Metalexicography as Knowledge Graph

David Lindemann
Universität Hildesheim, Germany
david.lindemann@uni-hildesheim.de

Christiane Klaes
Universität Hildesheim, Germany
Georg Eckert Institute for International Textbook Research, Braunschweig, Germany
klaesc@uni-hildesheim.de

Philipp Zumstein
Mannheim University Library, University of Mannheim, Germany
philipp.zumstein@bib.uni-mannheim.de

Abstract

This short paper presents preliminary considerations regarding LexBib, a corpus, bibliography, and domain ontology of Lexicography and Dictionary Research, which is currently being developed at University of Hildesheim. The LexBib project is intended to provide a bibliographic metadata collection made available through an online reference platform. The corresponding full texts are processed with text mining methods for the generation of additional metadata, such as term candidates, topic models, and citations. All LexBib content is represented and also publicly accessible as RDF Linked Open Data. We discuss a data model that includes metadata for publication details and for the text mining results, and that considers relevant standards for an integration into the LOD cloud.

2012 ACM Subject Classification Information systems → Resource Description Framework (RDF); Information systems → Document representation; Information systems → Ontologies; Information systems → Information extraction; Information systems → Web Ontology Language (OWL)

Keywords and phrases Bibliography, Metalexicography, Full Text Collection, E-science Corpus, Text Mining, RDF Data Model

Digital Object Identifier 10.4230/OASIcs.LDK.2019.19

Category Short Paper

Supplement Material `http://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/`

1 Introduction

Our goal is an online bibliography of Lexicography and Dictionary Research (i. e. metalexicography) that offers hand-validated publication metadata as needed for citations, that represents, if possible, metadata using unambiguous identifiers and that, in addition, is complemented with the output of a Natural Language Processing toolchain applied to the full texts. Items are tagged using nodes of a domain ontology developed in the project; terms extracted from the full texts serve as suggestions for a mapping to the domain ontology. Main considerations regarding the project have been presented in [7].

In this publication, we focus on the data model for LexBib items, its integration into the LOD cloud, and on relevant details of our workflow. In Section 2 we describe how publication metadata and full texts are collected and stored using Zotero, data enrichment and transfer to RDF format. Section 3 addresses the text mining toolchain used for the generation of additional metadata, that are linked to the corresponding bibliographical items. As shown in Fig. 1, an OWL-RDF file is the place where this merging is carried out. In Section 4 we describe the multilingual domain ontology that will be used to describe the full text content with keywords or tags.

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 19; pp. 19:1–19:8

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

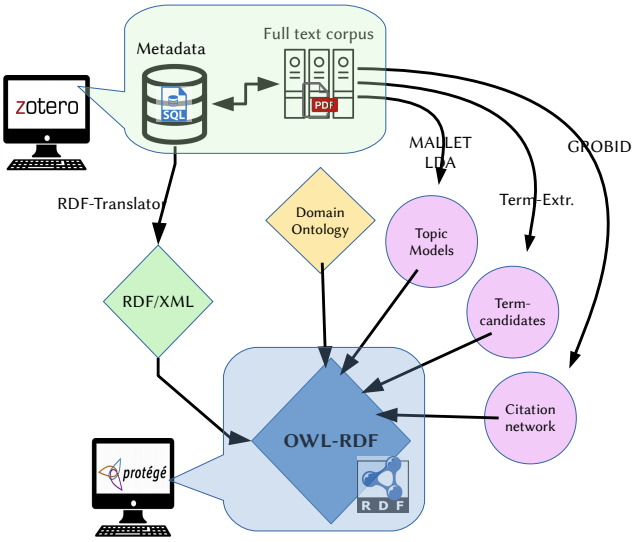

Figure 1 Workflow for combining publication metadata and additional metadata in LexBib.

2 Data Enrichment: Publication Metadata

For the task of web scraping and manually validating publication metadata, and for storing the corresponding full texts, the Zotero software application[1] offers state-of-the-art functions, such as one-click data ingestion from structured metadata as well as from general websites, keyword indexing, attaching of files, notes, and links, and duplicate detection.

