

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

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

L N B
DIGITĀLĀ BIBLIOTĒKA

Nosaukumu datu bāze

RDF/XML
Meklēšana

Korpusa meklēšana

Bonito pieeja

TEKSTA KORPUSI

[Visi](#) | [Grāmatas](#) | [Periodika: Pirmskara](#) | [Neatkarības](#) | [Padomju Savienības](#) | [Pēcpadomju](#) | [Darbi svešvalodās](#)


Meklēt: 🔍 ⌵ ⚙️ ?

< Lapa no 1 >
Atrasti 2 rezultāti (4.3924 sekundēs)
Konteksta izmērs: Rādīt lapā:

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" daļu klaja; atremontēts šaursliežu dzelzceļš, paceļot sliekšņus un piebīvējot zemi zem gulšņiem, papildinot balastu

< Lapa no 1 >
Atrasti 2 rezultāti (4.3924 sekundēs)
Konteksta izmērs: Rādīt lapā:

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



Nosaukumu datu bāze

Datu servisi
Meklēšana

Teksta korpusi

Korpusa meklēšana
Bonito pieeja

LABORATORIJA

[Visi](#) | [Persona](#) | [Vieta](#) | [Organizācija](#)

Atslēgvārds:

MEKLĒŠANAS REZULTĀTI

Atrasti: 34183 rezultāti Rādīt lapā: 40 ▼

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

- 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 the LABORATORIJA website interface. At the top, there is a navigation menu with categories: GRĀMATAS, PERIODIKA, KARTES, ATTĒLI, AUDIO & VIDEO, WWW ARHĪVS, KOLEKCIJAS, and LABORATORIJA. The user is logged in as 'Autorizēties'.

The main content area is titled 'LABORATORIJA' and includes a search bar with the text 'Atslēgvārds:'. Below the search bar, there is a filter for 'BRĪVĪBAS IEĻA' and 'LOC.ADDR'.

The main table displays the following data:

Nosaukums	Sastopamība	No	Līdz	Komentārs
Brīvības iela	7179	1989		
Brīvības iela	7179	1923	1942	
Ļeņina 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			

At the bottom of the table, there are two summary rows:

Saistīti objekti	Komentārs
Ārējie resursi	Komentārs

On the left side of the interface, there is a sidebar with the following sections:

- Nosaukumu datu bāze**
 - Datu servisi
 - Meklēšana
- Teksta korpusi**
 - Korpusa meklēšana
 - Bonito pieeja

- 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

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!