Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 8–15
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

On the Linguistic Linked Open Data Infrastructure

Christian Chiarcos[1], Bettina Klimek[2], Christian Fäth[1], Thierry Declerck[3], John P. McCrae[4]
[1] Goethe-Universität Frankfurt am Main, Germany
[2] Universität Leipzig, Germany
[3] DFKI GmbH, Multilinguality and Language Technology Lab, Saarbrücken, Germany
[4] Data Science Institute/Insight Centre for Data Analytics, NUI Galway, Ireland
[1]{chiarcos,faeth}@informatik.uni-frankfurt.de, [2]klimek@informatik.uni-leipzig.de
[3]declerck@dfki.de.de, [4]john@mccr.ae

Abstract

In this paper we describe the current state of development of the Linguistic Linked Open Data (LLOD) infrastructure, an LOD (sub-)cloud of linguistic resources, which covers various linguistic data bases, lexicons, corpora, terminology and metadata repositories. We give in some details an overview of the contributions made by the European H2020 projects "Prêt-à-LLOD" ('Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors') and "ELEXIS" ('European Lexicographic Infrastructure') to the further development of the LLOD.

Keywords: language resources, standards, interoperability, Linguistic Linked Open Data (LLOD)

1. Background

1.1. Interoperability and Collaboration

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 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 Semantic Web and Linked Data, 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 Natural Language Processing (NLP) have been a major factor that led to the formation of the Open Linguistics Working Group[1] as a working group of Open Knowledge Foundation (OKFN).[2] The OWLG adopted OKFN's principles, definitions and infrastructure as far as they are relevant for linguistic data. The OKFN defines standards and develops tools that allow anyone to create, discover and share open data. The Open Definition of the OKFN states that "openness" refers to: "A 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."[3] One of its primary goals is thus to attain openness in linguistics. This includes:

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.

One of the earliest activities of the OWLG was to compile a list of potentially relevant language resources, and by the end of 2011, it developed the idea of a Linked Open Data (sub-)cloud of language resources. Subsequently, developing this Linguistic Linked Open Data (LLOD) cloud has become one of the main activities of the group.
The LLOD cloud is a result of a coordinated effort of OWLG participants, but also supported by several broad-scale projects, mostly funded by the EU. This includes early support projects such as *LOD2. Creating Knowledge out of Interlinked Data* (FP7, 2010-2014), an EU-funded project that brought together 15 European partners

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

and one from South Korea, *MONNET. Multilingual Ontologies for Networked Knowledge* (FP7, 2010-2013), and *LIDER. Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe* (FP7, 2013-2015). A recently funded H2020 projet, *Prêt-à-LLOD. 'Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors* is extending the line of development of LLOD, including also new industrial use cases. And the H2020 infrastructure project *ELEXIS. 'European Lexicographic Infrastructure'* is having at its core the LLOD for the building of a dictionary matrix.

Along with these projects, a number of closely related W3C Community Groups emerged. The Ontology-Lexica Community (OntoLex) Group[4] was founded in September 2011, in parts as a continuation of the MONNET project (McCrae et al., 2012). OntoLex develops 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. The resulting OntoLex-Lemon vocabulary was published in 2016 as a W3C Community Report (Cimiano et al., 2016).[5]

In addition to its original application for ontology lexicalization, the OntoLex-Lemon model has also 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. This is reflected in the development of an accompanying OntoLex module for lexicography (OntoLex-Lexicog, (Bosque-Gil et al., 2019))[6] as well as the on-going development of modules for morphology (OntoLex-Morph, (Klimek et al., 2019)),[7] respectively frequency, attestation and corpus information (OntoLex-Frac).

Other notable W3C community groups include Linked Data for Language Technology (LD4LT) and Best Practices for Multilingual Linked Open Data (BPMLOD), both formed in 2013 in the context of the LIDER project. BPMLOD published a series of recommendations about using and creating linked language resources. LD4LT contributed to the development and dissemination of the NLP Interchange Format (NIF), an RDF vocabulary for linguistic annotations on the web, and continues its activities to this day. Another important community group is Open Annotation, a community that emerged in BioNLP with the goal to facilitate the annotation of web resources – albeit not specifically with linguistic annotation. The Open Annotation community report serves as the basis of the Web Annotation standard, published in 2017.

