NLL: Language resources

Baltic HLT, 07.010.2016

Jana Ķikāne,
Head of Digital Heritage Centre
National Library of Latvia

We have digitized

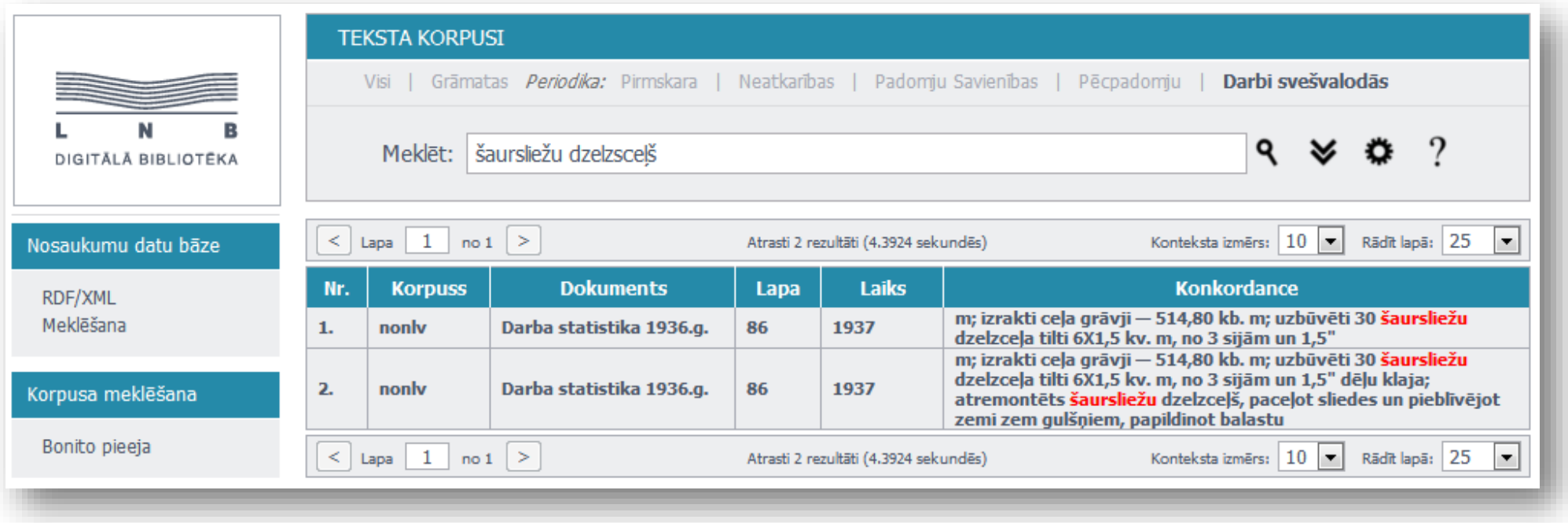
- Historical newspapers
  - More than 3 000 000 pages (200 000 issues)
  - Latvian, German, Russian
  - 80% of content is OCRed
  - periodika.lv, europeana-newspapers.eu
- Books
  - More than 1 500 000 pages (8 000 volumes)
  - Latvian etc.
  - 100% of content is OCRed
  - gramatas.lndb.lv

