Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), pages 61–69
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Using Ontolex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects

Yalemisew Abgaz
Adapt Centre, School of Computing
Dublin City University Ireland
Yalemisew.abgaz@adaptcentre.ie

Abstract

This paper describes the conversion of a lexicographic collection of a non-standard German language dataset (Bavarian Dialects) into a Linguistic Linked Open Data (LLOD) format within the framework of ExploreAT! Project. The collection is divided into three parts: 1) conceptual content for unique corpus collection - questionnaire dataset (*DBÖ_questionnaires*) which contains details of the questionnaires and associated questions, 2) metadata regarding the collection framework - including collectors and hierarchical system of localisations, and 3) lexical dataset (*DBÖ_entries*) - both unique data collections as answers to the questions and unique data collections as excerpts of already published sources. In its current form, the *DBÖ_entries* dataset is available in a TEI/XML format separately from the questionnaire dataset. This paper presents the mapping of the lexical entries from the TEI/XML into an LLOD format using the Ontolex-Lemon model. We present the resulting lexicon of Bavarian Dialect and the approach used to interlink the data collection questionnaires with their corresponding answers (lexical entries). The output complements *DBÖ_questionnaires* dataset, which is already in an LLOD format, by semantically interlinking the original questions with the answers and vice-versa.

Semantic publishing, Historical data, Linguistic Linked Open Data, *exploreAT_TEI* conversion

1. Introduction

With the adoption of open access policy, public institutions that deal with a large collection of language resources have shown a growing interest in the publication of resources as linked data using machine-readable lexical models available in the LLOD cloud (Chiarcos et al., 2013). Language resources collected over a long period, with wider geographic coverage and using traditional data collection methods are still in the process of transformation to make the data available in a machine-readable, interlinked and interoperable format. The process widely involves digitisation of both original data collection methods and the collected data from a physical medium such as paper slips, cards, recordings, etc. Semantically linking the questionnaires along with the collectors, time, medium, etc., opens new doors for rich and efficient exploration and reuse to support multidimensional analysis and exploration of the data. This multidimensional analysis uses features such as the question text, authors, collectors, place, and time in addition to the features of the lexical entries such as forms, Part Of Speech (POS), grammar, etc.

The Database of Bavarian Dialects in Austria [Datenbank der Bairischen Mundarten in Österreich] (DBÖ), a digitised non-standard German language resource (Österreichische Akademie der Wissenschaften, 2018), is one of the rich linguistic and lexicographic resources collected from 1913-1998 to document the Bavarian Dialect and rural life in present-day Austria, Czech, Slovakia, Hungary and northern Italy. This collection roughly contains 762 questionnaires with a total of 24,382 questions and 3.6 million paper slips comprising answers to individual questions. There has been a long process of digitisation of the collection including the conversion of the paper-based information to a digital format initially using an old text processing tool called TUSTEP (Barabas et al., 2010), followed by the subsequent conversion of the data into a relational database (dbo@ema)(Wandl-Vogt, Eveline, 2010) and then into TEI/XML formats (*exploreAT_TEI*) (Schopper et al., 2015; Bowers and Stöckle, 2018). A recent conversion of the dbo@ema database into an LLOD format is performed on the *DBÖ_questionnaires* including authors, collectors, places, sources and paper slips using OLDCAN ontology (Abgaz et al., 2018b; Abgaz et al., 2018a) in the framework of the project exploreAT!.

Despite several efforts made, so far the conversion did not include the *DBÖ_entries*. First efforts in dealing with LLOD were made by Wandl-Vogt and Declerck in 2014 to create a model for the conversion of the printed dictionary (Declerck and Wandl-Vogt, 2014). The *exploreAT_TEI* data efficiently supports the query and retrieval of the lexical entries, offers a well-established data model, yet is still not in a native RDF format and is not compatible with the latest *DBÖ_questionnaires* dataset. With the recent development in publishing linguistic data using widespread lexical models such as Ontolex-Lemon (Cimiano et al., 2020; Cimiano et al., 2016), several efforts are being made in curating, enriching, interlinking and publishing of the DBÖ data in the LLOD platform.

The Ontolex model is widely used to represent and publish lexical resources (Declerck, 2018; Tittel et al., 2018; McCrae et al., 2017; Tiberius and Declerck, 2017; Bosque-Gil et al., 2015). This paper presents an ongoing effort in the conversion of the current *exploreAT_TEI* entries into an LLOD format using the Ontolex-Lemon model and the OLDCAN ontology to link the entries to the corresponding questions. The core entities contained in the *exploreAT_TEI* files are identified and the relevant information is extracted for representing the lexical entries. Since the *DBÖ_entries* dataset contains diverse information extracted from the paper slips, only the relevant elements are included in the conversion.

The main contribution of this paper includes:

- the conversion of the *DBÖ_entries* dataset using the standard Ontolex-Lemon model and the linking of the *DBÖ_entries* with *DBÖ_questionnaires* dataset which is used to collect the original data. This semantic interlinking flourishes a bi-directional exploration of the data: from lexical entries to questions and questionnaires and vice-versa using aspects including topics, authors, collectors, places, paper slips, etc.

- the analysis of the data in its current form and the mappings from the *exploreAT_TEI* into LLOD and,
- the presentation of the challenges and the lesson learned while converting the data and publishing the resulting lexicon using the Ontolex-Lemon model.

The remaining sections are organised as follows: Section 2. presents the structure and the content of the current TEI/XML format. The mapping to Ontolex-Lemon model and the major design decisions are presented in Section 3. Section 4. discusses the process of interlinking the original questions with the lexical entries, and Section 5. further presents a systematic interlinking of concepts, generated by experts at the questionnaire level, and the lexical entries. Finally, Section 6. presents some of the data quality issues that need to be addressed before publishing the dataset to the public.