These W3C Community Groups differ from the Open Linguistics Working Group in their goals and their focus on specific aspects of, say, language resources or language technology. In particular, they aim to develop community reports on clearly delineated topics that can serve as a basis for future standardization efforts. At the moment, the OntoLex-Lemon vocabulary remains at the level of a community report, whereas Web Annotation has been published as a W3C recommendation. With the wider thematical scope and band-width that it provides, the OWLG serves as a platform to facilitate the flow of information between these W3C CGs, individual research projects and related efforts and thus serves an umbrella function.

1.2. 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)[8] are represented 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 database implementations for RDF data are available, and these can be accessed using SPARQL (Harris and Seaborne, 2013), 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

[4] `http://www.w3.org/community/ontolex`

[5] See also `https://www.w3.org/2016/05/ontolex/` and (McCrae et al., 2017).

[6] See also `https://www.w3.org/2019/09/lexicog/`.

[7] See also `https://www.w3.org/community/ontolex/wiki/Morphology`.

[8] 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.

the user 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.[9] 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.

1.3. Linked (Open) Data for Language Resources

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., 2013) 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 navigation along these connections, 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), 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 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, it may be added 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.

1.4. 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. Among these, the Open Linguistics

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

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, Fig. 1) cloud.

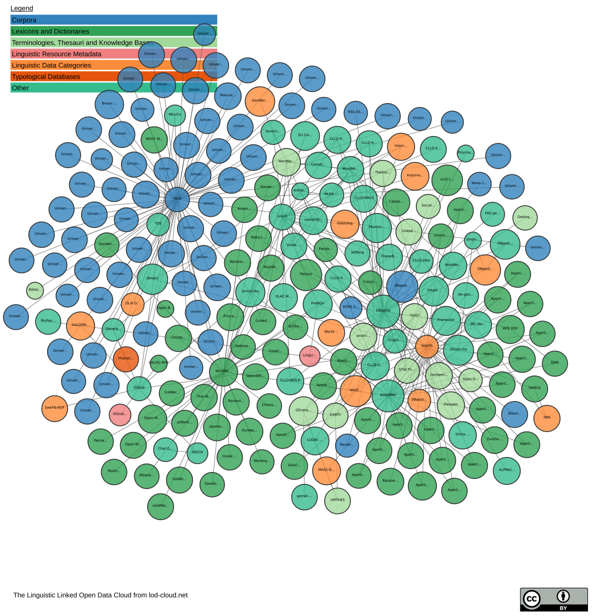
Legend
Corpora
Lexicons and Dictionaries
Terminologies, Thesauri and Knowledge Bases
Linguistic Resource Metadata
Linguistic Data Categories
Typological Databases
Other
The Linguistic Linked Open Data Cloud from lod-cloud.net

Figure 1: Linguistic Linked Open Data cloud as of March 2019.

With the increasing popularity of LLOD, 'linguistics' was recognized as a top-level category of the colored LOD cloud diagram in August 2014, with LLOD resources formerly being classified into other categories. In August 2018, a copy of the LLOD cloud diagram was incorporated into the LOD cloud diagram as a domain-specific addendum. Within the LOD cloud, Linguistic Linked Open Data is growing at a relatively high rate. While the annual growth of the LOD cloud (in terms of new resources added) in the last two years has been at 10.2% in average for the LOD cloud diagram, the LLOD cloud diagram has been growing at 19.3% per year, cf. Fig. 2.

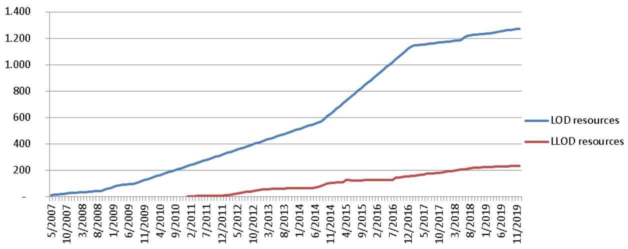

Figure 2: Number of resources in the LOD and LLOD cloud diagrams, 2007-2019, resp. 2011-2019

