Using RDFa to Link Text and Dictionary Data for Medieval French

Sabine Tittel,* **Helena Bermúdez-Sabel,◇** **Christian Chiarcos°**
*Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany
◇Laboratorio de Innovación en Humanidades Digitales, UNED, Madrid, Spain
°Goethe University Frankfurt, Frankfurt am Main, Germany
sabine.tittel@urz.uni-heidelberg.de, helena.bermudez@linhd.uned.es,
chiarcos@informatik.uni-frankfurt.de

Abstract

This paper presents an endeavor to transform a scholarly text edition (of a medical treatise written in Middle French) into a digital edition enriched with references to an on-line dictionary. Hitherto published as a book, the resulting digital edition will use RDFa to interlink its vocabulary with the corresponding lexical entries of the *Dictionnaire étymologique de l'ancien français* (DEAF). We demonstrate the feasibility of RDFa for the semantic enrichment of digital editions within the philologies. In particular, the technological support for RDFa excels beyond domain-specific solutions favored by the TEI community. Our findings may thus contribute to future technological bridges between TEI/XML and (Linguistic) Linked Open Data resources.

The original data of the edition is available in a LaTeX format that includes profound semantic markup. We convert this data into XML/TEI, and integrate RDFa-compliant attributes for every lexeme attested in the text. The HTML5 edition generated from the XML sources preserves the RDFa attributes and thus (a) embeds (links) its vocabulary within the overall system of the medieval French language, and that

(b) provides and displays linguistic features (say, sense definitions given in the original corpus data) along with the critical apparatus of the original book publication.

Keywords: TEI, RDFa, LaTeX, Digital Scholarly Text Edition, Lexicography, Old French, Middle French.

1. Introduction

In the field of textual philology of Medieval French, the traditional way to print books from a scholarly text edition is still the publication method of choice. The number of digitally published editions of Old and Middle French is growing at a pace that is rather slow compared to the text editions of modern texts. The results are embracing a wide spectrum of approaches from retro-digitization (Reisdoerfer, 1996a; Reisdoerfer, 1996b; Harsch, 1996) to digital-born on-line editions whose creation demands for a big investment of time and resources (Eley et al., 2005; Nichols and Choudhury, 2017; Laidlaw, 2015). However, regardless of their format, these publications are almost exclusively stand-alone products. To a greater or lesser degree, the reader of a stand-alone edition of Old and Middle French texts is thus left alone with the attempt to understand the text. On its own, an edition is often deprived of a means that helps to explain the meaning of the vocabulary in a broader context: Neither is the vocabulary related to a more comprehensive lexicon of Old and Middle French, nor does a mere glossary reveal the significance of its specific vocabulary within the history of the language. Only very few digital editions provide a more comprehensive access to such information, for example *The Online Froissart. A Digital Edition of the Chronicles of Jean Froissart*, Version 1.5, edited by Peter Ainsworth and Godfried Croenen, Sheffield (HRIOnline) 2013, `http://www.hrionline.ac.uk/onlinefroissart/` [accessed 01-03-2018]: With a specific preference (*Global Viewing Mode 'DMF'*), this on-line edition links its lexemes with the entries of the *Dictionnaire du Moyen Français* – DMF (Martin, 2015).

However, conventional web editions provide a human-readable view, optimized for a particular way of visual arrangement, whereas the underlying semantics are left implicit (and thus cannot be easily recovered for applications to be developed on top of these resources, say, tools for the automatically supported linguistic analysis of Medieval French). The Resource Description Framework (RDF)[1] is designed as a means to complement or to replace conventional hyperlinks with *semantically typed* links. However, the state of the art in this area is defined by the XML specifications of the TEI (TEI Consortium, 2017b), and these provide only limited support for integrating RDF data structures or references to Linked Data resources. An alternative which – to our best knowledge – has not previously been applied in our domain would be to extend conventional XML representations with RDFa (Herman et al., 2015), a formalism that provides specifications to extend XML formats with arbitrary RDF links. This way of application, however, entails that the TEI vocabulary is to be extended by RDFa attribute definitions *within its own namespace*.

Hence, we strive for a dual objective in this paper:

- We describe the transformation of a scholarly text edition (whose original data is encoded with LaTeX markup) into a high-quality on-line edition in a time-saving manner.
- We further evaluate the possibility of using RDFa to provide explicit semantics for the links between this on-line publication and the entries of a reference dictionary.

[1] RDF 1.1 Primer, W3C Working Group Note 24 June 2014, `https://www.w3.org/TR/rdf11-primer/` (work in progress) [accessed 01-06-2018].

As a concrete benefit of an RDFa-extended edition of the edited text, these links can be automatically processed by RDFa parsers such as `http://sparql.org` [accessed 01-06-2018]. Also, together with other resources pointing to the same dictionary (or, if available, an RDF edition of the dictionary itself), they can be integrated into larger data bases that combine lexical information from multiple editions.

With the interlinking of the lexemes in the text with the entries in the dictionary we are able to embed the vocabulary within the overall system of the medieval French language as it is established by our reference dictionary. This defines their place within the semantic framework given in the dictionary articles.

Our main contribution is to demonstrate the feasibility of RDFa as an instrument of semantically enriched digital editions within the philologies. In particular, the technological support for RDFa excels beyond domain-specific solutions favored by the TEI community. Our findings may thus contribute to the development of technological bridges between TEI/XML and (Linguistic) Linked Open Data resources.

