On Modelling Corpus Citations in Computational Lexical Resources

Anas Fahad Khan, Maxim Ionov, Christian Chiarcos, Laurent Romary, Gilles Serasset, Besim Kabashi

On Modelling Corpus Citations in Computational Lexical Resources

Anas Fahad Khan[1], **Maxim Ionov**[2], **Christian Chiarcos**[3], **Laurent Romary**[4], **Gilles Sérasset**[5], **Besim Kabashi**[6]

[1]CNR-ILC, Italy, [2]University of Cologne, Germany, [3]University of Augsburg, Germany,
[4]INRIA, France, [5]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, France,
[6]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[1]fahad.khan@ilc.cnr.it, [2]mionov@uni-koeln.de, [3]christian.chiarcos@uni-a.de
[4]laurent.romary@inria.fr, [5]gilles.serasset@imag.fr, [6]besim.kabashi@fau.de

Abstract

In this article we look at how two different standards for lexical resources, TEI and OntoLex, deal with corpus citations in lexicons. We will focus on how corpus citations in retrodigitised dictionaries can be modelled using each of the two standards since this provides us with a suitably challenging use case. After looking at the structure of an example entry from a legacy dictionary, we examine the two approaches offered by the two different standards by outlining an encoding for the example entry using both of them (note that this article features the first extended discussion of how the Frequency Attestation and Corpus (FrAC) module of OntoLex deals with citations). After comparing the two approaches and looking at the advantages and disadvantages of both, we argue for a combination of both. In the last part of the article we discuss different ways of doing this, giving our preference for a strategy which makes use of RDFa.

Keywords: Corpus Citations, Computational Lexical Resources, Electronic Lexicographic Resources, Linked Data, Text Encoding Initiative (TEI), RDFa, OntoLex & OntoLex-FrAC

1. Introduction

Corpus citations of source material featured in dictionary entries can be an important source of linguistic, historical and cultural information.[1] However, they are not represented in a standardised way across dictionaries, which makes it difficult to automatically extract the information contained in them. This is especially the case with legacy dictionaries, where often fairly important pieces of metadata, such as the title or the name of the author of a cited source, cannot be retrieved from the text of an entry and/or the front matter of a lexicographic work by themselves. Consequently, it would be useful to have a means of annotating this information (manually if necessary) and thus making it more readily available for extraction and querying. This would also facilitate the comparison and combination of corpus citation information from different lexicographic sources and assist in the creation of tools and interfaces allowing easy access to the full corpus texts themselves from a dictionary interface. However, in order to ensure the interoperability and re-usability of lexicographic resources annotated for such kinds of information, it is advisable to use standards[2] which are currently in use by the community or communities of reference. For this reason, in the current work we have chosen to look at two of the most popular standards for modelling and publishing lexical resources as structured data, namely, **TEI(-XML)** and **OntoLex-Lemon**. Both models distinguish between different levels of description: in particular, see the distinction between **typographical**, **editorial**, and **lexical** views in Chapter 9 of the TEI guidelines dealing with dictionaries.[3][4] In the case of TEI(-XML), we are dealing with an XML-based standard familiar to many digital humanists, and which has recently become more widely known among lexicographers too, thanks to TEI Lex-0,[5] a TEI customisation intended specifically for encoding dictionaries (Romary and Tasovac, 2018). OntoLex-Lemon (or OntoLex for short) is an RDF-based standard that, though less well known among (digital) humanists than TEI, does offer producers and consumers of

[1]Note that in what follows, we adopt the distinction found in e.g. Klosa (2015) between **corpus citations** representing authentic examples of past language usage from external corpora and **competence examples** which are example phrases devised by lexicographers based on native speaker intuition. In the former case we can say that (in case of a successful citation) some linguistic phenomenon or other is attested by the corpus citation in question.

[2]These can either be official standards published by internationally recognised standards bodies such as ISO or *de facto* standards adopted by a community without being endorsed by any particular standards body.

[3]https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DIMV

[4]Allowing for these separate levels of description becomes especially useful in cases where we combine information from diverse resources and where it is important to distinguish between the linguistic claims being made, their provenance, and how they are presented.

[5]https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

lexicographic datasets access to all of the advantages of linked data and the Semantic Web, including enhanced interoperability, ease of linking, and the use of a powerful query language, SPARQL, among many others. In this article, we will look at how both standards allow for the encoding of citations in lexical resources, particularly lexicographic resources. We will compare both approaches and argue that rather than choosing one over the other, we should play to the strengths of both of these standards. Finally, we will demonstrate how to do this.

More precisely, in this article we will: **detail the kinds of information which we can annotate** in a corpus citation with reference to an example taken from a legacy dictionary (Section 2); **outline a TEI based approach** to modelling corpus citations with relation to the preceeding example (Section 3.1); **discuss the FRequency Attestation and Corpus (FrAC) module of OntoLex** with a focus on the provision it offers for **modelling corpus citations** (note that the current article is the first extended discussion of this topic in the context of OntoLex) (Section 3.2); compare both of these approaches looking at the **strengths and weakness of each** (Section 3.3); look at how both of these approaches might be **combined** (Section 4), including via **XSLT transformations** (Section 4.1), **standoff** (Section 4.2) and **RDFa** (Section 4.3), arguing ultimately for the latter as the best strategy (Section 4.4.1).

