Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon

Sabine Tittel[*], **Christian Chiarcos**[◇]
[*]Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany
[◇] Goethe University Frankfurt, Frankfurt am Main, Germany
sabine.tittel@urz.uni-heidelberg.de, chiarcos@informatik.uni-frankfurt.de

Abstract

The adaptation of novel techniques and standards in computational lexicography is taking place at an accelerating pace, as manifested by recent extensions beyond the traditional XML-based paradigm of electronic publication. One important area of activity in this regard is the transformation of lexicographic resources into (Linguistic) Linked Open Data ([L]LOD), and the application of the OntoLex-Lemon vocabulary to electronic editions of dictionaries. At the moment, however, these activities focus on machine-readable dictionaries, natural language processing and modern languages and found only limited resonance in philology in general and in historical language stages in particular. This paper presents an endeavor to transform the resources of a comprehensive dictionary of Old French into LOD using OntoLex-Lemon and it sketches the difficulties of modeling particular aspects that are due to the medieval stage of the language.

Keywords: Linked Open Data, Ontolex-Lemon, Lexicography, Old French

1. Introduction

1.1. The Lexical Resource

The *Dictionnaire étymologique de l'ancien français* – DEAF (Baldinger, since 1971) is a longstanding dictionary compiled in Heidelberg under the aegis of the Heidelberg Academy of Sciences and Humanities. Its aim is to document and study the Old French language from its first resource 842 AD until ca. 1350 AD. To date, the publication channel of the outcome of the editorial process is twofold: The dictionary is traditionally published as a series of printed books (via LaTeX) and, since 2010, also as a versatile electronic dictionary (DEAF*él*) with on-line dictionary entries and elaborate research functions based on the XML and XHTML data exported from a MySQL database.[1]

However, DEAF*él* constitutes a data silo. The information stored can be accessed either by reading the articles or by using the research functions offered by the publication. This has the following shortcomings: Regardless of the high standard of the on-line publication, the accessibility and usability of the dictionary is to be improved. Using the dictionary may require a considerable knowledge of Old French in general and about the internal structure of the dictionary in particular. This is not necessarily given. To answer a research question (say, about the concepts of health and illness in medieval society based on Old French literature) is not an easy task for someone who is not familiar with the Old French terminology for the respective domain (here, medicine).

Also, the internal data format of such a data silo is proprietary and its publicly accessible serialization focuses solely on human consumption. It does not allow for queries that have not been foreseen a priori. Most importantly, the data format is not well suited for automatic processing.

Thus, by transforming the data into RDF and Linked Open Data (LOD), we want to emancipate the valuable dictionary outcome from the limits of such a data silo.

1.2. Facilitating Resource Interoperability with the Resource Description Framework

Following the emergence of the internet, the Resource Description Framework (Klyne et al., 2004, RDF) was developed as a standard to represent metadata, and to express relations between and statements about web resources as well as offline resources. The aim is to facilitate processability and interpretability of metadata entries, but, subsequently, also of web resources themselves. Beyond its original use case, RDF thus rose to importance as a cornerstone of the emerging Semantic Web and even beyond classical Semantic Web applications that involve reasoning, inference and formal knowledge bases. RDF established itself as a generic representation formalism for data on the web and, in particular, for the *integration* of data on the web. In this role, a rich technological ecosystem evolved and ultimately lead to the emergence of Linked Data and its adaptation in various fields, e.g., as Linguistic Linked Open Data (LLOD) in linguistics and natural language processing. Our objective here is to facilitate the usability, queriability and interpretability of DEAF data for *automated* consumption and transformation. On the basis of such automated processes, more advanced functionalities for the end user can then be developed, e.g., improved means of querying, exploring or integrating other lexical or textual data sets. Such services are our ultimate goal, and we address first steps towards the development of (L)LOD-based methodology and infrastructure for historical philologies.

RDF implements a (multi-)graph model, where nodes are connected via edges that point from a source node ('subject') to a target node ('object') and that have a particular semantic type ('property'). Source nodes, target nodes and properties are identified with URIs, e.g., objects accessible

[1] https://deaf-server.adw.uni-heidelberg.de/ [accessed 12-12-2017].

via HTTP. RDF is thus naturally suited to describe structured data on the web. In particular, this includes lexical data, as the (directed multi-)graph is generally recognized to be a generic formalism for the representation of dictionaries and machine-readable lexical resources. As such, already the Lexical Markup Framework (Francopoulo et al., 2006, LMF) built on *feature structures* (largely equivalent to directed multi-graphs, but serialized in XML), and the increasing popularity of OntoLex-Lemon (and RDF) for lexical resources mostly reflects a transition from traditional XML-based representations to RDF-based representations of the same underlying data structure (Gracia et al., 2018). In opposition to XML which provides validation on a syntactic level only, the RDF data model allows to formalize the semantics independently from constraints on their order of representation. It is thus more suitable to establish interpretability and semantic processability of the data by its subsequent users and downstream applications.

On a format level, RDF can be serialized in different ways. A common text-based representation is the Turtle format that allows to express statements in the form of *triples*, including the subject URI, the property URI and the object URI (or, alternatively, a literal value), followed by a dot. (Various shorthands are possible.) The W3C-standardized query language SPARQL basically follows the same notation for graph fragments to be retrieved but extends it with variables. In the examples below, we employ a Turtle serialization of RDF data because it is particularly well-suited for subsequent querying.