In our workflow, a predefined minimal metadata set is collected and hand-validated for every publication, including author(s), title, publishing year, name of the publication (e. g. the journal or the container volume), publication place, etc. This includes the metadata needed for citation, plus (1) the item type, and (2) a unique identifier (DOI, ISBN) for the publication, where available, and/or (3) a URL that leads to the item's landing page on the publishers' web platform, where the original full text can be accessed. The former are all stored as literal strings or integers, the latter two, i.e. DOI/ISBN, and the link to the full text (URL), are stored as Uniform Resource Identifiers (URI).

Zotero stores the metadata in a SQLite relational database, which then can be used for making citations and lists of references (main purpose of a reference management system) but also to export it in various formats. This second point in combination with the easy data extraction from various websites and platforms makes Zotero also interesting as a cataloguing tool for different purposes. For example, Zotero is used as cataloguing and automatic data ingesting tool in the project IndexTheologicus [5]. Moreover, as another example, the one-click option to add a reference in the graphical interface of Wikipedia is supported by Zotero.[2]

[1] See `http://zotero.org`.
[2] See `https://www.mediawiki.org/wiki/Citoid/Zotero%27s_Tech_Talk`

```
<rdf:RDF xmlns:...>
    <bibo:AcademicArticle rdf:about="https://academic.oup.com/ijl/article/25/4/398/923874">
        <bibo:pages>398-436</bibo:pages>
        <bibo:doi>10.1093/ijl/ecs026</bibo:doi>
        <dcterms:language>en</dcterms:language>
        <dcterms:abstract>Corpus-driven lexicography and the International Journal of Lexicography (IJL) made
  their first appearance in the world in 1987 and 1988 respectively. ... </dcterms:abstract>
        <dcterms:title>The Corpus Revolution in Lexicography</dcterms:title>
        <dcterms:creator rdf:nodeID="n5"/>
        <bibo:authorList>
           <rdf:Seq><rdf:li rdf:nodeID="n5"/></rdf:Seq>
        </bibo:authorList>
        <dcterms:isPartOf>
            <bibo:Issue>
                <bibo:volume>25</bibo:volume>
                <bibo:issue>4</bibo:issue>
                <dcterms:date>2012-12-01</dcterms:date>
                <dcterms:isPartOf>
                    <bibo:Journal>
                        <dcterms:title>International Journal of Lexicography</dcterms:title>
                        <bibo:issn>0950-3846</bibo:issn>
                        <bibo:shortTitle>Int J Lexicography</bibo:shortTitle>
                    </bibo:Journal>
                </dcterms:isPartOf>
            </bibo:Issue>
        </dcterms:isPartOf>
    </bibo:AcademicArticle>
    <foaf:Person rdf:nodeID="n5">
       <foaf:givenname>Patrick</foaf:givenname>
       <foaf:surname>Hanks</foaf:surname>
    </foaf:Person>
</rdf:RDF>
```

Figure 2 Sample metadata set exported from Zotero as Bibliontology RDF/XML.

In the LexBib project we use Zotero's Bibliontology RDF/XML translator[3] for exporting to linked data. The name bibliontology comes from the Bibliographic Ontology (BIBO) which is used as the main vocabulary in this translator. Besides BIBO the vocabularies Dublin Core (dcterms), Friend of a Friend (FOAF), and MARC Code List for Relators are mainly used. Some less used item types like software, blog post or audio/video recording are exported by also using vocabularies like DOAP Ontology, Programmes Ontology, SIOC Types Ontology. Finally, there is a special minted namespace within zotero.org which is used for everything which is then still unmapped. The current implementation will be checked and possibly improved by considering also the recently published recommendations for RDF-representation of bibliographic data by the Competence Centre on Interoperable Metadata (short KIM in German) [3].