2. The Structure of Lexical Entries: An Example

To illustrate some common features of citations in legacy dictionaries (and in particular of more demanding scholarly dictionaries which typically feature a high density of information encoding) and the modelling challenges which they pose[6], we will look at an example culled from a 19th century Old English dictionary, *An Anglo-Saxon Dictionary*, otherwise known as the Bosworth-Toller (Bosworth, 1898). Namely, we will look at the entry for *geár-dagas* 'days of yore' as given in Figure 1. In this entry, the definition for the word is followed by two variant multi-word phrases/collocates merged into one ('In [on] geardagum') and this latter is associated with a list of citations;[7] other phrases involving the headword (e.g., *ure geár-dagas dies*) are then successively cited. Finally, an Icelandic cognate of *geár-dagas* is given in square brackets -– we will ignore this part in what follows, though much of what we have to say about citations in terms of the modelling challenges they pose also applies to etymons and cognates. In the next few sections we will describe a partial encoding of the entry using both TEI and OntoLex models, comparing them and looking at the strengths and weaknesses of each. (As a caveat, we should point out that there is obviously more than one way to represent the entry using either standard; we have aimed for a fairly conservative encoding in each case.)

geár-dagas; *pl. m.* [geár, dæg] YORE-DAYS, *days of yore, days of years, time of life;* dies antīqui, annōrum dies:—In [on] geardagum *in days of yore,* Exon. 11 b; Th. 16, 11; Cri. 251: 77 a; Th. 289, 6; Wand. 44: Cd. 21; Th. 287, 16; Sat. 368: Beo. Th. 2; B. 1: 2712; B. 1354: 4458; B. 2233. In geárdagan, Menol. Fox 231; Men. 117. Úre geárdagas *dies annōrum nostrōrum,* Ps. Th. 89, 10. Scyle gumena gehwylc on his geárdagum georne biþencan *every man should in the days of his years well consider,* Exon. 19 b; Th. 51, 26; Cri. 822: 61 a; Th. 225, 4; Ph. 384: Elen. Grm. 1267: L. Eth. vii. 24: Th. i. 334, 21. [*Icel.* í árdaga *in days of yore.* Cf. Gen. 47, 9, 'The days of the years of my pilgrimage are an hundred and thirty years.']

Figure 1: A sample entry (*geár-dagas*) from the Bosworth-Toller Old English dictionary.

3. Encoding Dictionary Citations in TEI and OntoLex-FrAC

3.1. Encoding *geár-dagas* in TEI

We do not give the full listing of the TEI encoding of the entry here, although it is available at `TEIAtt/geardagas.xml`. Instead, we will give a more general description of this encoding, with the overall structure of the TEI entry represented in Figure 2.

In our encoding, we have closely followed the guidance set out in the dictionary chapter of the TEI guidelines (TEI Consortium) in using the <cit> element to encode "a quotation from some other document, together with a bibliographic reference to its source". In addition, the <quote> element is used to annotate the actual quotation string itself (as well as separately annotating translations of a preceeding citation). Moreover, under the <teiHeader> element for the entry we have included a list of bibliographic references using the <listBibl> element, with each individual reference encoded using the <bibl> element and featuring all the salient metadata information that we were able to glean from the original text and secondary sources (e.g., <title>, <editor>, <place>, <date>). See, for instance, the following listing for the title *the Anglo-Saxon Poems of Beowulf* edited by Benjamin Thorpe:

```
<bibl xml:id="Beo_Th" type="corpus">
    <title>The Anglo-Saxon Poems of Beowulf
        </title>
    <editor>Benjamin Thorpe</editor>
    <placeName>Oxford</placeName>
    <date>1855</date>
</bibl>
```

[6] Indeed, we have chosen to look at retrodigitized legacy dictionaries as a special use case of lexical resources precisely for the difficulty of the challenges they pose.

[7] These citations aren't always easy to fully interpret even after consulting the front matter (something which, as we have already mentioned, will cause problems for the automatic extraction of citations from legacy dictionaries).

```
<TEI
    xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>...</titleStmt>
            <publicationStmt>...</publicationStmt>
            <sourceDesc>...</sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <entry>
                <form>...</form>
                <gramGrp>...</gramGrp>
                <etym>[geár, dæg]</etym>
                <sense>
                    <def>YORE-DAYS...; </def>
                    <cit>...</cit>
                    <cit>...</cit>
                    <cit type="example">...</cit>
                    <cit>...</cit>
                    <cit>...</cit>
                    <cit type="example">...</cit>
                    <cit>...</cit>
                    <cit type="example">...</cit>
                    <cit type="example">...</cit>
                    <cit type="example">...</cit>
                    <cit type="example">...</cit>
                    <cit type="example">...</cit>
                    <cit type="example">...</cit>
                    [Icel. ....']
                </sense>
            </entry>
        </body>
    </text>
</TEI>
```

