eLexicon

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

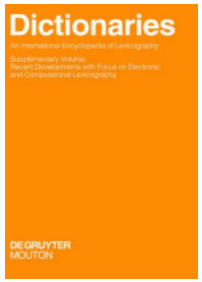
Krzysztof Nowak

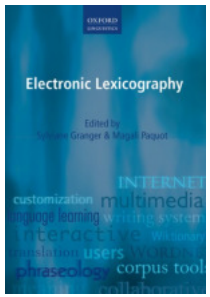
Lexicon Mediae et Infimae Latinitatis Polonorum
Institute of Polish Language
Polish Academy of Sciences

8/11/2017

Electronic Lexicography

Electronic Lexicography as a *champ scientifique* (Bourdieu 1976)

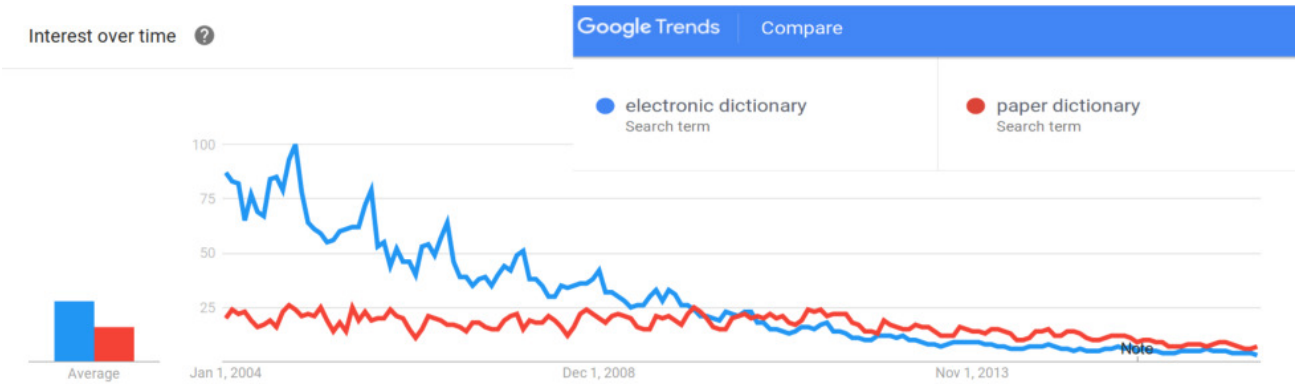




- conferences
- research problems
- research objects
- methods

handbooks monographs bibliography

Electronic Dictionary or Dictionary *tout court*?

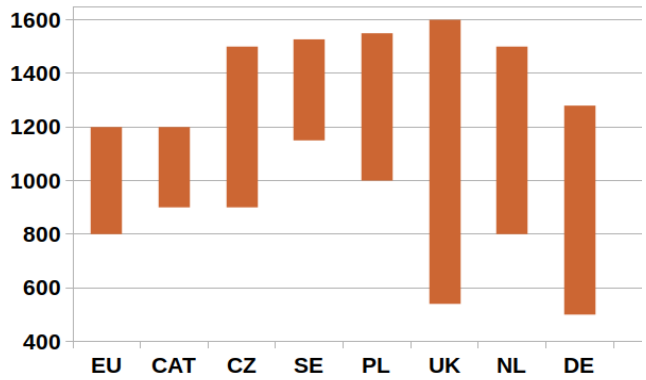


Outline

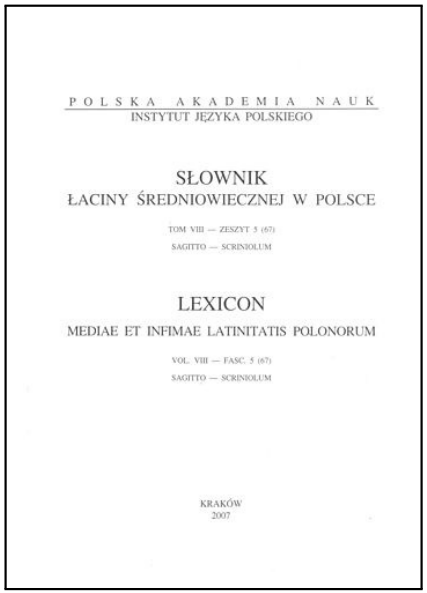
1. Medieval Latin and its dictionaries
2. Electronic Dictionary of Polish Medieval Latin: principles and workflow
3. micro-structure $\mapsto$ TEI (choices, challenges, lessons)
4. eXist-db Web App
5. Perspectives

