Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked Open Data

C. Chiarcos, C. Fäth, H. Renner-Westermann, F. Abromeit, V. Dimitrova
Goethe Universität Frankfurt am Main, Germany
{chiarcos@em, faeth@em, h.renner-westermann@ub, abromeit@em, v.dimitrova@ub}.uni-frankfurt.de

Abstract

This paper introduces a novel research tool for the field of linguistics: The *Lin|gu|is|tik web portal* provides a virtual library which offers scientific information on every linguistic subject. It comprises selected internet sources and databases as well as catalogues for linguistic literature, and addresses an interdisciplinary audience. The virtual library is the most recent outcome of the *Special Subject Collection Linguistics* of the German Research Foundation (DFG), and also integrates the knowledge accumulated in the *Bibliography of Linguistic Literature*. In addition to the portal, we describe long-term goals and prospects with a special focus on ongoing efforts regarding an extension towards integrating language resources and Linguistic Linked Open Data.

Keywords: virtual library, Bibliography of Linguistic Literature (BLL), linguistic terminology, Linguistic Linked Open Data (LLOD), Ontologies of Linguistic Annotation (OLiA)

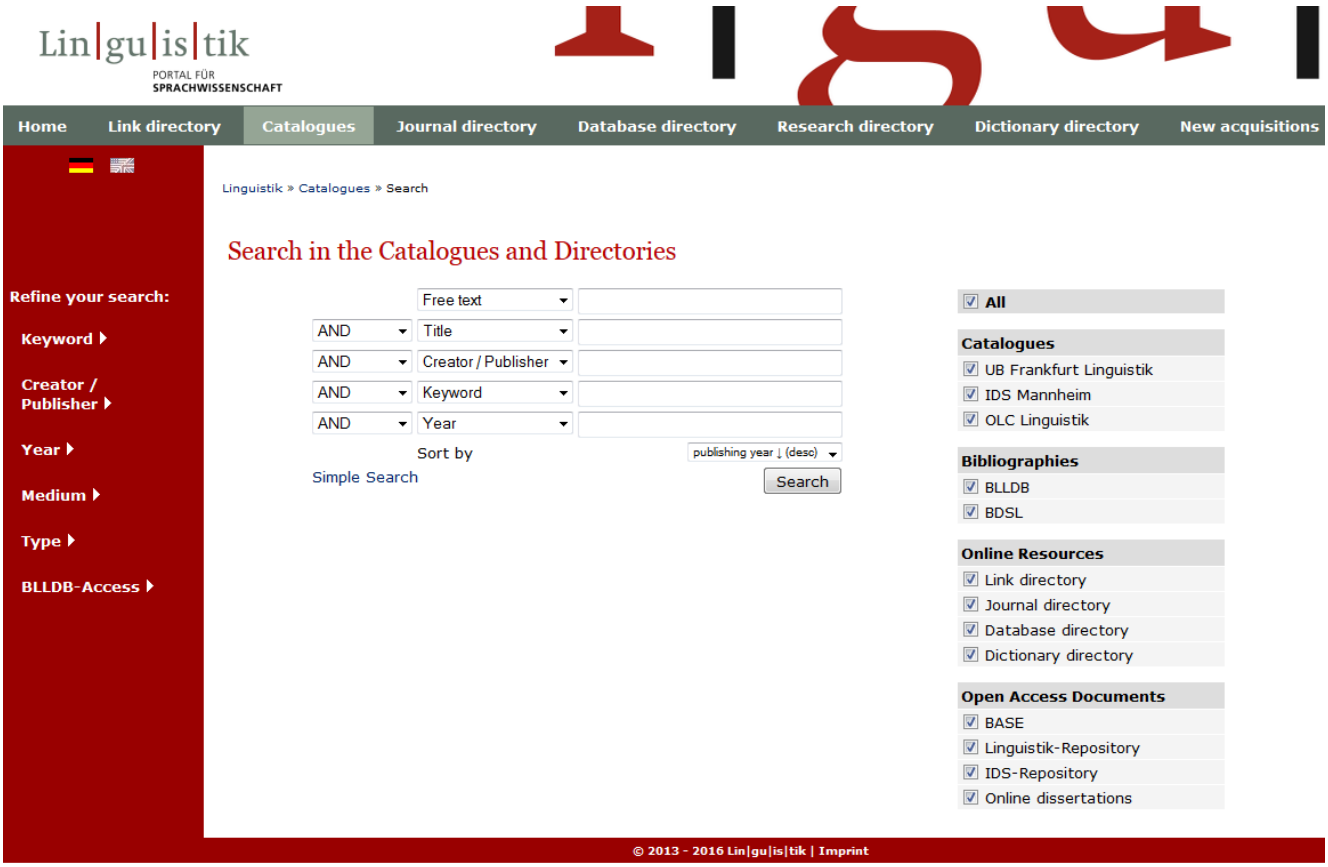

Figure 1: The extended catalogues search of the *Lin|gu|is|tik* portal in English

1. Introduction

Libraries have always been the basis for scientific progress and proliferation, and the interlibrary loan system (ILL) accelerated the scientific breakthroughs since 1900, even before the creation of online catalogues. After WWII, the German research community benefited from the combination of the ILL and the establishment of *Special Subject Collections* (SSG, Sondersammelgebiete) at different university libraries. These were designed to support the acquisition of the entire international literature for every specific field of research (Deutsche Forschungsgemeinschaft, 2015). In the digital age, the amount of information collected over more than 60 years is increasingly being made available over the web. The *Lin|gu|is|tik web portal*[1] (henceforth *Lin|gu|is|tik* portal) is an outcome of the *Special Subject Collection General Linguistics*[2] hosted since the beginning of the 50s till recently by the University Library J.C. Senckenberg in Frankfurt (Renner-Westermann, 2013).