Figure 2: The structure of the TEI Encoding of *geár-dagas*

The entry itself (encoded naturally enough using the <entry> element) is nested directly within the <body> element and includes information on the form of the word (in <form>), its grammar (in <gramGrp>), relevant etymological information (in <etym>, in this case the two words which *geár-dagas*, a compound, derives from), as well as information about the word's meaning as enclosed in the <sense> element. This latter element contains a definition (in <def>) and a succession of <cit> elements with associated quotes (in <quote>) and bibliographic information (in <bibl>) which references bibliographic elements in the header (via the @source attribute) – where, of course, this information is available. <cit> elements can be typed via the @type attribute, which allows us to, in our case, specify which citations are examples and which are translations. In the following listing, for instance, the translation of the original Old English phrase is included in a nested <cit> element within the original <cit>:

```
<cit type="example">
  <quote>Scyle gumena gehwylc on his geá
      rdagum georne bipencan</quote>
  <cit type="translation" xml:lang="en">
    <quote>every man should in the days of
        his years well consider</quote>
  </cit>
  <bibl source="#Exon_Th">Exon.19 b</bibl>
</cit>
```

The TEI encoding here is clearly faithful to the linear order of the relevant textual elements in the original. It's worth noting that in cases where we are interested in capturing the visual elements of the original entry, such as for instance the different typographical conventions used, this is also possible with TEI (c.f., the discussion on different dictionary views in Chapter 9 of the guidelines). Overall then, TEI provides us with the elements to mark-up the main (lexicographically-relevant) pieces of information in the original entry as well as demonstrating a certain degree of expressivity: for instance it allows us to mark out translations of the original Old English text, specify cases where a cited example is found in various different texts and to refer to the same bibliographic resource across different entries (though in our case there is only one entry).

3.2. Encoding *geár-dagas* in OntoLex-Lemon

Ontolex-Lemon (McCrae et al., 2017) is, currently the best known and most widely used vocabulary for the creation, publication and sharing of lexical resources as linked data. Since the publication of the original vocabulary in 2016, a number of further extensions have been proposed which have either already been published or are currently under development; among the latter we can cite the **Frequency, Attestation and Corpus information** module (FrAC), which deals principally with the inclusion of corpus-based information in lexical resources (Chiarcos et al., 2020). FrAC is currently at an advanced stage of completion (Chiarcos et al., 2022) and can be considered stable at the time of writing. Indeed, one of the purposes of the present article is to showcase the provision that FrAC offers for citations and attestations to a wider language resources community. Fig. 2 presents the current model in diagrammatic form. For reasons of space we will not give full definitions of FrAC classes and properties in this paper, they can be found in the GitHub repository for the module[8]. We will only mention those definitions that are relevant for this article. These are: the class `Attestation` constituting "a special form of citation that provide[s] evidence for the existence of a certain lexical phenomena"; the property `attestation` which associates an `Attestation` to the FrAC class `Observable`; and the property `locus` which points to the location in the text at which the relevant word(s) can be found.

As with the TEI encoding of *geár-dagas* we do not give the full entry here, instead, it can be found at `https://github.com/max-ionov/`

[8] https://github.com/ontolex/frequency-attestation-corpus-information

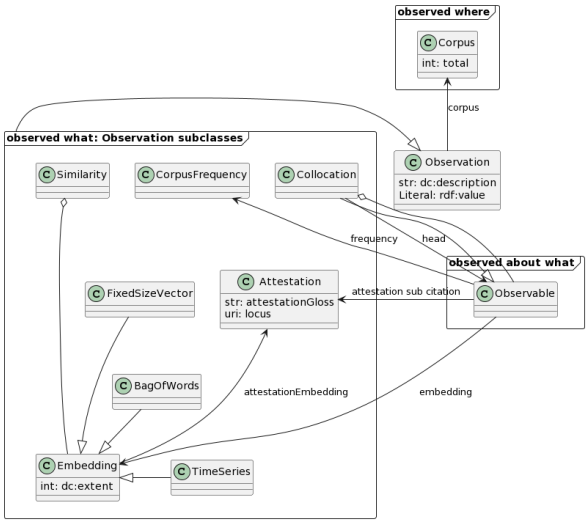

Figure 3: The FrAC model.

attestationExample/tree/main/TEIAtt.[9] The diagram in Figure 4 shows part of the OntoLex encoding. The main entry in our example is

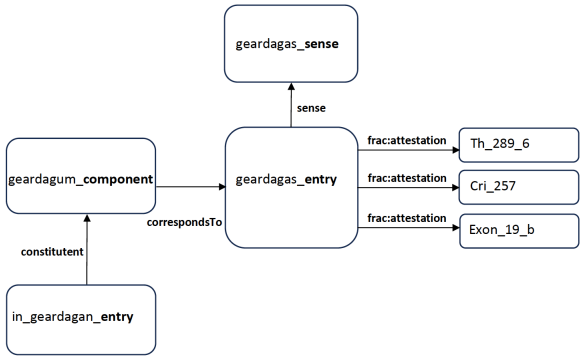

Figure 4: The OntoLex encoding.