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

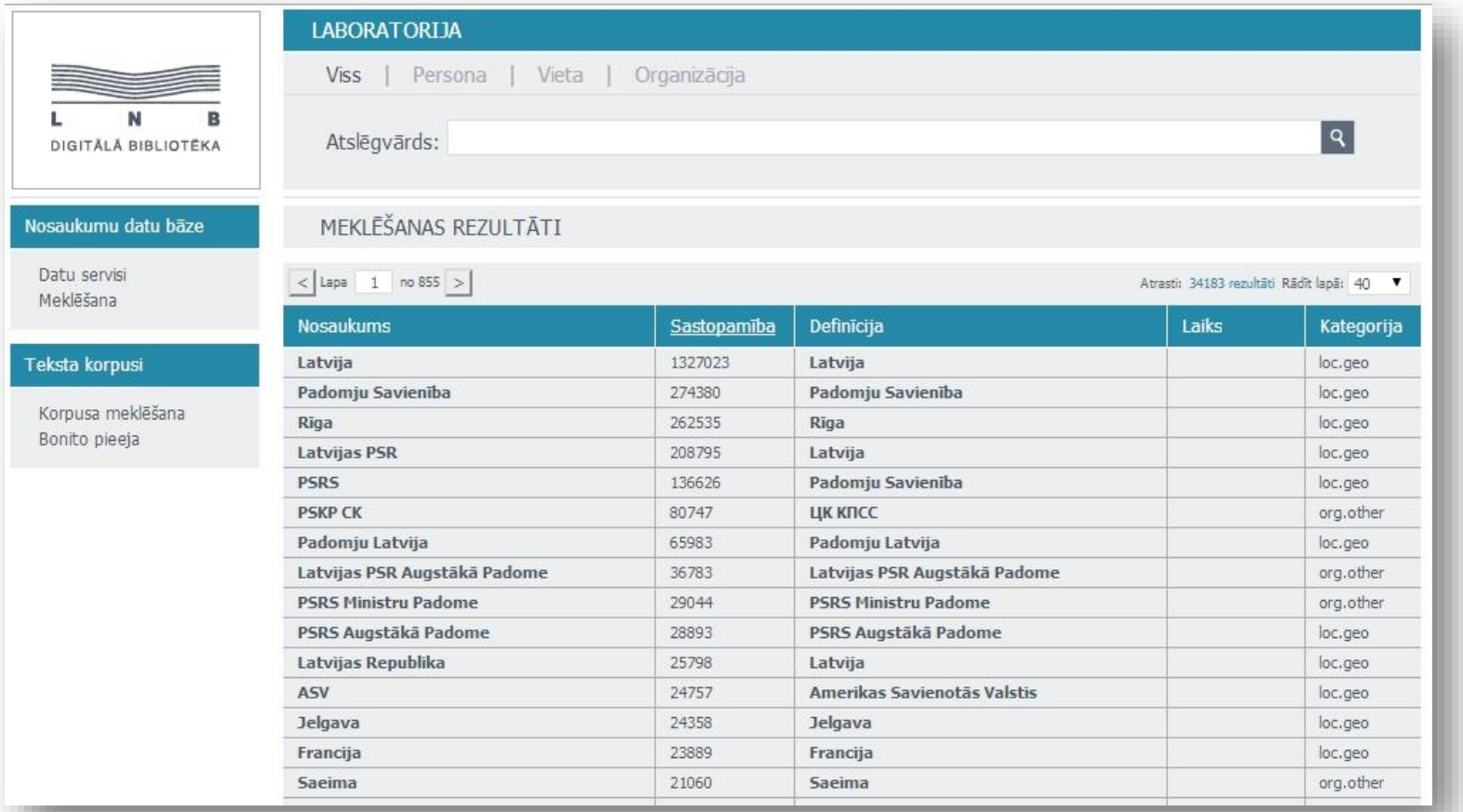
Building Digital Humanities Infrastructure (1)

- laboratorija.lndb.lv
  - Indexed corpora of newspaper and book texts



Building Digital Humanities Infrastructure (1)

- laboratorija.lndb.lv
  - Indexed corpora of newspaper and book texts
  - Named Entity database
    - Persons
    - Places
    - Institutions
    - Multi-language content



Nosaukums	Sastopamība	Definīcija	Laiks	Kategorija
Latvija	1327023	Latvija		loc.geo
Padomju Savienība	274380	Padomju Savienība		loc.geo
Rīga	262535	Rīga		loc.geo
Latvijas PSR	208795	Latvija		loc.geo
PSRS	136626	Padomju Savienība		loc.geo
PSKP CK	80747	ЦК КПСС		org.other
Padomju Latvija	65983	Padomju Latvija		loc.geo
Latvijas PSR Augstākā Padome	36783	Latvijas PSR Augstākā Padome		org.other
PSRS Ministru Padome	29044	PSRS Ministru Padome		org.other
PSRS Augstākā Padome	28893	PSRS Augstākā Padome		loc.geo
Latvijas Republika	25798	Latvija		loc.geo
ASV	24757	Amerikas Savienotās Valstis		loc.geo
Jelgava	24358	Jelgava		loc.geo
Francija	23889	Francija		loc.geo
Saeima	21060	Saeima		org.other

