NLL: Language resources

Baltic HLT, 07.010.2016

Jana Kikāne,
Head of Digital Heritage Centre
National Library of Latvia
We have digitized

• Historical newspapers
  • More than 3 000 000 pages (200 000 issues)
  • Latvian, German, Russian
  • 80% of content is OCRed
  • periodika.lv, europeana-newspapers.eu

• Books
  • More than 1 500 000 pages (8 000 volumes)
  • Latvian etc.
  • 100% of content is OCRed
  • gramatas.lndb.lv
Our plans in digitization 2020

• Historical newspapers
  • To reach 5 500 000 digitized pages
  • Latvian, Russian, German, Polish
  • 100% of content will be OCRed

• Books
  • To reach 2 000 000 digitized pages
  • Latvian
  • 100% of content will be OCRed
• laboratorija.lndb.lv
  • Indexed corpora of newspaper and book texts
Building Digital Humanities Infrastructure

- laboratorija.indb.lv
  - Indexed corpora of newspaper and book texts
  - Named Entity database
    - Persons
    - Places
    - Institutions
    - Multi-language content
• **laboratorija.lndb.lv**
  • Indexed corpora of newspaper and book texts
  • Named Entity database
    • Persons
    • Places
    • Institutions etc.
    • Multi-language content
  • Time-sensitive dictionary
    • Riga Street Name database
Building Digital Humanities Infrastructure

• Historical text modernization service
  • Correction of OCR errors
  • Modernization of orthography
  • «Translating» historical words into contemporary language

<table>
<thead>
<tr>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>m → w</td>
</tr>
<tr>
<td>f → s</td>
</tr>
<tr>
<td>w → v</td>
</tr>
<tr>
<td>ah → ā</td>
</tr>
<tr>
<td>tsch → č</td>
</tr>
<tr>
<td>ee → ē</td>
</tr>
<tr>
<td>ee → ie</td>
</tr>
<tr>
<td>ee → ee</td>
</tr>
</tbody>
</table>
Building Digital Humanities Infrastructure

- Linked data prototype collection «Rainis un Aspazija»
  - Content
    - Rainis’ and Aspazija’s literary works, letters, photography, documents, video
  - Annotations of works
  - Commentaries to letters
  - Linked data
Building Digital Humanities Infrastructure

- Linked data prototype collection «Rainis un Aspazija»
  - Content
    - Rainis’ and Aspazija’s literary works, letters, photography, documents, video
  - Annotations of works
  - Commentaries to letters
- Linked data
- The data network
To Do’s in DHI development

• To continue digitization

• To make a strategy for labarotorija.lndb.lv development to meet needs of DH researchers
  • Specialized text corpora
  • Free available text analysis tools developed by NLL and/or in corporation with language research organizations/industry
    • For creation of linked data
    • Historical text modernization service etc.
  • Wiki about using free available tools
    • Word frequency and other statistics in corpora
    • Stylomety etc.

• To create new version of academic repository academia.lndb.lv for aggregation and storage of research results, incl. data (RDM)
Thank you!