LiDo RDF: From a Relational Database to a Linked Data Graph of Linguistic Terms and Bibliographic Data

Bettina Klimek[1,2,a], Robert Schädlich[1,2,a], Dustin Kröger[2,b], Edwin Knese[2,b], Benedikt Elßmann[2,b]
[1]AKSW/KILT Research Group, InfAI and University of Leipzig, Germany
[2]Faculty of Mathematics and Computer Science, University of Leipzig, Germany
[a]{klimek, schaedlich}@informatik.uni-leipzig.de, [b]{dk33cuku, ek41cyvi, be57xocu}@studserv.uni-leipzig.de

Abstract

Forty years ago the linguist Dr. Christian Lehmann developed a framework for documenting linguistic terms, concepts and bibliographic data that resulted in the LiDo Terminological and Bibliographical Database (LiDo TBD). Since 2006 students and linguistic researchers benefit from the data by looking it up on the Web. Even though, the LiDo TBD is implemented as a relational database, its underlying framework aims at yielding a terminological network containing data nodes that are connected via specific relation edges in order to create an interrelated data graph. Now, with the emergence of Semantic Web technologies we were able to implement this pioneering work by converting the LiDo TBD relational database into a Linked Data graph. In this paper we present and describe the creation of the LiDo RDF dataset and introduce the LiDo RDF project. The goals of this project are to enable the direct use and reuse of the data both for the scientific research community and machine processing alike as well as to enable a valuable enrichment of already existing linguistic terminological and bibliographic data by including LiDo RDF in the LLOD cloud.

Keywords: Lido Terminological and Bibliographical Database, linguistic concepts, linguistic terminology, linguistic bibliography, Linguistic Linked Open Data (LLOD)

1. Introduction

The clarification of terminological problems is a prerequisite for methodologically and scientifically sound linguistic research (Lehmann, 1996). Resources providing and defining linguistic terminology are, therefore, of inestimable value because they practically assist the researcher in finding, understanding and reusing linguistic terms in the context of his work. Such a resource constitutes the *LiDo Terminological and Bibliographical Database*[1] (Lehmann, 1996; Lehmann, 1976). Having started in year 1976, Christian Lehmann[2] contributed his knowledge as a general and comparative linguist by manually collecting more than 4500 unique linguistic concepts, 15000 terms, 20000 books and 1200 journals. He structured this still ongoing data compilation effort within a relational database that interrelates concept, term and bibliographical data which resulted in the LiDo website that is available since 2006. While this browsable Web interface is usable as a look-up resource, it provides no means that enable a citation of the terms and concepts by Christian Lehmann because the single entries are not rendered with specified URLs which are usually required for citing Web resources.

Subsequently, the LiDo RDF project emerged in order to enable the direct citation of the terminological data for researchers but also to empower machine processing. This is achieved by converting the relational database that is the source of the LiDo website (cf. footnote 1) into an RDF dataset[3]. The choice of the RDF format is motivated by the advantages of Linked Data in general, i.e. the interoperability and integration of the LiDo data into the Linguistic Linked Open Data (LLOD) Cloud[4] which allows a direct reuse and interconnection to other (terminological and/or bibliographical) linguistic data resources.

The remainder of this paper is structured as follows. An overview of the scope and aims of the LiDo RDF project is given in Section 2. A delimitation of the LiDo RDF dataset in comparison to the existing related work is outlined in Section 3. Section 4. explains how the LiDo RDF dataset has been created. This includes the description of the source data (Section 4.1.), the presentation of the created ontology (Section 4.2.) and also an illustration of the used SQL to RDF mapping tool (Section 4.3.). This is followed by an overview of the resulted LiDo RDF dataset in Section 5. and an investigation of the quality of this data in Section 6. The paper concludes in Section 7. by giving a prospect of the future work.

2. The LiDo RDF Project

The LiDo RDF project evolved out of the initial aim to provide referenceable term and concept resources for the *LiDo Terminological and Bibliographical Database* (in short *LiDo TBD*). Therefore, the data that is available on the LiDo TBD website has been converted into the LiDo RDF dataset following the creation procedure as described in Section 4. While the resulting dataset (cf. Section 5.) constitutes the main effort of the LiDo RDF project, the following goals are pursued in addition to the mere dataset conversion task and are to a large extent achievable due to the underlying Linked Data format of LiDo RDF:

1. **Data evolution**: Because the LiDo TBD is still edited and updated by Christian Lehmann, the LiDo RDF dataset strives to evolve accordingly. Therefore, a new version of LiDo RDF will be generated from LiDo

[1] http://linguistik.uni-regensburg.de:8080/lido/Lido

[2] https://www.christianlehmann.eu/

[3] The authors thank Christian Lehmann for enabling this dataset creation by providing them with his database and allowing that the LiDo RDF dataset can be published and reused under an open license.

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

TBD twice a year. The previous versions will be kept available for download.