Building Digital Humanities Infrastructure (1)

- 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

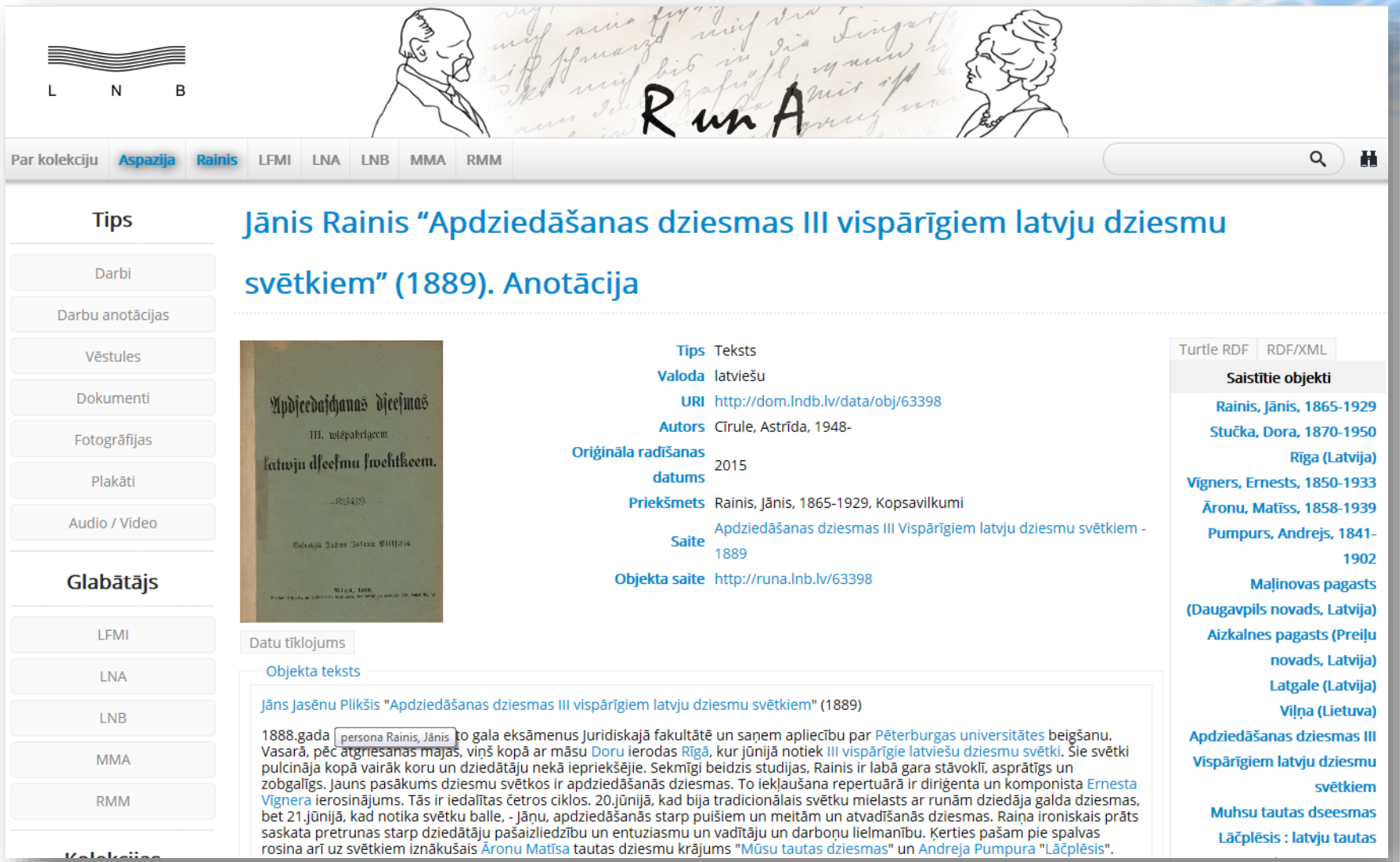
Building Digital Humanities Infrastructure (2)

