

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

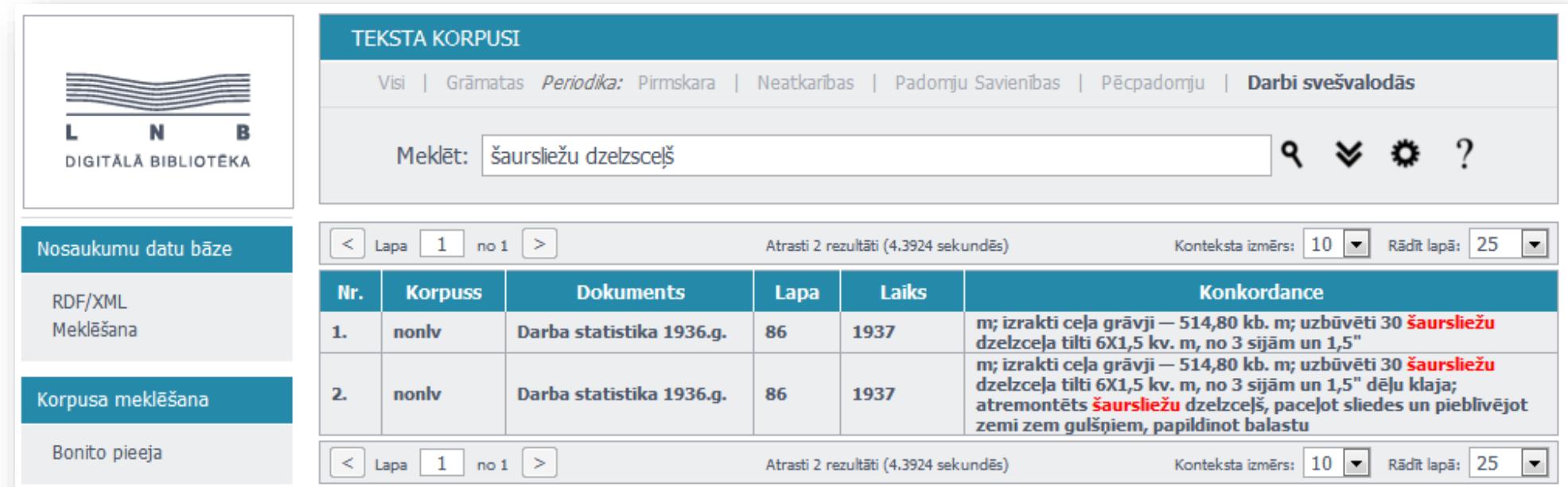
Our plans in digitization 2020

- 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



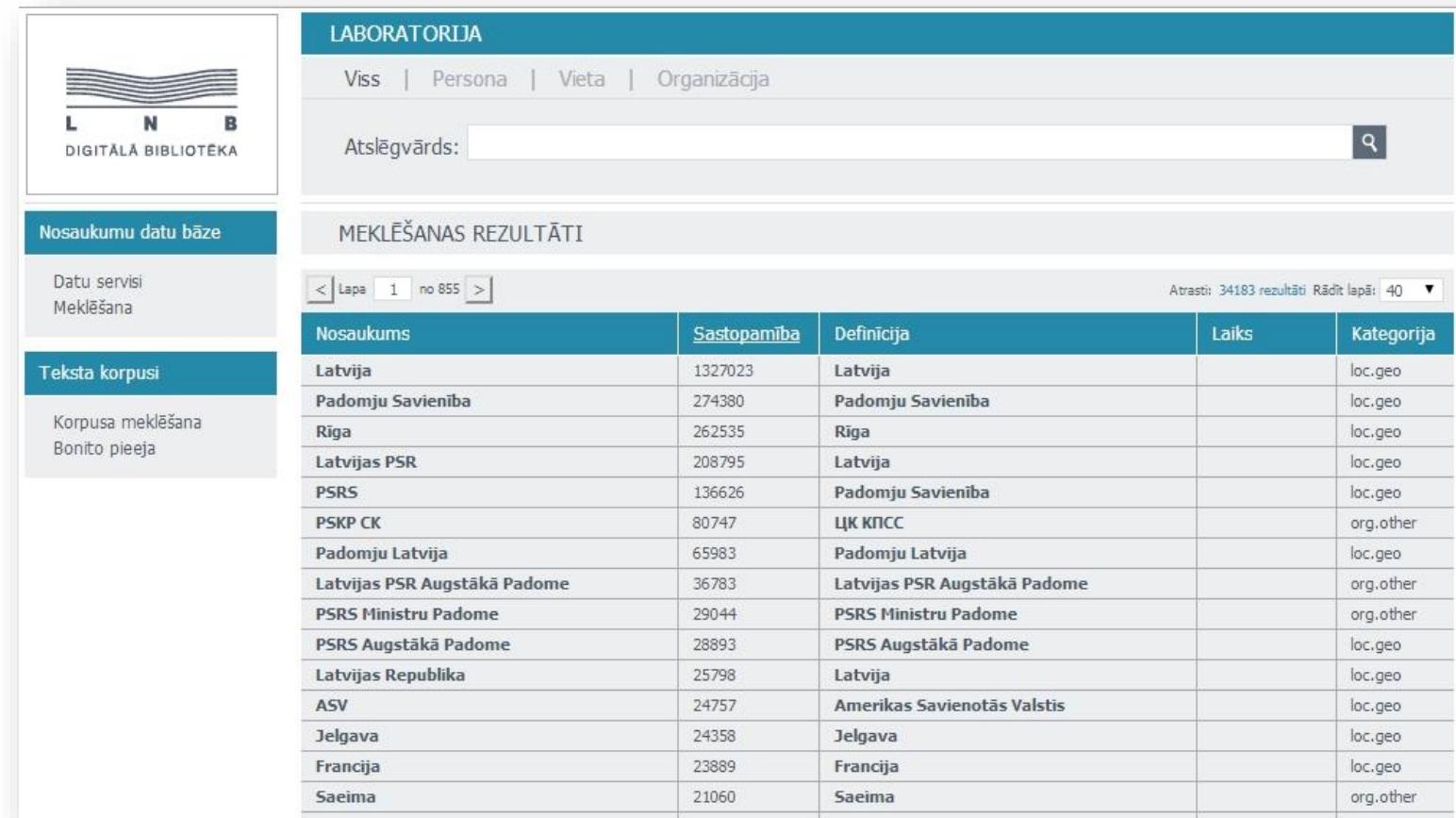
The screenshot shows a search interface for a digital library corpus. The top navigation bar includes links for Visi, Grāmatas, Periodika (with sub-links Pirmskara, Neatkarības, Padomju Savienības, Pēcpadomju), Darbi svešvalodās, and a search input field containing 'šaursliežu dzelzceļš'. Below the search bar, a message indicates 2 results found in 4.3924 seconds. The main content area displays two entries from the document 'Darba statistika 1936.g.' (page 86, year 1937) with matching snippets highlighted in red.

Nr.	Korpuss	Dokuments	Lapa	Laiks	Konkordance
1.	nonlv	Darba statistika 1936.g.	86	1937	m; izrakti ceļa grāvji — 514,80 kb. m; uzbūvēti 30 šaursliežu dzelzceļa tilti 6X1,5 kv. m, no 3 sijām un 1,5"
2.	nonlv	Darba statistika 1936.g.	86	1937	m; izrakti ceļa grāvji — 514,80 kb. m; uzbūvēti 30 šaursliežu dzelzceļa tilti 6X1,5 kv. m, no 3 sijām un 1,5" dēļu klaja; atremontēts šaursliežu dzelzceļš, paceļot sliedes un pieblīvējot zemi zem gulšņiem, papildinot balastu

Building Digital Humanities Infrastructure (1)

- laboratorija.lndb.lv

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



The screenshot shows a web application interface for a named entity database. The top navigation bar is teal with the text "LABORATORIJA". Below it is a header with links: "Viss", "Persona", "Vieta", "Organizācija", and a search input field "Atslēgvārds" with a magnifying glass icon. The main content area is titled "MEKLĒŠANAS REZULTĀTI" and displays a table of search results for the term "Latvija". The table has columns: "Nosaukums", "Sastopamība", "Definīcija", "Laiks", and "Kategorija". The results are as follows:

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



The screenshot shows a web-based application interface for a digital library. At the top, there is a navigation bar with links: GRĀMATAS >, PERIODIKA >, KARTES >, ATTĒLI >, AUDIO & VIDEO >, WWW ARHĪVS >, KOLEKCIJAS >, and LABORATORIJA >. On the far right of the navigation bar, there is a dropdown menu labeled "Autorizēties". Below the navigation bar, the word "LABORATORIJA" is prominently displayed in a teal header bar. Underneath this, there are four categories: Viss, Persona, Vieta, and Organizācija. A search input field labeled "Atslēgvārds:" is followed by a magnifying glass icon. The main content area is titled "BRĪVĪBAS IELA LOC.ADDR". It contains a table with columns: Nosaukums, Sastopamība, No, Līdz, and Komentārs. The table lists various street names along with their counts and ranges. At the bottom of the table, there are two additional teal header bars: "Saistīti objekti" and "Ārējie resursi", each with a "Komentārs" column.

Nosaukums	Sastopamība	No	Līdz	Komentārs
Brīvības iela	7179	1989		
Brīvības iela	7179	1923	1942	
Ķepina iela	2915	1945	1989	
Aleksandra iela	293	1818	1861	
Ādolfa Hitlera iela	169	1942	1945	
Adolf Hitler Strasse	0			
Ādolfa – Hitlera Aleja	0			
Alexanderstrasse	0			
Freiheits Strasse	0			
Grosse Alexanderstrasse	0			
Lielā Aleksandra iela	0	1861	1923	
Александровская улица	0			

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



The screenshot shows a digital interface for a collection of Rainis and Aspazija's works. At the top, there is a header with the library's logo and the title "Rainis un Aspazija". Below the header, a navigation bar includes links for "Par kolekciju", "Aspazija", "Rainis", "LFMI", "LNA", "LNB", "MMA", "RMM", a search bar, and a magnifying glass icon.

Tips

- Darbi
- Darbu anotācijas
- Vēstules
- Dokumenti
- Fotogrāfijas
- Plakāti
- Audio / Video

Glabātājs

- LFMI
- LNA
- LNB
- MMA
- RMM



Jānis Rainis "Apdziedāšanas dziesmas III vispārīgiem latvju dziesmu svētkiem" (1889). Anotācija

Objekta teksts

Jāns Jasēnu Plikšis "Apdziedāšanas dziesmas III vispārīgiem latvju dziesmu svētkiem" (1889)

1888.gada persona Rainis, Jānis to gala eksāmenus Juridiskajā fakultātē un saņem apliecību par Pēterburgas universitātes beigšanu. Vasarā, pēc atlīgvielas mājās, viņš kopā ar māsu Doru ierodas Rīgā, kur jūnijā notiek III vispārīgie latviešu dziesmu svētki. Sie svētki pulcināja kopā vairāk koru un dziedātāju nekā iepriekšējie. Sekmīgi beidzis studijas, Rainis ir labā gara stāvoklī, asprātīgs un zobgalīgs. Jauns pasākums dziesmu svētkos ir apdziedāšanās dziesmas. To iekļaušana repertuārā ir diriģenta un komponista Ernesta Vignera ierosinājums. Tās ir iedalitas četros ciklos, 20.jūnijā, kad bija tradicionālais svētku mielasts ar runām dziedājā galda dziesmas, bet 21.jūnijā, kad notika svētku balle, - Jānu, apdziedāšanās starp puišiem un meitām un atvadīšanās dziesmas. Rainis ironisks prāts saskata pretrunas starp dziedātāju pašāizledzību un entuziasmu un vadītāju un darboju lielmantību. Kerties pašam pie spalvas rosina arī uz svētkiem iznākušais Āronu Matīsa tautas dziesmu krājums "Mūsu tautas dziesmas" un Andreja Pumpura "Lāčplēsis".