Aside from maintaining the LLOD cloud diagram, the OWLG aims to promote open linguistic resources by raising awareness and collecting metadata, and aims to facilitate wide-range community activities by hosting workshops, through their mailing list, and through publications. In doing so, they facilitate exchange between and among more specialized community groups, e.g., the W3C community groups such as the Ontology-Lexica Community Group (OntoLex),[10] the Linked Data for Technology Working Group (LD4LT)[11], or the Best Practices for Multilingual Linked Open Data Community Group (BPMLOD).[12] At the time of writing, the most vibrant of these W3C community groups is the OntoLex group, which is developing specifications for lexical data in a LOD context, and this correlates with the high popularity of the OntoLex vocabulary (Cimiano et al., 2016) among LLOD resources. Whereas specifications for lexical resources are relatively mature, as are term bases for language varieties (Nordhoff and Hammarström, 2011; de Melo, 2015) or linguistic terminology (Chiarcos, 2008; Chiarcos and Sukhareva, 2015), the process of developing widely applied data models for other types of language resources, e.g., corpora and data collections in general, is still on-going.

2. Current State and Future Directions

2.1. Usability and practicality of LLOD

It seems that two initial goals of the LLOD community have been achieved. First, the creation of a considerable amount of language resources in the interoperable RDF data format and the involvement of researchers from non-computational but language-focused disciplines like linguistics and philology. Second, these accomplishments revealed new challenges that need to be considered in the future. The growing number of Linked Data language resources opens new questions about interoperability, such as interlinking, ontology usage and the creation of new ontology standards. At the same time the practical needs of researchers unfamiliar with but willing to use the Linked Data framework demand to focus more intensely on the utilization of LLOD by developing appropriate tools to create and exploit the amount of existing language data.

2.2. Selected Developments since 2018

Since 2018, a number of important developments in the Linguistic Linked Open Data community took place. This includes a number of novel, large-scale projects building on LLOD technology and resources, e.g., the H2020 Research and Innovation Actions *ELEXIS. European Lexicographic Infrastructure* (2018-2022)[13], *Prêt-à-LLOD. Ready-to-use multilingual linked language data for knowledge services across sectors* (2019-2021)[14] and the ERC Consolidator Grant *LiLa. Linking Latin* (2018-2023, Marco Carlo Passarotti, Università Cattolica del Sacro Cuore).[15] Equally important is that the Open Linguistics Working Group and related initiatives are being complemented by the new Cost Action *Nexus Linguarum. European network for Web-centred linguistic data science*.[16]

[10] https://www.w3.org/community/ontolex
[11] https://www.w3.org/community/ld4lt/
[12] https://www.w3.org/community/bpmlod
[13] https://elex.is/.
[14] https://www.pret-a-llod.eu/.
[15] https://lila-erc.eu/.
[16] https://www.cost.eu/actions/CA18209/.

2.3. Prêt-à-LLOD

In this section we describe briefly the contributions of the Prêt-à-LLOD project to the further development of the Linguistic Linked Open Data infrastructure. Prêt-à-LLOD aims to achieve this by creating a new methodology for building data value chains applicable to a wide range of sectors and applications. This methodology is based around language resources and language technologies that can be integrated by means of semantic technologies.

This is realised by providing data *discovery* tools based on metadata aggregated from multiple sources, methodologies for describing the licenses of data and services, and tools to deduce the possible licenses of a resource produced after a complex pipeline. Related with this is the development of a *transformation* platform that maps data sets to the formats and schemas that can be consumed by the LLOD. Finally, the project is developing an ecosystem to support the linked data-aware language technologies, from basic tools such as taggers to full applications such as machine translation systems or chatbots, based on semantic technologies that have been developed for LLOD to provide interoperable pipelines.

One of the key approaches of the project is the application of state-of-the-art semantic *linking* technologies in order to provide semi-automatic integration of language services in the cloud. This is the method to implement approaches for ensuring interoperability and for porting LLOD data sets and services to other infrastructures, as well as the contribution of the projects to existing standards.

The sustainability of language technologies and resources is a major concern. Prêt-à-LLOD aims to solve this by providing services as data, that is, wrapping services in portable containers that can be shared as single files. Language data also eventually becomes valueless as the documentation and expertise for processing esoteric formats is lost, and the project thus apply the paradigm of data as services, where services can be embedded in multi-service workflows, that demonstrates the service's value and supports long-term maintenance through methods such as open source software. Furthermore, Prêt-à-LLOD is building tools to measure and analyse the validity, maintainability and licensing of the data and services, with the objective of increasing the quality and coverage of language resources and technologies by ensuring that services are easier to archive and reuse, and thus remain available for longer.