2. The *exploreAT_TEI* Data

The main goal of the collection is to document the Bavarian dialects in Austria and publish it in the form of a dictionary (WBÖ) and an atlas. The digitisation of the data collection process and its various supporting materials (DBÖ) offered a knowledge base for a comprehensive, joint approach (dictionary + atlas), prototyped within dbo@ema (Scholz et al., 2008; Wandl-Vogt, 2010) and a cultural, Pan-European exploitation, prototyped within exploreAT!. The data is collected using questionnaires and paper slips distributed via mails and direct interaction with the respondents. The collection suffered several stages of evolution including the scanning and digitisation of paper slips using TUESTEP file format (Barabas et al., 2010), conversion to MySQL (Barabas et al., 2010) and TEI/XML formats (Schopper et al., 2015). The current version of the *exploreAT_TEI* data is TEI version 2 which significantly transformed the original data by reducing redundant data categories (Bowers and Stöckle, 2018). The *exploreAT_TEI* files are organised into folders with the corresponding labels from A-z matching to the physical drawers. Each file contains several elements representing lexical entries with unique identifiers.

The structure of the entries is not homogeneous. However, there are common elements shared among the majority of the entries. These major elements constitute entry, form, orthography, grammar group, POS, sense, etymology, usage, place and date. The entries further contain additional elements such as quotes, references, notes, bibliographies, etc. A snippet of the *exploreAT_TEI* file for an entry ("Oberhaus") is presented in Listing 1.

Each of the above major elements has distinct XML elements and attributes that describe the content of the elements. For keeping the discussion concise, we started from the `<entry>` element and subsequently move deep into the `<form>` element to introduce the detail information contained in each element. An entry contained in `<entry> ... </entry>` block represents a unique lexical entry. The `<entry>` element has `<form>` representing the different forms of the lexical entry. A lexical entry could have more than one `<form>` element identified by its attribute `'type'`. The type of a form could be one of the following five categories: Hauptlemma (Main lemma), Lautung (Pronunciation), Lehnwort (Loan word), Nebenlemma (Other lemmas) and Verweislemma (Additional related form). The form with the Hauptlemma also has the `<orth>` element representing the orthography of the main lemma. A typical form has one or two `<orth>` entries identified by the type attribute. The `<orth>` could be original (as it appeared on the original paper slip) or normalised (edited by a professional). An entry further has `<gramGrp>` representing the grammar group of the entry, `<sense>` representing the sense of the form, `<ref>` representing additional data such as archive, source, questionnaire number, etc. Finally, an entry has `<usg>` element representing the usage of the lexical entry. The usage type identifies how the lexical entry is used and in the majority of the cases, it is a geographic location.

Listing 1: A snippet of the *exploreAT_TEI* file

```
<entry xml:id="h385_qdb-d1e386" xml:lang="bar">
 <form type="hauptlemma">
   <orth type="orig">(Ober)haus</orth>
   <orth type="normalized">Oberhaus</orth>
 </form>
 <gramGrp>
   <pos>Subst</pos>
 </gramGrp>
 <form type="lautung" n="1">
   <pron notation="tustep">s -..ow˜An h&#xE2;;us
   </pron>
   <pron notation="ipa" resp="#JB" change="01">
     s -..ow˜An h&#xE2;;us
   </pron>
   <gramGrp>
        <gram>[n,sg+A]</gram>
   </gramGrp>
 </form>
 <sense corresp="this:LT1">
   <def xml:lang="de">Vorhaus im ersten Stock</def>
 </sense>
 <form type="nebenlemma">
   <orth type="orig">(Obern)haus</orth>
   <gramGrp>
      <pos>Subst</pos>
   </gramGrp>
         <orth type="normalized">Obernhaus</orth>
 </form>
 <ref type="archiv">
    HK 385, h3850131.pir, korr. E.V.
 </ref>
 <ref type="quelle">Strobl Flachg. Bauer (1972)</
      ref>
 <ref type="quelleBearbeitet">
      {4.5d06} s&#xF6;Flachg.:
         Sa. Aufn.BAUER&#xB7; (1972) [GaFb2; chTr]
 </ref>
 <usg type="geo">
   <placeName type="orig">Strobl Sa.</placeName>
   <listPlace ref="sigle:4.5d06">
     <place type="Bundesland">
       <placeName>Sa.</placeName>
              <idno>4</idno>
              <listPlace>
                   ...
        </place>
   </listPlace>
 </usg>
</entry>
```

Among these elements, the lexicographers who are working in this project have identified the elements that constitute the core of the lexicon. The following section presents a detailed discussion on how these core elements are mapped to Ontolex-Lemon model using R2RML mapping. An intermediate relational database is introduced to facilitate the conversion and to support compatibility with the *DBÖ_questionnaires* dataset. There are three user requirements that the conversion process needs to deliver.

- The use of standard, and widely used model for publishing the LLOD data. The final dataset should use existing models that are standardised and widely used by the lexicographic community.
- The resulting LLOD shall link the lexical entries with the questions used to collect the data explicitly. This will create the bridge between the questionnaire dataset and the lexical dataset.
- The selected method shall consider future semantic enrichment using resources including DBpedia [1], KBpedia [2] and BabelNet [3].

To achieve this, the prevalent Ontolex-Lemon model is used for publishing the lexical data on the LLOD platform. The OLDCAN ontology is also used to preserve the link between the entries and questions. This aspect is dealt with more detail in the following sections.

3. Mapping *exploreAT_TEI* to Ontolex-Lemon

A series of decisions are made to map the core elements of the *exploreAT_TEI* data into Ontolex-Lemon representation using an intermediate relational database and R2RML Mapping. The choice of including an intermediate relational database is to support backward compatibility with the *DBÖ_questionnaires* dataset, which is previously converted from MySQL database (dbo@ema) and also to interlink the lexical data with the questionnaire dataset which was also based on a relational data model (Abgaz et al., 2018b; Abgaz et al., 2018a).