The *Lin|gu|is|tik* portal's main modules, functions and objectives are described in Section 2. Section 3 outlines the role of the Bibliography of Linguistic Literature (BLL) within the portal. The benefits of the Linked Open Data (LOD) technology in the context of a virtual library are introduced in Section 4. The main part of the paper (Sect. 5) addresses the ongoing efforts towards an LOD interface. Starting with the general approach towards the connection of the *Lin|gu|is|tik* portal with the Linguistic Linked Open Data (LLOD) cloud, we describe the conceptual and technical implementation with special focus on the linking of the BLL Thesaurus with the Ontologies of Linguistic Annotation (OLiA), and the development of a search algorithm and data storage solutions.

2. A Virtual Library for Linguistics

The *Lin|gu|is|tik* portal, freely accessible under `www.linguistik.de`, represents a virtual library with an integrated access to scientific information on every subject of linguistics, ranging from general and comparative linguistics to larger European languages through to small, threatened and ancient languages.

Funded by the DFG, the *Lin|gu|is|tik* portal is an ongoing cooperation between Goethe University Frankfurt, represented by the University Library and the Applied Computational Linguistics lab, the Institute of German Language (IDS Mannheim), and the LINSE Linguistik-Server of the University Duisburg-Essen with its link database LinseLinks. After the end of the second funding period (see below) the *Lin|gu|is|tik* portal will be maintained by the University Library Frankfurt.

The main resources and functionalities of the *Lin|gu|is|tik* portal were established during the first funding period (May 2012 – August 2014). In April 2013, the portal was launched in its first instantiation. Currently, it comprises the following modules:

Link directory: circa 9,000 scientifically relevant resources covering different fields of general linguistics and the linguistics of single languages, including websites with practical orientation (corpora, tools, educational material, etc.)

Journal directory: over 2,000 linguistic online journals taken from the language related subject areas of the Electronic Journal Library (EZB)[3]

Database directory: more than 500 linguistic databases with approximately 300 databases originating from the language related subject areas of the Datenbank-Infosystem (DBIS)[4]

Research directory: information about research projects and groups, collaborative research centres as well as research reports.

Dictionary directory: circa 1,400 online dictionaries including 700 links contributed by the Online Bibliography of Electronic Lexicography of the IDS Mannheim (OBELEXdict)[5]

Catalogues: an integrated search function for numerous sources including

- the above-mentioned directories of online resources (links, journals, databases, dictionaries and research);
- the catalogues of the University Library Frankfurt and IDS Mannheim as well as the Online Contents Linguistik (a database with bibliographic descriptions of more than 280,000 journal articles);
- diverse open access documents: the linguistic repositories of the Goethe University[6] and IDS Mannheim[7], selections of the Bielefeld Academic Search Engine BASE[8], and online dissertations provided by the German National Library;
- the Bibliography of Linguistic Literature (BLL) with its online version BLLDB[9];
- a selection of linguistically relevant publications from the Bibliographie der deutschen Sprach- und Literaturwissenschaft (BDSL)[10].

Conceptualised as a hub for scientific information, the *Lin|gu|is|tik* portal continues aggregating linguistically relevant resources as extensively as possible (more than 1.2 million entries at present). To make relevant resources more easily accessible and meet the heterogeneous requirements of the different addressees a detailed indexing according to subject and language and specific possibilities of access (e.g., "Links for beginners", "Research", "Corpora") are provided.

[1] officially *Lin|gu|is|tik – Portal für Sprachwissenschaft*

[2] `http:\\www.ub.uni-frankfurt.de\ssg\ling_en.html`

[3] `http://ezb.uni-regensburg.de/`

[4] `http://rzblx10.uni-regensburg.de/dbinfo`

[5] `http://www.owid.de/obelex/dict`

[6] `http://www.ub.uni-frankfurt.de/ssg/ling.html#dokumentenserver`

[7] `http://ids-pub.bsz-bw.de/home`

[8] `http://www.base-search.net/`

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

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

Within the *Lin|gu|is|tik* portal, there are no restrictions regarding the language under study as long as the resource is linguistically relevant. Currently, the portal covers more than 1,600 natural languages. We use a classification based mainly on *Ethnologue*[11] with three-letter codes from ISO 639-3 where available and three levels of presentation: language family, language group, and language.
The main goals of the second funding period (September 2015 - December 2016) are to integrate additional catalogues and databases, and to implement a LLOD interface. The portal will be extended with an LOD-based search facility to immediately retrieve LLOD resources. The connecting point between *Lin|gu|is|tik* and LLOD will be the BLL.