For transforming the DEAF into RDF, we implemented the following workflow: We firstly selected one exemplary dictionary article as data for a proof of concept implementation. Using this data, we defined an application profile for the dictionary entries. Secondly, we transformed the XML data of the selected article into LOD with RDF/Turtle and the OntoLex-Lemon vocabulary in line with the application profile. This step was performed manually. Thirdly, we developed a set of XSLT scripts to automatically perform this transformation step and we evaluated problematic issues within this step. We then tested the scripts with the data of the respective article and also with the data of further dictionary entries. Finally, directions for future work have been identified.

1.3. Linked Data for Lexical Resources

Linked Data has emerged as a paradigm for publishing and interlinking datasets about ten years ago. It has been a success story, leading to many datasets being published following the four Linked Data principles (Bizer et al., 2009):

- Use URIs as (unique) names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using Web standards such as RDF, and SPARQL.
- Include links to other URIs, so that they can discover more things.

Applying Linked Data principles to modeling lexical data comes with important advantages (Chiarcos et al., 2013), most notably *structural interoperability* (same format, same query language), *conceptual interoperability* (shared vocabularies), *accessibility* (uniform access, data can be accessed using standard Web protocols without additional software, etc.), *resource integration* (linking resources) and *federation* (cross-resource access).

Most important for our use case is interoperability: By resorting to RDF as data model, one achieves structural interoperability as language resources following the Linked Data paradigm are provided according to a uniform data model (in different, equivalent and convertible serializations). Conceptual interoperability, i.e., the use of shared vocabularies, is encouraged in Linked Data since its nature encourages the reuse of existing vocabularies across datasets. Following this practice thus leads to more and more datasets using the same vocabulary to describe data. Hence, it facilitates to establish interoperability on both the syntactic (format / access) and the semantic (conceptual) level.

One vocabulary that rose to particular importance for lexical resources is the *Lexicon Model for Ontologies* (Lemon). The Lemon model has originally been developed in the Monnet project to augment ontologies with rich linguistic information in order to facilitate their automated rendering in natural language (Declerck et al., 2010). Since 2012, the Ontology-Lexicon W3C Community Group has been further developing this model towards a generic data model for lexical resources, and its application to the historical lexicography of a medieval language variety is the main contribution of our paper.

Despite the growing popularity of the Linked Data paradigm in application to lexicographic resources (Witte et al., 2011; Bouda and Cysouw, 2012; Declerck et al., 2015), and in particular, adaptations of Lemon (Borin et al., 2014; Klimek and Brümmer, 2015; Bosque-Gil et al., 2016; Gracia et al., 2018), the focus of current activities in this direction lies on the modern stages of the languages. Notable exceptions in this context include etymological dictionaries, e.g., on Germanic languages (Chiarcos and Sukhareva, 2014), and dictionaries of classical languages, e.g., on Ancient Greek (Khan et al., 2017). To our best knowledge, however, these approaches take a technological focus in that they aim to demonstrate the applicability of digital methods in the humanities, rather than being grounded in philological research or traditions. This gap in research is being addressed in this paper: We present an endeavor to transform the resources of a comprehensive dictionary of Old French into LOD using OntoLex-Lemon and we evaluate the difficulties of modeling particular aspects that are due the medieval stage of the language.

1.4. The OntoLex-Lemon Data Model

In its published version from May 2016, the OntoLex-Lemon model[2] is divided into five modules: The *OntoLex core model* describes the elements that are necessary for all instantiations of the model, including lexical entries, forms

[2] `https://www.w3.org/2016/05/ontolex/` [accessed 12-12-2017]

and senses of a word. The *syntax and semantics* module describes in more detail the interaction of the syntax of words and their interpretation in an ontology. The *decomposition module* is used to describe the composition of multi-word terms and compound words. The *variation and translation module* supports the description of relationships between words and senses including translation and cross-lingual links. Finally, the *metadata module* allows for high-level descriptions of a lexicon and the number of links between elements.

The primary class in the OntoLex model is the *lexical entry*, which represents a head word in the lexicon. The lexical entry groups all forms of a word together into a single element, e.g., it includes inflected forms. For example, the entry for the Old French verb *jogler* "to ridicule someone" (< Latin JOCULĀRE v.) would include inflected forms such as *joglant*, *joglot*, *joglé*. However, the Old French noun *jogler* m. "juggler" (< Latin JOCULĀRIS adj. "funny") with a different part of speech and a different etymology would logically represent a separate lexical entry. Lexical entries are further grouped into three classes: (single) words, multiword expressions and affixes (such as *anti-*).