`:geardagas_entry` which, as well as being associated with various kinds of morpho-syntactic information (part of speech, gender, etc.), is also associated with a number of FrAC `Attestation` individuals via the FrAC `attestation` property.

```
:geardagas_entry rdf:type ontolex:Word ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:plural ;
  frac:attestation :Th_16_11b, :Cri_257, :
      Th_289_6, :Wand_44, :Th_287_16, :
      Beo_Th, :Exon_19_b ;
  ontolex:sense :geardagas_sense .
```

The individuals of type `Attestation` can be further described and related to other individuals describing bibliographic entities; and we can do this by re-using specialised external vocabularies, as well as more generic ones (such as the Dublin Core[10]). In our case we describe relationships between parts of works/texts in a corpus using e.g. BIBO;[11] as well as modelling a variety of citation relationships and texts at different levels of abstraction using the SPAR ontology suite.[12] Furthermore, thanks to the fact that we are working within the linked data paradigm, we can easily link to (and therefore describe our data using) resources describing the authors, publishers, historical events, geographical regions, etc which are relevant to the corpus citations contained in the original work. For instance, take one of the first citations given in the text and written as "Th. 16, 11", which corresponds to `:Th_16_11b`, an entity of type `frac:Attestation`. Here we have also given the relevant text pertaining to the citation in the original entry using the property `rdf:value`.

```
:Th_16_11b a frac:Attestation ;
  frac:locus
    [dc:isPartOf :Codex_Exoniensis] ;
  rdf:value "In␣[on]␣geardagum␣in␣days␣of␣
      yore[...]␣Th.␣16,␣11"@en.
```

According to the front matter of the dictionary, this particular attestation can be found in an edition of the *Codex Exoniensis* edited by Benjamin Thorpe. We encode this information as follows using the external vocabularies BIBO and Dublin Core along with external resources/knowledge graphs such as VIAF and DBPedia.

```
:Codex_Exoniensis a bibo:CollectedDocument;
  dc:title "Codex␣Exoniensis";
  dc:description "A␣collection␣of␣Anglo-
      Saxon␣poetry␣from␣a␣MS.␣in␣the␣
      library␣of␣the␣Dean␣and␣Chapter␣of␣
      Exeter␣by␣Benjamin␣Thorpe,␣London,␣
      1842.";
  bibo:editor <https://viaf.org/viaf
      /66566155/#Thorpe,_Benjamin,_1782
      -1870> ;
  bibo:place <https://dbpedia.org/resource/
      London> ;
  bibo:time "1842".
```

OntoLex also permits us to encode multiword expressions[13], such as *in geardagum* in our example, as separate entries and to link them to their constitutent entries such as `:geardagas_entry`. We then have the option of associating individual attestations either to these multi-word expressions (e.g., *in geardagum*) or with the 'main' entries in our lexicon such as (*geár-dagas*) – by transitivity, these will also be attestations for the 'main' entries. In our case we associate all attestations with `geardagas_entry`, but this is a design preference only.

[9] Note that the version in the link is also only a partial encoding of the original dictionary entry. We did not encode the entire entry there because this would have made it difficult for the reader to appreciate the most significant parts of it.

[10] https://www.dublincore.org/

[11] https://www.bibliontology.com/

[12] http://www.sparontologies.net/

[13] As does TEI, but in the opinion of at least some of the authors of the current work, in a much less natural or obvious way.

If, instead, attestations for the same 'main' entry (such as *geár-dagas*) are directly associated (via the `attestation` property) with related entries for multi-word expressions (such as *in geardagum*) we can still easily retrieve them with other data for the main entry via an appropriate SPARQL query).

```
:in_geardagan_entry rdf:type
    owl:NamedIndividual ,
    ontolex:PrepositionPhrase ;
  <http://www.w3.org/ns/lemon/decomp#
      constituent>
    :geardagum_component;
  ontolex:sense
    [rdf:type ontolex:LexicalSense;
    dc:description "In [on] geardagum in
        days of yore"@en].
:geardagum_component rdf:type
    owl:NamedIndividual ,
    ontolex:Component ;
  <http://www.w3.org/ns/lemon/decomp#
      correspondsTo> :geardagas_entry .
```

In sum, the OntoLex encoding seems to favour a (re-)structuring of the original entry on the basis of its content/ meaning,[14] tending to a representation of the original entry in terms of the lexical view mentioned above (and cited in Chapter 9 of the TEI guidelines). In effect we are creating a knowledge graph from the information extracted from the original text (one that we can link to other Semantic Web knowledge bases such as VIAF and DBPedia) but which no longer follows the original ordering of that text. This makes it harder to retrieve the original visual structure and textual representation of the data, and although this can be done via the OntoLex *lexicog* module (Bosque-Gil et al., 2017), the result quickly becomes verbose, hard to work with and even harder to read. In general, then, it is easier to give a faithful, linear representation of the typographical and editorial views of a lexicographic work in TEI than with OntoLex in general. In compensation, however, OntoLex allows us to create a much more enriched version of lexical view of the same work, via the possibility of linking our initial dataset to other knowledge graphs containing information relating to both the original lexicographic resource itself – its editors, the sources it cites, any source manuscripts – as well as to the linguistic phenomenon being described in the resource – linking it with other linguistic entities belonging to this and other languages (such as in the present case an Icelandic cognate – enabling us to model dictionaries more effectively as heterogeneous resources that along with linguistic information include cultural, historical, geographical information. We do this, moreover, by leveraging standard, off-the-shelf languages, technologies and resources, such as OWL and SPARQL. This all comes in addition to general benefits of linked data (well-rehearsed by now) in making datasets more interoperable and more FAIR in general. It is true that the potential for linked data in publishing the kinds of very hetereogeneous datasets often dealt with by the humanities has not (up until now) been sufficiently explored, but projects like LiLa (Passarotti et al., 2019) and Mapping the Manuscript Migrations (Burrows et al., 2021), as well as the ever increasing use of CIDOC-CRM (Bruseker et al., 2017) in harmonising catalogues across different GLAM-sector institutions show that the potential is very much there. Leaving aside their differences, the two encodings which we have examined still have much in common. Both provide means extensive means of representing (structured) bibliographic references corresponding to the citations in the entry. In the TEI case, the entry links to a list of bibliographic entries in the TEI header (with relevant TEI elements used to structure each individual reference); in the OntoLex-Lemon encoding each bibliographic reference is an RDF resource described using (pre-existing) Semantic Web vocabularies and models.