2. **Data reuse for human users:** RDF datasets are generally not easily to be read by humans. In order to keep the lookup of the data as user friendly as possible, a browsable search interface similar to the LiDo TBD has been created based on LiDo RDF. It differs, however, in that every entry can be now cited with a unique URL.

3. **Data exploration:** For users who are more familiar with the RDF data format, a navigation through the data graph via the resource links will be available in a Linked Open Data view of LiDo RDF. Moreover, further insights into the data can be obtained by querying the data using the provided SPARQL endpoint.

4. **External data enrichment:** The LiDo RDF dataset is one among other datasets for linguistic terminology and bibliographic resources. Therefore, it is desirable to interrelate the LiDo RDF data with other similar resources. This can be achieved by integrating LiDo RDF into the LLOD Cloud and interlinking it with already existing terminological or bibliographical resources.

All in all, the LiDo RDF project aims at preserving the original LiDo TBD data and providing the means to enable the reuse of the data for humans and for machine processing alike. In contrast to the LiDo TBD which is only indirectly available as a dataset to search and view via a web interface, LiDo RDF additionally provides the actual data that can be cited in ongoing research but also reused, shared and interlinked by the linguistic research community. Furthermore, it can be directly integrated into applications that need to consume or process the data. The dataset versions and all the features described will be available from `http://lidordf.aksw.org/`. More technical details are contained in this Github repository: `https://github.com/AKSW/lido2rdf`.

3. Related Work

With regard to the models that are present for representing terminological data in linguistics as a Linked Data graph, the OntoLex-Lemon model[5] (McCrae et al., 2017) and the Ontology for Linguistic Terminology (OnLiT) (Klimek et al., 2017) have to be mentioned. OntoLex-Lemon is specified for representing lexical language data. This poses several difficulties which led to the conclusion that this vocabulary is not suitable as a modelling basis for the LiDo TBD into RDF. Most importantly what is defined as a *term* in LiDo TBD does not apply to the notion of *lexical entry* in OntoLex-Lemon. Also the concepts included in LiDo TBD cannot be understood as *lexical sense*, which would be the corresponding equivalent in OntoLex-Lemon. What is more, LiDo TBD contains a set of relations for which appropriate object and datatype properties do not exist in the OntoLex-Lemon vocabulary. It turned out that the modelling of the terminological data in LiDo TBD requires a more specific vocabulary. Such a vocabulary is OnLiT, that has been created for the purpose to model LiDo TBD in RDF. How it has been eventually used is explained in detail in Section 4.2.

Contentwise the LiDo RDF dataset is concerned with three different kinds of data: 1) linguistic concepts, 2) linguistic terms and 3) linguistic bibliography. Various Linked Data datasets exist that contain data of one or two of these domains. In the following, exemplary datasets are mentioned.

The General Ontology for Linguistic Description (GOLD)[6] (Farrar, 2010) provides a taxonomy of nearly 600 linguistic concepts for descriptive linguistics with corresponding term data encoded only as strings in RDF labels and not as independent resources.

The Bibliography of Linguistic Literature Thesaurus (BLL Thesaurus) (Chiarcos et al., 2016) includes around 7400 linguistic terms. However, only a part of the term data is semantically defined via links to corresponding OLiA[7] concepts. In this case, a comprehensive interlinking to LiDo RDF could provide appropriate concepts and their definitions for other yet unspecified term resources in the BLL Thesaurus.

The Grammis online resource[8] contains both linguistic concepts and terms specifically for German grammar (Suchowolec et al., 2017).

In the context of the Semantic Web and the existing LLOD Cloud, efforts emerged that aim at interlinking the aforementioned and other existing datasets in order to provide more comprehensive terminological and bibliographic data that is relevant for the linguistic sciences. Such an endeavour, for instance, is the BLL – Linguistic Linked Open Data Edition[9] (Chiarcos et al., 2016) that interlinks the BLL Thesaurus with the Bibliography of Linguistic Literature[10].

However, the framework that Christian Lehmann has built for the LiDo TBD pursued the same goal already forty years ago. In this respect, to the best of our knowledge, LiDo RDF is unique because it provides term, concept and bibliographic data —that is semantically more specifically and consistently interrelated than it is in existing datasets — joined into one single and multilingual dataset. Therefore, we believe, the ongoing task of interlinking existing resources about linguistic terminology would highly benefit from considering or at least discussing the reuse of the LiDo ontology (i.e. mainly OnLiT) in order to arrive at a more coherent and semantically richer Linked Data graph as a terminological basis for linguistic research.