3. Bibliography of Linguistic Literature (BLL)

The Bibliography of Linguistic Literature (BLL) is one of the most comprehensive linguistic bibliographies worldwide. It covers general linguistics with all its neighbouring disciplines and subdomains as well as English, German and Romance linguistics.
Dating back as far as 1971, BLL lists over 453,000 references covering monographs, dissertations, articles from periodicals, collective works, contributions to conferences, unpublished research papers, etc., with an annual growth of about 10,000 references.
BLL can be compared with mainly two international bibliographies: the International Bibliography of the Modern Language Association (MLA)[12], and the Linguistic Bibliography Online (LBO)[13]. With 2.3 million references, MLA exceeds BLL in size, but only a fraction of this is concerned with linguistics: MLA also includes modern languages, literature and folklore. With 380,000 citations, LBO is smaller than BLL, and orientated mainly towards the coverage of lesser-known Indo-European and non-Indo-European languages. Thus, we consider the BLL unique in focus and scope[14].
Within *Lin|gu|is|tik*, BLL is of twofold importance: It represents a significant source of bibliographic data, and it provides a hierarchically categorised bilingual thesaurus of domain-specific index terms in English and German. The subject terms used for indexing online resources (Sect. 2) are based mainly on the BLL Thesaurus. Furthermore, the connection between the *Lin|gu|is|tik* portal and the LLOD cloud will be implemented by linking the Thesaurus to LLOD terminological repositories.
The following sections describe the ongoing efforts to position the BLL and the *Lin|gu|is|tik* portal in the wider context of LLOD and thereby to generate synergies with resources and bibliographies created and used in this context.

4. Linguistic Linked Open Data (LLOD)

Linguistic Linked Open Data is a movement about publishing open language resources for different use cases in academic research, applied linguistics or natural language processing. A linguistically relevant resource is considered a LLOD resource if it adheres to the following principles: (1) published under an **open** licence, (2) its elements are uniquely identifiable in the web of data by means of **URIs**, (3) its URIs should **resolve** via HTTP, (4) it can be accessed using **web standards** such as RDF and SPARQL, and (5) it should include **links** to other resources to help users discover new resources and provide explicit semantics.
From metadata collected under `http://datahub.io`, an LLOD diagram is generated and regularly published under `http://linguistic-lod.org`. Currently, it comprises 126 resources, including lexical-conceptual resources (dictionaries, knowledge bases), corpora, terminology repositories (thesauri, ontologies and registries for linguistic concepts, features, and terms), and metadata collections (language resource metadata, bibliographies). Since its first instantiation in September 2012, it has been rapidly growing and continues to do so because of 7 primary benefits as compared to legacy formalisms (Chiarcos et al., 2013):

Representation: Represent linguistic data flexibly as linked graphs
Structural Interoperability: Integrate data easily using RDF
Explicit Semantics: Define RDF resources by linking to term bases
Conceptual Interoperability: Use and re-use shared vocabularies
Federation: Combine data from multiple, distributed sources
Dynamicity: Access the most recent edition live over the web
Ecosystem: Benefit from widely available open source tools for RDF and linked data

Using shared vocabularies is particularly fruitful in the context of a virtual library: By linking the BLL Thesaurus to LLOD terminologies, BLL records immediately **become interoperable** with other LLOD resources such as the World Atlas of Language Structures[15], the Phonetics Information Base and Lexicon[16], or the Glottolog/LangDoc bibliography[17]. These links and the use of shared vocabularies allow us to automatically access and index LLOD language resources and thereby to develop a **(linked) language resource search** as part of the *Lin|gu|is|tik* portal.
Initially, we focus on morphosyntactic and syntactic concepts, categories and features, and for these, the Ontologies of Linguistic Annotations OLiA[18] (Chiarcos and Sukhreva, 2015) represent the central terminology hub in the LLOD cloud. Designed to leverage the linguistic terminology used in corpus annotation, and as collected in community-maintained terminology repositories, OLiA introduces a 'Reference Model' to mediate between resource- or language-specific 'Annotation Models' and several 'External Reference Models'. Annotation Model concepts are modelled as OWL classes, and the linking is represented

[11] `http://www.ethnologue.com/`
[12] `https://www.mla.org/bibliography`
[13] `http://www.brill.com/publications/online-resources/linguistic-bibliography-online`
[14] We exclude the Linguistics and Language Behaviour Abstracts (LLBA, `http://www.proquest.com/products-services/llba-set-c.html`) from this comparison as its maintainers do not provide any statistics about the number of references.
[15] `http://wals.info`
[16] `http://phoible.org`
[17] `http://glottolog.org/langdog`
[18] `http://purl.org/olia`

by means of `rdfs:subClassOf` properties that assign a given class a superclass from the OLiA Reference Model. OLiA Reference Model classes are linked with externally provided terminology repositories[19], and accordingly, any resource provided with an Annotation Model and linked with the Reference Model can also be interpreted in terms of these 'External Reference Models'.
Beyond morphology, syntax and discourse, links with other LLOD vocabularies will be more appropriate, e.g., `lexvo.org` and `glottolog.org` for language identifiers, `phoible.org` for phonological features, etc.

5. Connecting *Lin|gu|is|tik* and LLOD

The main goal of the second funding period is to enhance the functionality of the *Lin|gu|is|tik* portal with an LOD interface and make LLOD resources accessible to the users of the portal. In this section, we describe the steps towards a connection between the *Lin|gu|is|tik* portal and the LLOD cloud; we present our methodological approach, give conceptual and technical details, discuss challenges and propose solutions.

