Chapter 3
Linguistic Linked Open Data Cloud

Abstract This chapter introduces the Linguistic Linked Open Data (LLOD) Cloud. In recent years, there has been increasing interest in publishing linguistic datasets following linked data principles. A number of community-driven activities, foremost organized by the Open Linguistics Working Group (OWLG), have fostered and supported the publication of open linguistic datasets and have defined criteria for when a dataset can be regarded as forming part of the so-called LLOD. The LLOD cloud represents an index and temporal snapshot of the linguistic datasets that have been published on the Web following Linked Open Data principles. The LLOD cloud is a result of a coordinated effort of the OWLG, its members and collaborating initiatives, most notably the W3C Ontology-Lexica Community Group (OntoLex), which focuses on modelling lexico-semantic resources as linked data. The LLOD cloud is visualized by means of a cloud diagram that displays all the resources with their relative sizes and their connections. In this chapter we describe the efforts by many community activities and groups that have fostered the creation of the Linguistic Linked Open Data Cloud. We also describe the methodologies and principles that allow anyone to publish a dataset that can be included in the Linguistic Linked Open Data Cloud.

3.1 Background and Motivation

Many fields of linguistics, applied linguistics and computational linguistics, build on empirical methodologies, producing insights by analysing and processing data. Example fields are quantitative typology [1], corpus linguistics [2] and computational lexicography [3]. The empirical grounding of these areas has led to the creation of a large number of linguistic datasets and resources. These resources are not only increasing in number, but are becoming more and more diverse in terms of data formats, metadata and categories. Thus, establishing interoperability between datasets as well as fostering data reuse are becoming increasing challenges.

P. Cimiano et al., *Linguistic Linked Data*,
https://doi.org/10.1007/978-3-030-30225-2_3

Ide and Pustejovsky [4] define interoperability as consisting of two principal aspects:

- **Structural Interoperability:** Structural interoperability is concerned with ensuring that datasets from different origins can be straightforwardly accessed, combined and queried. This requires the standardization of data formats, protocols for obtaining as well as querying and accessing the data.
- **Conceptual Interoperability** Conceptual interoperability requires that metadata and annotations use a common vocabulary in terms of data categories/vocabularies and ontologies used. This would allow, for instance, for automatic integration or reconciliation of different datasets.

In this book, we describe a set of principles that allow to publish linguistic datasets in a way that fosters structural and conceptual interoperability by building on linked data principles. Our focus lies in particular on datasets that fulfil the following requirements:

- **linguistic**: datasets should be useful linguistic sets consisting of annotations or similar that are relevant to researchers in linguistics and NLP.
- **linked**: datasets should include links to third-party category systems to ensure conceptual interoperability in addition to data-level links to other datasets.
- **open**: datasets should be open according to the following definition[1]

> Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).

In this chapter, we give a brief overview of the linked data principles. Further, we provide an overview over a number of community-driven activities that have fostered the adoption of linked data principles for the publication of language datasets, most importantly the Open Linguistics Working Group. We in particular describe the so-called *Linguistic Linked Open Data Cloud* that represents an index/snapshot of all the language datasets published on the Web following linked data principles.

3.1.1 Linked Data

The Linked Open Data Paradigm is based on four principles for the publication of data on the Web[2]:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.

[1] Following https://opendefinition.org/.

[2] From https://www.w3.org/DesignIssues/LinkedData.html.

3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs, so that they can discover more things.

These rules facilitate data interoperability in many ways. The use of URIs makes resources and data elements globally uniquely identifiable. If URLs are used in particular, these data elements become also resolvable. Using standards such as RDF and SPARQL ensures structural interoperability as all data sources can be accessed and queried uniformly. Finally, by including links to other datasets, in particular reusing third-party vocabularies and category systems, conceptual interoperability is established. Further, all the datasets published on the Web form a network that can be systematically accessed and browsed, navigating from one resource to a related resource and thus allowing to access all relevant datasets from one entry point.

The Resource Description Framework (RDF) as introduced in Chap. 2 represents the core model for publishing linked data.

The concept of linked data is closely coupled with the idea of **openness** (otherwise, the linking is only partially retrievable) and, in 2010, the original definition of Linked Open Data has been extended with a 5 star rating system for data on the Web.[3] The first star is achieved by publishing data on the Web (in any format) under an open license; 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 datasets to provide context.