4. Dataset Creation

4.1. LiDo TBD Source Data

The LiDo TBD is only available for lookup purposes as it is present on the website `http://linguistik.`

[5] `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[6] `http://linguistics-ontology.org/gold-2010.owl`

[7] `http://acoli.cs.uni-frankfurt.de/resources/olia`

[8] `http://www.ids-mannheim.de/grammis/`

[9] `http://data.linguistik.de/bll/index.html`

[10] `http://www.blldb-online.de`

uni-regensburg.de:8080/lido/Lido. In order to create LiDo RDF we used the Microsoft Access database that is the editing basis of Christian Lehmann. From this we generated a PostgreSQL database that served as the source data for the data conversion[11]. The PostgreSQL database is very similar to the Microsoft Access database, however, PostgreSQL was the required format for the mapping tool that was used for converting the tabular data to RDF data. The source data mainly consists of tables that contain entities and attributes describing them, e.g. a "term" table or a "book" table[12]. Additionally, the source data contains tables that interrelate the entities of other tables, e.g. concept entities with book entities or concept entities with term entities. The large part of the data centers around concepts, terms and bibliographic elements such as books and journals. For concept entities attributes like "delimitation", "analytic procedure" and "phenomenology" have been entered in addition to rather common attributes such as "definition" and "example". The attributes that describe the term entities are "abbreviation", "etymology" and "language". The latter is provided for every term, predominantly in English, German, Spanish and Portuguese. With regard to the bibliographic entities typical attributes like "author", "title", ect. exist but also the "text sort" is stated which facilitates the identification of dictionaries, questionnaires and more.
Furthermore, 18 dedicated relations are established that interrelate term entities, concept entities and concept entities with term entities. Out of these, 14 constitute coordinating and subordinating relations which interconnect concept entities. Very general hierarchical or meronymic relations like "is a (kind of)" or "is part of", but also more specific relations such as "is result of", "manifests" or "is operator of" contribute to the creation of a relational network which is already quite close to the semantic RDF graph structure. In summary, the data basis consisted of 11 tables containing entities with attributes and 7 tables that interrelated these entities by cross reference. Metadata such as the date of the last edit for an entry or the logbook table were not part of the RDF conversion.

4.2. LiDo Ontology

Every Linked Data dataset needs to be formally described by a specified model, i.e. its underlying vocabulary or ontology. In compliance to the best practices and the given Semantic Web standards, we reused existing vocabularies and extended them where necessary in order to create the LiDo ontology[13] which is the basis for the LiDo RDF dataset. While the modelling decisions for the LiDo ontology will be explained in what follows, it is recommended to refer to Figure 1 which indirectly also exemplifies the usage of the vocabulary.
The LiDo TBD adheres to the statement that "proper terminology is concerned with the relationship between concepts, and between them and their designations, rather than with designations alone or with the objects they represent"[14] and, thus, concomitantly distinguishes and interrelates linguistic concept and term resources. While the former are defined as language-independent mental objects (i.e. units of meaning) the latter are defined as language-specific linguistic objects. Consequently, within the LiDo TBD source data **a linguistic concept is unique and associated with a linguistic term that in turn hypostatizes the concept**. Within the LiDo TBD source data this has been realized by identifiers and unique labels, which are in many cases Latin expressions chosen by Christian Lehmann, to represent the concept entities within the tabular data. Similarly, the term entries are also represented by identifiers which are interrelated with their language specific expressions. All semantic interrelations occur between concept entities which are associated with the respective term entities. All of this has been already modelled within the Ontology for Linguistic Terminology (OnLiT)[15]. OnLiT was created in previous work mainly for the purpose of creating LiDo RDF and emerged from the same LiDo TBD data source (Klimek et al., 2017)[16]. Since this vocabulary contains the modelling of concept and term resources as well as their established interrelations according to the source data, it is included and reused as an OWL import within the LiDo ontology.
The other part of the data containing the bibliographic data could be modelled by reusing the Bibliographic Ontology (D'Arcus and Giasson, 2011)[17], because it contained the required classes and object properties for representing the tabular source data entities, e.g. "book", "journal", "author" and "publication year".
While the majority of the LiDo TBD could be represented by importing OnLiT and using a part of the Bibo ontology, a set of tables containing data such as languages, areas or text sorts which were interrelated with the "term" or "book" tables remained. In order to cover this data as well, we decided to create new classes and object properties to model this relational data within the LiDo ontology. The object property lido:hasBibRef, for instance, needed to be introduced to account for the interrelation of concept resources with bibliographic resources. Moreover, with regard to the Lido TBD tables "languages" and "areas", it has to be mentioned that Linked Data resources already exist, e.g. datasets for geographical and language data. However, we did not reuse these because it requires a large amount of manual mapping effort to retrieve the ca. 1100 language entities and ca. 160 area entities. Instead, the classes lido:Language, lido:Area and lido:Textsort, which were not included within the Bibo vocabulary, have been newly created and populated with the respective individuals from the tables. While the more accurate mapping task remains open as future work the source data is at its current state modelled exhaustively with the LiDo ontol-

[11] Neither this PostgreSQL database nor the underlying MS Access file are or will be publicly available. The authors have been, however, generously granted access for the undertaking of the LiDo RDF dataset creation.