Medieval Latin and its Dictionaries

Time Coverage



Polish Medieval Latin and its Dictionary


POLSKA AKADEMIA NAUK
INSTYTUT JĘZYKA POLSKIEGO

SŁOWNIK
ŁACINY ŚREDNIOWIECZNEJ W POLSCE

TOM VIII — ZESZYT 5 (67)
SAGITTO — SCRINIOLUM

LEXICON
MEDIAE ET INFIMAE LATINITATIS POLONORUM

VOL. VIII — FASC. 5 (67)
SAGITTO — SCRINIOLUM

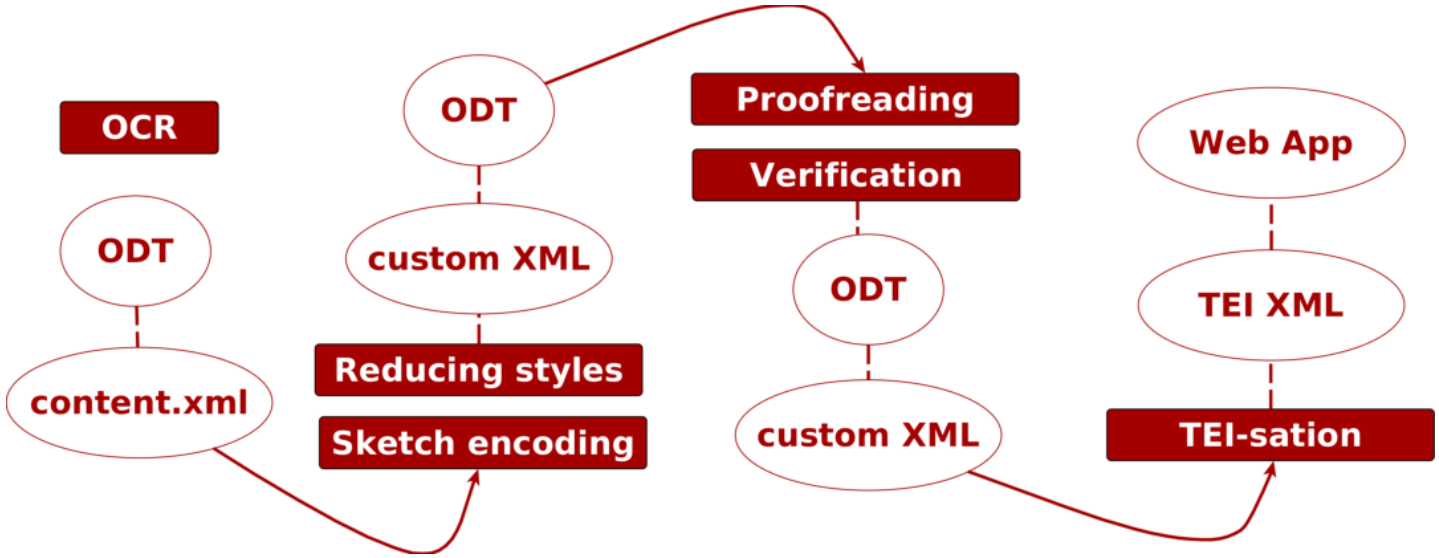
KRAKÓW
2007

- first fascicle: 1953
- time coverage: ca. 1000 - ca. 1550
- exhaustive: documents every word to be found in Polish sources