3.1.2 Linked Open Data

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 and conceptually interoperable. Chiarcos et al. [5] identified five main benefits involved in applying the linked data principles to the representation, modelling and publication of linguistic data:

- **Representation and modelling**: Linked data is based on the RDF graph model, which models data as a labelled directed graph. This represents a very versatile data model that can be used to represent stand-off annotations, feature structures, constituent parses, dependency parses, etc. directly without the need for encoding them as tabular structures.
- **Structural Interoperability**: The use of the HTTP protocol to retrieve data without the need for any proprietary protocols or services makes sure that all datasets can be accessed in the same way. The use of RDF eases the integration of datasets coming from different sources, while the use of URIs makes sure that

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

datasets can be uniquely referenced and provenance and governance is associated to the URL behind which data is published.

- **Conceptual Interoperability**: By requiring that datasets are linked, linked data fosters reuse of existing category systems/annotation schemas and thus fosters category reuse and conceptual interoperability. Conceptual interoperability is also fostered by the fact that metadata and functional descriptions can be retrieved by resolving URIs, providing access to definitions, OWL axioms and other constraints that allow to understand how to use the URI as intended by the owner of the corresponding domain.
- **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 by federation. Resources can be uniquely identified and easily referenced from any other resource on the Web through URIs. Similar to hyperlinks on the HTML web, the Web of Data created by these links allows to navigate along these connections and to retrieve these related resources. As such, it is not necessary to keep local copies of datasets, but instead data can be accessed remotely by means of APIs, which follow open standards.
- **Dynamicity**: When linguistic resources are interlinked by references to resolvable URIs instead of proprietary IDs (or static copies of parts from another resource), we always provide access to the most recent version of a resource. In community-maintained terminology repositories, new categories, definitions or examples can be introduced occasionally, and this information is available immediately. 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.
- **Ecosystem**: RDF as a data model for the Web is maintained by an interdisciplinary, large and active community, and it comes with a mature tool ecosystem 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. Further, the distributed approach of the linked data paradigm facilitates the distributed development of 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.

3.2 Linguistic Linked Open Data

Recent years have seen not only a number of approaches to model linguistic data as linked data, but also the emergence of larger initiatives that aim at interconnecting these resources. The LLOD is part of the LOD Cloud and is made available at the

LOD Cloud website[4] as well as independently on a distinct site for the subcloud.[5] The linguistic subcloud of the Linked Open Data cloud is maintained by the **Open Linguistics Working Group (OWLG)**,[6] 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),[7] 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). The group has spearheaded the creation of new data and the republishing of existing linguistic resources as part of an emerging Linked Open Data (sub-) cloud (see below) of linguistic resources. The LLOD cloud is thus a result of a coordinated effort of the OWLG, its members and collaborating initiatives, most notably the W3C Ontology-Lexica Community Group (OntoLex, see below), which focuses on modelling lexico-semantic resources as linked data.

3.2.1 The LLOD Cloud

The Linguistic Linked Open Data Cloud is an index of all linguistic datasets that have been published as LLOD. Furthermore, this is represented by a visualization (see Fig. 3.1), which shows the individual datasets and their connections. The *Linguistic* Linked Open Data is a subset of the larger linked open data cloud, shown in Fig. 3.2, and comprises of any kind of linked open dataset considered relevant for linguistic research or for natural language processing. In the main Linked Open Data Cloud Diagram, the linguistic resources are coloured in green. In March 2015, the OWLG proposed an operational definition to replace the earlier, informal use of the term **linguistically relevant**. In this context, a dataset is linguistically relevant if it provides or describes language data that can be used for the purpose of linguistic research or natural language processing. Besides *linguistic resources in a strict sense*, (1) this includes *other linguistically relevant* resources (2) that can be used for annotating, enriching, retrieving or classifying language resources.

1. **Linguistic resources in a strict sense** are resources which have been intentionally created for the purpose of linguistic research or natural language processing, and which contain linguistic classifications, annotations or analyses or have been used to provide such information about language data.
2. **Other linguistically relevant** resources include all other resources used for linguistic research or natural language processing, but not necessarily created for this purpose, e.g. large collections of texts such as news articles, encyclopedic

[4] https://lod-cloud.net.
[5] https://linguistic-lod.org.
[6] http://linguistics.okfn.org.
[7] http://okfn.org/.