[12] The source data is comprehensively described in (Lehmann, 1996), which is recommended for further reading.

[13] http://lidordf.aksw.org/ontology/

[14] http://www.computing.surrey.ac.uk/ai/pointer/report/section1.html

[15] http://lido.linguistic-lod.org/onlit.rdf

[16] Please consult this reference for more details and examples of the concept and term representation in the source data.

[17] http://purl.org/ontology/bibo/

ogy.

4.3. SQL to RDF Mapping

The actual data of the LiDo RDF dataset was created automatically by using the presented LiDo ontology and the Sparqlify tool[18] (Stadler et al., 2015) which enables a reliable and scalable transformation of relational data to RDF. Since the PostgreSQL source data and the desired RDF format differ fundamentally from each other, a manual mapping between both formats is required. Due to the supported Sparqlification Mapping Language (SML) that is integrated in Sparqlify we could easily express the necessary PostgreSQL to RDF mappings with SML[19]. As a result, the Sparqlify tool provides as output file the converted RDF data in N-Triples, a plain text serialization format for RDF graphs. In order to implement several post-processing possibilities, e.g. the set up of a Linked Open Data navigation view or the addition of links to other datasets, the output file has been loaded into a triplestore. Thereby we could also provide the SPARQL endpoint[20] for the Lido RDF dataset which enables a direct querying of the data for more detailed insights. Additionally, SPARQL queries are used to display LiDo RDF data dynamically while navigating through the human-readable search interface[21] of the LiDo RDF data. Because the search interface is based on the results of the queries for the current RDF concept and term resources, their URIs can be now used to cite the original LiDo TBD in scientific works.

5. LiDo RDF Dataset

5.1. LiDo as Linked Data Graph

By using the LiDo ontology all the data that is browsable within the LiDo TBD website could be converted into the LiDo RDF dataset. The namespace for the data is `http://lidordf.aksw.org/resource/` and its prefix is `lido`. Figure 1 illustrates a part of the data graph focusing especially on the semantically rich interrelation of linguistic concepts. The examples of the `lido:Book_13757`, `onlit:Concept_509` and `onlit:Term_511` instances show that the identifiers of the LiDo TBD source data have been reused to populate the `bibo:Book`, `onlit:Concept`, `onlit:Term` classes. The given designations for the concept identifiers (e.g. "denominatio") and the language-specific expressions corresponding to the term identifiers (e.g. "denomination" in English) in the tabular source data have been modelled by using the `rdfs:label` object property. The labels of all `onlit:Concept` instances in the LiDo dataset correspond to the list of entries that can be found under "Unique Designation" on the LiDo TBD website and are mostly suitable Latin expressions chosen by Christian Lehmann which makes it easier to refer to the concepts (in place of using the identifiers). The labels of `onlit:Term` instances, since encoding the concepts in different languages, are provided with a language tag.

In the following an overview will be given about the data that is created in LiDo RDF for the three main data types and how they are interconnected within the graph:

Bibliographic data: The main class within the LiDo ontology is the reused `bibo:Document` class, that contains the book and journal entries from the LiDo TBD source data. All book and journal resources are further specified for typical information about bibliographic entities, such as author, title, publisher, publication date but also for their text sort (not shown in Figure 1). What is remarkable, however, is that bibliographic works that can be consulted for more information about a certain linguistic concept are interlinked with respective concept resources, e.g. in Figure 1 the concept `lido:Concept_509` ('denominatio') is associated with the bibliographic reference `lido:Book_13757` ("Knobloch, Clemens and Schaeder, Burkhard (eds.) 1996, Nomination - fachsprachlich und gemeinsprachlich. Opladen: Westdeutscher Verlag (Sprachwissenschaft).").

Term data: Term instances can have additional etymological information or a given abbreviation (which are not shown in Figure 1). Furthermore, some terms are interrelated. I.e. it is explicitly stated whether a term is an abstract or concrete noun of another term, e.g. the German term *Semantik* is the abstract noun of the German adjective term *semantisch*. Regarding the interrelation of term and concept instances, LiDo RDF implemented the two possibilities of stating that a term is either the standard or non-standard expression of a concept according to the LiDo TBD as it is exemplified in Figure 1). To the whole dataset applies a one-to-one correspondence between term and concept resources. This meets the self-imposed requirement by Christian Lehmann for representing terminological data in order to ensure a disambiguated traceability of terms (and to which the LiDo TBD source data also applies). By that the LiDo RDF dataset represents a language-independent approach that enables the integration of multilingual terminological networks by defining a term in relation to the linguistic concept it encodes.