The Ontolex-Lemon model provides a rich semantics to represent linguistic resources by presenting morphological and syntactic properties of lexical entries, which are the core classes of the model. A lexical entry is a building block of a lexicon which consists of a set of forms and their associated meanings (Cimiano et al., 2016). The lexical entry is connected to a Lexical Concept via `evokes/isEvokedby` object property. Lexical entry further relates to Lexical Sense using `sense/isSenseOf` object property. The core Ontolex module is presented in Figure 2 (Cimiano et al., 2016).

A lexical entry represents a unit of analysis of the lexicon that consists of a set of grammatically related forms and a set of base meanings that are associated with all of these forms. Thus, a lexical entry is a `Word`, `Multiword Expression` or `Affix` with a single part-of-speech, morphological pattern, etymology and set of senses.

WBOLexicon is created using `ontolex:Lexicon` and the following namespaces are used to be defined throughout all the listings and examples in this paper. The TURTLE syntax is used to present the resulting data snippets.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-
     syntax-ns#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dct: <http://purl.org/dc/terms/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix lexinfo: <http://www.lexinfo.net/ontology
     /2.0/lexinfo#>.
@prefix lime: <http://www.w3.org/ns/lemon/lime#>.
@prefix oldcan: <https://explorations4u.acdh.oeaw.
     ac.at/ontology/oldcan#>.
@prefix ontolex: <http://www.w3.org/ns/lemon/
     ontolex#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema
     #>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix skos: <http://www.w3.org/2004/02/skos#>.
@prefix wbo: <https://exploreat.oeaw.ac.at/
     WBOLexicon/>.
```

3.1. Entries and Forms

Entries of the *exploreAT_TEI* dataset are the core elements of the collection. These entries are mapped to `ontolex:LexicalEntry` and are added to the WBOLexicon using `ontolex:entry`. This is a relatively simple mapping which defines all the entries as `ontolex:LexicalEntry` and lays the foundation for the rest of the elements. The following R2RML mapping creates instances of a lexical entry and associates each entry with the WBOLexicon.

```
<#LexiconEntryTriplesMap>
a rr:TriplesMap;
rr:logicalTable [ rr:sqlQuery """
Select 'WBOLexicon' as lexicon, e.id, e.lang from
     entry e; """ ];
rr:subjectMap [
  rr:template "https://exploreat.oeaw.ac.at/{
       lexicon}";
  rr:class ontolex:Lexicon ;
  rr:graph lexGraph: ;] ;
rr:predicateObjectMap [
  rr:predicate ontolex:language ;
  rr:objectMap [ rr:column "lang" ] ;
  rr:graph wbo:lexicon_graph;];
rr:predicateObjectMap [
  rr:predicate ontolex:entry;
  rr:objectMap [
    rr:template "https://exploreat.oeaw.ac.at/
         WBOLexicon/LexicalEntry/{id}" ;
  rr:graph wbo:lexicon_graph;]; ];.
```

The mapping retrieves all the entries in the database and represent them as lexical entries of the lexicon. The resulting lexicon and its lexical entries are presented below.