Publication creators (`dcterms:creator`) in Zotero correspond to the roles performed by persons or organisations, i.e. author, editor, series editor, contributor, translator, and, for reviews, reviewed author. While in Zotero, as for version 5.0, no data field is foreseen for the annotation of persons with identifiers such as ORCID or VIAF IDs, such mapping could nevertheless be done using FOAF element values found in the RDF/XML dump. For this task, in LexBib we propose a collaborative approach: The project team will ensure to find literals that refer to the same person and merge them (e.g. Patrick Hanks and P. Hanks). For each author, a profile page will be created. On dissemination events organized at central conferences of our discipline, and using communication channels used by the community, authors will be asked to attach an ORCID, VIAF or GND identifier to their profile page, along with other useful information, like personal homepages, etc. The advantage of that approach is two-fold: On the one hand, mismatches are avoided, and on the other hand, each person decides whether she wants to display an identifier next to her LexBib records that will link these to any other resource linked to the same identifier.

[3] See `https://github.com/zotero/translators/blob/master/Bibliontology%20RDF.js`.

Publication places are stored as literal strings in the Zotero database, and represented by the Bibliontology translator as `dcterms:publisher` / `address:localityName`. The `localityName` literals can be linked to instances of the GeoNames database, using the GeoNames API and related libraries.[4] [5]

Language names appear in the Zotero publication metadata in the "language" field, which refers to the language a publication is written in, and it is translated to `dcterms:language`. We propose to map all language names to instances of the LEXVO ontology,[6] a resource that contains languages and related information, such as the territory a language is spoken in, alternative names of the language, links to resources like ethnologue, etc. We will repeat the same process for the languages a publication is about (see Section 4). In LexBib, both language of publication and object language are relevant variables in the retrieval of bibliographic items, as filter options. At the same time, the LEXVO integration allows the language names to be displayed according to the users' preferred localisation, and, for example, a retrieval of items that refer to languages spoken in a given country.

3 Additional Metadata

In the LexBib project, computational methods are applied for obtaining term candidates, topic models, and citation references. The results shall be added to the items as additional metadata. Topic weights will be used for ranking bibliographic items with similar full text content. In the following, we explain our approach for generation and modeling of term candidates and citation relations.

3.1 Term Extraction

For term extraction, we use a variant of a tool suite developed at IMS, University of Stuttgart [11, 10], henceforth called "TrEx". It extracts the instances of part-of-speech patterns, e. g. (1) NN (single common nouns), (2) NN-NN (two common nouns), or (3) NN-NN-NN (three adjacent common nouns). Then, it ranks the extracted instances according to their termhood or keyness which is measured by dividing the relative frequency of the instance in a document by the relative frequency of the instance in a reference corpus (weirdness ratio, cf. [1]). We run this method twice for each document; for English,[7] once with the British National Corpus (BNC) as a reference corpus in general language in order to retrieve domain specific terms; and once with the whole LexBib English corpus as a reference corpus in order to identify document specific keywords. An example of term candidates extracted by this approach is shown in [7].

Term candidates will be stored in the LexBib database, linked to the corresponding item. Besides enhancing consistent subject indexing and retrieval, term candidates will be used for a mapping to instances of the LexBib domain ontology (see Section 4). Since we plan to display both term candidates and ontology concepts as metadata for LexBib items,

[4] Accessible at `https://www.geonames.org/`. Libraries for accessing GeoNames API are available at `https://www.geonames.org/export/client-libraries.html`.

[5] A similar mapping would be possible for author affiliation strings, that also can be mapped to places. Affiliations are not part of the standard publication metadata and have to be extracted from the full text, which is a non-trivial task of information extraction, that could be addressed using GROBID (see Section 3.3.); this, however, is not part of the workflow proposed here.

[6] Accessible at `http://www.lexvo.org/`.

[7] Our NLP toolchain in this preliminary stage is set up for English; other languages, starting with German and Spanish, will be considered during the lifetime of the project.

we need our RDF data model to distinguish between those two types of subject headings. Furthermore, by providing provenance metadata we will state, what agent (person, algorithm) generated a content descriptor according to what method (computational toolchain with a certain configuration, set of guidelines for manual validation). For term candidates extracted with TrEx, relevant metadata categories, including provencance, are listed in Table 1. The starting point for our RDF modeling is a proposal presented by German National Library (DNB),[8] that uses the W3C's PROV Data Model and PROV Ontology.[9]

Table 1 Points for provenance data for TrEx iterations and single term candidates.