Concept data: Consequently, not the term but the concept instances are specified with a definition. For these also examples and information on the analytic procedure, delimitation and history as well as the phenomenology of a concept resource can be given, which is exemplified in Figure2. All this information is stated in plain text for every single concept resource. In addition to that, the meaningful interrelation of concepts constitutes an added value for defining a linguistic concept within the broad domain of linguistics. I.e. in LiDo RDF are no loose concepts. Every concept has at least one direct relation to another concept. The underlying basis for these interrelations builds an hierarchical structure between the concepts that is created due to the 18 subordinating and coordinating relations and emerged from the concept-concept relations that are contained in the LiDo TBD source data. Figure 1 shows some of these re-

[18] `https://github.com/AKSW/Sparqlify`

[19] The created mapping file can be found here: `https://github.com/AKSW/lido2rdf/blob/master/SPARQLIFY%20SQL2RDF%20mapping/mapping.sml`

[20] To be found here: `http://lidordf.aksw.org/sparql/`.

[21] To be accessed here: `http://lidordf.aksw.org/glossary/`.

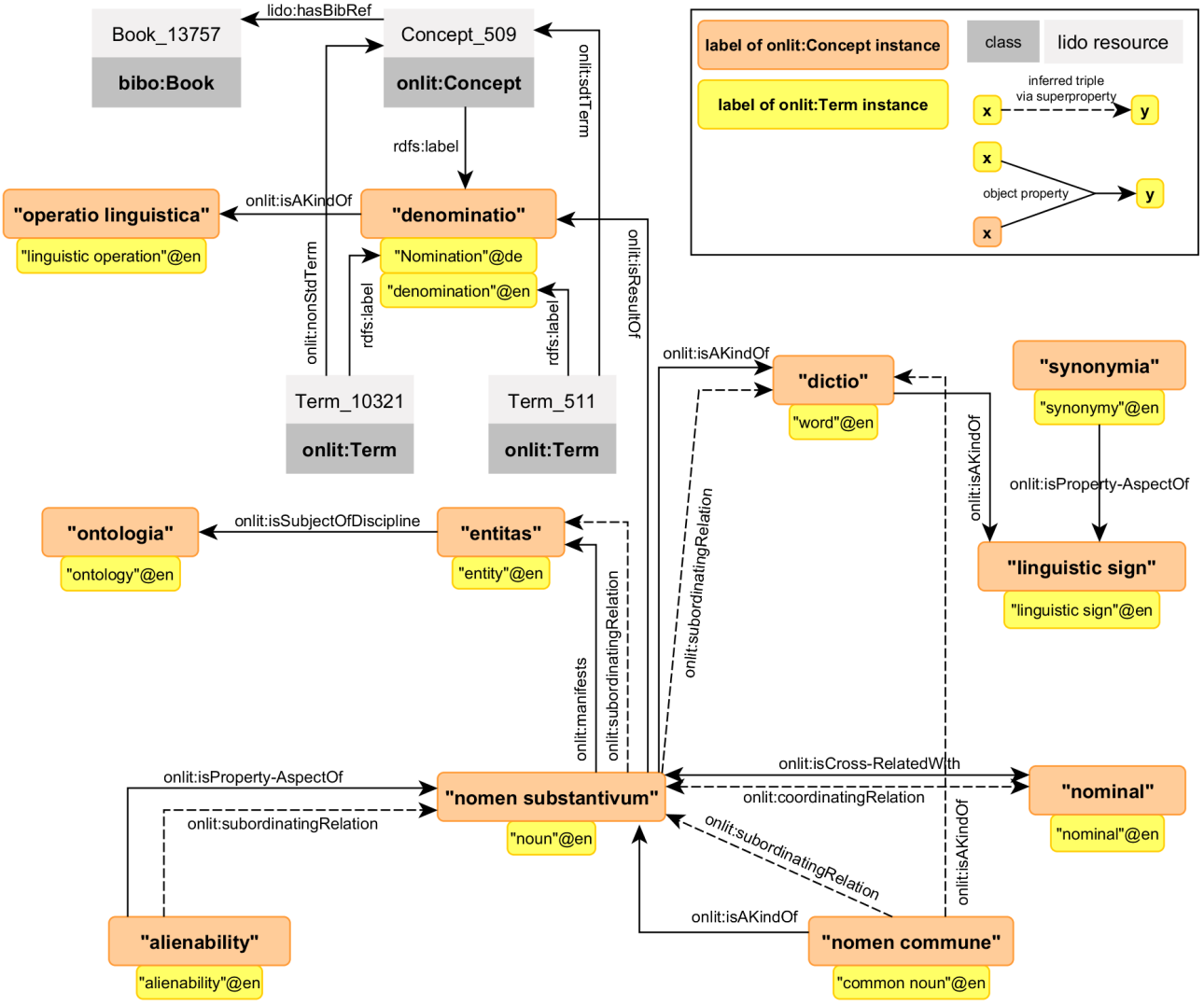

Figure 1: Example excerpt of the LiDo RDF data graph.

lations that hold between various concept instances[22] and have been modelled as OnLiT object properties. They are the cause for the emergence of the relational network of linguistic terms in the LiDo TBD source data and for the linked data graph structure of LiDo RDF respectively. To the best of our knowledge this manageable and semantically standardized set of relations is unique for a dataset of linguistic terminology because these more specific relations can be used to disambiguate the sub- and coordinating relations that are normally only formally defined and, hence, confound semantically different relations. As a result, the LiDo RDF data graph yields more informative insights into a linguistic concept, because the meaning of a linguistic concept is, next to a formulated definition, also defined by its semantic interrelation to other linguistic concepts.