```
wbo:WBOLexicon a ontolex:Lexicon ;
 ontolex:entry
  <https://exploreat.oeaw.ac.at/WBOLexicon/
       LexicalEntry/h385_qdb-d1e2>,
  <https://exploreat.oeaw.ac.at/WBOLexicon/
       LexicalEntry/h385_qdb-d1e108>,
  <https://exploreat.oeaw.ac.at/WBOLexicon/
       LexicalEntry/h385_qdb-d1e129>,
  ...
```

[1] https://wiki.dbpedia.org/
[2] http://kbpedia.org/
[3] https://babelnet.io/

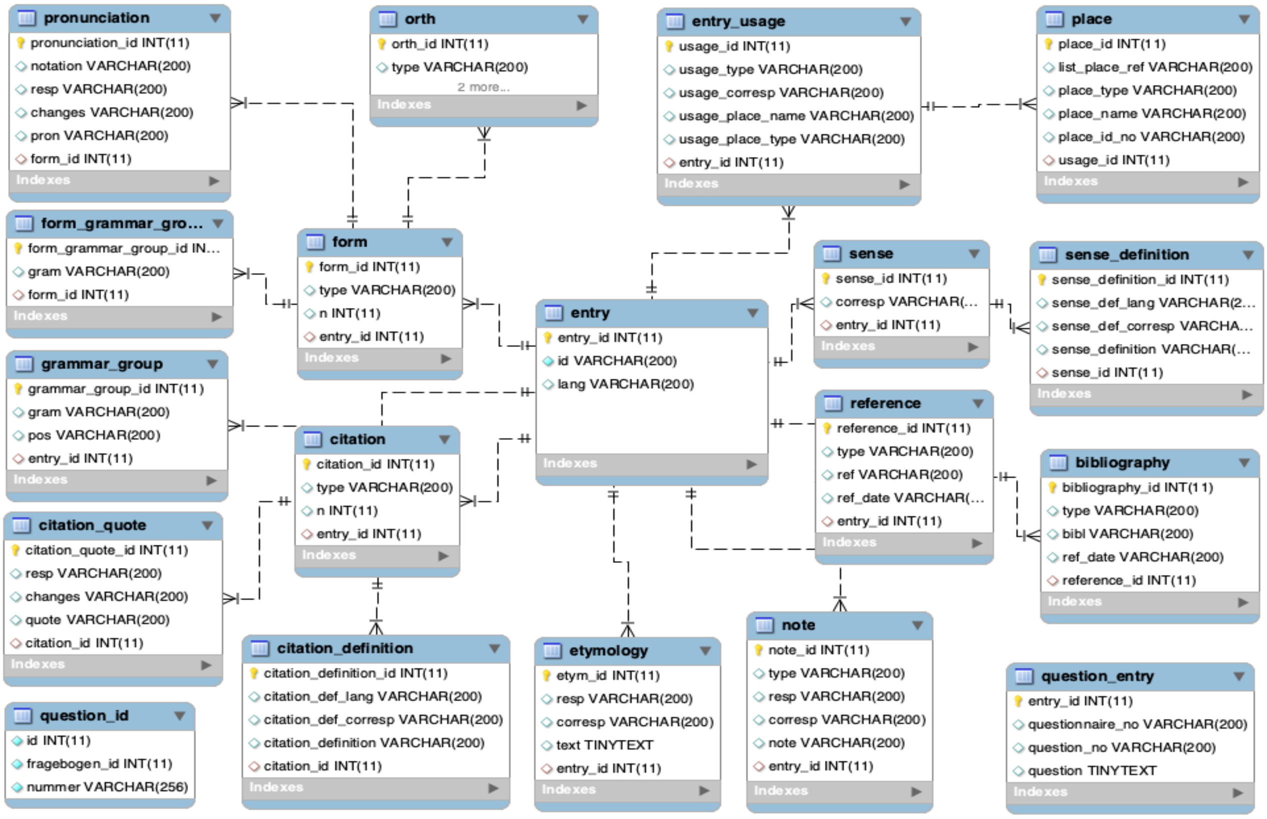

Figure 1: The *exploreAT_TEI* database schema

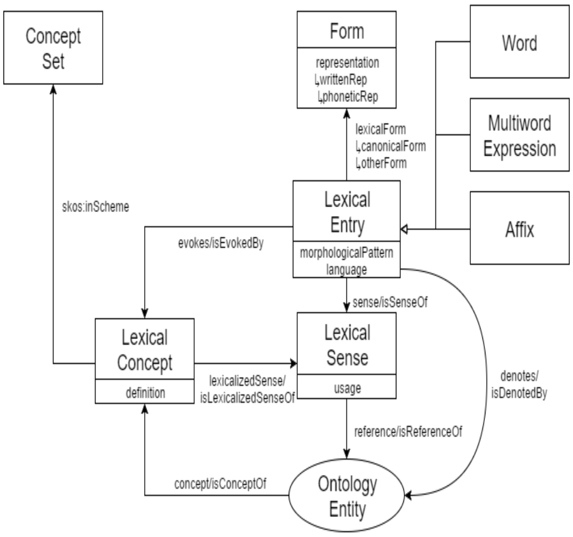

Figure 2: The Ontolex-Lemon model.

At this stage, the mapping does not distinguish between a word and Multiword Expression. However, it represents the entries using the general LexicalEntry class. This is done because the *exploreAT_TEI* dataset does not distinguish between word, MultiwordExpression and Affix. Furthermore, detecting German compound words and Affix from the dataset is complex and beyond the scope of this paper. Each lexical entry is represented using a unique URI generated from the unique id number of the entries in the *exploreAT_TEI* files. To support maximum interoperability with the legacy data, we stick to the existing id numbers following trends from similar conversions (Klimek and Brümmer, 2015) use the written representation of the entries.

3.2. Form, Canonical Form and Other Forms

A form is a grammatical realisation of a lexical entry (Cimiano et al., 2016). All the entries in the *exploreAT_TEI* data have at least one form which is represented using ontolex:Form. The form is linked to the lexical entry using ontolex:lexicalForm object property. We represent the forms with further details by distinguishing between canonical form and other forms.

In the Ontolex-Lemon model, there is only one canonical form allowed per entry. However, there are five different types of forms in the *exploreAT_TEI* dataset. 'HauptLemma' is the main lemma which is selected as a canonicalForm and the other four forms are treated differently. The so-called "Lautung" does not represent another form of the entry, but it represents the pronunciation of the entry in Tuestep and IPA notation. Thus, this is automatically excluded but later used to add pronunciation to the entry. "Nebenlemma" is treated as ontolex:otherForm, however, "Verweislemma" and "Lehnwort", are not considered important at this stage due to the quality of the data and the ambiguity of the meaning of the categories. Thus, all the forms with type='Hauptlemma' are represented as ontolex:canonicalForm, whereas, type=‘Nebellemma’ is represented as ontolex:otherForm. The mapping of the form and the canonicalForm is given below (Note that the mapping of the otherForm is also similar except the query used to extract the rows).

```
<#LexicalEntrycanonicalFormTriplesMap>
a rr:TriplesMap;
rr:logicalTable [ rr:sqlQuery """
select e.id, e.lang, f.entry_id, f.form_id from
entry e left join form f on e.entry_id =
f.entry_id where f.type ='hauptlemma'; """ ];
rr:subjectMap [
  rr:template "https://exploreat.oeaw.ac.at/
        WBOLexicon/
  LexicalEntry/{id}";
  rr:class ontolex:LexicalEntry ;
  rr:graph wbo:lexicon_graph ;] ;
rr:predicateObjectMap [
  rr:predicate ontolex:lexicalForm ;
  rr:predicate ontolex:canonicalForm ;
  rr:objectMap [
  rr:template "https://exploreat.oeaw.ac.at/
        WBOLexicon/Form/
               {form_id}";
  rr:graph wbo:lexicon_graph ;] ] ;
```

Based on the above mapping, a given form is represented using at least one ontolex:Form and ontolex:canonicalForm.