A lexical entry is composed of a set of lexical forms, each of which can be represented in different scripts by means of a string; one of the forms can be defined as the canonical form (i.e., the lemma). Thus, the simplest form of a lexical entry (e.g., Old French *flamesche* f. "spark") is as follows:

```
PREFIX ontolex:
       <http://www.w3.org/ns/lemon/ontolex#>

<flamesche>
  a ontolex:LexicalEntry, ontolex:Word ;
  ontolex:canonicalForm
    <flamesche#singular_form> ;
  ontolex:otherForm
    <flamesche#plural_form> .

<flamesche#singular_form> a ontolex:Form ;
  ontolex:writtenRep "flamesche"@fro .

<flamesche#plural_form> a ontolex:Form ;
  ontolex:writtenRep "flamesches"@fro .
```

The semantics of a lexical entry can be given by indicating that it *ontolex:denotes* an element in the ontology. The element in the ontology can be a class, a property or an individual. In many cases, this link to the ontology may need to be described in more detail. For this purpose, the model provides the class *lexical sense*, representing the connection between a lexical entry and its meaning in an ontology or knowledge graph. Unlike such 'semantic' entities provided by an external resource, lexical senses are specific to one particular lexical entry.

As a rule of best practice, lexical entries should be linked to ontologies via their respective lexical senses whenever an explicit definition or gloss is provided in the original dictionary. In this way, it is always possible to inspect their original definition regardless of possible (subsequent) updates of the definition (or usage patterns) of the ontological entity they refer to (Wang et al., 2011). Accordingly, lexical resources become more robust and verifiable in the face of concept drift in the Semantic Web. A simple example (extending *flamesche*) is the following:

```
PREFIX dbpedia:
       <http://www.dbpedia.org/resource/>

<flamesche> ontolex:sense
                    <flamesche#sense1> .

<flamesche#sense1> a ontolex:LexicalSense ;
   ontolex:reference dbpedia:Spark_(fire) .
```

As lexical senses are specific to individual lexical entries, lexical concepts have been added to the model to express groups of lexical senses that can be lexicalized in different ways. The exact definition of such *lexical concepts* is resource-specific, but one possiblity is to use them to represent sets of synonyms.[3] In particular, lexical concepts can be used for lexical entries that are defined with reference to (the definition of) another lexical entry, e.g., using conventional expressions such as *see also*, *cf.*, etc. However, in this case, also the definition of the referred lexical entry must be reflected as a lexical concept:

```
<flamesche> ontolex:sense
                    <flamesche#sense1> ;
  ontolex:evokes
            <flamesche#sense1_lexConcept> .

<flamesche#sense1_lexConcept>
  a ontolex:LexicalConcept ;
  ontolex:isConceptOf
                  dbpedia:Spark_(fire) ;
  ontolex:definition "petite parcelle ...
       ..., flammèche, braise légère"@fr ;
  ontolex:lexicalizedSense
                    <flamesche#sense1> .
```

2. Resource Modeling

To illustrate the modeling of a complete dictionary entry, we chose the Old French word *fiel* m. for it has an average complexity in terms of both its orthographic challenges and its semantic structure: *fiel* is the standard graphical representation of the Old French word (and is thus defined as the lemma of the entry) and it shows six more graphical realisations within the Old French literature, i.e., *fel*, *feel*, *fele*, *feil*, *fieil* and *fius*. Its semantic scope includes three main senses, i.e., "bile", "gall bladder" and, figuratively, "bitterness". The editor of the dictionary entry identified 13 sub-senses altogether, among which are collocations and metaphors (see the entry in its collapsed version in Fig. 1). Also, some of the lexical units (i.e., the entity of the lexeme *fiel* plus exactly one of its senses) are elements of the medical or the botanical terminology (e.g. in Fig. 2).

Following the core model of OntoLex-Lemon we defined the application profile for the DEAF entries. We visualize this in Fig. 3 (*fiel* with main sense n°1 "bile" [medical term]) and Fig. 4 (*fiel de la terre* "plant of the family of the common centaury, Centaurium erythraea Rafn." [botanical term], modeled as a multi-word term).

[3] This practice is not required by the model, and broader definitions are possible. In particular, a lexical concept cannot always be interpreted as a synset in the sense of WordNet (Miller, 1995).

Figure 1: DEAF*él* entry 'fiel', collapsed version.

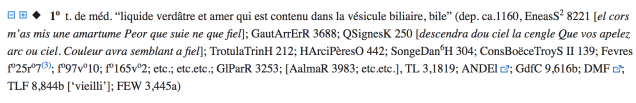

Figure 2: DEAF*él* entry 'fiel', main sense n°1, partly expanded version.

Beyond the OntoLex-Lemon core vocabulary we used classes and properties of the following ontologies: the OntoLex *decomposition module* (`decomp`[4]) to model the components of multi-word terms (`ontolex:MultiwordExpression` with `decomp:subterm`), and the OntoLex *variation and translation module* (`vartrans`[5]) to model their relations (`lexicalRel`). To model the part-of-speech categories we used the LexInfo ontology (`lexinfo`[6]), and to expressing linguistic features beyond LexInfo (e.g., referencing language registers with `TechnicalRegister`), we used OLiA (`olia`[7]). As for metadata, FOAF properties define the name and website of the editor (`name`, `homepage`), DublinCore properties refer to the extralinguistic reality (`subject`) and also facilitate non-linguistic annotation (`creator`, `publisher`, `license`, `date`). Also, we defined new classes and properties to meet particular requirements of our use case: `deaf:TechReg` (technical register) defines specialized terminology and `deaf:idem` models the case where a sub-sense 'B' of a main sense 'A' inherits A's definition (and then specifies it in a certain way). The entity `deaf:TechReg` is defined as an instance of the OLiA class `olia:TechnicalRegister`; for `deaf:idem`, we found no existing vocabulary to be applicable.[8]

For the modeling process, we prioritized the lexical information, that is, the Old French lexemes including their written representations and their senses. However, this is a first step and the modeling currently ignores other relevant information such as the information given in the etymological discussion of each DEAF entry (etymon and corresponding words in other Romance and non-Romance languages), the dating of each lexical unit, the quotations taken from the Old French texts, and more. We thus identified the modeling of the hitherto excluded data as future work.