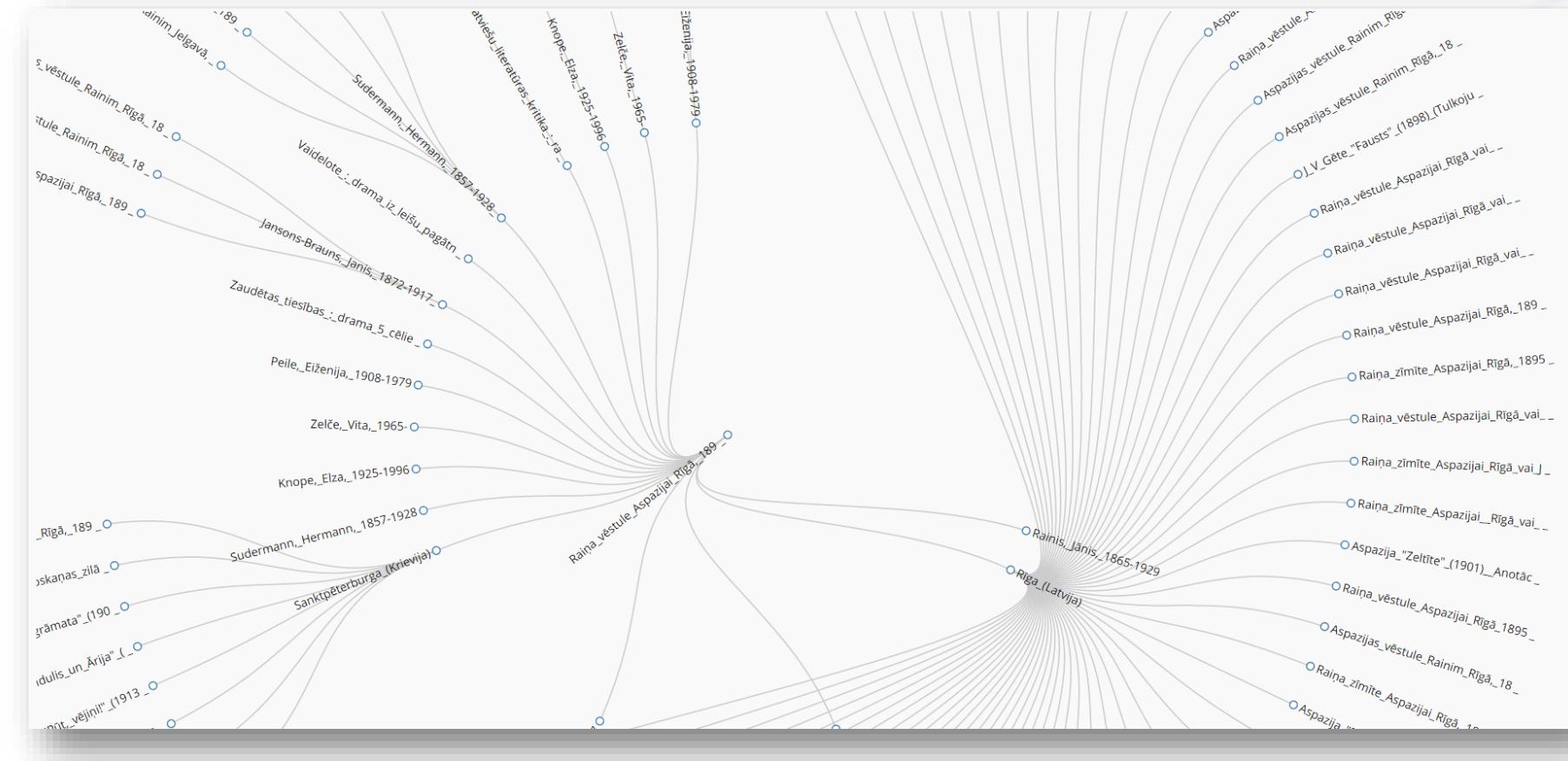
Turtle RDF **RDF/XML**

Saistītie objekti

- Rainis, Jānis, 1865-1929
- Stučka, Dora, 1870-1950
- Rīga (Latvija)
- Vigners, Ernests, 1850-1933
- Āronu, Matīss, 1858-1939
- Pumpurs, Andrejs, 1841-1902
- Malīnovas pagasts (Daugavpils novads, Latvija)
- Aizkalnes pagasts (Preiļu novads, Latvija)
- Latgale (Latvija)
- Viļņa (Lietuva)
- Apdziedāšanas dziesmas III
- Vispārīgiem latvju dziesmu svētkiem
- Muhsu tautas dziesmas
- Lāčplēsis : latvju tautas

