eLexicon. Dictionary of Polish Medieval Latin: from TEI encoding to an eXist-db application

Krzysztof Nowak
krzysztof.nowak@ijp-pan.krakow.pl
Polish Academy of Sciences Krakow, Poland

Introduction

From the *Lexicon* to the *eLexicon*

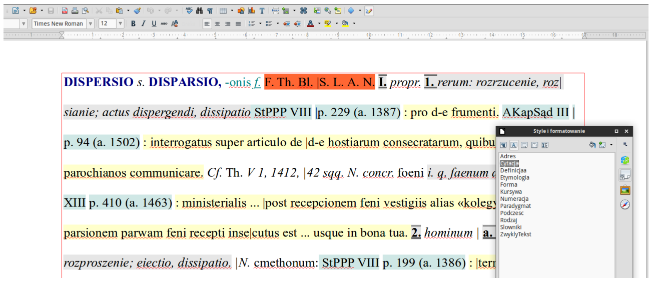
The first fascicle of the *Lexicon mediae et infimae Latinitatis Polonorum* (Dictionary of Polish Medieval Latin, henceforth LMILP) was published in 1953. The project aims at providing an exhaustive account of the Latin vocabulary used on Polish territory during the Middle Ages. Addressed to a scholarly public, the dictionary does not make many concessions to a less advanced user. Information is often conveyed only indirectly, by means of typographic devices or is left to be inferred by the reader. The project of retro-digitization of the LMILP started in the mid-2011 and was completed by the mid-2014. The web application, although completed, is still subject to modifications and refinements.

Dictionary Annotation

The XML encoding of the dictionary was by no means an ultimate goal of the project. Instead, the idea was to make the rich content of the LMILP fully searchable through a user-friendly interface. This objective, however, has deeply influenced the XML schema design. The TEI (TEI Consortium 2016) was chosen as an annotation standard because at the time the work started it had been already employed in major electronic lexicography projects (Lewis-Short by the Perseus Project; DuCange by the ENC). The popularity that the standard had gained among scholars contributed to emergence of lively community which produced documentation and use cases which supplemented the "Dictionaries" chapter of the TEI Guidelines. Also, the very fact that the TEI Guidelines offered a set of ready-to-use tags for the description of lexicographic content was not without significance. Finally, the TEI XML was supported by major software providers, an important factor for the project in which adaptation of existing rather than writing new software was planned.

Workflow

The paper dictionary was scanned and the output of the OCR program (Abbyy FineReader 11) was exported to ODT files; from each a *content.xml* file was extracted and then applied a series of XSL transformations. The main goal was to simplify styles that were automatically generated by the OCR software. Resulting XML files underwent second phase of XSL processing in which constitutive parts of the dictionary, such as entry, headword, sense definition *etc.*, were encoded. The output XML files were again retranslated into ODT format: entries were encoded as paragraph styles, other tags were represented as character styles. In the next phase of the project the lexicographers started to proofread OCR text and correct errors of automatic annotation. This task was performed with the help of LibreOffice Writer exclusively without annotators being actually conscious of the underlying XML structure. From the practical point of view, annotation consisted in verifying whether automatic XSL processing produced correct styles; if this was not the case correct style had to be applied, as in standard text processing task.



DISPERSIO *s.* **DISPARSIO,** -onis *f.* F. Th. Bl. |S. L. A. N. **I.** *propr.* **1.** *rerum: rozrzucenie, roz|*
sianie; actus dispergendi, dissipatio StPPP VIII |p. 229 (a. 1387) : pro d-e frumenti. AKapSąd III |
p. 94 (a. 1502) : interrogatus super articulo de |d-e hostiarum consecratarum, quibu
parochianos communicare. *Cf.* Th. *V 1, 1412, |42 sqq. N. concr.* foeni *i. q. faenum*
XIII p. 410 (a. 1463) : ministerialis ... |post recepcionem feni vestigiis alias «kolegy
parsionem parwam feni recepti inse|cutus est ... usque in bona tua. **2.** *hominum* | **a.**
rozproszenie; eiectio, dissipatio. |*N.* cmethonum: StPPP VIII p. 199 (a. 1386) : |terr

Fig. 1: "XML-unaware" annotation in the LibreOffice window

This approach allowed for reducing the learning curve to a minimum so that the team members could focus on the lexicographic content. However, it also has a serious drawback: annotation in the text editor cannot produce more complex hierarchies, since paragraph and character styles allow for representing at best two levels deep nesting.

TEI for the *eLexicon*

A guiding principle of the subsequent TEI annotation was to combine *editorial* and *lexical view* of the dictionary content by (1) preserving its original text and (2) storing normalized data in attributes and empty XML elements. Typographic properties of the text, on the other hand, were not generally encoded, they are easily reconstructible though.