2. Text and Dictionary

We demonstrate our approach for an edition of a part of the Middle French *Grande Chirurgie* composed by Gui de Chauliac. The reference dictionary is the comprehensive etymological dictionary of the Old French language, the *Dictionnaire étymologique de l'ancien français* – DEAF (Baldinger, since 1971) which is also relevant for our Middle French text. Its on-line publication, DEAF*él*, offers dictionary articles for the Old French lexis in open access, see `https://deaf-server.adw.uni-heidelberg.de/` [accessed 12-12-2017].[2]

Gui de Chauliac is one of the most widely known physicians and medical authors of the French Middle Ages. His opus, the Latin *Chirurgia magna* from 1363 AD (*Grande Chirurgie* in its French translation), is considered the most profound compendium of the medical knowledge of its day. In seven treatises, Gui de Chauliac describes the human anatomy, tumors and cancer, wounds and fractures, the plague, eye, ear, and dental pathology, etc. The work was translated into French presumably soon after 1363 and the oldest French manuscript transmitting this medical milestone dates from the 2nd third of the 15th century.

The *Grande Chirurgie* is of particular relevance for the history of the French language as well as for both the history of medicine and mentality: On the one hand, the content of the *Grande Chirurgie* passes on the medical beliefs propagated by the authorities of Classical Antiquity, e.g., the concept of humoralism by Galen of Pergamon. On the other hand, it is a valuable witness of a critical re-examination and cautious overcoming of antique explanatory models and movement towards a modern understanding of the body.

The impact of the text motivates our scheme: To enable the reader of the text edition to better understand its valuable content we relate the vocabulary of the *Grande Chirurgie* to the respective dictionary entries of the DEAF and to their etymological, semantic and encyclopedic discussions.

3. Original LaTeX data

The critical text edition of the first treatise of the French translation, the *Anathomy*, serves as our corpus data. The text edition is hitherto published as a book (Tittel, 2004); in the following, we cite this edition with its siglum GuiChaulMT.[3] We have access to the original format of the data, which comes as a text file with LaTeX markup (example 1). This markup has been the basis for a fully automated generation of the text edition (PDF) in 2004 including a critical apparatus (we used the Edmac package for typesetting scholarly critical editions[4]) and an index of the vocabulary (with approx. 1,350 entries). Also, the markup includes the semantic information to create an index of personal names, i.e., of the authorities of classical and medieval medicine that are invoked by Gui de Chauliac to underpin his expertise, e.g., Galen of Pergamon and Avicenna. The typesetting has been executed by a tailor-made LaTeX class, see Fig. 1 and Fig. 2 for the published result.

30 guemant. Et c'est ce que Henry de Mondeville argue ou premier de sa ‹Cirurgie› par ceste maniere ci: tout mestre doit savoir et cognoistre le subjet en quoy il fait son eovre, car autremant, en ouvrant, il erre. Mais le cirurgien est mestre de santé de corps humain, donc le cirurgien doit savoir la nature et la composicion de corps humain et, par consequent, il doit savoir la anathomie. Vecy la seconde
35 rayson par similitude: car c'est comparacion semblable d'ung aveugle qui coupe et tranche bois – ainsi que por fere une ymage – et d'ung cirurgien qui veult coper ou tranchier en corps humain quant il ne scet la anathomie. Car l'aveugle

12 curer] Erstes *r* über der Zeile nachgetragen. 16 car… d'icelles] Unvollständig übersetzt (cp. GuiChaulJL 19,21s. *quia curam oportet diversificare secundum differentias ipsarum*). 16 les parties] Ms. *laes parties*. 18 ont] *ont* l. *sont*? 32 le cirurgien] Ms. *les cirurgien*, *s* expungiert. 34 il] Über der Zeile nachgetragen. 36 por] *o* überschreibt *a*.

Figure 1: Detail of the text edition, GuiChaulMT p. 78.

parler v.intr. 'articuler les sons d'une langue naturelle; parler' 592 ▪ ***parler de*** v.tr.indir. 's'entretenir de; parler de' 2; 3; 5; 172; 293; 897; 1236 *parlle* 3.p.sg. ind.prés. 482 v.intr.: TL 7,286; GdfC 10,278a; FEW 7,606a sub PARABOLARE 'sprechen'; AND 497a. – ***parler de*** v.tr.indir. : TL 7,288; GdfC 10,278c; FEW 7,606a; AND 497a.

Figure 2: Detail of the glossary, GuiChaulMT p. 371.

The first step towards our aim was thus the conversion of the original data format into an XML representation. The source material provides excellent preconditions because it contains valuable information that we can use for automated processing: The LaTeX annotation comprises the lemmatization of the vocabulary (i.e., the alignment of the individual graphical realization of a lexeme with the spelling that is accepted as the Old French standard (Möhren, 2015)), word-category disambiguation (part-of-speech tagging) and also word-sense disambiguation (semantic tagging). The annotation is performed on the level of the lexical unit, i.e., the entity of one lexeme plus ex-

[2] A series of articles of a specific article type (Tittel, 2010) within the letters D–F underlies temporary access restrictions defined by the publishing house De Gruyter (Berlin). The articles in question are available in open access with a delay of two years after their book publication.

[3] See the dictionary's bibliographical supplement DEAFBibl*él* at `http://www.deaf-page.de/bibl/bib99g.php#GuiChaulmT` [accessed 12-12-2017].