3. Converting DEAF to RDF

3.1. Manual Transformation

Preparing the transformation, we identified the following issue: The original XML data of a DEAF entry includes information that is not modeled by the application profile. We therefore isolated the data that is relevant for the transformation into RDF. The result is as follows (extract with only two graphical forms and one sense):

```
<?xml version="1.0"?>
<xsd:schema
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:m="http://www.deaf-page.de/ns/markup"
targetNamespace="http://www.deaf-
 page.de/ns/markup">

<article author="Sabine Tittel">
<title><lemma developed="false"
 language="afr.">fiel</lemma>
 <pos>m.</pos></title>

<variant type="standard">fiel</variant>
<variant>fel</variant>
[...]

<sense><description>
 <m:terminology type="medecine">
 t. de m&#xE9;d.</m:terminology>
 <m:definition>liquide verd&#xE2;tre et
 amer qui est contenu dans la
 v&#xE9;sicule biliaire,
 bile</m:definition></description>
</sense>
```

We then manually transformed the data of the entry *fiel* into RDF/Turtle. Finally, we reviewed the data using standard validation tools.

3.2. Automated Conversion

The application profile and the RDF data of *fiel* then served as a model for the creation of a set of XSLT scripts. In

[4] `http://www.w3.org/ns/lemon/decomp`.

[5] `http://www.w3.org/ns/lemon/vartrans`.

[6] `http://www.lexinfo.net/ontology/2.0/lexinfo`, Cimiano et al. (2011).

[7] `http://purl.org/olia/olia.owl`, Chiarcos and Sukhareva (2015).

[8] In particular, the `skos:broader` property of the Simple Knowledge Organization Scheme (Miles and Bechhofer, 2009) does not seem to be applicable as it should hold between SKOS concepts rather than between individuals. Accordingly, the former use of `skos:broader` within Monnet-Lemon has been considered deprecated and removed from the Ontolex-Lemon community report.

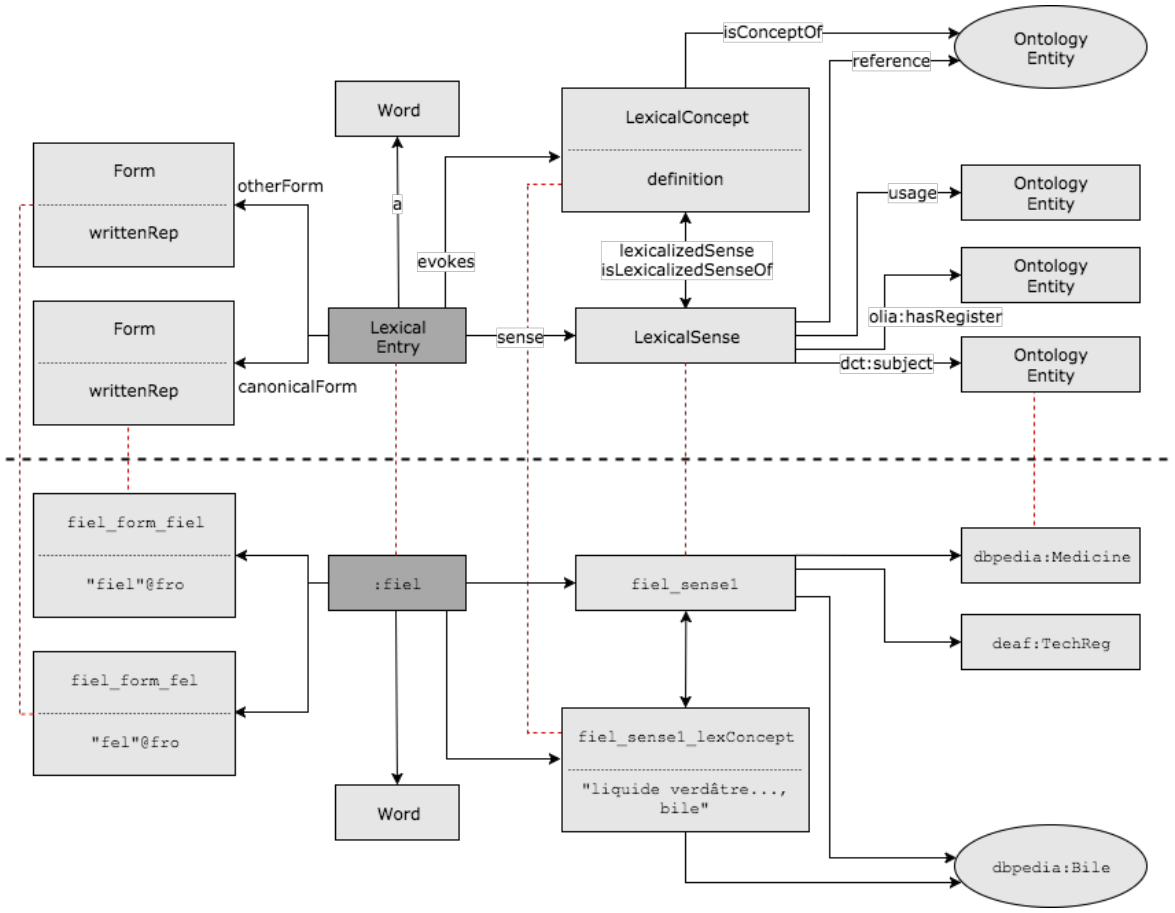

Figure 3: Model of DEAF entry 'fiel' with main sense n°1.