```
<https://exploreat.oeaw.ac.at/WBOLexicon/
      LexicalEntry/h385_qdb-d1e2>
    lexinfo:partOfSpeech lexinfo:noun ;
    a ontolex:LexicalEntry ;
    ontolex:lexicalForm <https://exploreat.oeaw.ac.
         at/WBOLexicon/Form/1> ;
    ontolex:canonicalForm <https://exploreat.oeaw.ac
         .at/WBOLexicon/Form/1> ;
<https://exploreat.oeaw.ac.at/WBOLexicon/
      LexicalEntry/h385_qdb-d1e108>
    lexinfo:partOfSpeech lexinfo:noun ;
    a ontolex:LexicalEntry ;
    ontolex:lexicalForm <https://exploreat.oeaw.ac.
         at/WBOLexicon/Form/4> .
    ontolex:canonicalForm <https://exploreat.oeaw.ac
         .at/WBOLexicon/Form/4> ;
```

3.3. Form Written Representation and Pronunciation

The *exploreAT_TEI* data contains the <orth> element embedded inside the form element. The <orth> element represents the orthography of the “Hauptlemma” or “NebenLemma”.

3.3.1. Written Representation

The orthography of a lexical entry is represented by ontolex:writtenRep. The *exploreAT_TEI* dataset also uses a type attribute to distinguish between the original written representation and the normalised representation. The normalised representation transforms the original orthography which contains several diacritic marks and special characters into a normalised representation. I took the normalised representation as a written representation of the form. This is done for two reasons. First, the character encoding of the original representation is not human readable, and the second, search and retrieval with such representation will pose a difficulty.

The written representation is further enriched by skos:prefLabel and rdfs:label. The content of the original written representation is also captured using skos:altlabel which will serve as an alternate label for the form and enable the representation of the standard form.

```
<#FormTriplesMapNormalised>
a rr:TriplesMap;
rr:logicalTable [ rr:sqlQuery """
Select o.orth_id, o.type, o.orth, f.form_id from
form f inner join orth o on f.form_id =o.form_id
where o.type<>'orig'; """ ];
rr:subjectMap [
  rr:template "https://exploreat.oeaw.ac.at/
        WBOLexicon/Form/
   {form_id}";
  rr:class ontolex:Form ;
  rr:graph wbo:lexicon_graph ;] ;
rr:predicateObjectMap [
  rr:predicate ontolex:writtenRep ;
  rr:objectMap [ rr:column "orth" ;rr:language "bar
       ";];
  rr:graph wbo:lexicon_graph ;];
rr:predicateObjectMap [
  rr:predicate rdfs:label;
  rr:predicate skos:preflabel;
  rr:objectMap [ rr:column "orth" ;rr:language "bar
       ";];
  rr:graph wbo:lexicon_graph ;] ;
---------------------------------------------
<#FormTriplesMapPronunciationIPA>
a rr:TriplesMap;
rr:logicalTable [ rr:sqlQuery """
SELECT pron,notation, form_id FROM pronunciation
where notation='ipa'; """ ];
rr:subjectMap [
    rr:template "https://exploreat.oeaw.ac.at/
          WBOLexicon/Form/
    {form_id}" ;
    rr:class ontolex:Form ;
    rr:graph wbo:lexicon_graph ;] ;

rr:predicateObjectMap [
    rr:predicate ontolex:phoneticRep ;
    rr:objectMap [ rr:column "pron";
    rr:language "ipa"; ] ;
    rr:graph wbo:lexicon_graph ;];
```

3.3.2. Pronunciation

The pronunciation of the “Hauptlemma” is included in a separate form element with type “Lautung”. All the variant pronunciations with IPA notation and the so called Tustep notation are also included inside <pron> element with notation attribute. This information about the pronunciation is represented using ontolex:phoneticRepresentation. Below, We demonstrate the result of the mapping of both Tustep and IPA notations.

```
<https://exploreat.oeaw.ac.at/WBOLexicon/Form/1>
    a ontolex:Form ;
    ontolex: rdfs:label "Oberhaus"@bar ;
    skos:altlabel "(Ober)haus"@bar ;
    skos:preflabel "Oberhaus"@bar ;
    ontolex:phoneticRep "'s Oberhaus"@ipa,
                 s"'s Oberhaus"@tustep ;
    ontolex:writtenRep "Oberhaus"@bar .
```

3.4. Part of Speech (POS) and Grammatical Groups

The POS of an entry which applies to all the forms within an entry is provided inside the <gramgrp> element. This POS applies to all the forms except those forms which have their grammar group. If a form has its grammar group and if the POS is defined there, this form will get its POS instead of inheriting the entry-level POS. Whenever a POS

TEI	Lexinfo	TEI	Lexinfo
Verb	verb	Verb Verb	verb
Subst	noun	Subst Subst	noun
Pron	pronoun	Adj Adj	adjective
Adv	adverb	Adv Adv	adverb
Adj	adjective	Adj Subst	?
Interj	interjection	Verb Subst	?
Num	numeral	Subst Prep	?
Conj	conjunction	Affix	?
Prep	fusedPreposition		

Table 1: Mapping of POS between *exploreAT_TEI* and Lexinfo.

Gram	Gram	Gram	Gram
[P2/1+A]	[sg3+5P3]	[sg3]	[D1,n+A]
[P2/1]	[pl1+5P1]	[pl2+5P2]	[I/1,n+A]
[P1/1,n+A]	[imp,sg2]	[kj,pl3+5P3]	[P1/1,f+A]
[P2]	[pl3+5P3]	[+7]	[P1/1,f]
[P2/1+U]	[kj]	[pl3]	[sg3+0]
[sg+U]	[P1]	[sg3+5P3]	
[sg2+5P2]	[kj,sg1+5P1]	[kj,sg3+5P3]	[m+A]
[sg1+5P1]	[imp]	[kj,sg2+5P2]	[m+U]
[sg2]	[+5P1]	[kj,pl1+5P1]	[il] [m+A]
[D1]	[sg1]	[kj,pl2+5P2]	[+A]

Table 2: Sample grammar group observed in the dataset

information is identified inside the `<form>` element, it is mapped to `lexinfo:pos` in addition to the POS associated with the entry. In the *exploreAT_TEI*, there are 17 different POS used whereas in lexinfo there are only 13 (Buitelaar et al., 2011). A partial mapping of the POS from the *exploreAT_TEI* to lexinfo is implemented during the mapping process shown in Table 1. There are also POS instances (with question marks) which are not mapped to lexinfo due to ambiguous POS elements.