3.3. The Choice between TEI and OntoLex-Lemon

When it comes to deciding between the two standards, there are a number of issues to consider, aside that is from those discussed in the last section. Chief among these is the popularity of TEI amongst humanists (lexicographers and linguists being no exception) and the fact that many of them (at least those who have some background in the Digital Humanities) will already be familiar with TEI, in addition to many of the pipelines and tools which work with that standard; this situation is helped along by the existence of numerous courses and training materials which are targeted towards such users. TEI's relative popularity amongst users with a humanities background can be put down to the fact that it is, as its name suggests, very much text-centric and without the same clear tendency towards concept-driven abstraction as RDF and other Knowledge Engineering languages. Moreover, in the case of retrodigitized legacy dictionaries, the preference will usually be for a more linear, text-based representation (i.e., to give more weight to editorial and typographical views), something that comes much more naturally with TEI than RDF (both from the encoder's and the end-user's point of view). Given all this and given the desirability of taking advantage of the many affordances offered by the Semantic Web, we would recommend that rather than choosing between the two, TEI and RDF, the best option, in a large number of, cases is to take advantage of both.[15] And indeed in the

[14] Or rather – if this invocation of 'meaning' seems too vague for the reader's preferences – according to a more general (purpose) conceptual model for describing lexical semantics, i.e. OntoLex-Lemon.

[15] This for instance is the strategy which is being pursued by the ongoing MORDigital project (Costa et al.,

next section we look at several different strategies for doing this.

4. Bringing TEI and OntoLex together

4.1. Transformations of TEI into RDF

The first strategy which we discuss here is the use of an XSLT transformation in conjunction with XLST-based technologies such as GRDDL with the aim of converting a TEI dictionary to OntoLex, something which effectively creates a default interpretation of TEI in RDF. There exist several XSLT solutions for transforming TEI to OntoLex. A very generic approach is provided by `https://github.com/acoli-repo/LLODifier/tree/master/tei`, but here, no semantic encoding is performed, but only a 1:1 reconstruction of XML data structures in RDF, which is practically too verbose and can only serve as pre-processing step for further scripts that perform semantic extraction with RDF technology (Fäth and Chiarcos, 2022). As a domain-specific solution, `https://github.com/elexis-eu/tei2ontolex` has been designed with lexical data in mind, but it covers only a small subset of TEI (and not even all of the specifically dictionary-oriented parts of TEI described in Chapter 9 of the guidelines and dealt with in TEI Lex-0). One general issue with a domain-specific XSLT approach is the semantic ambiguity of many TEI elements, so that a single TEI element can have more than one interpretation in OntoLex. This is less of an issue with the TEI Lex-0 customisation of TEI since this is much more focused on the dictionary domain (which is why we recommend using TEI Lex-0 for dictionary projects). Another issue arises from the need to synchronise the RDF and the TEI representation of the same work if one is derived from the other, but edited independently. This requires the XSLT scripts to either be systematically bundled with the data or to be published under persistent URIs and with guaranteed accessibility for an infinite future.

4.2. The Stand-off Approach

Another approach for bringing TEI together with a Linked Data representation is to use explicit standoff annotations over TEI documents, as implemented by the Recogito tool,[16] which uses JSON-LD standoff (i.e., RDF data in a JSON serialisation, using W3C WebAnnotation vocabulary[17]) to annotate TEI/XML and text documents. The main difficulty here is that the XML document and its stand-off annotation can get too easily detached from each other, e.g. the case where a WebAnnotation `oa:TextPositionSelector` refers to an offset in a file that has gone even just a marginal change such as whitespace normalisation. Similar problems exist with the `oa:XPathSelector` if modifications in the DOM tree occur. A more robust way to address XML elements in TEI is the usage of explicit URIs, e.g., with the TEI `xml:id` attribute. Although this may seem equivalent to using the RDFa attribute `about` (we will discuss RDFa in more detail below), one main difference is that `xml:id` attributes can use a TEI-specific short-hand notation, so that they cannot be directly interpreted / resolved by standard RDF tools, while `about` (and `href`) can. A related approach would be to use the TEI element <xenoData>[18] which can be used to insert non-XML, non-TEI metadata into the header of TEI documents, and thus provide explicit RDF data in the RDF serialisation of choice, for example RDF-XML or JSON-LD. The problem with this approach is that this element is limited to the TEI header, and aside from the advantage of being provided in a single file and thus being less prone to the potential effects of desynchronisation from the original source data, it still comes with the synchronisation challenge between TEI and non-TEI annotations.[19]