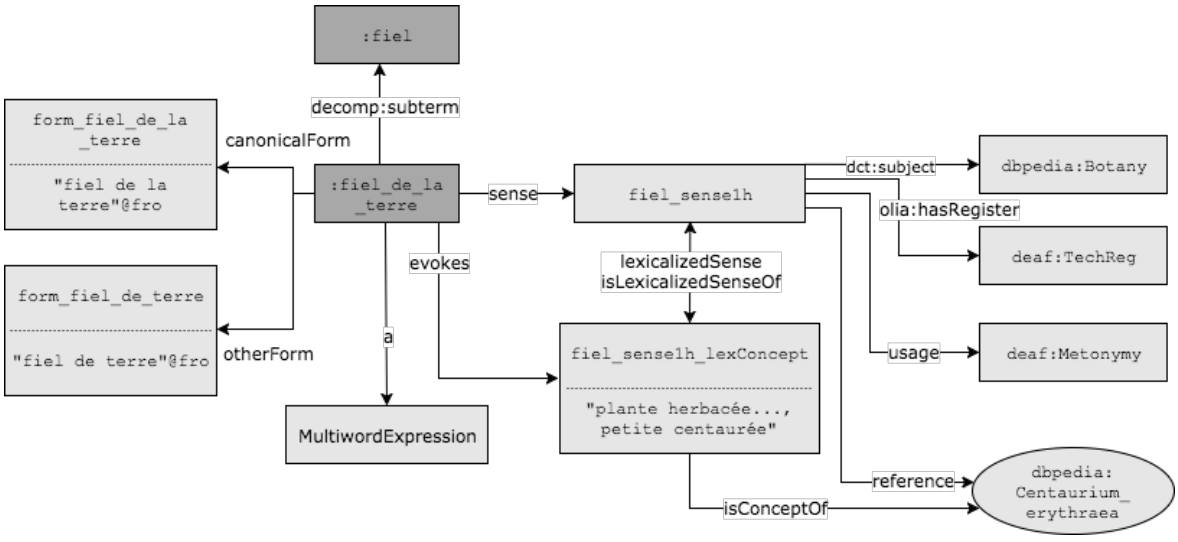

Figure 4: Model of multi-word term 'fiel de la terre'.

order to be able to eventually convert the total of approx. 83,000 dictionary entries, these scripts not only cover the specific use cases provided by our proof of concept article *fiel* but also all valid XML elements with their attributes and values defined by the XML schema of the dictionary. For example, we implemented a specific template for the automatic conversion of a given list of technical domains like medical, astronomical, musical terminology, etc. This template inserts links to the respective entity of DBPedia to define the type of terminology (using `dct:subject`, `olia:hasRegister`, `deaf:TechReg`, also literals). A fragment of the conversion template is illustrated in Fig. 5. An example of the outcome is:

```
:fiel_sense1
  a ontolex:LexicalSense ;
  dct:subject dbpedia:Medicine ,
    "medicine"@eng ;
  olia:hasRegister deaf:TechReg ,
    "t. de méd."@fr .
```

It should be noted that this representation aims for a middle ground between human and machine interpretability: We provide both the original information from DEAF (as a string value) and its semantic interpretation (with a URI that references external terminology repositories and knowledge bases), and in order to preserve their association, both are assigned as objects to the same property.

While this representation is lossless and allows to trace entity links in a (relatively) user-friendly, compact and unambiguous fashion that particularly facilitates their debugging, this is semantically valid only in the context of the general RDF data model. Nevertheless, it should be avoided in more strictly formalized Semantic Web languages such as OWL. However, a subsequent SPARQL Update script can easily eliminate literal values for OWL object properties such as `olia:hasRegister`, thereby deriving a more compact and semantically valid representation of DEAF that is suitable for consumption by downstream applications and users.

After conversion, we evaluated the validity of the XSLT scripts against the manually produced RDF data of *fiel*. In addition, a random sample of five further DEAF entries, resp. their Linked Open data conversion has been manually inspected and verified, indicating the applicability of our converter to DEAF data structures. However, this conclusion must be taken with a grain of salt when it comes to linking with external resources.