TrEx run	Term candidate
Source corpus description	TrEx run
Reference corpus description	Weirdness ratio
Retrieved part-of-speech patterns	Term status (manual evaluation)
Weirdness and rank thresholds	Mapping to ontology concept
Timestamp	

3.2 Citation Network

Scientific publications usually contain a reference section at the end. The LOC-DB project [6] developed a software application,[10] that wraps all of the following steps in a single GUI: (1) Optical Character Recognition of the full text item for scanned print publications, (2) the information extraction tools GROBID[11] and ParsCit[12], (3) scripts for queries to external publication metadata collections, and (4) a module for defining and storing citation relations. In the LexBib project we will use this Open Source Software for our text corpus and adapt the steps for our needs.

The GROBID tool works on a plain text version of the PDF full text content (or, if this is not available, on the output of the OCR engine) and isolates the block of bibliographic references, the entries of which are then parsed and converted into a structured format compliant to the TEI guidelines (element `<listBibl>`). GROBID uses Conditional Random Fields (CRF), a supervised machine learning method which learns a model based on annotated training data [9]. Problematic citation styles, i. e. formats that are not properly parsed by the tool, will require further annotated training data. Metadata extracted by GROBID are compared to items found in the LexBib collection,[13] or, if not found, sent to an API of external resources containing OpenCitations, Crossref, and library catalogues such as WorldCat, in order to obtain mapping candidates. Then, one (or several) candidate(s) can be manually chosen and thereby connect the LexBib item to an already online existing item. On the one hand, this mapping is used for updating the `<listBibl>` from the metadata in citation style independent format found in the external source, and for enriching it with URI, as done in LOC-DB project. In addition, we plan to use that output for GROBID's CRF training, and also for updates of the citation relations available at the OpenCitations

[8] See `https://wiki.dnb.de/pages/viewpage.action?pageId=146383331`.
[9] See `https://www.w3.org/TR/prov-dm/` and `https://www.w3.org/TR/prov-o/`.
[10] See `https://github.com/locdb`.
[11] See `https://github.com/kermitt2/grobid`.
[12] See `https://github.com/knmnyn/ParsCit`.
[13] Preliminary experiments related to that are explained in [8].

database.[14] We aim at implementing these features during the duration of the LexBib project.

Based on the extracted references, a citation network is visualised and publication clusters can be identified based on citation relations, as it has been proposed in related work (e.g. [4]). The item relations obtained from the analysis of the reference sections in the full texts include (i) the publications cited in a publication, (ii) the publications citing a publication, and (iii) the membership of a publication in a cluster in a citation network.

4 Domain Ontology

The term "Lexicography" is present in controlled vocabularies used for text content description, such as the Library of Congress Subject Headings (LCSH) or Gemeinsame Normdatei (GND). In these general (i.e., not domain-specific) ontologies, but also in a domain-specific keyword collection, such as the one used for indexing publications at BLLDB,[15] a database for linguistic literature, we find a maximum of one level of hyponym terms linked to that term. However, many relevant concepts in the field of lexicography such as "lemmatization" or "neologism" can already be found in these existing ontologies, along with additional semantic information or even mappings to other vocabularies and classifications, but without a defined relation to the term "Lexicography".

Specific thematic indices of Lexicography and Dictionary Research have been proposed (see [7] for reference), isolated from each other. Most proposals are a flat list of keywords, while some define hierarchical relations between them. It is our aim to create a Domain Ontology for Metalexicography (henceforth, DOME), that consists of a multilingual thesaurus, i.e. a tree-like structure of subject headings, each of which is connected to labels (i.e. lexicalizations) in multiple languages, listing possibly more than one synonym in each language. The root element of this thesaurus, "Lexicography", is linked to the same term in the above mentioned widely used general ontologies. DOME will thus constitute a branch, a further ramification of the latter, adding new concepts but also extending relations between existing concepts; we also plan to map LCSH nodes labeled "Lexicography" that are child elements to languages or disciplines to DOME. In order to provide a highly reusable and interconnected resource, we aim to contribute DOME to various existing infrastructures, such as the Linguistic Linked Open Data Cloud (LLOD), Wikidata, or to the ongoing project coli-conc, a resource for managing and sharing concordances between library knowledge organization systems [2].