5.1. Remodelling the BLL Thesaurus in RDF

As a first step towards an interface with the LLOD cloud, the BLL Thesaurus is being remodelled as an ontology and linked with LLOD terminology repositories, e.g. the OLiA Reference Model.
At present, the BLL Thesaurus comprises 7,481 hierarchically organised index terms. 2,141 terms are available for the indexing of languages including dialects, reconstructed or artificial languages. The main branches *Levels*[20] (including the levels of language description, e.g., *Syntax*, *Phonology*) and *Domains* (covering the subdisciplines of linguistics, e.g., *Psycholinguistics*, *Sociolinguistics* *Pragmalinguistics*) consist of 1,983 and 3,050 subject terms respectively.
The Thesaurus evolves over time through continuous accommodation to the ongoing development in the field of linguistics. This happens mainly by inclusion of new subject terms: In 2014, for example, the total number of new terms was 235 including *Argument sharing*, *Parasitic participle* and *Whispered interpreting*. Deletions happen extremely seldom, but are not completely ruled out. In such cases, related subject terms are merged into a new category. For example, in 2014 the subject terms *Geography (technical language)* and *Geodesy (technical language)* were combined to form the new subject term *Earth sciences (technical language)*. The fact that the internal representation of the BLL subject terms is based on unique, stable IDs favours an ontological remodelling. In case of subject term deletion, the respective IDs are blocked and cannot be reused.
Due to the specificity of the Thesaurus, our approach differs from the general methodology for ontology building as introduced by Farrar (2007) and Farrar and Langendoen (2010). The BLL Thesaurus provides a list of domain specific subject terms and presents them in a hierarchical tree structure, but the existing hierarchical relations only partially fulfill the criteria of an ontology.
The structure of the BLL Thesaurus, internally represented in OCLC PICA[21], has semantics based on lexical associations rather than the object-oriented model underlying OWL and the `rdfs:subClassOf` property. Figure 2 shows the subject term *Adjective* with its BLL parent, siblings and subcategories: While its subcategories can indeed be regarded as ontological subclasses, the interpretation of *Adjective syntax* and the relation of *Adjective* and its sibling concepts is problematic.

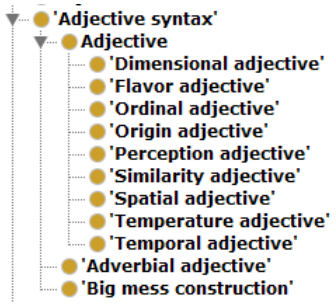

Figure 2: Hierarchical structure of the BLL Thesaurus.

Because of the nature of the hierarchical relations, the outcome of a "naïve" automated conversion to OWL wherein the hierarchy is represented by `rdfs:subClassOf` relations will not be a valid ontological model. Instead, the BLL hierarchy is expressed by less rigidly defined `skos:broader` relations as recommended for modelling thesauri in Pastor et al. (2009). The automatically created SKOS file is then imported into an OWL editor[22] and all BLL concepts are manually classified and organised to build the actual BLL Ontology. Since the URIs are based on stable IDs, future conversions of the Thesaurus produce identical concept URIs, and previously classified concepts maintain their ontological rendering. Newly added subject terms need an ontological rendering, i.e. manual classification. In case of subject term deletion, the ID of the deleted term is blocked permanently, and any references to it in the BLL Ontology are marked `deprecated`.
An experimental ontological model for the BLL branches *Syntax* and *Morphology* and its linking with the OLiA Reference Model is currently under development. Thus, BLL concepts become interoperable with OLiA, GOLD, ISOcat, TDS, etc.
The automated conversion, the manual remodelling process and the linking of the BLL Thesaurus results in a three-layer RDF-model:

- The BLL Thesaurus' internal hierarchy in an automatically generated `bll.skos` file. Current triple count: 55,048 (18.02.2016).

[19] E.g., GOLD (`http://linguistics-ontology.org`) (Farrar and Langendoen, 2010), ISOcat (`http://isocat.org`) (Kemps-Snijders et al., 2009), TDS (`http://languagelink.let.uu.nl/tds/`)
[20] Thesaurus subject terms are represented in italics.
[21] `http://www.oclc.org`
[22] We employ Protégé 5.0 for remodelling the Thesaurus as ontology.

```
:133075826 a owl:Class ;
  rdfs:subClassOf :BLLConcept ;
  skos:broader :133073629 ;
  rdfs:label
      "Adjective"@en
    , "Adjektiv"@de ;
  skos:prefLabel
      "Adjective (lex.)"@en
    , "Adjektiv (lex.)"@de ;
  skos:altLabel "lex."@en ;
  skos:altLabel "lex."@de .
```

The labels of the BLL subject terms can consist of two parts: main name and addition in brackets that specifies the context of usage or the perspective of analysis and helps to avoid homonyms (e.g., *Adjective (lex.)*). For better machine interpretation both parts are included separately (`rdfs:label` and `skos:altLabel` resp.). The combination of the parts is represented by `skos:prefLabel` in order to secure better readibility in ontology editors.

- The BLL Ontology with its manually remodelled class hierarchy. Currently containing 1,328 triples reorganising 775 BLL index terms (work in progress 18.02.2016).

```
:133075818 rdf:type owl:Class ;
  owl:equivalentClass
      :133075826
    , :133075850 ;
  rdfs:subClassOf
      :MorphosyntacticCategory .
```

- The BLL Linking Models (currently a Linking Model to the OLiA Reference Model using the `rdfs:subClassOf` property is under construction)

Thus, we preserve the original BLL structure (`skos:broader`), and its ontological model (BLL Ontology), and clearly separate both from their interpretation in terms of OLiA (etc.).

The remodelling of the BLL Thesaurus starts with the levels of linguistic description and more exactly with the branches *Syntax* and *Morphology*, consisting of 289 and 191 subject terms respectively.