```
<https://exploreat.oeaw.ac.at/WBOLexicon/
     LexicalEntry/h385_qdb-d1e2>
   lexinfo:partOfSpeech lexinfo:noun ;
<https://exploreat.oeaw.ac.at/WBOLexicon/
     LexicalEntry/h385_qdb-d1e689>
   lexinfo:partOfSpeech lexinfo:noun ;
<https://exploreat.oeaw.ac.at/WBOLexicon/
     LexicalEntry/h385_qdb-d1e72>
   lexinfo:partOfSpeech lexinfo:adverb ;
```

This work looks into the grammar group represented at the form level. The grammar group identifies between gender, number and case. Here again, an attempt is made to map the grammar groups at the form level using `lexinfo:gender`, `lexinfo:number` and `lexinfo:case`. However, in the collection, there are more than one million rows of data related to the grammar group. What makes it worse is that there are 5,720 unique combinations of pos, number, gender and case. Supporting a mapping of this grammatical information to the respective representation required significant effort and knowledge. Some of the complexity of the data is shown in the following table where the possible combinations are presented. Due to this complexity, this work does not include details of the form in the current conversion process and this task is left for future work (see the lexinfo entries with "?" in Table 1).

3.5. Sense, Definition, and Etymology

The entry has `ontolex:Sense` information which specifies the context in which the given entry is used. The `<sense>` element also has the `<def>` element which provides the definition of the word. The sense further contains the ISO 639-21 language tag which specifies the language of the definition. Whenever the entry has more than one sense, additional `<sense>` element containing the definition is added. These elements are identified using a number attribute @n.

```
<#SenseTriplesMap>
a rr:TriplesMap;
rr:logicalTable [ rr:sqlQuery """

select s.sense_id, sense_definition from Sense s
left join sense_definition sd
on s.sense_id =sd.sense_id; """ ];
rr:subjectMap [
  rr:template "https://exploreat.oeaw.ac.at/
      WBOLexicon/Sense/
  {sense_id}" ;
   rr:class ontolex:LexicalSense ;
   rr:graph wbo:lexicon_graph ;] ;
rr:predicateObjectMap [
   rr:predicate dct:description ;
   rr:objectMap [ rr:column "sense_definition";
   rr:language "de"; ] ;
   rr:graph wbo:lexicon_graph ;];
```

At this stage, sense is mapped to `ontolex:Sense` and is associated the definition of the sense using `skos:definition` and `dct:description` together with the language in which the definition is given.

```
<https://exploreat.oeaw.ac.at/WBOLexicon/Sense/1>
   dct:description "das obere Stockwerk"@de ;
   a ontolex:LexicalSense ;
   skos:definition "das obere Stockwerk"@de .
<https://exploreat.oeaw.ac.at/WBOLexicon/Sense/10>
   dct:description "Dachbodenraum; Dachboden"@de ;
   a ontolex:LexicalSense ;
   skos:definition "Dachbodenraum; Dachboden"@de .
<https://exploreat.oeaw.ac.at/WBOLexicon/Sense/11>
   dct:description "Vorhaus im ersten Stock"@de ;
   a ontolex:LexicalSense ;
   skos:definition "Vorhaus im ersten Stock"@de .
<https://exploreat.oeaw.ac.at/WBOLexicon/Sense/12>
   dct:description "Husl bei Strengberg"@de ;
   a ontolex:LexicalSense ;
   skos:definition "Husl bei Strengberg"@de .
```

This paper further presents the etymology of the lexical entries whenever they are available. The etymology of the lexical entries represents the origin of the word and a proposed module for representing details of the etymology is presented in (Khan, 2018). Since our etymology collection is not complex, it is represented using the `lexinfo:etymology` object property linked to the lexical entry. A careful investigation of the etymology data in the collection shows that a further expert analysis of the content of the etymological data is crucial for the efficient utilisation by non-expert users.

```
https://exploreat.oeaw.ac.at/WBOLexicon/
     LexicalEntry/h385_qdb-d1e2>
   lexinfo:etymology "s.a. TSA 3,53"@de ;
   lexinfo:partOfSpeech lexinfo:noun ;
   a ontolex:LexicalEntry ;
   ontolex:canonicalForm <https://exploreat.oeaw.ac
        .at/WBOLexicon/Form/1>;
   ontolex:lexicalForm <https://exploreat.oeaw.ac.
        at/WBOLexicon/Form/1>;
   ontolex:sense <https://exploreat.oeaw.ac.at/
        WBOLexicon/Sense/1>.
```

4. Interlinking Lexical Entries to the Original Questions

One of the requirements is to create a meaningful relationship between the different stages of the collection. In (Abgaz et al., 2018b), the data collection method is represented with OLDCAN ontology. The subsequent task which interlinks the original questions used to collect the data to the answers is also covered included in the model. OLDCAN models the answers initially as lemma and subsequently, they are represented as lexical entries using Ontolex-Lemon. This has not been done initially due to the absence of information to represent the answers in a detailed form. However, once the *exploreAT_TEI* data is converted into LOD, the next step is to link the questionnaire with the lexical entries.

Each entry in the *exploreAT_TEI* file contains a `<ref>` element with a pointer to the question number (fragebogen-Nummer) that combines the questionnaire and the question number to identify the corresponding question for the lexical entry. This provides crucial information, however, the raw data itself is not represented accurately and it poses a challenge to directly create the required link. To address this problem, the scope is narrowed down to the Systematic, Additional and Dialectographic questionnaires (1-120)(Abgaz et al., 2018b) and link the questions of these questionnaires with the lexical entries. For the rest of the questionnaire, currently, it is not possible to resolve the links from the data provided in the *exploreAT_TEI* dataset.