Prêt-à-LLOD is also concerned with the issue of detecting and "chaining" licensing conditions for the language resources and services, which can be combined in complex pipelines. So that in addition to the three basic methodologies concerned with delivery, transformation and linking, the project also deals with the automated execution of smart policies for language data transactions. In particular, part of this work is based on the ODRL specifications.[17]

Since all those steps need to be carefully designed and integrated in a workflow, Prêt-à-LLOD is therefore designing a protocol, based on semantic mark-up, that aims at enabling language services to be easily connected into multi-server workflows.

Sustainability of such an infrastructure can in the end only be warranted if it can prove its usability, in different academic and industrial scenarios. Prêt-à-LLOD involves four pilot projects, lead by industry partners, that are especially designed to demonstrate the relevance, transferability and applicability of the methods and techniques under development in the project to practical problems in the language technology industry and their solutions. While Prêt-à-LLOD workflows and methodologies cut across many potential application domains and sectors, the pilots showcase potentials in the context of the following sectors: technology companies, open government services, pharmaceutical industry, and finance. As overarching challenges, all pilots are addressing facets of *cross-language transfer* or *domain adaptation*, in varying degrees.

2.4. ELEXIS

The ELEXIS infrastructure (Krek et al., 2018) has its main aim, the creation of a virtuous cycle of lexicography that consists of the following steps:

1. The creation of digital-native (Gracia et al., 2017) lexicographic resources by lexicographers
2. The linking of these resources into a single dictionary matrix allowing sharing of information
3. The application of these linked dictionaries in natural language processing application
4. The development of tools utilizing natural language processing to help lexicographers develop and improve their dictionaries

As such, linguistic linked data is a key part of this architecture and provides the second step in this virtuous cycle. The project is developing new methods for linking dictionaries, in particular using the architecture of the Naisc system (McCrae and Buitelaar, 2018), which approaches the task of linking in the following steps: first the entries are grouped together and it is analyzed which senses may link taking into account any restrictions such as part-of-speech; at this stage entries with single senses are also linked. Secondly, the entries are examined and key textual facts such as the definition, translation or examples are extracted. Thirdly, textual similarity methods are used to estimate the similarity between the senses of each entry. Next, if there is a graph in the dictionary, such as in a wordnet, graph analytics are used to analyse similarity between senses. Then, machine learning based methods are used to combine all the features into a single probability that a sense is related. Finally, global constraints (Ahmadi et al., 2019) are applied to limit the number of senses and find the most likely overall matching.

The project has recently developed a new benchmark for this "monolingual word sense alignment" task (Ahmadi et al., 2020), which is available for 15 languages and enables evaluation of the approach. This system will then be made available as part of the ELEXIS infrastructure and offered to users through its dictionary matrix.

[17] ODRL stands for "Open Digital Rights Language" and is a W3C specification (see `https://www.w3.org/TR/odrl-model/`).

3. Summary and Outlook

Ten years after the formation of the OWLG, the situation of linked data in language technology and linguistics changed drastically. In 2012, when the first book dedicated solely to the topic was published (Chiarcos et al., 2012), the community was largely building on small-scale experiments and a bright vision of the future. Since then, providers of existing infrastructures and existing platforms are becoming increasingly involved in the process and the discussion, documented, e.g., in Pareja-Lora et al. (2019), and a clear set of community standards and conventions has emerged that facilitate creating and using Linguistic Linked Open Data.

In the ten years of existence so far, the OWLG has engaged in developing and advancing Linguistic Linked Open Data and provided an umbrella for numerous more specialized activities. A constantly pursued activity has been the organization of a long-standing series of international workshops, collocated with representative conferences, esp. the series of international workshops on Linked Data in Linguistics (LDL). The topics of LLOD have also been presented in Summer Schools and a series of Datathons.

In parallel, the LLOD cloud has grown considerably. Since 2014, linguistics is recognized as a top-level category of the LOD diagram, and since 2018, the LLOD diagram is also provided as an official 'sub-cloud' of the LOD diagram. As of March 2019, the diagram features 222 resources, i.e., it constitutes about a fifth (222/1239 resources) of the LOD cloud.