The establishment of a basic class structure and top-level concepts happens by means of grouping the BLL subject terms around the notions linguistic **category**, linguistic **feature**, linguistic **process**, and linguistic **relation**. The entities that can be clearly defined as one of those are categorical by nature and constitute an ontological class. So, *Verb* and *Adverb* are defined as morphosyntactic categories, *Case* and *Tense* as morphosyntactic features and *Word formation* and *Inflection* as examples of a morphological process.

Since many cases of subordination within the BLL Thesaurus cannot be regarded as ontological subclass relations, a complete adoption of the BLL tree structure is not an option. We start with a verification of the definitions and an in-depth examination of the existing hierarchical relations, which often leads to reorganisation of taxonomies and building of new ones.

The reorganisation of the taxonomic structure is facilitated by the addition of ontological classes without a corresponding BLL subject term, e.g., `MorphologicalCategory`, `MorphosyntacticFeature`. A partial preservation of the existing hierarchical relations is possible in many cases (see Figure 2, subclasses of *Adjective*). Most of the subcategories of the BLL *Word Formation*, e.g., *Apheresis*, *Contamination*, and *Derivation*, also fulfil the requirements for an ontological subclass.

Generally, the requirements for a consistent ontological structure can be met by a name change or a change in the hierarchical position. So, the BLL subject term *Syntax* is renamed `SyntacticTerm`, and *Embedding*, a BLL subcategory of *Subordinate clause*, becomes a subclass of the newly created `SyntacticProces`. The disambiguation of some subject terms, however, requires different formal and conceptual solutions. A few ambiguous BLL subject terms (e.g., *Compounding*) denote a linguistic process as well as the result of that process. Others refer to an opposition (e.g., *Mass noun/count noun*) that has to be resolved.

Compounding is defined as a subclass of `AmbiguouslyDefinedConcept` and also set equivalent to the disjunction of the newly introduced classes `Composition` (a subclass of `MorphologicalProcess`) and `Compound` (a subclass of `Morpheme`). For *Mass noun/count noun* a similar approach is followed, and it is equated by an *EquivalentClasses* axiom to the disjunction of the newly created classes `MassNoun` and `CountNoun`. To capture the inherent nature of the opposition, `MassNoun` and `CountNoun` are disjointed by a *DisjointClasses* axiom.

The Thesaurus' nature inhibits general solutions for the challenging cases: We treat them individually by means of scrutinising the indexed bibliographic entries and choose to stay close to the primary BLL meaning in case of doubt.

5.2. Crawling the LLOD Cloud

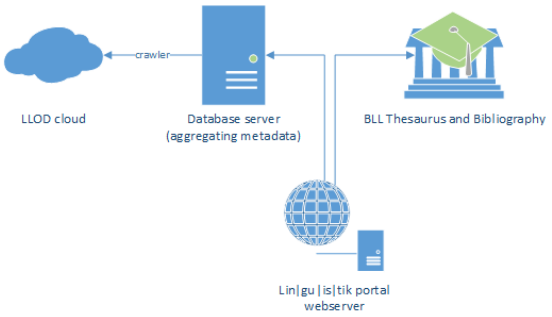

Figure 3: *Lin|gu|is|tik* LLOD interface architecture.

Because of the vast and complex nature of the LLOD cloud it will not be possible to search through it directly "on the fly". Instead, a means of indexing the information available