```
<https://exploreat.oeaw.ac.at/WBOLexicon/
     LexicalEntry/h385_qdb-d1e108>
   lexinfo:partOfSpeech lexinfo:noun ;
   a ontolex:LexicalEntry ;
   ontolex:canonicalForm <https://exploreat.oeaw.ac
        .at/WBOLexicon/Form/4> ;
   ontolex:lexicalForm <https://exploreat.oeaw.ac.
        at/WBOLexicon/Form/4> ;
   oldcan:isAnswerOf <https://exploreat-
        questionnaireexplorer.hephaistos.arz.oeaw.
        ac.at/Question/13225>.

 <https://exploreat-questionnaireexplorer.
      hephaistos.arz.oeaw.ac.at/Question/13225>
   oldcan:isQuestionOf <https://exploreat-
        questionnaireexplorer.hephaistos.arz.oeaw.
        ac.at/Questionnaire/92> ;
   oldcan:originalQuestion "Wohnhaus/Dachboden:
        Dachboden (Speicher, Unterdach, Diele); Ra.
         wie: auf der hoh’ Diel’"@de ;
   a oldcan:SyntacticQuestion;
   oldcan:number "F13";
   oldcan:shortQuestion "Dachboden (Speicher,
        Unterdach, Diele); Fg./Ra.*"@de .
```

Thus, this work implements the link using the `oldcan:hasAnswer` object properties with the question as a domain and the lexical entry as a range of the object property along with its inverse oldcan:isAnswerOf property. The Previous example shows the details of a question linked with its answers.

5. Interlinking of Questionnaire Concepts to Lexical Entries

In previous efforts, the questionnaires were linked to DBpedia concepts via a semi-automatic extraction of fine-grained questionnaire topics. These topics in combination with the questionnaire titles were used to extract potential concepts using DBpedia Spotlight[4]. Further, the suggested concepts with greater than 99% accuracy were evaluated and selected by subject matter experts. Even if these concepts are a bit generic, they are very useful in representing the main concepts that are covered by the questionnaires. This gives us the starting point to link the lexical entries to DBpedia concepts using `ontolex:denotes` relationship. At this stage, an experiment is conducted on some selected questionnaire concepts to see whether it is appropriate to use these suggested DBpedia concepts for lexical entries. The result shows that the concepts at the questionnaire level are too generic and can not be used meaningfully to represent the concepts of the lexical entries. As the assumption is evaluated, the topics in the questionnaires provide only high-level concepts, whereas the lexical entries provide very detailed concepts. The gap is created because the concepts in the questionnaires are further specialised in the questions and subquestions. The lexical entries are collected in response to the questions and due to this, they represent very specific concepts. To effectively resolve this problem, both bottom-up and top-down approach should be used. The bottom-up approach seeks to retrieve a matching concept for the lexical entry from DBpedia and the top-down approach will provide a mechanism to disambiguate the results of the bottom-up approach. With this idea in mind, this paper demonstrates the potential of the interlinking process to support further enrichment to the collection. Thus, we decide to relate these questionnaire concepts indirectly via `oldcan:isAnswerOf` relation (Section 4.), which link the lexical entries with the questions.

6. Data Quality Issues for Further Improvement

The resulting LLOD data represents the lexicographic collection with rich information using the standard ontolex model. Sample *exploreAT_TEI* file, the database structure, the R2RML mapping and some resulting dataset in a TTL format is available at github[5]. Since the final data size is large, it not available for public use at this stage. The entries are represented using the core classes defined in the Ontolex-Lemon model. The dataset in its current form, however, needs further quality checks before it is made available to the public. Some of the data quality issues and potential remedies are outlined below.

[4] https://www.dbpedia-spotlight.org/

[5] `https://github.com/yalemisewAbgaz/TEI-XML_Mapping.git`

6.1. Word, MultiwordExpression and Affix

The current conversion of the lexical entries does not use the subclasses of the ontolex:LexicalEntry. The entries are not classified as Word, MultiwordExpression or Affix. In its current form, it is not a trivial task to classify the lexical entries into the subclasses. However, by combining the grammatical information with external resources such as GermaNet, BabelNet and DBPedia entries, it is possible to classify the entries with their respective subclasses. This will improve the quality of the final dataset by incorporating useful details about the entries.

6.2. Part of speech, Grammar and Etymology

The conversion represents a significant portion of the POS of the lexical entries. However, there are some POS entries that are not mapped to lexinfo:partOfSpeech. There are two options to address this problem. First, involving experts to map the parts of speech that are not mapped to lexinfo:partOfSpeech and provide the complete mapping. The second option is to use the parts of speech in the *exploreAT_TEI* files and include them in the OLDCAN ontology to represent them, which is a less preferable option. The first option will keep the data compatible by using standardised POS used elsewhere, however, it requires a deeper expert analysis of the cases. This will improve the quality of the resulting LLOD data.

The grammar group is also another area of investigation to deliver a rich lexicon with the grammatical information already available in the *exploreAT_TEI* file. It requires a deeper analysis of the combinations of the grammar groups and a method to decipher the grammatical data and map it to the standard grammatical groups, for example, lexinfo:case, lexinfo:number, lexinfo:gender, etc.