Finally, with LiDo RDF an openly accessible version of the LiDo TBD dataset is available for further reuse. The underlying graph structure of the data does not only satisfy the requirements of representing the relational network of the LiDo TBD source data, but also enables the gain of deeper insights into the dataset by means of inference and reasoning. I.e. obtaining all concepts that are interlinked to one bibliographic resource, or extracting the whole hierarchy tree for a concept at once is now possible but also incomplete concept entries (e.g. with missing definitions) for future editing of the dataset can be easily retrieved. The generated LiDo RDF dataset is finally completed after it has passed a data quality check-up.

5.2. LiDo within the LLOD Cloud

While the LiDo TBD in its current state as a searchable Web interface is isolated in terms of data reusability, the availability of the LiDo RDF dataset entails the possibility to interrelate it with other existing datasets containing terminological and/or bibliographic data for the domain of linguistics. Such an interconnection of datasets including similar content is desirable because it provides the research community with an overview of various data sources that

[22] Note that for better readability the labels of the `onlit:Concept` instances and their corresponding English `onlit:Term` instances are displayed in place of their respective instances.

semantically refer and, consequently, also enrich each other by pointing to similar data entries that can be also consulted for further information. For linguistic data in the Semantic Web landscape such an environment provides the LLOD Cloud. In order to enrich the LiDo RDF dataset with external data sources and simultaneously provide the possibility for other dataset creators to point to linguistic concepts or terms within LiDo RDF, an integration of the data into the LLOD Cloud is considered a worthwhile undertaking.

Therefore, a manual set of 50 links adhering to the requirements of publishing datasets within the LLOD Cloud[23] has been created as a starting point to add LiDo RDF to the already existing collection of linguistic data within the Cloud. For the linking the Bibliography of Linguistic Literature (BLL) Thesaurus[24] (Chiarcos et al., 2016) has been chosen to be a suitable dataset because it also contains terminological and bibliographic data. In this case the linguistic concepts between LiDo RDF and the BLL Thesaurus have been mapped by using the object property `skos:exactMatch`. The linking for the linguistic concept 'circumfix', for instance, is as follows:

```
<http://lidordf.aksw.org/resource/
Concept_290>

skos:exactMatch

<http://data.linguistik.de/
bll/bll-thesaurus#bll-133083225>.
```

It has to be noted that linguistic concepts in LiDo RDF are represented as OWL individuals, while they are represented as OWL classes and are also of the type `skos:Concept` within the BLL Thesaurus. A valid statement is, however, created because the skos object property `skos:exactMatch` entails and, therefore, automatically creates a type assertion for the `lido:Concept_290` instance to be also of the `rdf:type skos:Concept`, which is necessary to yield a valid statement. The 50 created links have been added to the LiDo RDF dataset and further include, for example, the linguistic concepts 'dative', 'number', 'subject' and 'juxtaposition'.

The identification of matching linguistic concepts required a detailed study of both datasets. I.e. in order for a LiDo RDF and a BLL Thesaurus concept to be considered as an exact match, the following two requirements had to be fulfilled: 1) The English `skos:prefLabel` of the `bllt:bll-133083225` class and the label of the standard English term instance that corresponds to the `lido:Concept_290` instance had to be identical; 2) the textual definition of the LiDo concept instance and the textual comment defining the OLiA[25] class that corresponds to BLL Thesaurus class had to convey a close to equivalent content. Even though an entire linking by using also less strict mapping properties, such as `skos:broader` or `skos:narrrower` seems to be promising, the manual linking just explained showed that this task will be time-consuming and requires human judgment about the similarity of two concepts. An automated linking process due to the amount of data is, hence, favourable but should also implement some kind of quality assurance.

In conclusion, the contribution of LiDo RDF to the LLOD Cloud shall rise awareness of the dataset itself and will ideally result in collaborative work with the creators of similar existing or future datasets concerning the realization of a more extensive interlinking of terminological or bibliographic data that benefits the whole linguistic research community.

5.3. LiDo RDF Web Interface

As has been already mentioned, one of the main goals for creating the LiDo RDF dataset was to enable the citation of the term and concept data. Since not all users are, however, familiar with navigating through a Linked Data graph and using the resource URIs for citation purposes, a web interface that is intuitive and easy to browse had to be provided as well. In order to create such an interface, the URIs of the resources that are contained within LiDo RDF could be reused, as has been stated earlier in Section 4.3. The interface is accessible under the URL `http://lidordf.aksw.org/glossary/` which retrieves the respective resource identifier from LiDo RDF whenever a specific term or concept entry is selected[26]. The screenshot of the English term entry *quantifier* in Figure 2 illustrates the layout of the web interface. Only the data that is transformed from the LiDo TBD source data into RDF can be searched, while additional links to external datasets are only accessible via the SPARQL endpoint or the Linked Data view of Lido RDF.