[4] The latest version is Eledmac, see `https://ctan.org/pkg/eledmac` [accessed 12-12-2017].

actly one of its senses (Blank, 1997, 6).[5] The information is coded as follows:

(1)
```
1  Ou nom\wdx{nom}{m. 'mot servant à
2  désigner les êtres, les choses qui
3  appartiennent à une même catégorie
4  logique'}{\textbf{au nom de}
5  \emph{'en vertu de'}} de Dieu
6  misericord\wdx{misericort}{adj.
7  'qui a de la miséricorde;
8  miséricordieux'}%
9  {misericord \emph{m.sg.}}.
10 [...]
11 de l'anathomie\wdx{⋆anatomie}{f.
12 1\hoch{o} 'structure et
13 composition du corps humain et
14 animal, et, en parlant dans un
15 sens abstrait, science de cette
16 structure'}{anathomie}
17 [...]
18 les passions\wdx{passïon}{f.
19 'souffrance physique'}{}
20 d'icelles, \text{car selon la
21 difference d'icelles}%
22 \lemma{car{\dots} d'icelles}%
23 \fnb{Unvollständig übersetzt
24 (cp. GuiChaul\textsc{jl} 19,21s.
25 \emph{quia curam oportet
26 diversificare secundum
27 differentias ipsarum}).}/%
28 \wdx{car}{conj. qui unit à une
29 proposition une proposition
30 suivante qui donne la raison de
31 ce qu'affirme la première}{}
32 [...]
33 si co\emph{m}me
34 Galen\adx{Galien}{}{Galen} le dit
```

In the preceding example, the `\wdx{}{}{}` command produces the index entry and consists of three parameters: #1 contains the lemma, #2 contains the part-of-speech and the sense definition including information about technical terminology, #3 contains, as a rule, the attested graphical realization of the lexeme in case it differs from the lemma (often together with additional part-of-speech information, as in `misericord \emph{m.sg.}`, line 9); an exception is made to note multi-word terms also in #3 (as in `\textbf{au nom de}`, line 4-5). The `\adx{}{}{}` command produces the index of personal names, parameter #1 containing the normalized form and #3 the graphical realization in our text (#2 being idle).

The Edmac package offers three commands to organize the critical apparatus: `\text{}` marks the word or text passage in the main text that is repeated in the apparatus, `\lemma{}` is used whenever the word or text passage of the main text shall not be congruent with the one repeated in the apparatus, e.g., when it is abbreviated as in `\lemma{car{\dots} d'icelles}`, line 22, and `\fnb{}` contains the note printed in the apparatus.

[5] The so-called small words with a marginal status in the Middle French lexis (some pronouns, prepositions, articles: *le*, *la*, *de*, *pour*, *a*, etc.) are not annotated. It is thus not a full but a partial semantic disambiguation (Habert et al., 1997, 74). Also, only the first ten attestations of each lexical unit are annotated. For many lexemes, this is equal to a complete annotation, but very commonly used lexemes occur more than ten times within the text.

4. Conversion to XML/TEI

Through a PERL script, we first transformed the LaTeX edition into an XML representation that renders the original data structures in XML. In a second step, an XSLT script is applied to restructure this intermediate XML representation in compliance with TEI P5 (TEI Consortium, 2017b). The main challenge of this step was to transform the presentational, LaTeX-based markup into a descriptive, TEI-compliant one.

As explained in Sec. 3., the implementation of the Edmac package provided the LaTeX edition with semantic markup. Even if the goal of that package is to produce an accurate presentation of the critical apparatus, it forces the editor to explicitly formalize the different parts of the apparatus. As a result, the conversion of the Edmac markup to the TEI Apparatus Module (TEI Consortium, 2017a) was straightforward.

In a similar manner, the commands that enable the indexation of the lexemes in the LaTeX source file were used to add the linguistic layer of analysis. Nevertheless, this step provided more challenges than the previous one since it engages more acutely with presentational markup. For instance, each command to display the script in italics or in bold type respectively had to be semantically interpreted in order to be mapped with the correct TEI element. Inferences based on the context had to be made: For example, the text in bold type inside the third parameter of the command `\wdx{}{}{}` should be interpreted as a collocation, and the following text in italics, together with single quotation marks, as the sense definition of this very collocation. In `\wdx{nom}{...}{\textbf{au nom de} \emph{'en vertu de'}}` (in example 1, line 4-5), the LaTeX class defines the third argument as being typeset in italics; thus, the `\textbf`-command simply set the italics in bold type but the sense definition needed an additional italics command to be set in roman type as a result. The result is coded as follows:

(2)
```
1  <seg about="http://www.deaf-
2       page.de/guichaul.html/#1"
3       property="rdfs:seeAlso"
4       resource="https://deaf-server.
5       adw.uni-heidelberg.de/lemme/nom">
6    <w property="rdfs:label"
7       lemma="nom" type="m.">nom</w>
8    <gloss property="skos:definition">
9    mot servant à désigner les
10   êtres, les choses qui
11   appartiennent à une même
12   catégorie logique</gloss>
13   <note>
14   <span type="collocation">
15   au nom de</span>
16   <gloss property=
17     "skos:definition">en vertu
18     de</gloss></note>
19 </seg>
```

5. Integrating RDFa Links

We then introduced RDFa-compliant attributes (Herman et al., 2015) to represent semantically typed links between text and dictionary. RDFa is a means of expressing RDF-style relationships, i.e., a subject-predicate-object model, using simple attributes in existing markup languages. RDFa requires its attributes to be included into the namespace of its host language. A syntactically sound (and valid) TEI/XML+RDFa representation thus requires an extension of the current TEI vocabulary, which – given its current expressive wealth which fosters a well-justified conservatism regarding possible additions – can only be justified by a concrete use case that demonstrates an added value. We argue that this value can be seen in the existing infrastructural support for RDFa.

For our endeavor, the main focus of the RDFa attributes is to link the recording of each graphical realization of each Middle French lexeme and its corresponding lemma to the DEAF dictionary (see Fig. 3). In order to add the RDFa layer we implemented a small number of modifications to the TEI schema: We added `@about`, `@property` and `@resource` to the attribute class `att.global.linking`.[6] Those three attributes suffice to describe our data through the RDFa syntax. A subject IRI reference is indicated using `@about`. For its part, the attribute `@property` works as a predicate. Finally, the objects which are IRI references are represented using `@resource`, while objects that are literals consist on the textual node of the element.