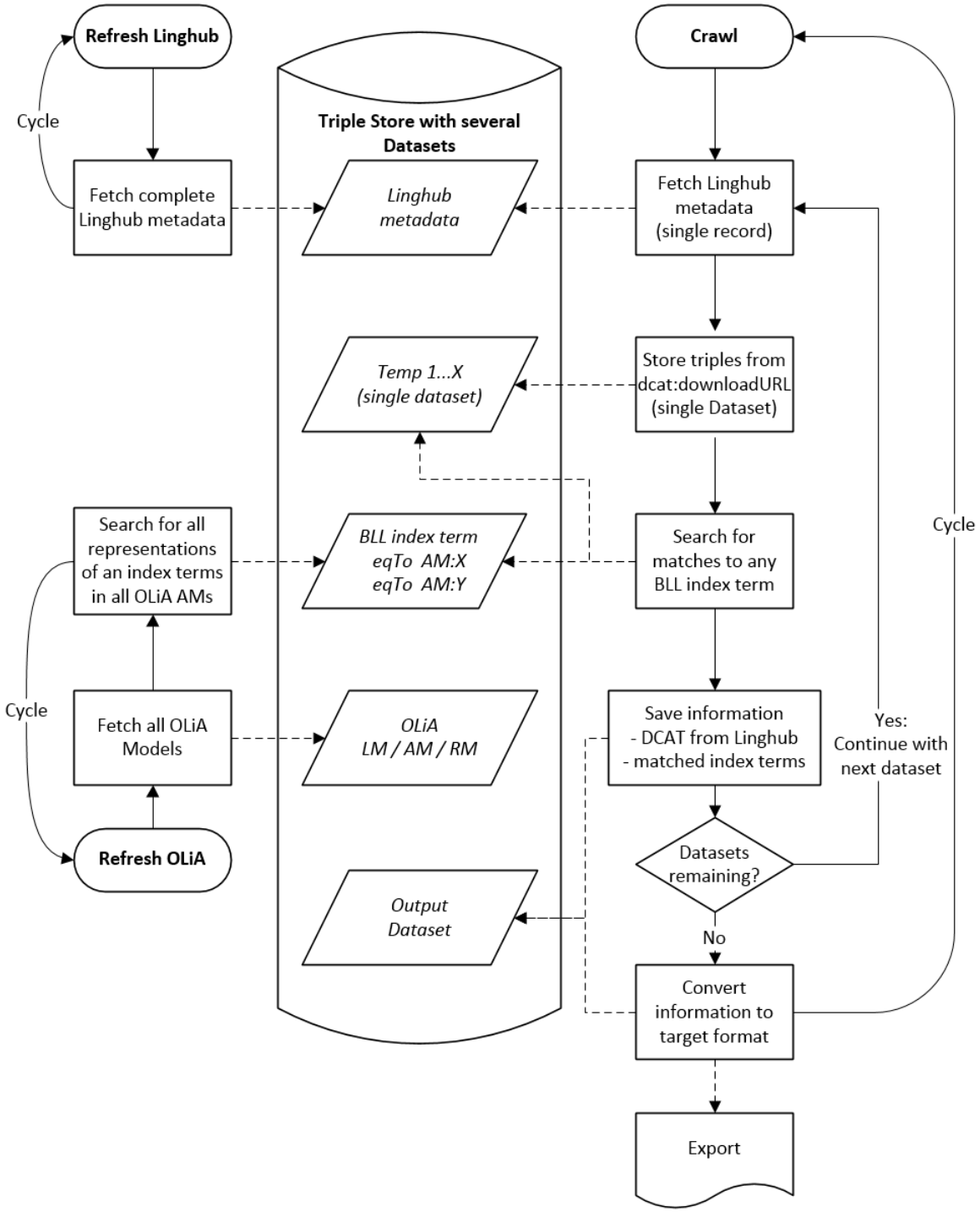

Figure 4: Concept of the LLOD crawler

through Linghub[23] has to be implemented (Figure 3).
This *LLOD crawler* is currently being developed by the Applied Computational Linguistics lab and will be maintained by the University Library in the future. It is situated on a separate Database server VM. Its architecture comprises four interdependent components all using a central triple store for gathering information (Figure 4):

- The complete set of OLiA Linking and Annotation Models alongside the BLL Ontolgy are cyclically cached. Then, a set of SPARQL queries uses the Linking Models to find any possible equivalent concept of all BLL index terms in all other OLiA Annotation Models. The resulting equivalencies are stored in a search cache dataset.
- The linghub RDF dump[24] is also cached within the same cycle.

[23] `http://linghub.org/`

[24] `http://linghub.lider-project.eu/linghub.nt.gz` (Current triple count 5,918,686 containing 196,307 Datasets and 229,586 `dcat:accessURL` as of 18.02.2016)

- The actual crawler searches then through the linghub cache for links to available ressources using the `dcat:accessURL` and `dcat:downloadURL` respectively. Available ressources are scanned for matching subjects in the aforementioned search cache dataset. In order to reduce traffic, the algorithm takes change dates and checksums into consideration. In a separate output dataset, matching BLL index terms are stored alongside the resources' metadata.
- The output dataset is then used for exporting the crawling results to the *Lin|gu|is|tik* portal. There, the resources are indexed internally in order to provide a fast and reliable search engine for end users.

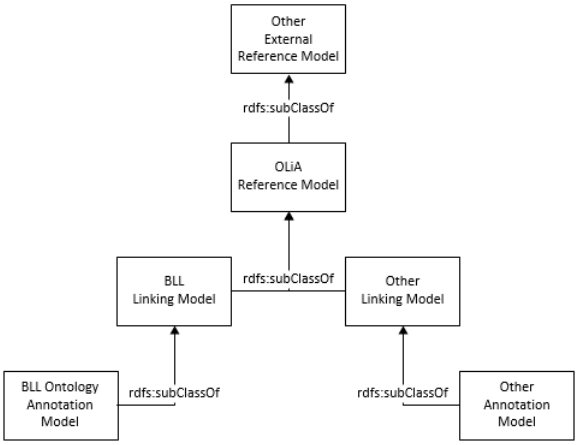

Figure 5: OLiA architecture

The BLL index term search algorithm uses all available OLiA models in order to find equivalent terms or subcategories in other ontologies. It is necessary, however, to distinguish between regular Annotation Models (e.g. MULTEXT East[25]) and External Reference Models such as ISOcat / GOLD. While the linking of the former is represented by `lm:term rdfs:subClassOf olia:term`, the latter work in the opposite direction establishing OLiA terms as subclasses of the External Reference Model terms (Figure 5).
Therefore, a set of SPARQL queries is created in order to find possible representations of OLiA terms in linghub resources. Namespace discrimination is used to differentiate between Annotation Models and External Reference Models. The following sample query shows in a simplified way how to get all qualified subcategories of `olia:Adjective` in the ISOCat Ontology. In order to maintain better readability, long prefixes and base URLs are omitted:

```
select DISTINCT ?oliarm1 ?isorm2
{
# find all Morphosyntactic Categories
# in OLiA reference model (RM)
?oliarm1 rdfs:subClassOf*
  olia:Adjective .

# find all superclasses of RM term
# in any external reference model (ERM)
?oliarm1 rdfs:subClassOf ?isorm1 .

# find all subclasses of ERM term
# within ERM.
?isorm2 rdfs:subClassOf* ?isorm1 .

FILTER regex(str(?oliarm1)
  , "((http://purl.org/olia).*)"
  , "i") .
FILTER regex(str(?isorm1)
  , "((.*dcr.owl#).*)"
  , "i") .
FILTER regex(str(?isorm2)
  , "((.*dcr.owl#).*)"
  , "i") .
}
```