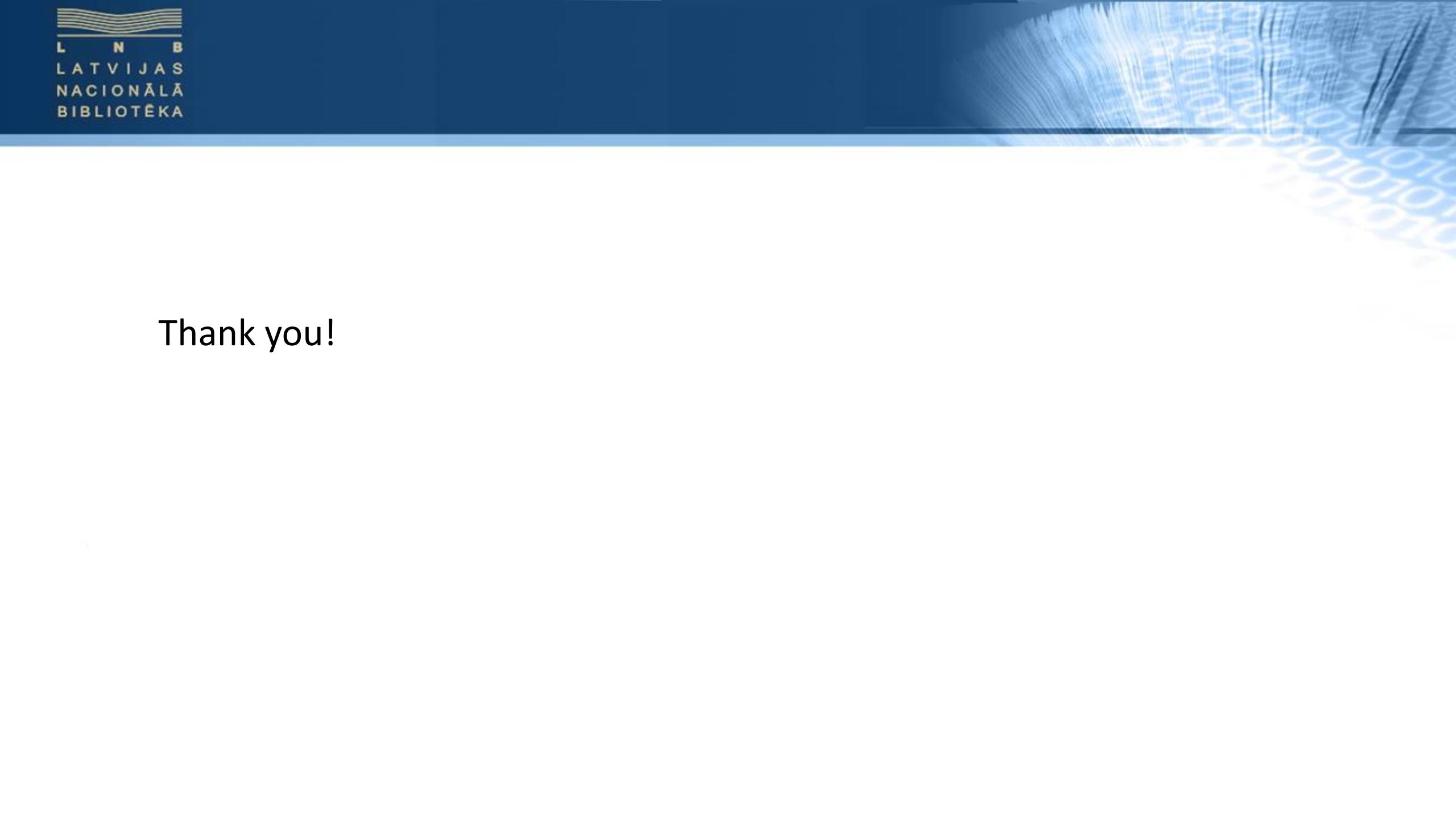
Building Digital Humanities Infrastructure (3)

- 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
 - Stylometry etc.
- To create new version of academic repository [academia.lndb.lv](#) for aggregation and storage of research results, incl. data (RDM)



Thank you!