v. Hahn (University of Hamburg, Germany)
Eva Hajicova (Charles University Prague, Czech Republic)
Harald Hammarström (Radboud Universiteit Nijmegen, The Netherlands)
Yoshihiko Hayashi (Osaka University, Japan)
Sebastian Hellmann (Universität Leipzig, Germany)
Dominic Jones (Trinity College Dublin, Ireland)
Lutz Maicher (Universität Leipzig, Germany)
Pablo Mendes (Open Knowledge Foundation Deutschland, Germany)
Steven Moran (Universität Zürich, Switzerland/Ludwig Maximilian University, Germany)
Sebastian Nordhoff (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany)
Antonio Pareja-Lora (Universidad Politécnica de Madrid, Spain)
Maciej Piasecki (Wroclaw University of Technology, Poland)
Adam Przepiorkowski (IPAN, Polish Academy of Sciences)
Laurent Romary (INRIA, France)
Felix Sasaki (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany)
Andrea Schalley (Griffith University, Australia)
Marco Tadic (University of Zagreb, Croatia)
Marieke van Erp (VU University Amsterdam, The Netherlands)
Daniel Vila (Universidad Politécnica de Madrid, Spain)
Menzo Windhouwer (Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands)

Data Challenge Organizers

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

Data Challenge Committee

Christian Chiarcos (Goethe-Universität Frankfurt am Main, Germany)
Philipp Cimiano (Universität Bielefeld, Germany)
Thierry Declerck (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany)
Jorge Gracia (Universidad Politécnica de Madrid, Spain)
Sebastian Nordhoff (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany)
John McCrae (Universität Bielefeld, Germany)
Steven Moran (Universität Zürich, Switzerland/Ludwig Maximilian University, Germany)
Petya Osenova (University of Sofia, Bulgaria)