All in all the interface is similar to the LiDo TBD website. The main difference consists in the arrangement of the term, concept and bibliographic data boxes. While the left side of the term data box in principle shows the semantic relations between linguistic concept resources within the dataset, these are displayed, however, by using their corresponding term expressions in the selected language. The right side of the term data box then shows the corresponding expressions of the chosen concept in other languages. Further, the bibliographic references that correspond to a concept can be selected as another view within the concept data box while the term data above stays visible. However, a detailed search interface for the bibliographic resources has not been implemented. Finally, the novelty and only additional feature of the LiDo RDF based web interface are the two "cite" buttons, which provide a pre-formulated reference for a selected term or concept entry that can be directly copied from the pop-up window (cf. Figure 3).

[23] `https://wiki.okfn.org/Working_Groups/Linguistics/How_to_contribute`

[24] `https://old.datahub.io/dataset/bll-thesaurus`

[25] Please cf. to (Chiarcos et al., 2016) for more details of the usage of the OLiA Ontologies within the BLL Thesaurus development.

[26] Note that the difference between for example `http://lidordf.aksw.org/resource/Term_1651` and `http://lidordf.aksw.org/glossary/Term_1651` lies only in the way the LiDo RDF data is displayed, i.e. as Linked Data view or within the web interface.

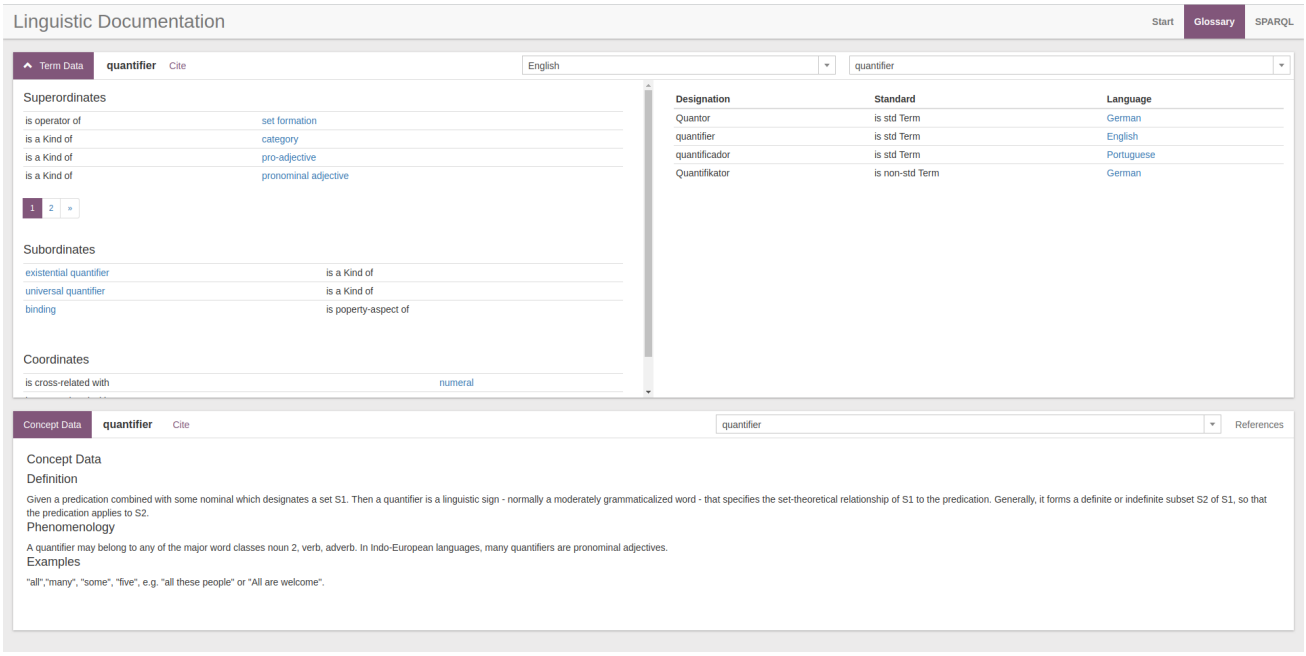

Figure 2: Screenshot of the LiDo concept 'quantifier' as displayed within the provided web interface.