In the following excerpt of our TEI edition (example 3), we present the encoding of the lemmatized words of the corpus. Each analyzed segment (`seg[@about]`) contains an instance of a word with its gloss (`w` and `gloss` respectively). For each word we present its graphical realization in the text (that is, the object of the property `rdfs:label`), its lemma and its part-of-speech categorization (`@lemma` and `@type`). The gloss is presented with the property `skos:definition`. For every lemma that is available as an entry in the DEAF, we introduced the property `rdfs:seeAlso` with the link to DEAF as the value of the attribute `@resource`. The same information is available in the HTML version (example 4).

(3)
```
<seg about="http://www.deaf-
     page.de/guichaul.html/#8"
     property="rdfs:seeAlso"
     resource="https://deaf-
     server.adw.uni-
     heidelberg.de/lemme/anatomie">
  <w property="rdfs:label"
     lemma="*anatomie"
     type="f.">anathomie</w>
 <gloss property="skos:definition">
  structure et composition du corps
  humain et animal, et, en parlant
  dans un sens abstrait, science de
  cette structure</gloss>
</seg>
```

(4)
```
<span about="http://www.deaf-page.de/
guichaul.html/#8"
property="rdfs:seeAlso"
resource="https://deaf-server.adw.uni-
heidelberg.de/lemme/anatomie">
<span property="rdfs:label">
<a href="https://deaf-server.adw.uni-
heidelberg.de/lemme/anatomie"
target="_blank">anathomie</a></span>
<span property="skos:definition">
structure et composition du corps
humain et animal, et, en parlant dans
un sens abstrait, science de cette
structure</span>
</span>
```

Our concern was to provide a TEI dataset that is useful also for other scholars. Therefore, our encoding strategy went about deviating as little as possible from the most common TEI conventions to present a readable file for the TEI community. For instance, to lighten the amount of information embedded we use `<prefixDecl>`. This enables the use of prefixes of the vocabularies implemented following the TEI recommendation.

Together with the linguistic layer, we also converted the commands that create the index of personal names in the LaTeX source file: To mark the authorities of classical and medieval medicine mentioned in the text we employed the FOAF vocabulary (Brickley and Miller, 2014) (see example 5). We provided an URI to all named entities linking our edition to the VIAF dataset.[7]

(5)
```
<persName typeof="foaf:Person"
about="http://viaf.org/viaf/
44299175">
<name property="foaf:name">Galien
</name></persName>
```

Finally, with regard to a more detailed set of metadata, we enriched the TEI-header by implementing the Dublin Core vocabulary (DCMI Usage Board, 2012).

6. HTML and RDF Publication

6.1. HTML5+RDFa

Due to the integration of RDFa into the TEI edition, the conversion to HTML5+RDFa via XSLT was very straightforward. In fact, as long as the original XML embedding (with TEI markup elements) of the TEI data is rendered by a corresponding embedding in HTML (with HTML markup elements, e.g., `<span>`), RDFa attributes only need to be copied (preserved) during the transformation. In particular, it is not necessary to parse a complex, TEI-specific or non-standard RDF rendering to yield valid RDFa markup (and via RDFa, other RDF serializations).

The result is an on-line edition that includes the links to the reference dictionary. It also provides both the sense definitions given for every lexical unit and the apparatus notes of the source edition in an unobtrusive manner, see Fig. 4 and 5.[8]

[6] See `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.global.linking.html` [accessed 12-16-2017].

[7] *VIAF (`http://viaf.org/`): The Virtual International*

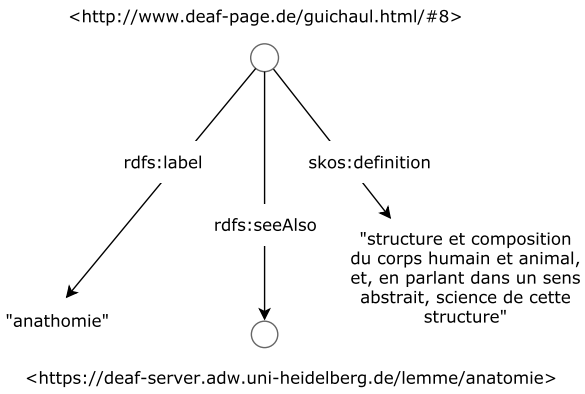

Figure 3: RDF-like visualization of an annotated word of the edition.

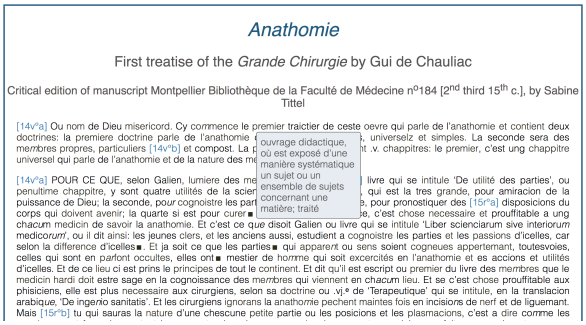

Figure 4: Detail of the on-line edition of GuiChaulmT with the display of the sense definition of *traitier* m.

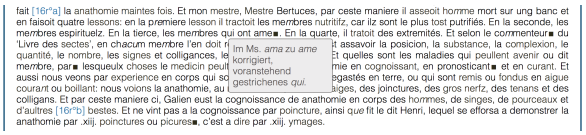

Figure 5: Detail of the on-line edition of GuiChaulmT with the display of an apparatus note.