Dictionary Digitization

- goal
  - making lexicographic data available to research community
  - widening the audience
  - foundation for NLP tools
- principles
  - open source tools
  - standardization
  - primary version: XML, derivatives: database etc.
  - applicable in other domains: corpus, digital edition

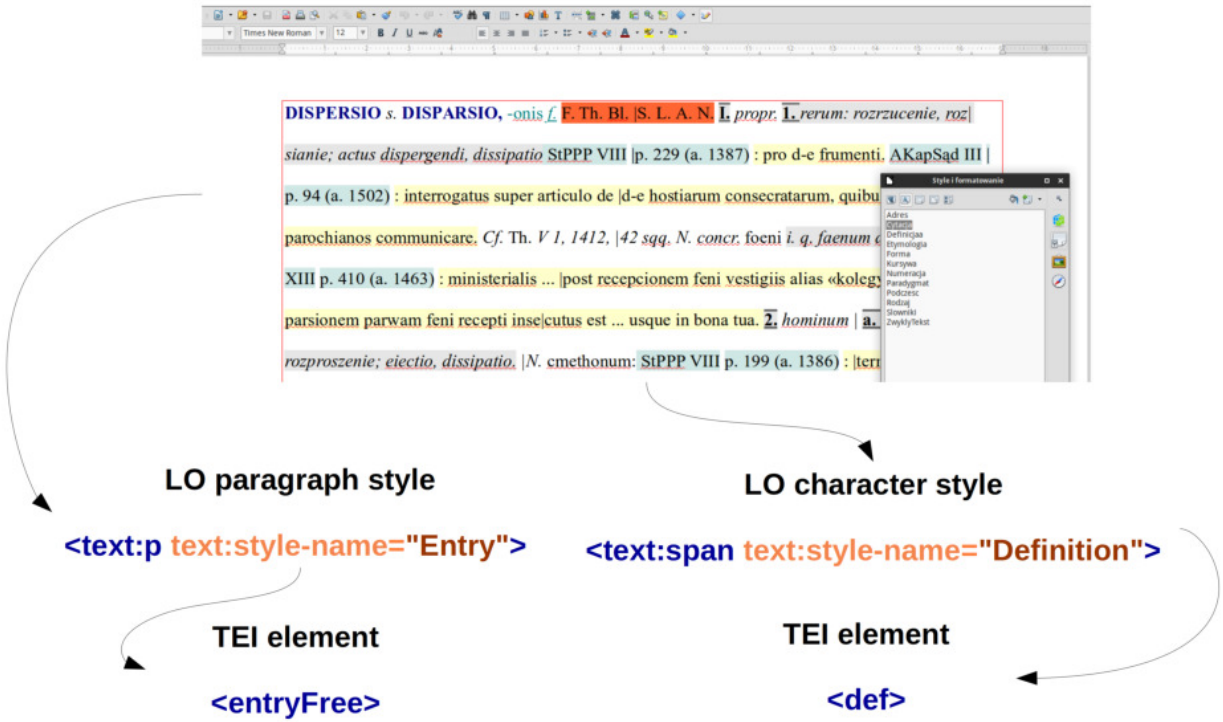
OCR and XSLT pre-processing



Typographic and string hints for automatic processing

font style	labels	indentation
RHINOCEROS *s.* **RINOCEROS** *s.* **RINOCERUS** ...	*locut.* = starts a collocation list, *N.* = starts a note block	␣**RETRACTABILIS**... ␣**RETRACTATIO**...

LibreOffice encoding ...



... that turned out to be a bad idea

- flat structure results in painstaking post-processing
- the team's XML skills didn't improve
- maybe appropriate in less structured resources?