The etymology data and other related data also needs some improvement. There are several abbreviations, mnemonics and acronyms that are included in the data. The presence of such data without the corresponding interpretation will make the data less usable both by humans and machines. To address this problem, a scripting language with some expert assistance can be used to transform the abbreviations, mnemonics and acronyms into their corresponding definitions.

7. Conclusion

This paper presents the results of ongoing conversion of a huge lexicographic dataset from *exploreAT_TEI* format to a LLOD format to digitally publish the RDF version of the dictionary of the Bavarian dialects. In the conversion process, the core elements of the *exploreAT_TEI* data are transformed into Ontolex-Lemon classes and properties. As the data is not homogeneous, the mapping process is not always straightforward, however, the implementation tries to identify the best mappings for each of the selected data. This is the first stage of the transformation of the *exploreAT_TEI* data by focusing on the core elements of the dataset. Future work will include the enrichment of the LOD data with additional information including fine-grained DBpedia concepts for each lexical entry, enrichment of the lexical entries into Word, Multiword Expression and Affix and integration of the resulting data into the visualisation system (Rodríguez Díaz et al., 2019) developed for the exploreAT! project.

Acknowledgements: This research is funded by the Nationalstiftung of the Austrian Academy of Sciences under the funding scheme: Digitales kulturelles Erbe, No. DH2014/22. as part of the exploreAT! and the ADAPT SFI Research Centre at Dublin City University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106.

8. Bibliographical References

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., and Way, A. (2018a). A semantic model for traditional data collection questionnaires enabling cultural analysis. In John P. McCrae, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Abgaz, Y., Dorn, A., Piringer, B., Wandl-Vogt, E., and Way, A. (2018b). Semantic modelling and publishing of traditional data collection questionnaires and answers. *Information*, 9(12).

Barabas, B., Hareter-Kroiss, C., Hofstetter, B., Mayer, L., Piringer, B., and Schwaiger, S. (2010). *Digitalisierung handschriftlicher Mundartbelege. Herausforderungen einer Datenbank. In Fokus Dialekt.* Analysieren-Dokumentieren-Kommunizieren; Olms Verlag, Hildesheim, Germany.

Bosque-Gil, J., Gracia, J., Aguado-de Cea, G., and Montiel-Ponsoda, E. (2015). Applying the ontolex model to a multilingual terminological resource. In Fabien Gandon, et al., editors, *The Semantic Web: ESWC 2015 Satellite Events*, pages 283–294, Cham. Springer International Publishing.

Bowers, J. and Stöckle, P. (2018). Tei and bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In Andrew U. Frank, et al., editors, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, pages 45–54. Gerastree proceedings.

Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9:29–51, 03.

Chiarcos, C., Cimiano, P., Declerck, T., and McCrae, J. P. (2013). Linguistic linked open data (LLOD). introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i – xi, Pisa, Italy, September. Association for Computational Linguistics.

Cimiano, P., McCrae, J. P., and Buitelaar, P. (2016). Lexicon model for ontologies: Community report. Technical report, W3C Ontology-Lexicon Community Group.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J., (2020). *Modelling Lexical Resources as Linked Data*, pages 45–59. Springer International Publishing, Cham.

Declerck, T. and Wandl-Vogt, E. (2014). Cross-linking Austrian dialectal dictionaries through formalized meanings. In Andrea Abel, et al., editors, *Proceedings of the XVI EURALEX International Congress*. EURAC research, July.

Declerck, T. (2018). Towards a linked lexical data cloud based on ontolex-lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 7–12, Miyazaki, Japan.

Khan, A. F. (2018). Towards the representation of etymological data on the semantic web. *Information*, 9(12).

Klimek, B. and Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman DICTIONARY News*.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: Development and applications. In *Proceedings of the 5th Biennial Conference on Electronic Lexicography (eLex 2017)*, page 587–597, Leiden, The Netherlands.

Rodríguez Díaz, A., Benito-Santos, A., Dorn, A., Abgaz, Y., Wandl-Vogt, E., and Therón, R. (2019). Intuitive ontology-based sparql queries for rdf data exploration. *IEEE Access*, 7:156272–156286.

Scholz, J., Bartelme, N., Fliedl, G., Hassler, M., Kop, C., Mayr, H., Nickel, J., Vöhringer, J., and Wandl-Vogt, E. (2008). dbo@ema. a system for archiving, handling and mapping of heterogeneous dialect data for dialect dictionaries. In *Proceedings of the XIII euralex International Congress*, pages 1467–1472. Documenta Universitaria.

Schopper, D., Bowers, J., and Wandl-Vogt, E. (2015). dboe@tei: Remodelling a data-base of dialects into a rich lod resource. In *Proceedings of the 9th International Conference on Tangible, Embedded, and Embodied Interaction (TEI 2015)*, Stanford, CA, USA.

Tiberius, C. and Declerck, T. (2017). A lemon model for the ANW dictionary. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference. Biennial Conference on Electronic Lexicography (eLex-17), Lexicography from scratch, September 19-21, Leiden, Netherlands*, pages 237–251. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (2018). Using rdfa to link text and dictionary data for medieval french. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 7–12, Miyazaki, Japan.

Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis*, pages 35–53, 02.

9. Language Resource References

Wandl-Vogt, Eveline. (2010). *Datenbank der bairischen Mundarten in Österreich electronically mapped [Database of the Bavarian Dialects in Austria electronically mapped] (dbo@ema)*. Österreichische Akademie der Wissenschaften.

Österreichische Akademie der Wissenschaften. (2018). *Datenbank der bairischen Mundarten in Österreich [Database of Bavarian Dialects in Austria] (DBÖ)*. Österreichische Akademie der Wissenschaften.