6.2. RDFa ↦ Turtle

One of the advantages of RDFa is the ease with which one can retrieve the triples in a variety of formats, e.g., using the *RDFa 1.1 Distiller and Parser*,[9] directly from HTML+RDFa.

However, for obtaining an RDF view on a philological data set, it should not be necessary to convert a machine-readable (XML) form to a format which is primarily intended for human consumption (HTML). With RDFa already used in TEI and being simply copied into the generated HTML, also the original XML data can be fed into a RDFa processor and converted for downstream applications as needed.

Together with the TEI edition and its HTML rendering, we provide our dataset in a Turtle serialization.[10]

7. Querying Philologically Edited Text with SPARQL

With a machine-readable view established along a human-readable view, the question remains what the actual benefit of such a machine-readable version may provide for the philologist.

One advantage can be seen in the immediate benefit in queriability that this approach provides. Without the need to set up a local database infrastructure, it is possible to parse RDFa from a web page using an RDFa parser like pyRDFa.[11] Moreover, it is equally possible to query this data using a web service such as `http://sparql.org` [accessed 01-06-2018].

As an example, pyRDFa can be configured with default parameters to process a particular HTML file, say, `http://www.deaf-page.de/guichaulmTel/edition.html`.

From the Turtle file obtained in this way, one can inspect its download link to get more insight into the coding of the relevant parameters, say, `https://www.w3.org/2012/pyRdfa/extract?uri=http://www.deaf-page.de/guichaulmTel/edition.html&format=turtle`.[12] This URL may now be used locally with SPARQL Update using the `LOAD` keyword:

```
CREATE SILENT GRAPH
      <http://www.deaf-page.de/guichaulmTel>;

LOAD <https://www.w3.org/2012/pyRdfa/
      extract?uri=http%3A%2F%2Fwww.deaf-page.
      de%2FguichaulmTel%2Fedition.html&
      format=turtle>
INTO <http://www.deaf-page.de/guichaulmTel>;
```

Such queries may require to set up a local triple store, but with a one-click installation such as supported, for example, by BlazeGraph, this should not represent a hurdle.[13]

Then, it is possible to query, for example, for all attestations of a particular *dictionary* lemma:

```
FROM <http://www.deaf-page.de/guichaulmTel>
SELECT ?attestation ?form ?dictEntry
WHERE {
  ?attestation rdfs:label ?form.
  ?attestation rdfs:seeAlso ?dictEntry.
}
```

Authority File [accessed 12-28-2017].

[8] For the complete edition, see `http://www.deaf-page.de/guichaulmTel/edition.html` [accessed 01-08-2018].

[9] `https://www.w3.org/2012/pyRdfa/` [accessed 12-28-2017].

[10] Available at `http://www.deaf-page.de/guichaulmTel/edition.ttl`.

[11] See `https://www.w3.org/2012/pyRdfa/` [accessed 01-06-2018].

[12] URI encoding may apply, with the source URL escaped like `http%3A%2F%2Fwww.deaf-page.de%2F...`.

[13] Alternatively, web services for querying RDF data *without a data base* can be used, e.g., `http://sparql.org/sparql.html`.

The following sparql.org hyperlink calls a SPARQL webservice to run this query against the (dynamically generated) result of the webservice that performs the RDFa parsing of our HTML edition.

If the underlying dictionary is also provided as RDF and directly references, it is possible to load this data or to run federated queries against it. Likewise, integrating attestations from other edited texts becomes possible. The actual attestation is a URI and if this resolves against, say, an HTML anchor in the HTML+RDFa edition, the result table can be directly used to access the content of the edition.

Despite this potential, RDF technology is known to have a certain acquisition bias, so if philologists will encounter SPARQL at any point in their careers, it is likely that they eventually return to more traditional ways of philological research. Thus, the role of RDF technology and RDFa is not so much in providing a means of querying, but rather in a mechanism that allows to integrate information *within a portal* tailored to the needs of philologists. This portal – whether it provides aggregation, faceted browsing, or, indeed, a user-friendly query language – may internally use SPARQL and/or RDF, be it as a means to access its own and federated databases, or just as a formalism to populate its internal database from web resources.

Having this in mind, queriability of philological editions by SPARQL is an important achievement of our approach. On the one hand, this is relatively easy to achieve, as we can benefit from broad technological support by the Semantic Web community. On the other hand, SPARQL is not meant to be used by philologists directly, but it can be an integral component of infrastructure solutions specifically targeting philological research questions and resources.

8. Summary and Discussion

We described the transition from (the source code of) a philological edition of a Middle French text available in LaTeX to a state of the art TEI representation. We further extended this XML edition with RDFa attributes that provide a structured view on parts of the information of the edition. More importantly, the RDFa attributes furnish semantically typed links with a dictionary of Old French.

8.1. TEI+RDFa

The benefit of this approach is that the HTML(+RDFa) rendered from the TEI/XML can trivially maintain the RDFa links and complement the human-readable digital edition with a machine-readable access. Both from the published HTML and from the source XML, RDF triples can be extracted using RDFa parsers or web services such as pyRDFa. These triples can be used to run SPARQL queries directly against the digital edition or to populate a database for the purpose. Such a database can be used to integrate information from multiple editions or from a machine-readable edition of the dictionary, thereby providing additional information about the lexemes found in a particular text and allowing to interrelate them with the linguistic and historical context of a particular expression, term or usage. If the URIs resolve against the digital edition (either directly against the HTML or a content negation service catering HTML), this information can be used to provide a link list to different fragments of the edited text.