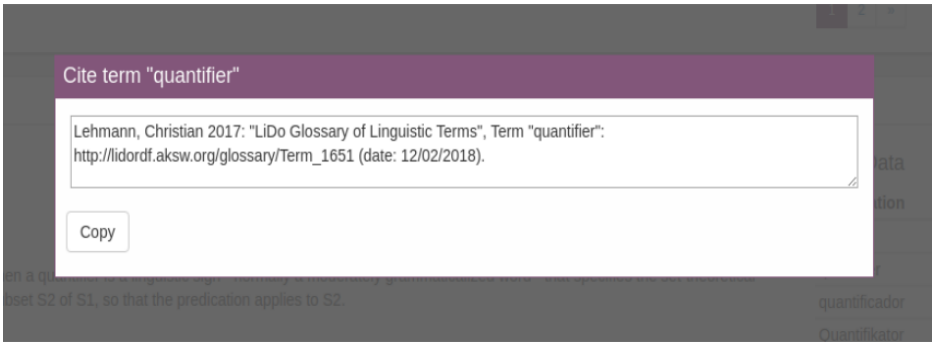

Figure 3: Screenshot of the citation function for the English term *quantifier* provided within the web interface.

The different layout presents just another option next to the existing LiDo TBD website. The design and features are still under development and might change if the user feedback demands refinements or other additional functionalities.

6. Data Quality

Automatically generated data is often subject to data quality issues. In order to preserve the high quality of the thoroughly compiled LiDo TBD, we conducted two different kinds of quality checks. The first constitutes a syntax validation, that ensures that the data is free of formal errors and can be processed by machines without any problems. The second check concerns the completeness of the data and assures that no data has been lost during the transformation process. Therefore, specific SQL and SPARQL queries have been created that count and compare the entries of both datasets. This validation process gives an immediate feedback about passing or not passing entries[27]. The completeness of the generated LiDo RDF data is then verified when all queries in the LiDo PostgreSQL source dataset and the LiDo RDF dataset produce the same output results. Only after both data quality checks are successfully applied the LiDo RDF dataset and its future versions will be published.

7. Conclusion and Future Work

In this paper we described LiDo RDF, i.e. the Linked Data version of the LiDo TBD. The primary goal of enabling the citation of single term and concept entries of the LiDo TBD could be achieved by providing a similar search interface that reuses the LiDo RDF resource URIs in a slightly altered way. Following the aim to contribute to the provision of this linguistic resource we also presented the LiDo RDF project that maintains the human-readable search interface as well as the SPARQL endpoint.

[27] See the results at `http://lidordf.aksw.org/validation/`.

As a result, the LiDo TBD is not an isolated dataset on the Web anymore, but reusable for human users and machine processing alike. Furthermore, the underlying modelling framework of LiDo TBD and LiDo RDF that represents linguistic terminology as an interrelation of concepts and corresponding terms by means of a consistent and manageable set of relations can serve as a starting point for the discussion of standards for modelling linguistic terminology as Linked Data in the future. This is especially interesting with regard to already existing standard vocabularies like OntoLex-Lemon for lexical language data, for which another module for specifically representing terminological data might be developed.
Finally, the LiDo RDF dataset could be integrated into the LLOD Cloud and is, thus, visible to the broader linguistic Linked Data research community. By that, we hope to initiate the collaboration with the creators of other datasets, such as the BLL Thesaurus or Grammis, in order to conduct a high quality and far-reaching interlinking task. This will consequently benefit linguistic research in general by contributing to a large and interconnected knowledge graph for linguistic terms, concepts and bibliographic data.

Acknowledgment This paper‘s research activities were funded by grants from the H2020 EU projects ALIGNED (GA-644055) and the Smart Data Web BMWi project (GA-01MD15010B).

8. Bibliographical References

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 *LREC*.

D'Arcus, B. and Giasson, F. (2011). *Bibliographic ontology specification. Specification document, 4 November 2009.* Retrieved August 10 (2009).

Farrar, S. (2010). *General Ontology for Linguistic Description (GOLD).* Department of Linguistics (The LINGUIST List), Indiana University.

Klimek, B., McCrae, J. P., Lehmann, C., Chiarcos, C., and Hellmann, S. (2017). Onlit: An ontology for linguistic terminology. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*, volume 10318, pages 42–57. Springer.

Lehmann, C. (1976). *Um sistema de documentação para a lingüística*. Pontifícia Universidade Católica do RGS, Instituto de Letras e Artes.

Lehmann, C. (1996). Linguistische Terminologie als relationales Netz. In *Nomination—fachsprachlich und gemeinsprachlich*, pages 215–267. Springer.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: Development and applications. In *Proceedings of eLex 2017 conference*.

Stadler, C., Unbehauen, J., Westphal, P., Sherif, M. A., and Lehmann, J. (2015). Simplified rdb2rdf mapping. *Proceedings of the 8th Workshop on Linked Data on the Web (LDOW2015), Florence, Italy*.

Suchowolec, K., Lang, C., Schneider, R., and Schwinn, H. (2017). Shifting complexity from text to data model. adding machine-oriented features to a human-oriented terminology resource. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings*, volume 10318. Springer.