4. Linking DEAF Data

While data *structures* can be seamlessly converted to RDF, the generated outcome cannot always easily be put into relation with external knowledge bases. In particular, we find that the *Historical Semantic Gap* prohibits an unreflected and fully automated transition of philological resources of historical language stages with concept stores developed for present day applications and data: The mapping of a lexical unit to the correct entity in an ontology is a difficult task that cannot be automated for the Old French lexis. The reason for this is the historical dimension and the semantic gap lying therein: The extralinguistic concept of medieval reality denoted by a word in Old French oftentimes differs from the extralinguistic concept of modern reality denoted by the same word in modern French, e.g., because certain medical coherences were not yet known: For a 13th century doctor, *function of the brain* does not mean the same as for a 21st century one.

To overcome this problem, we implemented a semi-automatic process: This includes an automatic pre-processing as a time-saving preparation for a manual post-processing. The XSLT scripts place a wildcard (a simple `XXX`) where the entity of an ontology then needs to be specified by a linguist specialized in Old French lexical semantics. His expertise assures the correct mapping.

We believe it is possible to further enhance the automatic part of the procedure. For example, the sense definition of a botanical term is by default given in modern French but also includes the scientific Latin term of the plant. This term is usually taken from the *Systema naturae* by Carl von Linné (abbr. 'L.') or, less commonly, from the taxonomy by Carl Gottlob Rafn (abbr. 'Rafn', see above for *fiel de la terre*). We foresee an automatic mapping of these definitions to the entity in, e.g., DBPedia based on the scientific Latin term.

5. Discussion and Outlook

So far, we described the application of the OntoLex-Lemon model to modeling a reference resource for Old French lexicography as RDF, resp. its automated conversion to Linked Data – as well as limitations of a fully automated approach. To our best knowledge, this is the first broad-scale application of the Linked (Open) Data paradigm to a standard resource for medieval lexicography. We are aware of related activities on lexicographic resources for other language families, but we understand that these operate on the level of pilot studies, at the moment. Notable related work on medieval French beyond lexicography includes the Syntactic Reference Corpus on Medieval French (SRCMF[9]) use an RDF database as a backend for annotation graphs, albeit as an internal representation only, and without links to LOD resources. In fact, the actual data of the SRCMF is disseminated in a conventional XML format (Brants et al., 2004).[10]

The development of a LOD edition for the DEAF is conducted with the more general aim to transform the dictionary data into a sustainable and more easily re-usable format. The publication of the RDF edition of the full DEAF under an *open* license is foreseen by the first author, yet, it requires clarification about possible restrictions on use, dissemination and licensing – for these aspects, legal confirmation has been requested but is pending. The solution proposed is to model the role of the Heidelberg Academy of Sciences and Humanities using `dct:rightsHolder`.

With the LOD edition, we pave the way for the DEAF to become a part of the LLOD cloud in general and as a potential center within a net of linguistic resources of medieval French in particular. Beyond providing a novel set of philological lexical data in compliance with Linked Data principles, we also used this data to enrich a digitally published scholarly text edition of a medical treatise written in medieval French with references to the DEAF dictionary, as further described in Tittel et al. (accepted). This emphasizes the role of the DEAF as a standard reference also for other scholarly editions of Old French and Middle French texts. The conversion of the dictionary data into RDF and its publication within the LLOD cloud shows great capability of promoting the DEAF's role as a focal point of historical French text philology.

Apart from the afore-mentioned modeling of hitherto excluded data we identified two major issues that are yet to be

[9] `http://srcmf.org`, Mazziotta (2010).

[10] The distributed SRCMF RDF data is defective in the sense that '[t]he RDF file can be used to correct the annotation in NotaBene, but you need to pair it with the XML text source file.' (`http://srcmf.org` [accessed 03-02-2018]).

```
<xsl:template name="terminology_extern">
  <!-- the subject URI has been spelled out before -->
  <xsl:choose>
    <!-- when medicine -->
    <xsl:when test="./description/m:terminology/@type='medicine' or
      ./description/m:idem/m:terminology/@type='medicine'">
      dct:subject
        dbpedia:Medicine ,
        "<xsl:value-of select="./description/m:terminology/@type"/>
         <xsl:value-of select="./description/m:idem/m:terminology/@type"/>"@eng ;
      olia:hasRegister
        deaf:TechReg ,
        "<xsl:value-of select="./description/m:terminology"/>
         <xsl:value-of select="./description/m:idem/m:terminology"/>"@fr ;
    </xsl:when>
    <!-- when astronomy -->
    <xsl:when test="./description/m:terminology/@type='astronomy' or
      ./description/m:idem/m:terminology/@type='astronomy'">
      dct:subject
        dbpedia:Astronomy ,
        "<xsl:value-of select="./description/m:terminology/@type"/>
         <xsl:value-of select="./description/m:idem/m:terminology/@type"/>"@eng ;
      olia:hasRegister
        deaf:TechReg ,
        "<xsl:value-of select="./description/m:terminology"/>
         <xsl:value-of select="./description/m:idem/m:terminology"/>"@fr ;
    </xsl:when>
    <!-- etc. -->
  </xsl:choose>
</xsl:template>
```

Figure 5: XLST fragment for automated DEAF conversion.

addressed: language identification and sense hierarchies.