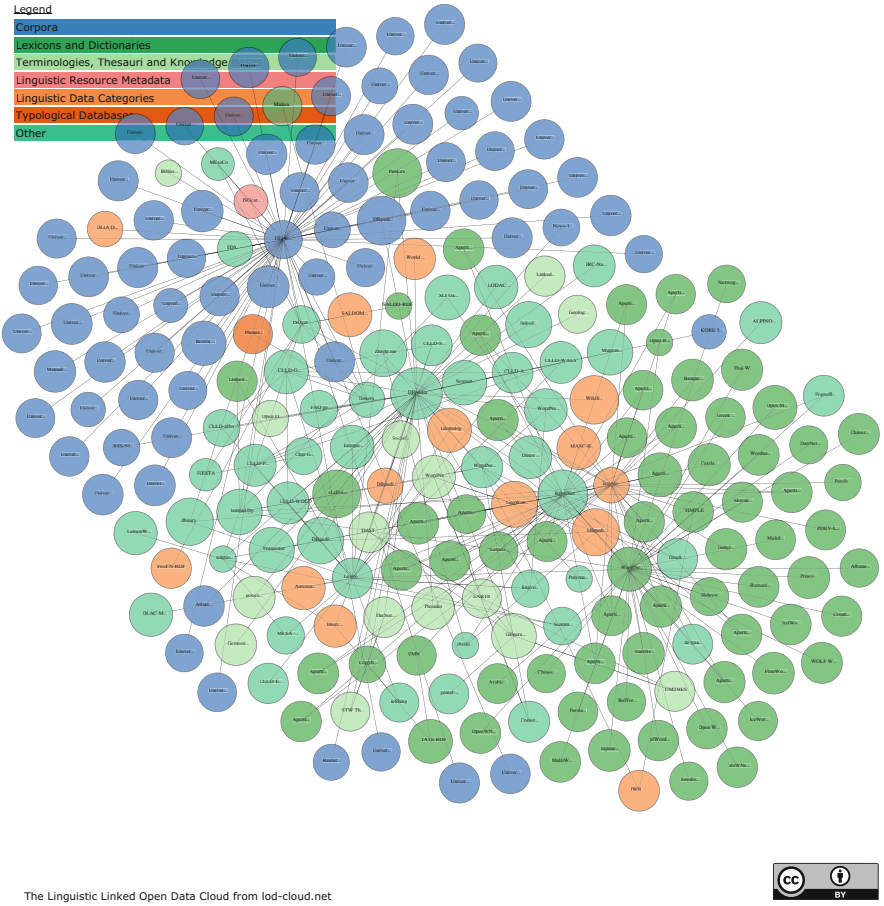

Fig. 3.1 The Linguistic Linked Open Data Cloud as of June 2018

or terminological knowledge or general knowledge bases such as DBpedia, or metadata collections, but only if they include incoming or outgoing links with at least one linguistic resource in a strict sense.

This definition is designed to provide clear-cut criteria as to whether a LOD resource can be included in the LLOD diagram, and in condition (2), it is specific to this purpose: Condition (1) can be verified by associated publications at linguistic/NLP conferences, journals or inclusion in metadata collections such as the LRE Map. Condition (2) can be verified by the existence of links between a resource (whose linguistic relevance is to be confirmed) and resources fulfilling condition (1).

A prototypical example for condition (1) would be a linguistics-/NLP-specific vocabulary, a dictionary with rich grammatical information or an annotated corpus.