?oliarm1	?isorm2
Http://purl.org/olia/olia.owl#Attributive Adjective	Http://www.isocat.org/ns/dcr.owl#attributive Adjective
Http://purl.org/olia/olia.owl#Present Participle Adjective	Http://www.isocat.org/ns/dcr.owl#present Participle Adjective
Http://purl.org/olia/olia.owl#Possessive Adjective	Http://www.isocat.org/ns/dcr.owl#possessive Adjective
Http://purl.org/olia/olia.owl#Substantive Adjective	Http://www.isocat.org/ns/dcr.owl#substantive Adjective
Http://purl.org/olia/olia.owl#Qualifier Adjective	Http://www.isocat.org/ns/dcr.owl#qualifier Adjective
Http://purl.org/olia/olia.owl#Ordinal Adjective	Http://www.isocat.org/ns/dcr.owl#ordinal Adjective
Http://purl.org/olia/olia.owl#Past Participle Adjective	Http://www.isocat.org/ns/dcr.owl#past Participle Adjective
Http://purl.org/olia/olia.owl#Participle Adjective	Http://www.isocat.org/ns/dcr.owl#participle Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#adjective
Http://purl.org/olia/olia.owl#Participle Adjective	Http://www.isocat.org/ns/dcr.owl#present Participle Adjective
Http://purl.org/olia/olia.owl#Participle Adjective	Http://www.isocat.org/ns/dcr.owl#past Participle Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#possessive Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#participle Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#present Participle Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#qualifier Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#ordinal Adjective
Http://purl.org/olia/olia.owl#Adjective	Http://www.isocat.org/ns/dcr.owl#past Participle Adjective

Figure 6: Qualified equivalencies and subconcepts of `olia:Adjective` within the ISOCat Ontology

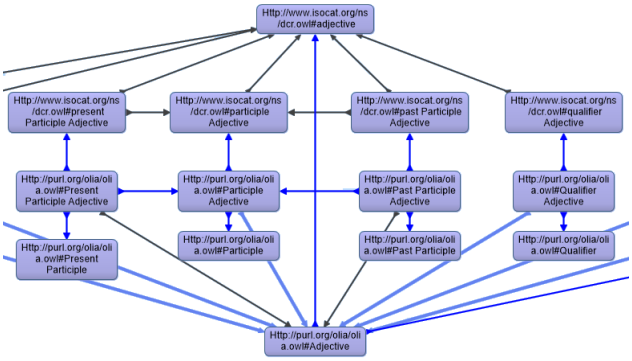

Figure 7: Graphic representation of the results of the query (Figure 6) with `subClassOf` relations

Figures 6 and 7 show the results of the query executed on Allegrograph using the graphical tool **gruff**[26]. The dataset contains the OLiA Reference Model alongside the current experimental build of the ISOcat External Reference Model[27] (Chiarcos, 2010).

[25] `http://nl.ijs.si/ME/`

[26] `http://franz.com/agraph/gruff/`

[27] Experimental OLiA Builds are available open source

Extrapolating the sample query on other Linking Models while using the BLL Linking Model to find fitting BLL index terms first before going to OLiA Reference Level would result in a dataset of the following form:

```
bll:Adjective
  owl:equivalentClass olia:Adjective ;
  owl:equivalentClass dcr:adjective ;
### ...
.
```

The so found equivalencies can then be used to map the BLL index terms to linghub resources by the algorithm depicted on the right-hand side of Figure 4.

5.3. Implementation Considerations

We employ Linghub (McCrae and Cimiano, 2015) as a starting point for retrieving LLOD data as it provides a uniform way to access LLOD resources. The Linghub portal stores metadata about roughly 250,000 linguistic resources. Metadata is modelled using DCAT, Dublin Core and META-SHARE standards (McCrae et al., 2015).

The portal allows for browsing its online catalogue, and also supports SPARQL queries on the site as well as a service. As such linghub metadata is also available as an RDF dump which we will exploit instead of using its online SPARQL service. The main reason being that the service is limited because it supports only a small fragment of the SPARQL standard (YuzuQL). Furthermore, querying over a network would lead to increased query times and overloading of the linghub service.

In order to identify resources relevant for the *Lin|gu|is|tik* portal, the metadata can be queried for up to 400 properties. Information about a resource not included on Linghub may be extracted directly from the data of the resource. In the latter case, the data of the resource has to be downloaded first and then searched for metadata. Linguistic resources listed on Linghub have different data file formats. For us, the following types are most relevant:

- RDF data as text (e.g., turtle, n3, rdf-xml, etc.)
- RDF file containing links to other datasets
- SPARQL-endpoint URL that allows for SPARQL queries

Since LLOD resources are designed by Linked Open Data principles, the data of a resource is not always included in a single file: It might also be distributed (linked) and located in different places. As a simple example consider the RDF version of the Brown-Corpus[28] [29] which is distributed as many small files. So, gathering data is the normal use case. For this purpose we utilise the *LDspider* LOD data crawler [30](Isele et al., 2010) which is freely available as a Java library.