Recent changes to OWLG and LLOD infrastructures include the following:

- The LLOD cloud diagram was originally generated from DataHub.io. Since 2016, it had been generated from LingHub.org, initially populated from Datahub and a number of language resource metadata providers. The diagram version provided as part of the LOD cloud diagram uses the same mechanism as the LOD cloud diagram, i.e., an online form. An update of Datahub is currently under development and will represent the basis for future versions of both LOD and LLOD diagrams.
- The Open Knowledge Foundation has been restructuring their services. This includes the OWLG wiki and mailing list. In parts as a reaction to European GDPR, they have been discontinuing their mailing lists. After a long discussion, the Open Linguistics mailing list is now being continued as a Google Group. This is the result of a vote among the participants, and a compromise between stability and simplicity. Unfortunately, a number of providers that we would have preferred as hosts, could not offer a migration, again, in parts due to GDPR concerns. At the same time, we introduce and maintain the catgeory"Open Linguistic" at the Open Knowledge Forum.
- A GitHub organization for the OWLGdata and documentation was created.
- The website originally hosted by the Open KnowledgeFoundation, is now maintained via GitHub and hosted by NUI Galway.

On this basis, the community continues the work and welcome contributors. Upcoming events include the Fourth Summer Datathon on Linguistic Linked Open Data (SD-LLOD 2021) and the Third Conference on Language, Data and Knowledge (LDK-2021).

The general situation is that a remarkable amount of Linguistic Linked Open Data is already available and that this amount continues to grow steadily, so that in the longer perspective, we can expect more data providers to offer an L(O)D view on their data, and to support RDF serializations such as JSON-LD as interchange formats. However, further growth and popularity depends crucially on the development of applications that are capable of consuming this data in a linguist-friendly fashion, or to enrich local data with web resources.

At the time of writing, working with RDF normally requires a certain level of technical expertise, i.e., basic knowledge of SPARQL and at least one RDF format. The authors' personal experience in university courses shows that linguists *can* be trained to acquire both successfully. However, this not normally done, and unlikely to ever be part of the linguistics core curriculum. This may change once designated text books on Linked Open Data for NLP and linguistics are becoming available,[18] but for the time being, a priority for this effort and the community remains to provide concrete applications tailored to the needs of linguists, lexicographers, researchers in NLP and knowledge engineering.

Promising approaches in this direction do exist: Existing tools can be complemented with an RDF layer to facilitate their interoperability. Likewise, LLOD-native applications are possible, e.g., to use RDFa (RDF in attributes) (Herman et al., 2015) to complement an XML workflow with SPARQL-based semantic search by means of web services (Sabine Tittel and Chiarcos, 2018), to provide aggregation, enrichment and search routines for language resource metadata (McCrae and Cimiano, 2015; Chiarcos et al., 2016), to use RDF as a formalism for annotation integration and data management (Burchardt et al., 2008; Chiarcos et al., 2017), or to use RDF and SPARQL for manipulating and evaluating linguistic annotations (Chiarcos et al., 2018b; Chiarcos et al., 2018a).

While these applications demonstrate the potential of LOD technology in linguistics, they come with a considerable entry barrier and they address the advanced user of RDF technology rather than a typical linguist. Even though concrete applications to exist, a long way is still to go to achieve the level of user-friendliness expected by occasional users of this technology.

A notable exception in this regard is LexO (Bellandi et al., 2017), which is a graphical tool for the collaborative editing of lexical and ontological resources natively building on the OntoLex vocabulary and RDF, designed to conduct lexicographical work in a philological context (i.e., creating the *Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*). Other projects whose objective is to provide LLOD-based tools for specific areas of application have been recently approved, so that progress in this direction is to be expected within the next years.

[18] A first step being realised by (Cimiano et al., 2020).

4. Acknowledgements

This paper is supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreeement no. 731015). It is also partially based upon work from COST Action CA18209 - NexusLinguarum "European network for Web-centred linguistic data science".
We also thank the anonymous reviewers for their helpful comments.

5. References

Ahmadi, S., Arcan, M., and McCrae, J. (2019). Lexical Sense Alignment using Weighted Bipartite b-Matching. In *Proceedings of the Poster Track of LDK 2019*, pages 12–16.