Language identification: The first issue concerns the modeling of the lemma and the (ortho)graphical variants of the respective word. We identify the Old French language in line with the International Standard for Language Codes ISO 639, i.e. with the ISO 639 code 'fro'.[11] We thus modeled the lemma and the variants using the OntoLex-Lemon vocabulary in the following way (*fiel* is the lemma = `canonicalForm`, *fel* is one variant = `otherForm`):

```
:fiel ontolex:canonicalForm
  :fiel_form_fiel .
:fiel_form_fiel a ontolex:Form ;
  ontolex:writtenRep "fiel"@fro .
:fiel ontolex:otherForm :fiel_form_fel .
:fiel_form_fel a ontolex:Form ;
  ontolex:writtenRep "fel"@fro .
```

However, it must be noted that – similar to the medieval stage of other Romance languages – Old French does not have a consistent orthographic norm. Each scribe of a manuscript realized the sound of a word in his own fashion, influenced by random circumstances but also by his dialect that could differ significantly from what we now consider the standard Old French language. As a consequence, we find a great variety of spellings for the same word.[12] Whenever a graphical variant is characteristic of a particular Old French scripta (i.e., the written form of a spoken dialect), the editor of the dictionary entry explicitly annotates it within the XML data of the entry. As a result, e.g., the entry *faisse*, designating a sort of ribbon or strap, lists Lorraine *faixe*, Anglo-Norman *fees*, and Picard *fasse* among the graphical variants.[13] Unfortunately, ISO 639 does not provide codes for Old French dialects,[14] and therefore, we provisionally identified all Old French dialects as standard 'fro'. But this is an intermediate solution because it ignores information that is very valuable for the research of Old

[11] `https://www.iso.org/iso-639-language-codes.html` [accessed 12-12-2017].

[12] The word with the highest number of attested variant spellings to date is the Old French adverb *iluec* "there" with more than 120 variants, see `https://deaf-server.adw.uni-heidelberg.de/lemme/iluec` [accessed 12-12-2017].

[13] `https://deaf-server.adw.uni-heidelberg.de/lemme/faisse` [accessed 12-12-2017].

[14] Varieties of historical language variants have been within the focus of ISO 639-6, which was, however, withdrawn as a standard in 2014, cf. `https://www.iso.org/standard/43380.html`. One possible alternative would be Glottolog `http://glottolog.org`, which does, however, take a focus on language documentation and is not appropriate for the needs of philologists. As an example, it conflates diachronic and dialectal criteria within a single hierarchy: The Romance language family is considered a subclass of Imperial Latin (as is, for example, Classical Latin), where – in fact – it evolved from it. Yet, this defective kind of modeling is not systematic, as Old Latin is a cousin of Imperial Latin rather than its ancestor/superclass.

French dialects. This information is given in the XML data but is lost in the LOD version. The solution to this shortcoming of the ISO 639 standard is to define the code 'fro' as a macrolanguage and to register the Old French dialects as varieties associated to 'fro'. A valid list of dialects is provided by the XML schema of the DEAF.

Sense relations: The second issue concerns the complex semantic relations between main senses and associated subsenses within the sense tree of a DEAF article. The hierarchical structure and the order of the sub-senses mirrors the semantic change the lexeme has undergone: It considers all figures of speech, e.g., metaphor, metonymy, irony, image, hyperbole, allegory, euphemism, etc. For each lexical unit of the respective lexeme the semantic relationship is explicitly expresses by, e.g., 'par métaph.', 'par méton.', 'par ironie'. This information is of great value for the study of semantic shift. We therefore attempt to model the semantic relationships expressed in the semantic tree. However, the properties of established vocabularies seem insufficient to do so. SKOS[15], for example, only offers two properties to model sense restriction and sense enlargement respectively: `narrower` and `broader`. In default of a more detailed range of properties we modeled the sense relations using the information contained in the XML data: 'par métaph.', etc. We implemented a template that automatically reads this information and transforms it into the respective RDF data using the OntoLex-Lemon property `usage` and a link to DBPedia. In the following we present an extract of this template:

```
<xsl:template name="usage_extern">
<xsl:choose>
<xsl:when test="./description/m:usage/
  @type='metaphor' or
  ./description/m:idem/m:usage/
  @type='metaphor'">
  ontolex:usage dbpedia:Metaphor ,
  "<xsl:value-of select="./description/
  m:usage"/>
  <xsl:value-of select="./description/
  m:idem/m:usage"/>"@fr ;
</xsl:when>

<xsl:when test="./description/m:usage/
  @type='irony' or
  ./description/m:idem/m:usage/
  @type='irony'">
  ontolex:usage dbpedia:Irony ,
  "<xsl:value-of select="./description/
  m:usage"/>
  <xsl:value-of select="./description/
  m:idem/m:usage"/>"@fr ;
</xsl:when>
</xsl:template>
```

An example of the outcome is:

```
:fiel_sense1.d a ontolex:LexicalSense ;
  ontolex:usage dbpedia:Metaphor ,
  "métaph."@fr .
```

[15] `http://www.w3.org/TR/skos-reference/#semantic-relations` [accessed 12-12-2017].

Aside from addressing the aforementioned shortcomings of established community standards, one direction of future research is to improve the linking with other lexical resources. We have to note, however, that the philological perspective entails that first-class citizens for such a linking would be dictionaries of historically or linguistically related language varieties. Such a linking requires also historical resources to become increasingly available within the LLOD cloud. Our own research represents a step in this direction, and, by demonstrating the feasibility, we hope to encourage others to work in this direction as well. In particular, we expect similar challenges to arise on other datasets from historical philologies, so that in the immediate future, a focus should be laid on developing rules of best practice and specifications for this particular community before we can expect a greater degree of convergence.