Translating Dictionary Micro-structure to TEI

Entry block: typical

```
<entryFree n="digitus.1" type="main">
...
</entryFree>
```

Entry block: less typical

```
REVERA cf. supra 399,12 sqq. (anomale!)

<entryFree xml:id="entry-N" n="REVERA"
    type="xref">
        <xr>
          <lbl>cf.</lbl>
          <ref target="#right(element
       (//pb[@n='399'/following-sibling::lb[@n='12']))"
       type="intra">supra 399,12sqq</ref>
          <lbl>(anomale!)</lbl>
        </xr>
</entryFree>
```

Extensive grouping

- grammar

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

- sense and citation

```
<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>
```

TEI: normalization and explicitness

Unifying labelling with @norm

from the point of view of dictionary search:

in textibus philosophicis
('in philosophical texts')
≈
phil.
(a label preceding sense definition)

Handling inexact dating

```
'15th c.'
<time notBefore="1401" notAfter="1500">saec. XV</time>

'before 1120'
<time notAfter="1120">ante 1120</time>

'beginning of the 15th c.'
<time notBefore="1401" notAfter="1450">saec. XV in.</time>
```

Explicitness

By convention **ROSA,** -ae *f.* 'rose' entry doesn't contain a PoS label: it is meant to be deduced by a user.

```
<gramGrp>
    <iType norm="1-a-ae-f">-ae </iType>
    <gen norm="f">f.</gen>
    <pos norm="subst"/>
</gramGrp>
```

TEI: challenges, problems

Element restriction

quantifying collocations

```
<usg type="colloc">
     ...
     <usg type="plev">
     frequent
     </usg>
</usg>
```

Element misuse

orig. incert.
'of uncertain origin'

```
<certainty cert="0">
     orig. incert.
</certainty>
```

Extension: language codes

- Classical Latin: **la-x-cla** (usually implicit)
- Medieval Latin: **la-x-med** (usually implicit)
- **lang-x-med** other medieval varieties (without going into more detail)
  - French (exp. as *Fr.*)
  - Italian (exp. as *It.*)
  - Polish (exp. as *Pol.*)
  - German (exp. as *Germ.*)

Lessons learnt (sometimes too late)

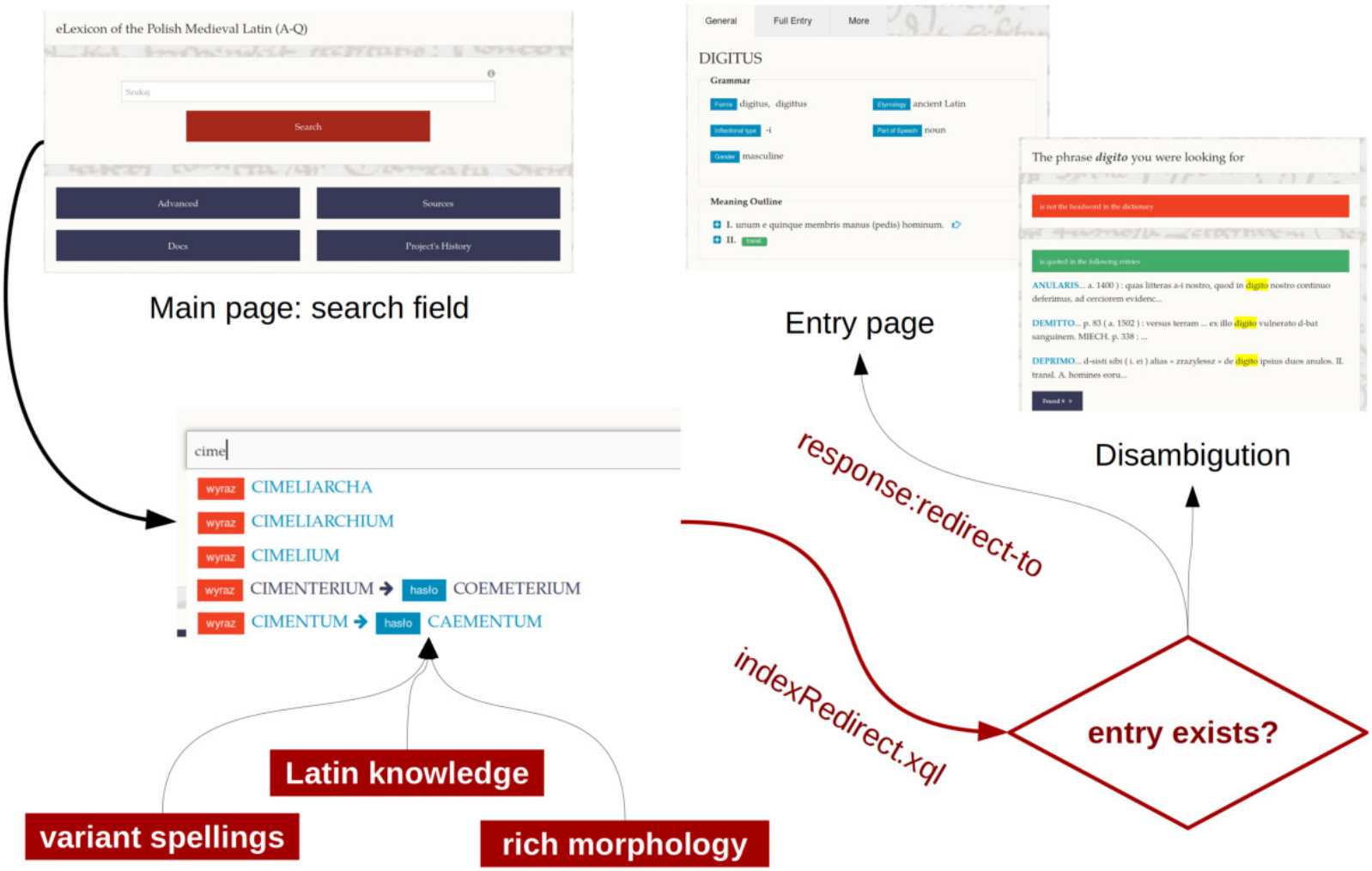
1. TEI can be customized.
2. Some encoding can wait.
3. Choosing a tag doesn't mean subscribing to linguistic theory.
4. Think about data storage, retrieval, and presentation.
5. There are usual more than one way to do things with TEI.
6. You have to do it on your own.