4.3. The RDFa Approach

4.3.1. RDFa: An Introduction

RDFa is an acronym for *Resource Description Framework in attributes* and refers to a language designated for extending markup languages such as HTML or XML with RDF triples. The original application of RDFa was to add a semantic layer to web sites in HTML or XML, but it can in principle be applied to any XML data. So that it is used, for example, for metadata definition in SVG graphics (Andersson et al., 2008, Section on metadata). RDFa 1.1 is defined in a set of W3C recommendations (Adida et al., 2015; Herman et al., 2015; Sporny, 2015; Sporny et al., 2015). An important characteristic of RDFa is that it requires a host lan-

2021).

[16] `https://recogito.pelagios.org/`

[17] `https://www.w3.org/TR/annotation-vocab/`

[18] `https://tei-c.org/release/doc/tei-p5-doc/de/html/HD.html#HD9`

[19] Yet another way of embedding explicit RDF data into TEI/XML is by using one of the TEI legacy approaches that have been applied for the purpose in the past. Overall, these are based on re-using vocabulary originally introduced for other purposes, so neither of these have an exact RDF interpretation – nor can they be reliably identified as RDF data in a fully automated fashion, as they can come with non-RDF semantics. However, more problematic is that this is not unambiguous, either. We will not go into details here, but we refer the interested reader to Cimiano et al. (2020).

guage to which it can add attributes. For an XML- or SGML-based language, RDFa encodes triples in attributes and exploits the structure of the document, i.e. *the Document Object Model (DOM) tree*, to identify RDF subjects, properties and objects. RDFa defines a number of attributes to describe RDF triples. As such, the attribute `@about` defines its argument as an RDF subject URI, the attribute `@property` defines its argument as an RDF property URI, and the attribute `@resource` defines its argument as an RDF object URI. These attributes can encode a single triple in one XML element, say

```
<span about=":myuri1" property=":myprop1"
    resource=":myuri2"/>
```

In RDF/Turtle, this corresponds to `:myuri1 :myprop1 :myuri2`. This interacts with the DOM tree structure in that `@resource` also introduces a novel subject for `@property` attributes along the descendant axis. So, the subject of `@property` is defined by the closest `@resource` or `@about` attribute along the ancestor axis. Likewise, the objects of properties can also be put into child or descendant elements. If the descendant axis contains multiple properties or multiple objects referring to the same subject, one triple will be created for each. Data properties are modelled analogously, but using `@content` instead of `@resource` for identifying the object literal. If no explicit object URI or literal is provided, `@property` will be interpreted as a data property that takes the CDATA content of the current element as literal value. The type and language of literals can be declared with `@datatype` and `@lang`.

Beyond these core elements, there are some additional attributes which eventually provide short notational hands (`@typeof`, `@inlist`, `@prefix`, `@vocab`), or some syntactic sugar (`@src` or `@href` instead of `@resource`; `@rel` or `@rev` instead of `@property`; `@xml:lang` instead of `@lang`).

4.4. Encoding the lexical entry *geár-dagas* in TEI+RDFa

In order to extend the TEI modelling described in Section 3.1, we have added attributes corresponding to the modelling presented in Section 3.2.[20] For the most part, this process is straightforward. Consider, for example, a modified version of the bibliographical reference introduced earlier:

```
<bibl xml:id="Beo_Th" type="corpus"
    resource=":Beo_Th" typeof="bibo:Book">
  <title property="dc:title">The Anglo-
      Saxon Poems of Beowulf</title>
  <editor rel="bibo:editor" src="https://
      viaf.org/viaf/66566155/#Thorpe,
      _Benjamin,_1782-1870">Benjamin Thorpe
      </editor>
  <placeName rel="bibo:place" src="http://
      dbpedia.org/resource/Oxford">Oxford</
      placeName>
  <date rel="bibo:time">1855</date>
</bibl>
```

Now, in addition to the TEI semantics, this snippet contains a link to external resources with additional information about the author of the bibliographical reference. A problem arises when dealing with TEI elements that should correspond to several RDF triples with different predicates. Consider a fragment modelling one of the examples, corresponding to a TEI fragment introduced earlier in the paper:

```
<sense resource=":geardagas_sense"
  typeof="ontolex:LexicalSense">
  <cit type="example">
    <quote>Scyle gumena gehwylc on his geá
        rdagum georne bipencan</quote>
    <cit type="translation" xml:lang="en">
      <quote>every man should in the days
          of his years well consider</quote
          >
    </cit>
    <bibl resource=":ge%C3%A1r-
        dagas_sense_1_Att_6_1"
      rel="frac:attestation"
      typeof="frac:Attestation"
      source="#Exon_Th">
      <span property="rdf:value">
          Scyle gumena gehwylc on his
          geárdagum georne bipencan
          every man should in the days of
          his years well consider,
          Exon 19 b
      </span>
      <span rel="frac:locus">
        <span property="dc:isPartOf"
          resource=":Exon_Th"/>
      </span>
    </bibl>
  </cit>
</sense>
```