Regarding the object language or languages of the contribution, i.e. the language(s) the features presented in the article apply to, LexBib-DOME follows an alternative approach: Instead of having thematic keywords as child elements to language names, as in LCSH or existing metalexicographical keyword indices (cf. [7]), or defining language chapters as dependent to every topic, items will be indexed with the instances in LEXVO that correspond to the object language(s), independently of their thematic classification. As a consequence, DOME avoids redundancy, and bibliographic search queries that combine an object language with a topic can be answered in a straightforward way.

[14] The LOC-DB software output follows an adaptation of the OpenCitations linked data model, cf. `https://opencitations.wordpress.com/2019/01/02/opencitations-enhancement-project-final-report/`.
[15] Accessible at `http://www.blldb-online.de`.

5 Summary and Conclusion

We have presented some details of the data model and workflow proposed for the LexBib project, focusing on aspects that are relevant for the representation and availability of publication metadata as RDF Linked Open Data. An overview of the item relations inside LexBib, and of links to external resources is given in Figure 3.

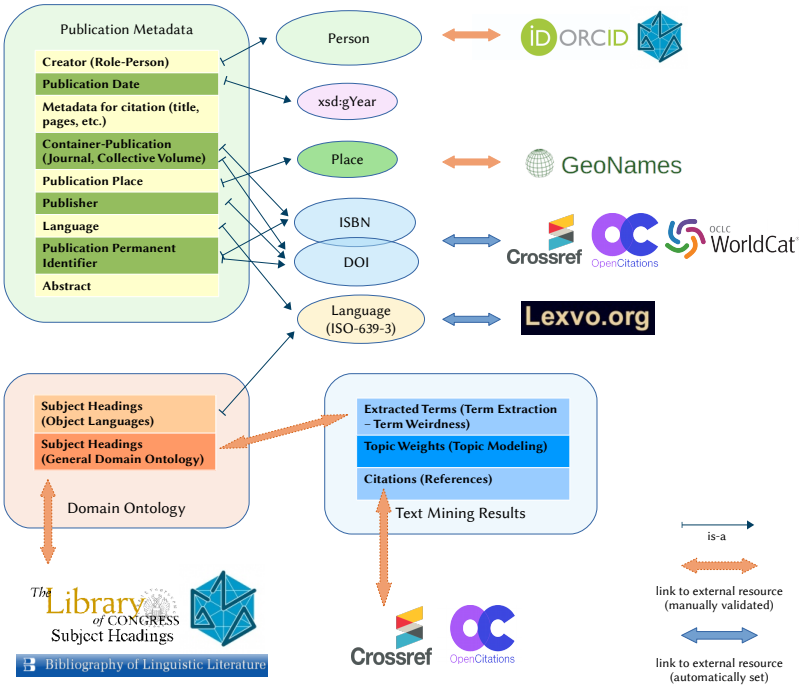

Figure 3 Data Model: Relations inside LexBib and links to external resources.

For the field of Lexicography and Dictionary Research, a domain-specific bibliography with the described features and a thematic index represented as an ontology are an innovation. But we believe that beyond that interest, some questions addressed here are relevant also from a broader or even general perspective.

The LexBib project foresees manual validation and editing effort at several points in the workflow: (i) aggregating and completing the publication metadata and full text collection, (ii) processing and enriching them as linked data, and (iii), the generation of additional content-describing metadata through a combination of computational and manual means. We track and analyse manual work performed for the different tasks as process metadata. This allows then, on the one hand, to evaluate the performance of different combination settings for computational tools and manual validation, and, on the other, to make predictions about the manual work to be foreseen in similar workflows for broader domains.