Ahmadi, S., McCrae, J. P., Nimb, S., Troelsgård, T., Olsen, S., Pedersen, B. S., Declerck, T., Wissik, T., Monachini, M., Bellandi, A., Khan, F., Pisani, I., Krek, S., Lipp, V., Váradi, T., Simon, L., Győrffy, A., Tiberius, C., Schoonheim, T., Moshe, Y. B., Rudich, M., Ahmad, R. A., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, T., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J. L., na Ruiz, R.-J. U., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stanković, R., Krstev, C., Lazić, B., Marković, A., Perdih, A., and Gabrovšek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*.

Bellandi, A., Giovannetti, E., Piccini, S., and Weingart, A. (2017). Developing LexO: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France, September. Association for Computational Linguistics.

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

Bosque-Gil, J., Lonke, D., Gracia, J., and Kernerman, I. (2019). Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference.*, pages 726–746, Brno, Czech Republic, October. Lexical Computing CZ s.r.o.,.

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 *Proc. of the 3rd International Joint Conference on NLP (IJCNLP)*, pages 389–396, Hyderabad, India.

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.

Christian Chiarcos, et al., editors. (2012). *Linked Data in Linguistics*. Springer Berlin Heidelberg.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In A. Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources*. Springer, Heidelberg, Heidelberg, Germany.

Chiarcos, C., Fäth, C., Renner-Westermann, H., Abromeit, F., and Dimitrova, V. (2016). Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may. European Language Resources Association (ELRA).

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C., Schreur, J. W., and Nevskaya, I. (2017). LLODifying Linguistic Glosses. In *International Conference on Language, Data and Knowledge*, pages 89–103, Cham. Springer, Springer.

Chiarcos, C., Khait, I., Pagé-Perron, É., Schenk, N., Fäth, C., Steuer, J., Mcgrath, W., and Wang, J. (2018a). Annotating a low-resource language with LLOD technology: Sumerian morphology and syntax. *Information*, 9(11):290.

Chiarcos, C., Kosmehl, B., Fäth, C., and Sukhareva, M. (2018b). Analyzing middle high german syntax with rdf and sparql. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan.

Chiarcos, C. (2008). An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16. Foundations of Ontologies in Text Technology, Part II: Applications.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report. W3C community group final report, World Wide Web Consortium.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data - Representation, Generation and Applications*. Springer.

de Melo, G. (2015). Lexvo.org: Language-related information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, 6(4):393–400, August.

Dostert, L. (1955). The Georgetown-IBM experiment. In W. Locke et al., 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.

Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

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

Harris, S. and Seaborne, A. (2013). SPARQL 1.1 query language. W3C recommendation, World Wide Web Consortium.

Herman, I., Adida, B., Sporny, M., and Birbeck, M. (2015). RDFa 1.1 primer - third edition. W3C working group note, World Wide Web Consortium.

Ide, N. and Pustejovsky, J. (2010). What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. of the 2nd International Conference on Global Interoperability for Language Resources (ICGL*, 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.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., and Chiarcos, C. (2019). Challenges for the representation of morphology in ontology lexicons. In *Proceedings of eLex 2019. Electronic lexicography in the 21st century: Smart lexicography*.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C., and Wissik, T. (2018). European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

McCrae, J. P. and Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1):109–123.

McCrae, J. P. and Cimiano, P. (2015). Linghub: a Linked Data based portal supporting the discovery of language resources. In *Proc. of the 11th International Conference on Semantic Systems*.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–719.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proc. of the 5th Biennial Conference on Electronic Lexicography (eLex)*.

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

Nordhoff, S. and Hammarström, H. (2011). Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science 2011*, Bonn, Germany.

Antonio Pareja-Lora, et al., editors. (2019). *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*. MIT Press.

Sabine Tittel, H. B.-S. and Chiarcos, C. (2018). Using RDFa to link text and dictionary data for medieval French. In *Proc. of the 6th Workshop on Linked Data in Linguistics (LDL-2018): Towards Linguistic Data Science*, Paris, France. European Language Resources Association (ELRA).

Windhouwer, M. and Wright, S. E. (2012). Linking to linguistic data categories in ISOcat. In *Linked Data in Linguistics*, pages 99–107. Springer.

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.