What is important here is that this integrated search functionality is achieved by means of off-the-shelf RDF technology, and that the technological support for RDFa excels beyond domain-specific solutions currently favored by the TEI community, as discussed in Sect. 8.2. Our findings may thus contribute to the development of technological bridges between TEI/XML and (Linguistic) Linked Open Data resources.

8.2. The State-of-the-Art TEI Approach

The Text Encoding Initiative (TEI), founded in 1987, is the authoritative body that develops and maintains an XML-based interchange format for textual data, in particular for the electronic edition of printed (or printable) publications. Beyond its historical focus on literary science and linguistics, the current edition of the TEI guidelines, P5 (proposal 5), represents a de facto standard for the entire field of Digital Humanities.

The TEI aims to provide a compromise between a formal description of layout elements (e.g., *italics*) and their abstract function (e.g., *emphasis*): Its markup elements are given interpretable names, but the provided definitions are informative only, not normative, as the TEI standardizes only *their form and structure*. Accordingly, the TEI guidelines are traditionally implemented and validated by a set of modular schemas. For practical applications, the TEI takes a text-driven approach: The form, content and structure of the underlying text are preserved, and are enriched by markup elements, only. Despite considerable overlap in their intentions (i.e., to facilitate interoperability and explicate semantics), this is an important difference in comparison with RDF and the (Linguistic) Linked Open Data (LLOD) world: LLOD pursues a semantics-driven approach to text and linguistic annotations, and – unless explicitly coded in designated RDF properties – surface characteristics of the text are lost. In particular, this includes sequential order and hierarchical structure of elements in the text, which is obligatory (and implicit) in TEI/XML but needs to be explicitly asserted in RDF graphs if it is to be preserved.

TEI conventions, on the other hand, are not very rigorous about the semantics of markup elements. In fact, a frequently applied approach to represent a novel phenomenon is *tag abuse*, meaning that a markup element originally intended for one particular use case is applied for another use-case in a way that contradicts its original definition.

As such, the TEI provides several kinds of pointer structures. According to TEI Consortium (2017b), this includes

xr (cross-reference phrase) contains a phrase, sentence, or icon referring the reader to some other location in this or another text,

ref (reference) defines a reference to another location, possibly modified by additional text or comment,

ptr (pointer) defines a pointer to another location,

lbl (label) contains a label for a form, example, translation, or other piece of information, e.g. abbreviation for, contraction of, literally, approximately, synonyms, etc.

However, none of these pointer structures are LOD-compliant as they do not require their targets to be URIs, nor do they enforce the use of URIs to define the type of relation. For both aspects, free text content can be used, so that an RDF interpretation of any of these pointer structures cannot be asserted from the XML alone.

Accordingly, alternative directions to represent RDF triples in an unambiguous way have been explored, most notably in the context of SAWS project (Jordanous, 2015) whose approach to tackling RDF is featured in examples given in the TEI P5 documentation:[14]

```
<relation resp="http://viaf.org/...36/"
 ref="http://...#isVariantOf"
 active="http://../cts/urn:cts:gr..."
 passive="http://data.perseus.org/
          citat..."/>
```

The SAWS solution is apparently being favored by the TEI community. It has the great advantage of reusing established TEI markup, it is capable to express arbitrary triples, and the `ref` element (used to represent the property) requires its value to be a (sequence of) URI(s). However, the terminology is not very transparent, and, in particular, describing triple subjects and objects in terms of being 'active' and 'passive' (motivated from another use case) is highly confusing in an RDF context. More severe, however, is the intended function of `relation` within TEI P5: It is syntactically restricted to be a child node of a list of named entities (events, nyms, organizations, persons, places) or a `listRelation`. Despite its generic name, the latter is also restricted to prosopography, i.e., it "provides information about relationships identified amongst people, places, and organizations".[15]

The TEI definition of `relation` is thus highly constrained and limited to applications related to prosopography and entity linking. This represents a frequent area of application of RDF, indeed, but not the linking between and among philologically edited text and lexical resources. Nevertheless, it has been used for this purpose, as the example above illustrates, already.

According to the TEI P5 guidelines,[16] "[t]his example records a relationship, defined by the SAWS ontology, between a passage of text identified by a CTS URN, and a variant passage of text in the Perseus Digital Library, and assigns the identification of the relationship to a particular editor (all using resolvable URIs)."

As a workaround, the SAWS specifications redefine the syntactic constraints for `relation` and allow it to occur freely in semantically empty wrapper elements such as `ab` and `seg` (roughly corresponding to HTML `span` elements).[17] However, as this violates TEI specifications, this is a clear instance of tag abuse.

Finally, this solution requires a relatively complicated transformation process to produce processable information, i.e., either a TEI-specific RDF parser is to be developed or this information is to be converted into a more easily processable representation as part of the serialization into published end formats (e.g., as HTML+RDFa). This conversion is complicated by the fact that `relation` permits multiple 'active' and 'passive' URIs, symmetric relations and triple groups with, e.g., different properties connecting the same (groups of) subjects and objects. Generating triples from `relation` thus represents a considerable effort in triple detection (not every `relation` is intended to be RDF) and parsing. Moreover, this effort in triple detection and parsing needs to be replicated for every serialization routine that preserves RDF information.

Furthermore, object literals cannot be easily represented in this way without redundancy. Where RDFa permits to declare CDATA content as an object literal, it needs to be repeated in the SAWS/TEI approach.

In comparison, RDFa has the advantages of semantic clarity, it is immediately processable by off-the-shelf technology, it is an established W3C standard, sufficiently expressive and trivially transferable to output formats such as HTML.

9. Outlook

While we demonstrated that TEI+RDFa represents a viable bridge between the TEI-dominated world of computational philology and the (L)LOD world, the obvious benefit is that now, links with LOD resources can be established, resp., links with philological resources. For these resources, we can now argue that LOD provides technological advantages, as existing data sets can be easily set into relation. This includes both legacy databases that reside in data silos[18], as well as TEI-edited works designed originally only for (generating a) human-readable electronic (and print) representation.