A corresponding set of RDF triples is the following:

```
:geardagas_sense a ontolex:LexicalSense ;
  frac:attestation :ge%C3%A1r-
      dagas_sense_1_Att_6_1 .
:ge%C3%A1r-dagas_sense_1_Att_6_1 a frac:
    Attestation ;
  frac:locus [dc:isPartOf :Exon_Th] ;
  rdf:value "Scyle␣gumena␣gehwylc␣on␣his␣ge
      árdagum␣georne␣bipencan␣every␣man␣
      should␣in␣the␣days␣of␣his␣years␣well␣
      consider,␣Exon␣19␣b" .
```

In order to represent both `frac:locus` and `rdf:value` for the same attestation, we had to introduce additional elements which do not correspond to the semantic structure of the TEI document. In practice, this does not create any additional problems, and the added value of RDF semantics justify making the TEI structure a bit more verbose.

4.4.1. Advantages of RDFa

It is notable that RDFa semantics and processing rules are W3C-standardised and supported by off-

[20] As in the previous sections, the full example is available at https://anonymous.4open.science/r/attestationExample-9471/TEIAtt/geardagas-rdfa.xml

the-shelf parsers. Thus, compared to an implicit default interpretation of TEI elements in RDF (where a transformation needs to be explicitly provided for downstream applications, say, by means of XSLT/-GRDDL) and to the full, explicit, and potentially redundant specification of all RDF triples within a TEI document (by means of JSON-LD standoff, ‹xenoData› or TEI legacy formalisms), RDFa provides another advantage, i.e., it comes with a default interpretation of XML data structures in terms of RDF: the element with the `about` attribute will serve as subject of all triples specified along its descendant axis (unless another `about` URI is introduced); moreover, RDFa allows for the declaration of an element as a datatype property that takes its CDATA content as its value – without the need to repeat that content in an explicit triple. In that regard, RDFa is a compromise between fully specified, explicitly RDF-interpreted data and underspecified TEI data with an implicit RDF interpretation also in terms of verbosity. Nevertheless, it shares the great advantage of explicit RDF encodings that it can be directly processed with off-the-shelf RDF technology. Unlike currently dominant stand-off solutions based on WebAnnotation and JSON-LD, direct linking with OntoLex is both less verbose (in terms of triples) and more robust (since it is inline).

Beyond that, we argue that TEI+RDFa has a number of further advantages. For a start, it allows us to specify the semantics of TEI elements beyond those dealt with in TEI Lex-0. It also permits users to flexibly incorporate other RDF information (beyond OntoLex core) into TEI (as an example, consider the modelling of translations in DBnary) and facilitates a semantically explicit interpretation of TEI elements. Additionally, we can bundle markup and RDF data in a single file, thus making it self-contained and more robust than approaches operating with remote scripts (GRDDL/XSLT) or standoff annotations (e.g., WebAnnotation/JSON-LD).

5. Conclusion and Future Work

In this paper, we examined different ways to encode citations in digital lexical resources according to two standards: TEI and OntoLex. Showing how to encode citations in both, we argued for using the combination of the two, specifically, using the RDFa standard to enrich TEI modelling with RDF statements via attributes in XML tags. In fact, a formal bridge between RDF and TEI technologies has been under discussion within the TEI community for more than a decade, with RDFa long being proposed as a possible midway between the two.[21] However, prior to the formal standardization of RDFa in November 2015, RDFa was widely rejected because the TEI community did not want to introduce dependencies to a standard still under development. Instead, TEI-native solutions were introduced at that time. However, these have the downside that they are not well supported beyond the DH and XML communities, and even worse, alternative, incompatible modellings had been proposed and even found their way into the TEI P5 guidelines (Chiarcos and Ionov, 2019). Since then, however, RDFa development has led to a consolidated standard that has been stable for 8 years, so that it is worth reconsidering. Indeed, a number of projects that are already using this approach for different applications. For instance, Tittel et al. (2018) have demonstrated how RDFa can be used to link a digital edition in TEI with an attached glossary, how to query its RDF data by stacking web services, and how to cater this glossary in a way that is compatible with machine-readable dictionaries on the web, i.e., using OntoLex-Lemon (McCrae et al., 2017). An immediate consequence of this approach is that the Middle French vocabulary they describe can be immediately linked (and queried together with) external knowledge graphs and other lexical resources provided in OntoLex. And such datasets have been and are currently developed for a number of related languages, including Old Occitan (Weingart and Giovannetti, 2016), Latin (Passarotti et al., 2019), or for medieval French, Italian and Occitan in a conjoined fashion (Prifti et al., 2023). A different line of research is represented by Gelumbeckaitė et al. (2022), who use TEI+RDFa to encode intertextual relations between multiple digital editions of different texts. Here, 16th century Lithuanian sermons are linked with their Latin sources – and the same mechanism can also be used to link the edited text with lexical sources for the respective languages, i.e., machine-readable dictionaries of (Old) Lithuanian and Latin.

Another application of TEI+RDFa are the TEI corpora produced by the project Poetry Standardisation and Linked Open Data (POSTDATA), where a corpus of medieval Spanish poetry was linked with a detailed ontology describing aspects of literary analysis, structural, metrical and prosodic information, associated music, aspects of the transmission and additional information (Ruiz Fabo et al., 2021).