Regarding LOD integration, we have pointed out for which elements existing vocabularies can be re-used. There are no established standards for the representation of content descriptors, including provenance metadata, as we need it. For us, it is necessary to be able to annotate and, as users of LexBib, to identify content descriptors as, for example, as manually validated keywords that belong to a certain controlled vocabulary, or as term candidates extracted with a certain method, or as set of topic weights relative to a corpus

of publications. With LexBib, we can make a substantial contribution to ongoing work on developing such provenance standards, thus improving transparency and reproducibility of content metadata.

References

1 Khurshid Ahmad, Andrea Davies, Heather Fulford, and Margaret Rogers. What is a term?: The semi-automatic extraction of terms from text. In Mary Snell-Hornby, Franz Pöchhacker, and Klaus Kaindl, editors, *Benjamins Translation Library*, volume 2, page 267. John Benjamins Publishing Company, Amsterdam, 1994. doi:10.1075/btl.2.33ahm.

2 Uma Balakrishnan. DFG-Projekt: Coli-conc. Das Mapping Tool "Cocoda". *o-bib. Das offene Bibliotheksjournal / herausgegeben vom VDB, Bd. 3, Nr. 1 (2016)*, 3(1):11–16, March 2016. doi:10.5282/o-bib/2016H1S11-16.

3 AG KIM Gruppe Titeldaten DINI. *Empfehlungen zur RDF-Repräsentation bibliografischer Daten*. Deutsche Initiative für Netzwerkinformation (DINI), version 2.0 edition, November 2018. URL: https://edoc.hu-berlin.de/handle/18452/2153.3.

4 Nees Jan van Eck and Ludo Waltman. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, 111(2):1053–1070, May 2017. doi:10.1007/s11192-017-2300-7.

5 Thimotheus Chang-Whae Kim and Philipp Zumstein. Semiautomatische Katalogisierung und Normdatenverknüpfung mit Zotero im Index Theologicus. *LIBREAS*, 29:47–56, July 2016. doi:10.18452/9093.

6 Anne Lauscher, Kai Eckert, Lukas Galke, Ansgar Scherp, Syed T. R. Rizvi, Sheraz Ahmed, Andreas Dengel, Philipp Zumstein, and Annette Klein. Linked Open Citation Database: Enabling Libraries to Contribute to an Open and Interconnected Citation Graph. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18*, pages 109–118, Fort Worth, Texas, USA, 2018. ACM Press. doi:10.1145/3197026.3197050.

7 David Lindemann, Fritz Kliche, and Ulrich Heid. Lexbib: A Corpus and Bibliography of Metalexicographcal Publications. In *Proceedings of EURALEX 2018*, pages 699–712, Ljubljana, 2018. URL: http://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/.

8 David Lindemann, Fritz Kliche, and Kristin Kutzner. Lexikographie: Explizite und implizite Verortung in den Digital Humanities. In Georg Vogeler, editor, *DHd 2018 - Kritik der Digitalen Vernunft, Konferenzabstracts*, pages 257–261, Köln, 2018. Universität zu Köln.

9 Laurent Romary and Patrice Lopez. GROBID - Information Extraction from Scientific Publications. *ERCIM News, Scientific Data Sharing and Re-use*, 100, January 2015. URL: https://hal.inria.fr/hal-01673305/document.

10 Ina Rösiger, Julia Bettinger, Johannes Schäfer, Michael Dorna, and Ulrich Heid. Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 41–51, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL: http://aclweb.org/anthology/W16-4706.

11 Ina Rösiger, Johannes Schäfer, Tanja George, Simon Tannert, Ulrich Heid, and Michael Dorna. Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries. In Iztok Kosem, Miloš Jakubíček, Jelena Kallas, and Simon Krek, editors, *Proceedings of the eLex 2015 conference*, Ljubljana; Brighton, 2015. Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd. URL: https://elex.link/elex2015/conference-proceedings/paper-33/.