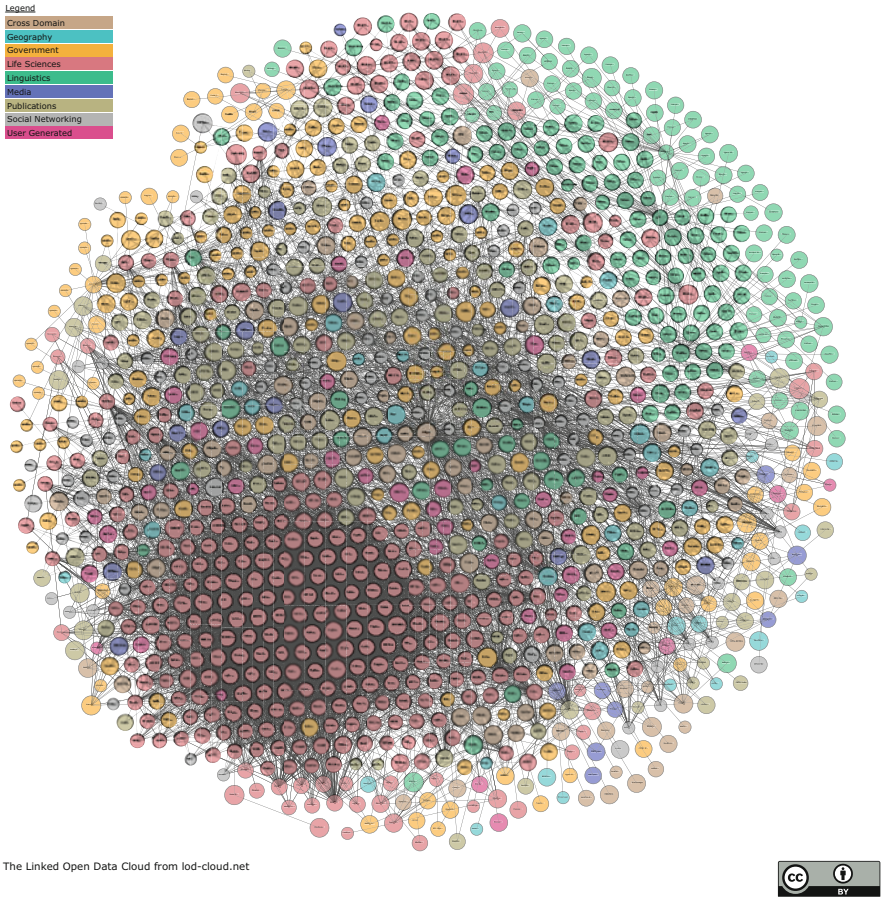

Fig. 3.2 The Linked Open Data Cloud as of June 2018. The linguistic resources are coloured green

Prototypical examples for condition (2) are resources which are frequently used in NLP and linguistics, but which have neither been created within these communities nor contain specifically linguistic information, e.g. the DBpedia.

It should be noted that condition (1) does not extend to all kinds of language resources, but limits our scope to those with *annotated or analysed data*. A corpus with vast amounts of primary data, even if created for linguistic research, published as such and automatically converted to linked data, e.g. using the NLP Interchange Format is certainly a valuable *language resource*, but does not necessarily constitute a linguistic resource in a strict sense according to condition (1). Nevertheless, it may be a linguistic resource by merits of condition (2).

Linguistic Linked Open Data, then, comprises datasets that are provided under an open license and are published in conformance with the linked data principles as

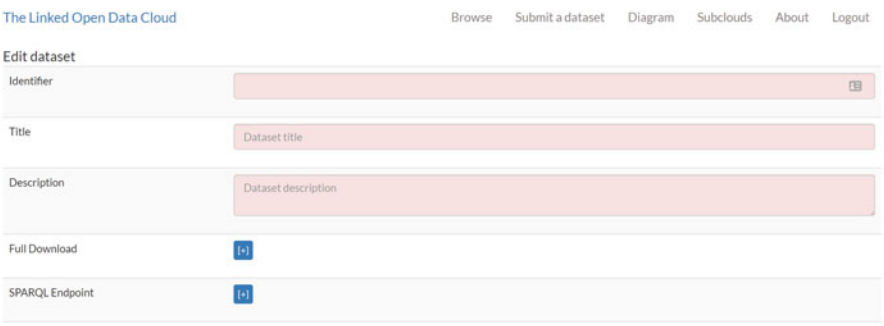

Fig. 3.3 A screenshot of the LOD Cloud Editor interface

stated above. Typically, these do not represent resources which are RDF-native, but resources that have been transformed into RDF.

3.2.2 Infrastructure and Metadata

The official LLOD cloud is hosted at http://linguistic-lod.org and is also available from the main LOD Cloud site at http://lod-cloud.net. The metadata of a LLOD resource can be updated via the LOD cloud editor interface shown in Fig. 3.3. In order for the datasets to be included in the LLOD cloud, it is necessary to select 'Linguistics' under the option 'Domain'. In addition, it is recommended that a keyword be added to describe the dataset type, which should be one of 'corpus', 'lexicon', 'metadata' or 'typology'. These are used to classify LLOD resources into three broad groups:

Corpora (blue resources in Fig. 3.1) are collections of language data, e.g. examples, text fragments and entire discourses. It should be noted here that—in accordance with condition (1)—a 'corpus' is always understood as a linguistically analysed resource, the defining element are annotations. The notion of 'corpus' thus extends both to classical RDF-only approaches where annotations *and* primary data are modelled in RDF [6], as well as to hybrid models where only annotations are provided as linked data, but the primary data is stored in a conventional format [7]. According to our definition, it does *not* extend to collections of (unanalysed) primary data. While it can be seen that corpora are less numerous than in general in linguistic data, there are still many corpus resources available showing that LLOD is as suitable for corpus resources as lexical resources.