We present another application of TEI+RDFa, demonstrating the genericity of the approach, and solidifying the case for officially embracing RDFa and developing best practises for using RDFa in the context of the TEI. We expect the standoff approach also to be pursued in the future – as it is currently well-supported by tools such as Recogito,[22] or Hypothes.is[23] and annotation libraries such

[21] See, for example https://github.com/TEIC/TEI/issues/311.

[22] https://recogito.pelagios.org/

[23] https://web.hypothes.is/

as Annotorius[24] –, especially in conjunction with using Web Annotation and JSON-LD to provide an annotation layer over TEI editions, but in circumstances where the underlying edition is by itself still under development and is occasionally updates, this approach is not sufficiently robust. For such cases, where there is no technically viable alternative to the in-line annotation of RDF data in TEI documents (or documentation generated from them), we expect to see more integration in the future, and as it currently stands, RDFa seems to be the only way to do that in accordance with W3C standards, such that this data can be readily consumed by web clients and downstream applications.

6. Acknowledgements

This work has been made possible in COST Action CA18209 Nexus Linguarum, supported by COST (European Cooperation in Science and Technology). Fahad Khan was supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

7. Bibliographical References

Ben Adida, Mark Birbeck, Shane McCarron, and Ivan Herman. 2015. RDFa Core 1.1. Syntax and processing rules for embedding RDF through attributes. Technical report, W3C Recommendation. Third edition.

Ola Andersson, Robin Berjon, and Erik Dahlström etal. 2008. Scalable Vector Graphics (SVG) Tiny 1.2 Specification. Technical report, W3C Recommendation.

Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. Towards a Module for Lexicography in OntoLex. In *LDK Workshops*, pages 74–84.

Joseph Bosworth. 1898. *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth*, volume 2. Clarendon Press.

George Bruseker, Nicola Carboni, and Anaïs Guillem. 2017. Cultural heritage data management: The role of formal ontology and CIDOC CRM. *Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data*, pages 93–131.

Toby Burrows, Doug Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Andrew Morrison, Kevin Page, Lynn Ransom, et al. 2021. A New Model for Manuscript Provenance Research: The Mapping Manuscript Migrations Project. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies*, 6(1):131–144.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christian Chiarcos and Max Ionov. 2019. Linking the TEI: Approaches, Limitations, Use Cases. In *DH2019*, Utrecht, The Netherlands.

Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. Modelling frequency and attestations for ontolex-lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data in Digital Humanities*, pages 229–262. Springer International Publishing, Cham.

Rute Costa, Ana de Castro Salgado, Anas Khan, Sara Carvalho, Laurent Romary, Bruno Almeida, Margarida Ramos, Mohamed Khemakhem, Raquel Silva, and Toma Tasovac. 2021. MORDigital. In *eLex 2021, 7th biennial conference on electronic lexicography*, pages 1–14.

Christian Fäth and Christian Chiarcos. 2022. Spicy Salmon: Converting between 50+ Annotation Formats with Fintan, Pepper, Salt and Powla. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 61–68.

Jolanta Gelumbeckaitė, Øyvind Eide, Mortimer Drach, Christian Chiarcos, and Maxim Ionov. 2022. The Postil Time Machine: “God help those

[24] https://annotorious.github.io/

who have begun writing down these books in Lithuanian" (BP1591 II 77,2–4). In *Abstracts of the 26th International Conference on Historical Linguistics (ICHL26)*.

Ivan Herman, Ben Adida, Manu Sporny, and Mark Birbeck. 2015. RDFa 1.1 Primer. Rich Structured Data Markup for Web Documents. Technical report, W3C Recommendation. Third edition.

Annette Klosa. 2015. On corpus citations in monolingual general dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 36(1):72–87.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Marco Carlo Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019. The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. In *LDK (Posters)*, pages 6–11.

Elton Prifti, Wolfgang Schweickard, Maria Selig, and Sabine Tittel. 2023. Sprachdatenbasierte Modellierung von Wissensnetzen in der mittelalterlichen Romania (ALMA): Projektskizze. *Zeitschrift für romanische Philologie*, 139(2):301–332.

Laurent Romary and Toma Tasovac. 2018. TEI Lex-0: A target format for TEI-encoded dictionaries and lexical resources. In *TEI Conference and Members' Meeting*.

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2021. The Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*, 36(Supplement_1):i68–i80.

Manu Sporny. 2015. RDFa Lite 1.1. Technical report, W3C Recommendation. Second edition.

Manu Sporny, Shane McCarron, Ben Adida, Mark Birbeck, Gregg Kellogg, Ivan Herman, and Steven Pemberton. 2015. HTML+RDFa 1.1. Support for RDFa in HTML4 and HTML5. Technical report, W3C Recommendation. Second edition.

Sabine Tittel, Helena Bermúdez Sabel, and Christian Chiarchos. 2018. Using RDFa to Link Text and Dictionary Data for Medieval French. In *6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science*, pages 30–38. European Language Resources Association (ELRA).

Anja Weingart and Emiliano Giovannetti. 2016. A Lexicon for Old Occitan Medico-Botanical Terminology in Lemon. In *SWASH@ ESWC*, pages 25–36.