Automatic and manual annotation consisted in three major procedures:

a. translation: custom ODT styles (corresponding to elements of dictionary structure) were "translated" into respective TEI elements or attributes;
b. grouping: deeply nested XML structure was produced from flat annotation;
c. enrichment: implicit information was made explicit.

Translation

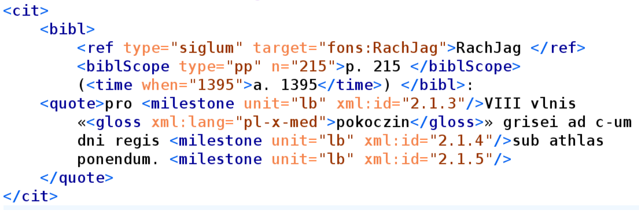
The paper justifies some of the annotation choices. Special attention is given to the peculiarities of encoding a scholarly lexicographic work.

1. **<entryFree>** element was chosen as a container for dictionary entries.
2. Essential features of the dictionary macro- and microstructure are encoded as: **<form>, <orth>; <gramGrp>, <gen>, <iType>, <pos>; <etym>, <lang>, <mentioned>; <cit>, <bibl>, <biblScope>, <date>, <quote>; <sense>, <usg>, <def>, <gloss>; <xr>, <ref>; <lbl>; <re>, <certainty>, <oVar>, <note>.**
3. Content and form peculiarities of the LMILP are reflected in respective attributes. So, for example, functional variation of the entries is represented in the **@type** attribute of the **<entryFree>** and can take one of the following values: **main, xref, hom.**
4. The TEI schema was only lightly customized: unused elements were deleted; a few content restrictions were overcome.

Grouping: adding depth

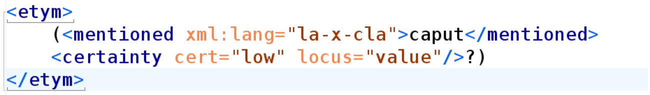
The flat entry structure had to undergo heavy XSL processing, so deep nesting typical of scholarly dictionaries could eventually emerge. Relative ease of the XML-unaware manual annotation resulted in time-consuming post-processing. The **xsl:for-each-group** XSL function was employed in order to structure:

1. citation groups:



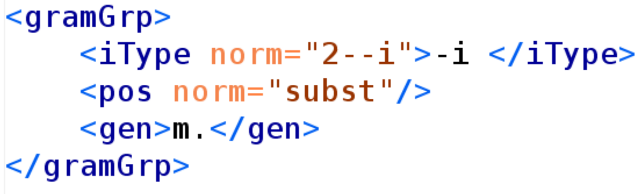
```
<cit>
    <bibl>
        <ref type="siglum" target="fons:RachJag">RachJag </ref>
        <biblScope type="pp" n="215">p. 215 </biblScope>
        (<time when="1395">a. 1395</time>) </bibl>:
    <quote>pro <milestone unit="lb" xml:id="2.1.3"/>VIII vlnis
        «<gloss xml:lang="pl-x-med">pokoczin</gloss>» grisei ad c-um
        dni regis <milestone unit="lb" xml:id="2.1.4"/>sub athlas
        ponendum. <milestone unit="lb" xml:id="2.1.5"/>
    </quote>
</cit>
```

2. etymological groups:



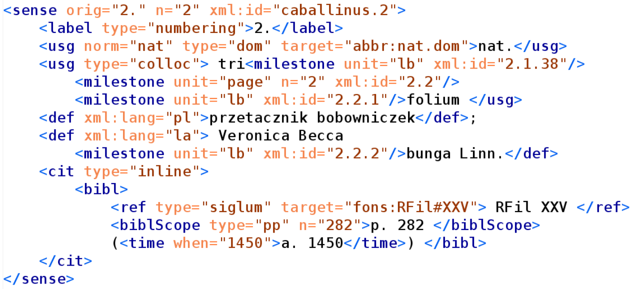
```
<etym>
    (<mentioned xml:lang="la-x-cla">caput</mentioned>
    <certainty cert="low" locus="value"/>?)
</etym>
```

3. PoS and grammar groups:



```
<gramGrp>
    <iType norm="2--i">-i </iType>
    <pos norm="subst"/>
    <gen>m.</gen>
</gramGrp>
```

4. sense groups:



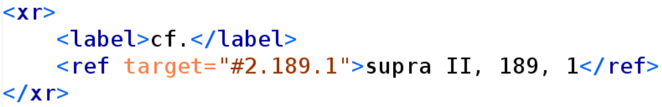
```
<sense orig="2." n="2" xml:id="caballinus.2">
    <label type="numbering">2.</label>
    <usg norm="nat" type="dom" target="abbr:nat.dom">nat.</usg>
    <usg type="colloc"> tri<milestone unit="lb" xml:id="2.1.38"/>
        <milestone unit="page" n="2" xml:id="2.2"/>
        <milestone unit="lb" xml:id="2.2.1"/>folium </usg>
    <def xml:lang="pl">przetacznik bobowniczek</def>;
    <def xml:lang="la"> Veronica Becca
        <milestone unit="lb" xml:id="2.2.2"/>bunga Linn.</def>
    <cit type="inline">
        <bibl>
            <ref type="siglum" target="fons:RFil#XXV"> RFil XXV </ref>
            <biblScope type="pp" n="282">p. 282 </biblScope>
            (<time when="1450">a. 1450</time>) </bibl>
    </cit>
</sense>
```

Enrichment: expanding the dictionary content

Considerable effort was put into enriching the original content of the dictionary, namely: (1) resolving references, (2) normalizing strings, (3) adding redundant and/or inferred information.

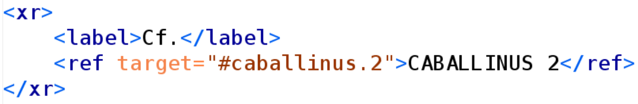
Resolving references

A typical reference to an exact location in the dictionary text was encoded as follows:



```
<xr>
    <label>cf.</label>
    <ref target="#2.189.1">supra II, 189, 1</ref>
</xr>
```

References to a specific entry or sense relied on the **@xml:id** attribute:



```
<xr>
    <label>Cf.</label>
    <ref target="#caballinus.2">CABALLINUS 2</ref>
</xr>
```

The encoding of most frequent type of references (pointing to a source of a language use example) is illustrated in the section II B 4 above.

String normalization

By string normalization, we mean a set of various procedures applied in order to generate a *lexical view* of the dictionary content. Standardized strings are usually stored in **@norm** attributes of such elements as language or usage labels, prepositional and inflectional patterns *etc.* Their primary goal is to enable unified search that would be agnostic of the exact formulation of the paper dictionary. For example, when looking up philosophy-related terms one should be able to retrieve them no matter whether they have been marked with a *phil.* label or with more verbose formula *in textibus philosophicis* "in philosophical texts", as both are annotated as **@norm="phil".** The second major goal of the normalization was to render chronological information consistent and machine-readable. Its proper annotation should reflect the fact that many medieval texts cannot be dated but only approximately. Apart from some obvious cases (`@when` attribute stores a year date, for example **<time when="1450">a. 1450</time>)** the LMILP employs:

1. century dates **(<time notBefore="1401" notAfter="1500">saec. XV</time>)**
2. imprecise dates in year **(<time notAfter="1120">ante 1120</time>**) or century format (**<time notBefore="1401" notAfter="1450">saec. XV in.</time>**).

<u>Making information explicit</u>

Finally, substantial effort has been devoted to making explicit what is not expressed directly in the paper dictionary, but left to be inferred by an expert user. In the LMILP, this is the case, for example, of a part of speech label which is provided for adverbs or conjunctions, but is normally omitted from verb or noun entries. Empty elements have therefore been created and their attributes filled with the inferred content. So, in a typical case, an element **<pos norm="subst"/>** would be appended to a **<gramGrp>** group whenever the paper dictionary informs about a word's part-of-speech only indirectly, by means of a gender label (*f.* for Lat. *femininum*) or inflectional ending typical of nouns (-*ae*):

The Dictionary Web Application

The last part of the paper briefly presents the overall architecture of the dictionary web application, its user interface having been already described elsewhere (Nowak 2014). Written entirely in XQuery, the application is served directly from the eXist-db instance with HTML and JavaScript code being equally stored in a database or generated on the fly. The presentation focuses on those features available in the eXist-db which are of critical importance for dictionary application design:

1. Various types of indexes available in the eXist-db allow for efficient retrieval of content from deeply nested dictionary files and dispersed textual data.
2. A templating system allows for fine-grained web presentation of the XML content.
3. A URL rewriting engine supports a logical system of dictionary content access.
4. An out-of-the-box RESTful API exposes lexicographic content to external applications.

In the conclusion, I will also point to some difficulties that I have encountered and which have mainly to do with handling application's state, a crucial feature for multi-language tools which require storing user search results.

Bibliography

Nowak, K. (2014). 'The eLexicon Mediae et Infimae Latinitatis Polonorum. The Electronic Dictionary of Polish Medieval Latin'. In *The User in Focus. Proceedings of the XVI EURALEX International Congress: 15 - 19 July 2014, Bolzano/Bozen*, edited by Andrea Abel, Chiara Vettori, and Natascia Ralli, 793–806. Bolzano: EURAC Research.

TEI Consortium. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 3.0.0. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.