Our approach has the advantage of being supported by off-the-shelf RDF technology (unlike the SOTA TEI approach described in Sect. 8.2.) and by enabling an integrated handling of XML and RDF data.

Directions for further research include the development of deeper and richer linkings with other resources, in particular, lexical resources, as well as the integration of estab-

[14] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-relation.html` [accessed 01-06-2018]: "This example records a relationship, defined by the SAWS ontology, between a passage of text identified by a CTS URN, and a variant passage of text in the Perseus Digital Library, and assigns the identification of the relationship to a particular editor (all using resolvable URIs)."

[15] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-listRelation.html` [accessed 01-06-2018].

[16] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-relation.html` [accessed 01-06-2018].

[17] `http://www.ancientwisdoms.ac.uk/media/documents/Markup_Guidelines_for_Gnomologia.html#TEI.relation` [accessed 01-06-2018].

[18] Relational databases can be augmented with RDF wrappers, e.g., using R2RML, `https://www.w3.org/TR/r2rml/` [accessed 01-18-2018].

lished or emerging LLOD vocabularies which are relevant for philological research.

9.1. Linking all Forms to DEAF

Our conversion is intended to be a proof of concept, with numerous natural extensions yet to be addressed. As such, we automatically established links between the lemmata of the edition and the corresponding dictionary entries by comparing the lemmata with the headwords of the dictionary entries. A perfect match of character strings thus created a link. As a consequence however, no link was created for every lemma of GuiChaulMT which does not coincide with the headword of the dictionary entry. For instance, this is the case for the lemma *humeur* f. (GuiChaulMT f°16r°b; f°23r°a; b; f°23v°a; etc.) vs. *umor* f. (DEAF*él* `https://deaf-server.adw.uni-heidelberg.de/lemme/umor` [accessed 01-06-2018]) and also for *distribuer* v. (GuiChaulMT f°18v°b; f°31v°a, f°32r°b; etc.) vs. *distribüer* v. (DEAF*él* `https://deaf-server.adw.uni-heidelberg.de/lemme/distrib%C3%BCer` [accessed 01-06-2018]).

At the time of the creation of the edition and thus the lemmatization of its vocabulary, the DEAF material was not yet completely lemmatized and published. This explains the discrepancies. A future step to eradicate this flaw entails the implementation of fuzzy matching.

9.2. Linking to the *Dictionnaire du Moyen Français*

For our endeavor, we chose the DEAF as our reference dictionary for a number of reasons: (a) The DEAF covers the Old French lexis and also illustrates its continued existence in Middle French and until today (if this is the case), (b) it integrates a certain range of Middle French words, (c) the articles include a comprehensive linguistic discussion of the lexemes, and (d) its editorial team has assembled a significant expertise in the understanding of the technical language of medicine which is reflected in the respective dictionary articles. However, the DMF (Sect. 1.) is the lexicographical work that represents the state of the language written by Gui de Chauliac. Its publication is on-line (Martin, 2015). A step further towards a network of text edition and lexicography will thus be to establish links from the vocabulary of GuiChaulMT also to the entries of the DMF.

9.3. Linking Lexical Units

We automatically established the linking of a lexeme in GuiChaulMT with the corresponding entry in the DEAF. This means that the linking operates on the word level.

However, many of the lexemes in GuiChaulMT are polysemous and so are the ones in the dictionary: Words have several meanings. E.g., GuiChaulMT documents two meanings of *humeur* f., both belonging to medical terminology: Firstly, *humeur* designates any fluid in the human or animal body (GuiChaulMT f°23r°a; b; f°23v°a; etc.), and secondly, it designates one of the four humors of the classical humoralism that govern the equilibria of a healthy body, i.e., blood, phlegm, yellow bile and black bile (ib. f°16r°b; f°32r°a).

Also, a considerable number of lexemes are attested in GuiChaulMT with only one sense but the respective dictionary entry shows several. E.g., *plaie* f. only designates the wound in the text edition (GuiChaulMT f°15r°b; f°18v°b; f°22v°a; etc.) though DEAF*él* gives two more meanings, i.e., a sort of slot and the same with a sexual connotation (`https://deaf-server.adw.uni-heidelberg.de/lemme/plaie1` [accessed 01-06-2018]).

Thus, the goal that we need to pursue next is to raise the linking from the word level to the sense level. GuiChaulMT fulfills the prerequisites: Its markup contains the necessary semantic disambiguation as explained above. The XML data of the dictionary does the same: Its structure models the semantic scope of each entry as a tree with main and sub-senses including their definitions. Nevertheless, the problematic issue will be to insert the linking. To be able to perform this step automatically the character string of a sense definition of a lexeme in GuiChaulMT must coincide somehow with the character string of the corresponding sense definition in the dictionary. However, in most of the cases it does not. E.g., *fermer* v.tr. "faire tenir (à une chose) au moyen d'une attache, d'un lien; attacher" (GuiChaulMT f°29r°a) vs. *fermer* v.tr. "attacher solidement, fixer" (DEAF*él* `https://deaf-server.adw.uni-heidelberg.de/lemme/fermer` [accessed 01-06-2018]); *fievre* f. "état maladif caractérisé par l'augmentation de la chaleur du corps" (GuiChaulMT f°32r°b) vs. *fievre* f. "élévation pathologique de la température habituelle du corps et, par ext., état maladif caractérisé par cette élévation (chez l'homme et chez l'animal), fièvre" (DEAF F 402,22; DEAF*él* `https://deaf-server.adw.uni-heidelberg.de/lemme/fievre` [accessed 01-06-2018]). This needs to be addressed.