Lexicons (green resources in Fig. 3.1) focus on the general meaning of words and the structure of semantic concepts. These represent by far the most established and flourishing type of linguistic resources in the linked data context. There is a long tradition and interest in applying Semantic Web data models to modelling lexical resources, going back to early attempts to integrate WordNet into the Semantic Web

world [8]. In the diagram, we distinguish two types of lexical-conceptual resources, i.e. *lexical resources* which also provide grammatical information (lexicons and dictionaries), and *term bases* which focus on vocabulary rather than linguistics (terminologies, thesauri and knowledge bases such as YAGO and DBpedia) and 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.

Metadata (orange resources in Fig. 3.1) includes resources providing information about language and language resources, i.e. typological databases (collections of features and inventories of individual languages, e.g. from linguistic typology), linguistic terminology repositories (e.g. grammatical categories or language identifiers) and metadata about language resources (linguistic resource metadata repositories, incl. bibliographical data).

Typologies (pink resources in Fig. 3.1) While bibliographical data and terminology management represent classical linked data applications, *typological databases* describe features of individual languages and are a particularly heterogeneous group of linguistic resources as 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.

For resources with missing tags, the classification is made in an automatic fashion.

Another type of metadata is concerned with licensing information. Among LLOD data sets, we encourage the use of **open** licenses. 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.'[8] At the moment, this condition is not yet enforced for the diagram. In fact, of the 86 resources that are declared as linguistic and for which a license is available,[9] all but 4 of these resources (4.7%) are licensed under open licenses, and these 4 are open but for non-commercial restrictions on the re-use of the data. However, legal metadata is also classified and a specific visualization can be generated with the dynamic edition of the diagram.[10] In the longer perspective, we expect a growth of linked data resources, so that our continuous move towards increasingly rigid quality criteria may eventually exclude non-open resources from the LLOD diagram. At the same time, however, we expect a growing number of licensed linked data resources, which may give rise to a 'Licensed Linked Data cloud' diagram which then takes the (pruned) LLOD

[8] http://opendefinition.org.

[9] 42 resources have no declared license at `lod-cloud.net`.

[10] http://linguistic-lod.org/llod-cloud.

diagram as its core—but extends it to other resources relevant for academic research as well as industry partners.

Furthermore, extensions and limitations of established vocabularies are being noted and an active development cycle has been started, e.g. pertaining possible extensions of lemon (see Chap. 4). It should be noted that such proposals are a sign of maturity and wider adoption, as new use cases and needs are being identified that were not foreseen when the vocabularies where originally developed. As such, NIF (see Chap. 5) was designed as a format for NLP annotations generated on the fly, not linguistic corpora, and lemon was not designed as a generic vocabulary for lexical resources, but for the specific task of adding lexical information to an existing ontology. Recently, however, NIF has been increasingly used to model and represent corpora, while lemon is being more and more used to model non-ontological lexical resources. The extension of vocabularies and the development of downward-compatible extensions will be one of the key issues of the future development of LLOD and the communities behind.

3.3 LLOD Community

The *Open Linguistics Working Group* (OWLG) of the Open Knowledge Foundation has grown steadily since its foundation in October 2010. One of its 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
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 150 people from more than 20 different countries. As the group is continuously growing, it also remains heterogeneous, and includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology. The ground for fruitful interdisciplinary discussions has been laid out. One concrete result emerging out of collaborations between a large number of OWLG members is the LLOD cloud as already sketched above.

The **Ontology-Lexica Community (OntoLex) Group**[11] was founded in September 2011 as a W3C Community and Business Group. As of May 2019, it featured 119 members. 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 includes the representation of morphological and 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

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

important issue herein will be to clarify how extant lexical and language resources can be leveraged and reused for this purpose. As a by-product 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 **Linked Data in Linguistics** (LDL) workshops have been the major focal point for the community and have had successful editions in 2012, 2013,[12] 2014,[13] 2015,[14] 2016[15] and 2018.[16] Here, the community has gathered to discuss and to convert data sets, and the OWLG has continued to refine the classification of language resources and encouraged others to contribute, e.g. by organizing various events on linked data and LLOD. These efforts have met with success such that the number of candidate resources for the cloud has increased substantially. Along with this growth, we continue to enforce increased quality constraints imposed on resources in the cloud diagram. In addition, there have been a number of other workshops on topics related to LLOD, including NLP&DBpedia, the Multilingual Semantic Web (MSW), Multilingual Linked Open Data for Enterprise (MLODE), Knowledge Extraction and Knowledge Integration (KĒKI) and Linked Open Data Resources for Collaborative Data-Intensive Research (LLOD-LSA).