In order to efficiently query RDF data of a resource (e.g., corpus, dictionary) for BLL index terms (see Figure 4), the resource has to be stored temporally in a local RDF store. We currently evaluate different data base solutions. Candidate RDF stores include Jena TDB[31], Blazegraph[32], Allegrograph[33], Openlink Virtuoso[34] and RDF-HDT[35]. With performance results becoming available we will publish our experiences in upcoming publications.

6. Summary and Outlook

The *Lin|gu|is|tik* portal is a hub for linguistically relevant scientific information. It provides catalogues for linguistic literature, online resources and open access documents as well as the BLL, one of the most comprehensive international bibliographies in linguistics. We continue aggregating linguistically relevant information as extensively as possible and enhancing the functionalities and the target-oriented offers.

Beyond this, connecting the *Lin|gu|is|tik* portal with the LLOD cloud will facilitate the accessibility and visibility of open data language resources to current users of the portal, resulting in mutual benefits for both platforms. On the one hand, using LLOD vocabularies and term bases, the *Lin|gu|is|tik* portal will gain access to an ever-growing pool of linguistic resources on the web. On the other hand, the LLOD cloud will not only benefit from a new, significant source of linguistically relevant data, but will also become accessible on a modern platform which is targeting optimised usability for less technically oriented linguists. The LOD search will be integrated in the Catalogues module so that no additional technological expertise will be required to use it.

Finally, the extension of the portal with an LOD interface facilitates the prospective integration of the *Lin|gu|is|tik* portal with other sources of bibliographical information available as RDF, such as the German National Library[36], WorldCat[37], OpenLibrary[38], etc.

The SKOS export of the BLL Thesaurus is already available as a dataset. Its current edition covers 5,340 subject terms, 2,141 language identifiers, and consists of a total of 55K SKOS triples. The BLL Ontology built on top of it and its OLiA linking are currently under development. We are in the process of clarifying details of a persistent hosting service and plan to publish the linked BLL Ontology under a Creative Commons licence in mid-2016, both for practical use in the *Lin|gu|is|tik* portal and inclusion in the LLOD cloud.

on SourceForge https://svn.code.sf.net/p/olia/code/trunk/owl/experimental

[28] http://linghub.org/datahub/brown-corpus-in-rdf-nif

[29] http://brown.nlp2rdf.org/lod/

[30] http://www.aifb.kit.edu/web/LDspider

[31] http://jena.apache.org/documentation/tdb/index.html

[32] https://www.blazegraph.com/

[33] http://franz.com/agraph/allegrograph/

[34] http://virtuoso.openlinksw.com/

[35] http://www.rdfhdt.org/

[36] http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

[37] https://www.oclc.org/developer/develop/web-services/worldcat-registry/rdf-interface.en.html

[38] https://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Open_Library_Data

7. References

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

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards Open Data for Linguistic: Linguistic Linked Data. In A. Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources. Theory of Applications of Natural Language Processing*, pages 7–25. Springer, Heidelberg.

Chiarcos, C. (2010). Grounding an Ontology of Linguistic Annotations in the Data Category Registry. In *LREC 2010 Workshop on Language Resource and Language Technology Standards (LR&LTS)*, pages 37–40, Valetta, Malta.

Deutsche Forschungsgemeinschaft. (2015). Richtlinien zur überregionalen Literaturversorung der Sondersammelgebiete und Virtuellen Fachbibliotheken. Bonn: DFG. Version 02/15. Web. 23.10.2015. <http://www.dfg.de/foerderung/programme/infrastruktur/lis/veroeffentlichungen/index.html>.

Farrar, S. and Langendoen, D. T. (2010). An OWL-DL Implementation of Gold. An Ontology for the Semantic Web. In A. Witt et al., editors, *Linguistic Modelling of Information and Markup Languages: Contributions to Language Technology*, pages 45–66. Springer, Dordrecht.

Farrar, S. (2007). Using 'Ontolinguistics' for language description. In A. C. Schalley et al., editors, *Ontolinguistics. How Ontological Status Shapes the Linguistics Coding of Concepts*, pages 175–191. Mouton de Gruyter, Berlin + New York.

Isele, R., Umbrich, J., Bizer, C., and Harth, A. (2010). LDSpider: An open-source crawling framework for the Web of Linked Data. In *Proceedings of 9th International Semantic Web Conference (ISWC 2010) Posters and Demos*, Shanghai, China.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., and Writh, S. E. (2009). ISOcat: remodelling metadata for language resources. *International Journal of Metdata, Semantics and Ontologies*, 4(4):261–276.

McCrae, J. P. and Cimiano, P. (2015). Linghub: a Linked Data based portal supporting the discovery of language resources. In *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS 2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15)*, pages 88–91, Vienna, Austria.

McCrae, J. P., Cimiano, P., Rodriguez-Doncel, V., Vila Suero, D., Gracia, J., Matteis, L., Navigli, R., Abele, A., Vulcu, G., and Buitelaar, P. (2015). Reconciling Heterogeneous Descriptions of Language Resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)*, pages 39–48, Beijing, China.

Pastor, J. A., Martinez, F. J., and Rodriguez, J. V. (2009). Advantages of thesaurus representation using Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, 14(4):Paper 422.

Renner-Westermann, H. (2013). Lin|gu|is|tik - Portal für Sprachwissenschaft. Webis. Aktuelles über Sammelschwerpunkte an deutschen Bibliotheken. Web. 7.3.2016. <http://blogs.sub.uni-hamburg.de/webis/2013/08/01/linguistik-portal-fuer-sprachwissenschaft/>.