9.4. OntoLex-Lemon

At this stage of our endeavor, we use the properties `rdfs:label` for the representation of the graphical realizations of the lemmata in the text and `rdfs:seeAlso` for the reference of the latter to the respective entries in the DEAF. However, the exploration of a scholarly text edition as a resource for the diachronic study of the language will be much more efficient if it incorporates a more specialized vocabulary. Particularly promising in this sense seems to be the OntoLex-Lemon model, an ontology-lexicon interface that provides a vocabulary for the modeling of lexical resources[19]. For a more detailed assessment of the use of OntoLex-Lemon in our domain, see Tittel and Chiarcos (accepted).

[19] *LExicon Model for ONtologies* (Lemon), published as the OntoLex-Lemon model in May 2016 by the Ontology-Lexicon W3C Community Group, see `https://www.w3.org/2016/05/ontolex/` [accessed 2017-12-12]).

10. 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 ERC-funded project *Poetry Standardization and Linked Open Data: POSTDATA* (ERC-2015-STG-679528).
The contribution of the third 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 conversion of the printed text edition into XML+RDFa was supported by the organizers and participants of the 2nd Summer Datathon on Linguistic Linked Open Data (SDLLOD 2017), June 2017, Cercedilla, Spain. There, the linking of the lexemes in the edition to the entries in the dictionary was explored and conducted 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), Mariana Curado Malta (Polytechnic University of Porto, Portugal), Frances Gillis-Webber (University of Cape Town) and Maxim Ionov (Goethe University Frankfurt, Germany) for their valuable input and contributions.
Finally, we would like to thank the anonymous reviewers for helpful comments and insightful feedback.

11. 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 [Accessed 12-24-2017]].

Blank, A. (1997). *Einführung in die lexikalische Semantik*. Niemeyer, Tübingen, Germany.

Brickley, D. and Miller, L. (2014). *FOAF Vocabulary Specification 0.99*. Available at: http://xmlns.com/foaf/spec/ [Accessed 12-22-2017].

DCMI Usage Board. (2012). *DCMI Metadata Terms*. Available at: http://dublincore.org/documents/dcmi-terms [Accessed 12-22-2017].

Eley, P., Simons, P., Longtin, M., Hanley, C., Shaw, P., and McLaughlin, J. (2005). *Partonopeus de Blois: An Electronic Edition*. HriOnline, Sheffield. http://www.hrionline.ac.uk/partonopeus [Accessed 01-04-2018].

Habert, B., Nazarenko, A., and Salem, A. (1997). *Les linguistiques de corpus*. Armand Colin, Paris, France.

Harsch, U. (1996). *Passion de Clermont*. Bibliotheca Augustana, Augsburg, Germany. http://www.hs-augsburg.de/~harsch/gallica/Chronologie/10siecle/Passion/pas_text.html [Accessed 01-04-2018].

Herman, I., Aside, B., McCarron, S., and Birbeck, M. (2015). *RDFa Core 1.1 – Third Edition*. Available at: https://www.w3.org/TR/rdfa-core [Accessed 12-22-2017].

Jordanous, A. (2015). Enhancing information retrieval and resource discovery from data using the Semantic Web. In *4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS-2015)*, pages 105–110. IEEE.

Laidlaw, J. (2015). *The Making of the Queen's Manuscript*. Bibliotheca Augustana, Edinburgh, Great Britain. http://www.pizan.lib.ed.ac.uk [Accessed 01-04-2018].

Martin, R. (2015). *Dictionnaire du Moyen Français – DMF*. ATILF – CNRS & Université de Lorraine, Nancy, France. [version 2015 (DMF 2015), http://www.atilf.fr/dmf [Accessed 01-04-2018]].

Möhren, F. (2015). L'art du glossaire d'édition. In David Trotter, editor, *Manuel de la philologie de l'édition*, pages 397–437. De Gruyter.

Nichols, S. and Choudhury, G. (2017). *Roman de la Rose Digital Library*. Johns Hopkins University / Bibliothèque Nationale de France, Baltimore / Paris. http://romandelarose.org [Accessed 01-04-2018].

Reisdoerfer, J. (1996a). *Serments de Strasbourg*. Centre Universitaire du Grand-Duché de Luxembourg (Project BABEL), Luxembourg. http://w3.restena.lu/cul/BABEL/T_SERMENTS.html [Accessed 01-04-2018].

Reisdoerfer, J. (1996b). *Séquence de sainte Eulalie*. Centre Universitaire du Grand-Duché de Luxembourg (Project BABEL), Luxembourg. http://w3.restena.lu/cul/BABEL/T_CANTILENE.html [Accessed 01-04-2018].

TEI Consortium. (2017a). Critical Apparatus. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0. Last updated on 10th July 2017*. Available at: http://www.tei-c.org/release/doc/tei-p5-doc/es/html/TC.html [Accessed 12-22-2017].

TEI Consortium. (2017b). *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0. Last updated on 10th July 2017*. Available at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ [Accessed 01-05-2018].

Tittel, S. and Chiarcos, C. (accepted). Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In *Proceedings of the LREC-2018 GLOBALEX workshop (GLOBALEX-2018)*, Miyazaki, Japan, May.

Tittel, S. (2004). *Die Anathomie in der Grande Chirurgie des Gui de Chauliac: Wort- und sachgeschichtliche Untersuchungen und Edition*. Niemeyer, Tübingen, Germany.

Tittel, S. (2010). Le « DEAF électronique » – un avenir pour la lexicographie. *Revue de Linguistique Romane*, 74:301–311.