A linking with language resources for modern varieties, on the other hand, would be technologically more feasible, but the theoretical and philological implications of such a linking requires a theoretical reflection in order to avoid mislinkings and incorrect interpretations arising from the Historical Semantic Gap.

We intend to publish the converted dictionary under an open license. However, we have to admit that legal clearance is still underway. Unfortunately, this situation is symptomatic for many valuable resources in the historical philologies, which are characterized by massive collaboration, long-term projects, often involving several institutions and complicated publication agreements for the underlying print edition.

6. Acknowledgements

Sabine Tittel is a full time redactor of the dictionary DEAF, Heidelberg Academy of Sciences and Humanities. The contribution of the second author was supported by the project "Linked Open Dictionaries" (LiODi), an Early Carreer Research Group funded by the eHumanities programme of the German Federal Ministry for Education and Research (BMBF).

The OntoLex-lemon edition of the data was supported by the organizers and participants of the 2nd Summer Datathon on Linguistic Linked Open Data (SD-LLOD 2017), June 2017, Cercedilla, Spain, where the linking of DEAF data was explored in a collaborative effort. This paper elaborates and builds on these experiments. In particular, we would like to thank Yifat Ben-Moshe (K Dictionaries, Tel Aviv), Helena Bermúdez-Sabel (Universidad Nacional de Educación a Distancia, Madrid), Mariana Curado Malta (Polytechnic Unicersity of Porto, Portugal), Frances Gillis-Webber (University of Cape Town) and Maxim Ionov (Goethe-University Frankfurt, Germany) for their input and contributions.

Furthermore, we would like to thank the anonymous reviewers for helpful comments and insightful feedback.

7. Bibliographical References

Baldinger, K. (since 1971). *Dictionnaire étymologique de l'ancien français – DEAF*. Presses de L'Université Laval / Niemeyer / De Gruyter, Québec, Canada /

Tübingen / Berlin, Germany. [Kurt Baldinger (founder), continued by Frankwalt Möhren, published under the direction of Thomas Städtler; electronic version DEAF*él*: `https://deaf-server.adw.uni-heidelberg.de`].

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22.

Borin, L., Dannells, D., Forsberg, M., and McCrae, J. (2014). Representing Swedish Lexical Resources in RDF with Lemon. In Matthew Horridge, et al., editors, *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, pages 329–332, Aachen, Germany. CEUR-WS.org.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado de Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries Case. In Ilan Kernerman, et al., editors, *Proceedings of GLOBALEX'16 Workshop at LREC'15, Portoroz, Slovenia*, pages 65–72, Aachen, Germany. European Language Resources Association.

Bouda, P. and Cysouw, M. (2012). Treating Dictionaries as a Linked-Data Corpus. In Christian Chiarcos, editor, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 15–23. Springer, Berlin / Heidelberg, Germany.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Chiarcos, C. and Sukhareva, M. (2014). Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41.

Chiarcos, C. and Sukhareva, 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 Linguistics: Lexical Linked Data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cimiano, P., 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(1):29–51.

Declerck, T., Buitelaar, P., Wunner, T., McCrae, J., Montiel-Ponsoda, E., and de Cea, A. (2010). lemon: An ontology-lexicon model for the Multilingual Semantic Web. In *W3C Workshop: The Multilingual Web - Where Are We?*, Madrid, Spain, Oct.

Declerck, T., Wandl-Vogt, E., and Mörth, K. (2015). Towards a Pan European Lexicography by Means of Linked (Open) Data. In Iztok Kosem, et al., editors, *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 342–355, Ljubljana/Brighton, 8. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 233–236, Genoa, Italy.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. (2018). The Apertium Bilingual Dictionaries on the Web of Data. *SWJ (Semantic Web Journal)*, 9(2):1–10.

Khan, F., Bellandi, A., Boschetti, F., and Monachini, M. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using OntoLex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*, pages 43–50, Galway, Ireland.

Klimek, B. and Brümmer, M. (2015). Enhancing Lexicography with Semantic Language Databases. *Kernerman DICTIONARY News*, 23:5–10.

Klyne, G., Carroll, J., and McBride, B. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation.

Mazziotta, N. (2010). Building the Syntactic Reference Corpus of Medieval French using NotaBene RDF annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW-2010)*, pages 142–146, Uppsala, Sweden, August. Association for Computational Linguistics.

Miles, A. and Bechhofer, S. (2009). SKOS Simple Knowledge Organization System reference. W3C Recommendation. Technical report, World Wide Web Consortium.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November.

Tittel, S., Bermúdez-Sabel, H., and Chiarcos, C. (accepted). Using RDFa to link text and dictionary data for Medieval French. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018)*, Miyazaki, Japan, May.

Wang, S., Schlobach, S., and Klein, M. (2011). Concept drift and how to identify it. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265.

Witte, R., Kappler, T., Krestel, R., and Lockemann, P. C. (2011). Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management. In Caroline Sporleder, et al., editors, *Language Technology for Cultural Heritage*, pages 213–230. Springer, Berlin / Heidelberg, Germany.