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

Rule
m → w
f → s
w → v
ah → ā
tsch → č
ee → ē
ee → ie
ee → ee

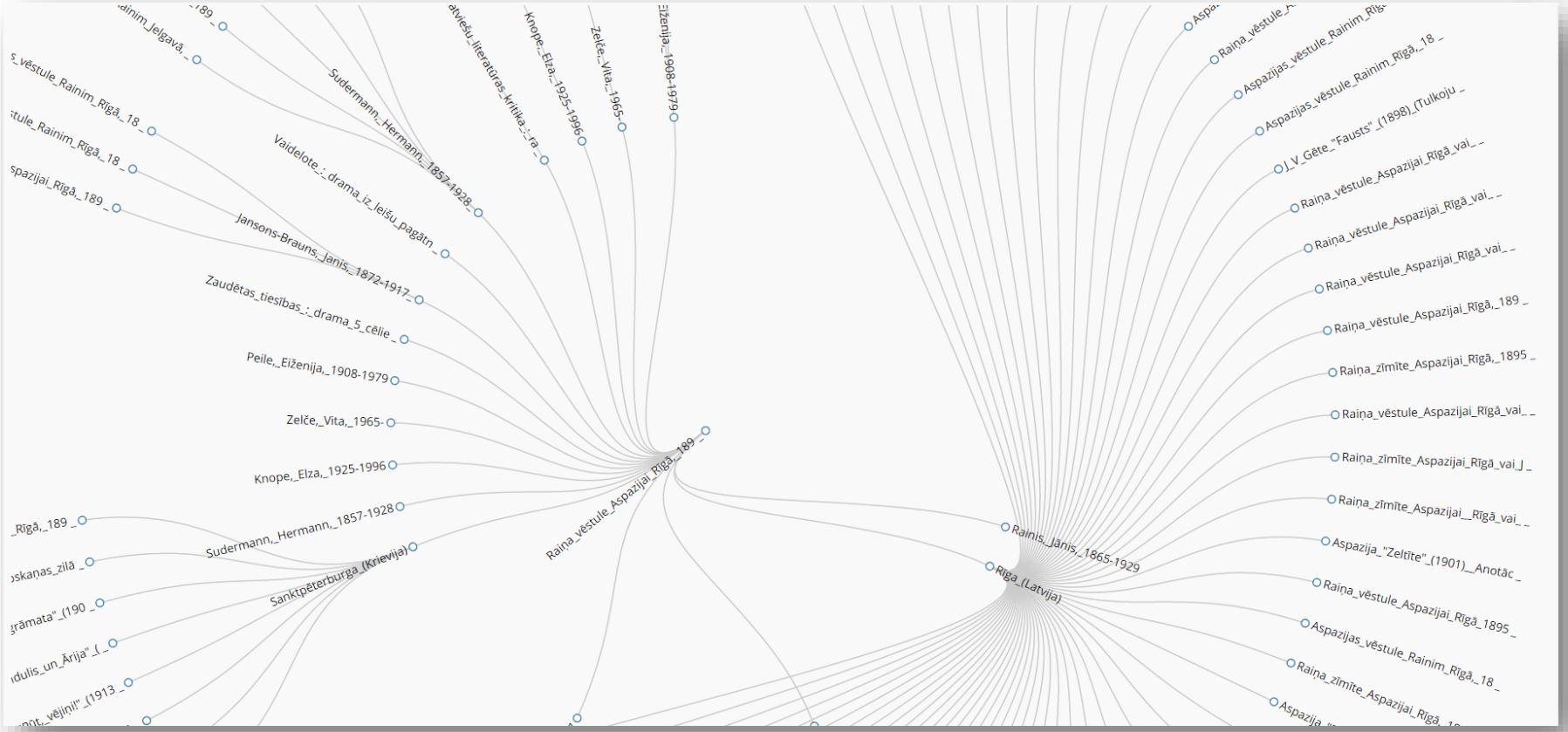
Building Digital Humanities Infrastructure (3)

- 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 (3)

- 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!