In addition, there have been a number of summer schools and hackathons that support LLOD and encourage people to learn about and adopt these technologies. In particular, the **Summer Datathons on Linguistic Linked Open Data** (SD-LLOD) have been organized in several events and have encouraged a range of participants to make their resources available as linked data. In addition, linguistic linked data has appeared as a topic in a number of other events, including being a theme of the 2014 EUROLAN Summer School and appearing in the program of the 2018 European Summer School on Logic, Language and Information (ESSLLI).

Of particular importance for supporting LLOD is the World Wide Web Consortium (W3C), where a number of groups have been formed for activities in relation to open linguistic resources as well as providing a forum for interested researchers, data providers and user communities from linguistics who would not normally work in the context of the W3C. In addition to the aforementioned Ontolex community group, there has been activity in the **Linked Data For Language Technologies** (LD4LT) and the **Best Practices for Multilingual Linked Open Data** (BPM-LOD) community groups.

[12] http://ldl2013.linguistic-lod.org/.
[13] http://ldl2014.linguistic-lod.org/.
[14] http://ldl2015.linguistic-lod.org/.
[15] http://ldl2016.linguistic-lod.org/.
[16] http://ldl2018.linguistic-lod.org/.

3.3.1 Summary and Further Reading

The Linguistic Linked Open Data Cloud is the main visualization to represent the current status of the linguistic datasets that are available as linked data. The steady growth in the size of the cloud clearly corroborates the wider adoption of the linked data principles for the publication of linguistic datasets. Furthermore, this has led to an active community that is continuing to organize events to support the growth of the cloud. The cloud has and will continue to be developed by a principle of slowly increasing requirements to be fulfilled by datasets so that the data is not only increasing in quantity but also increasing in quality.

The main advantages of linguistic linked open data were first described in [5] and then a summary of the developments has been given by McCrae et al. [9]. The LIDER Project[17] has developed many guidelines for publication and linking of linguistic datasets on the Web. A comprehensive review of models, ontologies and their extensions to represent language resources as LLOD, by focusing on the nature of the linguistic content they aim to encode, was developed by Bosque et al. [10].

References

1. J. Greenberg, A quantitative approach to the morphological typology of languages. Int. J. Am. Linguist. **26**, 178 (1960)
2. W.N. Francis, H. Kucera, Brown Corpus manual. Technical Report, Brown University, Providence, Rhode Island (1964). Revised edition 1979
3. W. Morris (ed.), *The American Heritage Dictionary of the English Language* (Houghton Mifflin, New York, 1969)
4. N. Ide, J. Pustejovsky, What does interoperability mean, anyway? Toward an operational definition of interoperability, in *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL)*, Hong Kong, 2010
5. C. Chiarcos, J. McCrae, P. Cimiano, C. Fellbaum, Towards open data for linguistics: lexical linked data, in *New Trends of Research in Ontologies and Lexical Resources* (Springer, Berlin, 2013), pp. 7–25
6. A. Burchardt, S. Padó, D. Spohr, A. Frank, U. Heid, Formalising multi-layer corpora in OWL/DL – lexicon modelling, querying and consistency control, in *Proceedings of the 3rd International Joint Conference on NLP (IJCNLP)*, Hyderabad, 2008, pp. 389–396
7. S. Cassidy, An RDF realisation of LAF in the DaDa annotation server, in *Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5)*, Hong Kong, 2010
8. A. Gangemi, R. Navigli, P. Velardi, The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet, in *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (Springer, Berlin, 2003), pp. 820–838

[17] http://lider-project.eu/.

9. J.P. McCrae, C. Chiarcos, F. Bond, P. Cimiano, T. Declerck, G. de Melo, J. Gracia, S. Hellmann, B. Klimek, S. Moran, P. Osenova, A. Pareja-Lora, J. Pool, The Open Linguistics Working Group: developing the Linguistic Linked Open Data cloud, in *Proceedings of the 10th Language Resource and Evaluation Conference (LREC)* (2016)
10. J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda, A. Gómez-Pérez, Models to represent linguistic linked data. Nat. Lang. Eng. **24**(6) (2018), https://doi.org/10.1017/S1351324918000347