What would help: centralizing interpretations

TEI practitioners as an interpretive community (Fish 1976)

- the Book: Guidelines
- interpretation tradition: TEI Journal > conference proceedings > project documentation > internet fora *etc.*
- authoritative > non-authoritative > heretic readings

Dictionary App: Basic Search



Dictionary App: disambiguation page

indexes

- full text

  `<text match="//tei:def"/>`
- range

  `<create qname="gen" .../>`

search

- Lucene

  `ft:query(def,"digito")`
- range

  `gramGrep[gen = "femininum"]`

simple suggestions

- fuzzy search

  `<fuzzy min-similarity="0.6"/>`

Dictionary App: advanced search

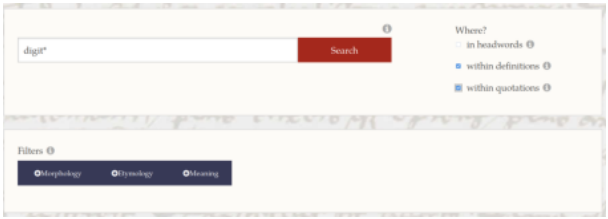


Figure: Search with faceting

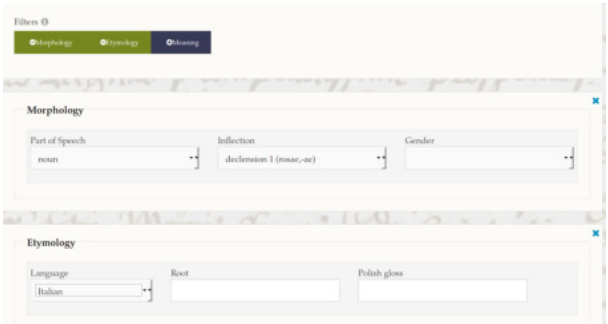


Figure: AdvancedBrowsing

indexes

- full text

```
<text match="//tei:def"/>
```

- range

```
<create qname="gen" .../>
```

search

- Lucene

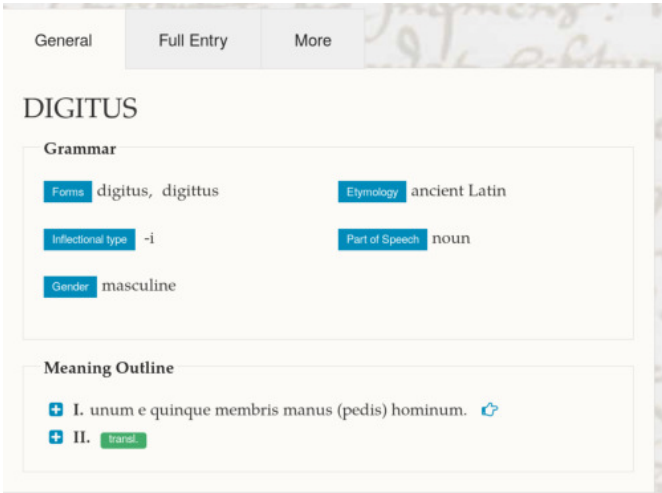
```
ft:query(def,"digito")
```

- range

```
gramGrep[gen = "femininum"]
```

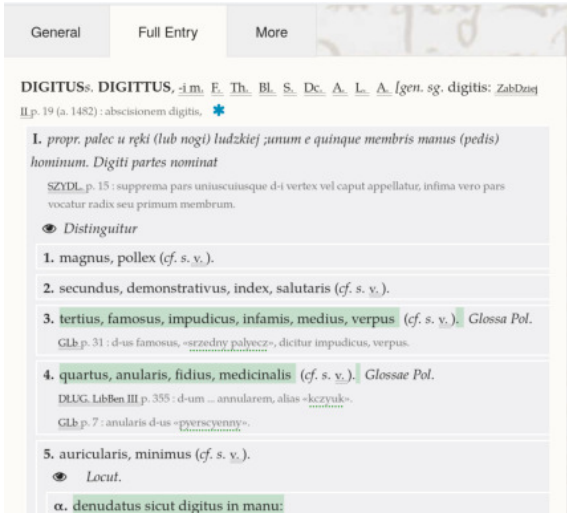
Dictionary App: displaying entry with XSLT

Figure: basicView.xsl



- perfect for quick browsing
- addressed to beginning users
- presents only selection of an entry

Figure: advView.xsl



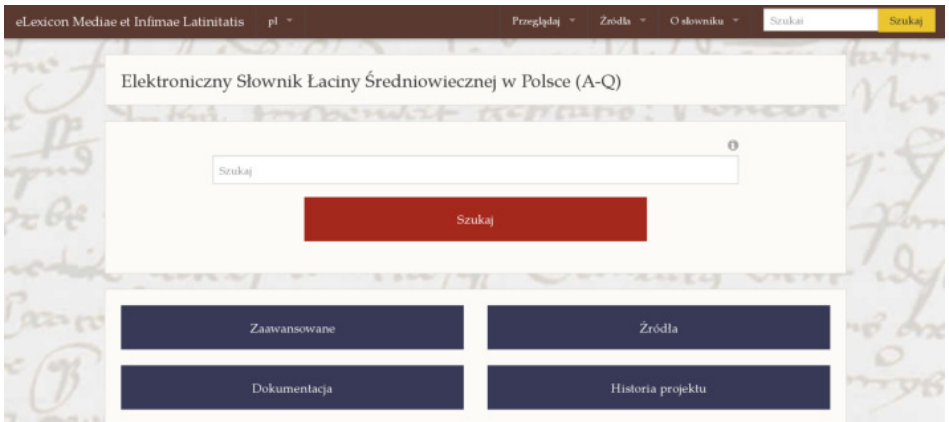
- perfect for in-depth browsing
- addressed to expert users
- a (rather desperate) attempt to make the dictionary user-friendly

Perspectives

What now?

- XML review
- app code review
- linking to other resources
- generalizing approach for easy deployment of the TEI-encoded dictionaries

eDictionary



scriptores.pl/elexicon

eCorpus

scriptores.pl/en/efontes