Eesti Keeleressursside Keskus

Keeleandmestiku eripära avatud andmete kontekstis

Kadri Vider (kadri.vider@ut.ee)

Eesti Keeleressursside Keskuse juht

Keeleandmestiku eripärad

- Mis on keeleandmestik ja meta-andmed?
- Millised digitaalse keelevara arhiivid ja repositooriumid on‘
- Millised on seisukohad andmete avatuse kohta?

Keeleandmestiku (nagu ka paljude muude humanitaar-valdkondade teadusandmestiku) **sisu** on peamiselt inim-tekkeline ja sageli kellegi looming ja/või seotud isiku tuvastamist võimaldavate andmetega

⇒ Andmete avatus sõltub **autoriõiguse** ja **isikuandmete kaitse** regulatsioonide piirangutest

EKRK

Mõisteid

Keeleressurss

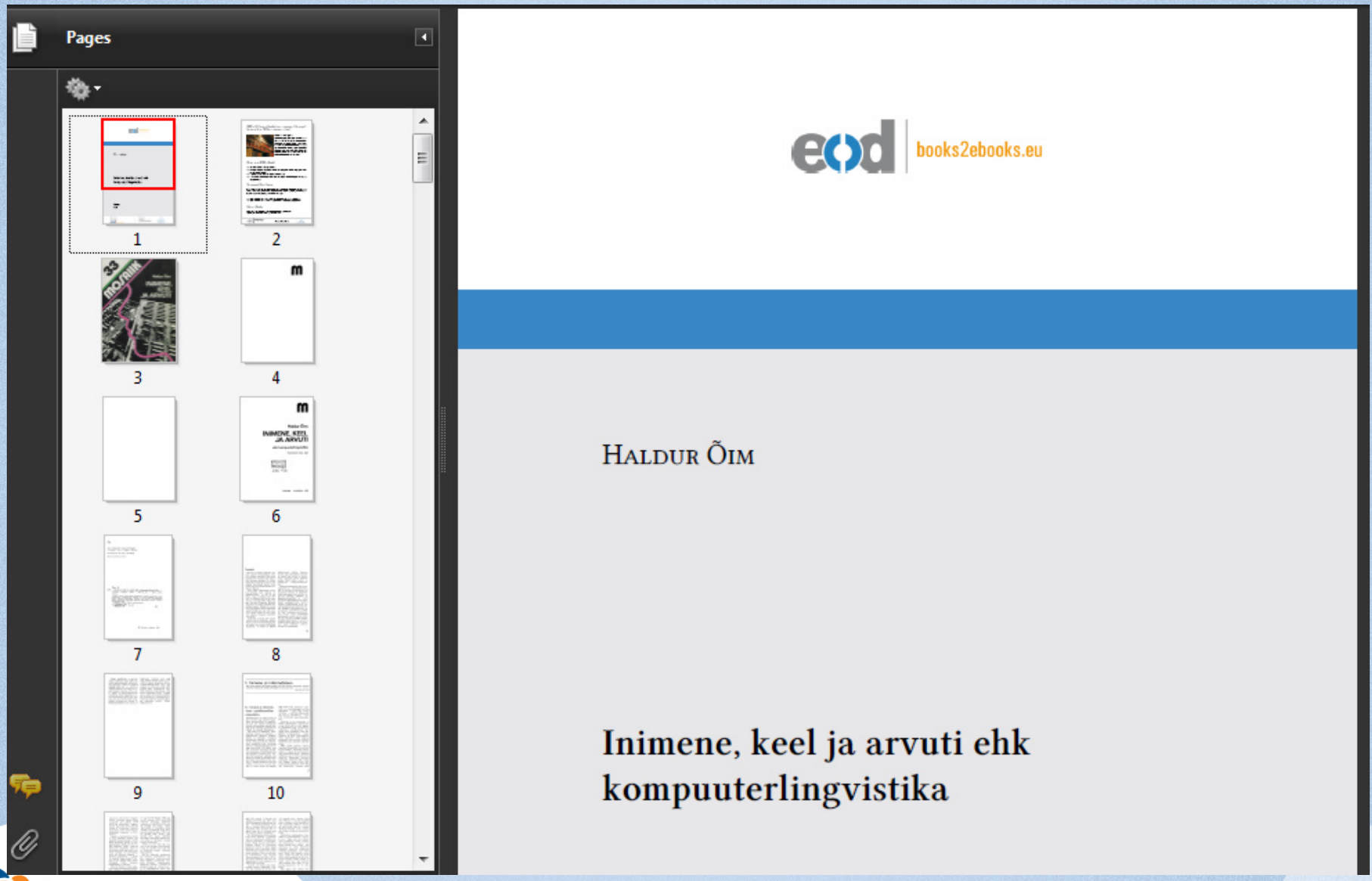
- andmestik keele kohta või vahend/abinõu sellise andmestiku esitamiseks või töötlemiseks
- näiteks tekstid, salvestised, keelekirjeldused ja –annotatsioonid, välitööde materjalid, tarkvara, protokollid, andmemudelid, veebiarhiivid ja -indeksid
- digitaalne või mitte, avaldatud või käsikirjas

Meta-andmed (Metadata, MD)

- Kirjeldav struktureeritud info ressursi kohta, näiteks info teavikute kohta kataloogikaardil raamatukogus
- Sisaldab kokkulepitud elementide komplekti (näiteks DCMI, IMDI, IsoCAT), mis võimaldab ka otsingut filtreerida

EKRK

Digiteeritud andmed



EKRK

Mis sellest on keeleandmestik?

TEKST: Meie eesmärgiks ei ole muidugi raamatukogude töötamispõhimõtete kirjeldamine ega ka neis seni kasutusel olnud infootsingumeetodite analüüs.

www.filosoft.ee/html_morf_et/html_morf.cgi

Most Visited | Täna ülikoolis - ut.ee | Introducing

olnu+d //_S_ pl n, //

infootsingumeetodite
info_otsingu_meetod+te //_S_ pl g, //

analüüs
analüüs+0 //_S_ sg n, //

Kuid
kuid+0 //_J_ //
kuu+id //_S_ pl p, //

raamatukogude
raamatu_kogu+de //_S_ pl g, //

kui
kui+0 //_D_ //
kui+0 //_J_ //

näite
näi+te //_V_ te, //
näide+0 //_S_ sg g, //
näit+e //_S_ pl p, //

varal
vara+l //_S_ sg ad, //
varal+0 //_K_ //

6im_inimekeelarvuti.pdf - Adobe Reader

File Edit View Document Tools Window Help

33 / 149 100%

infootsing

2. Tekstide otsing ja töötlus automatiseeritud infosüsteemides

Erinevalt inimese ajust ei unusta arvuti kunagi, mida talle on öeldud, ning võib vastava signaali saamisel kõik jälle välja laduda. Seepärast ongi ehk kõige kasulikum arvutile üldse mitte midagi öelda.

C. Ford, «Mõtlemise õpetus»

2.1. Automatiseeritud infootsisüsteemid — probleemid ja ehitus

Vajamineva informatsiooni ülesotsimine kui omaette probleem sündis ilmselt ühel ajal esimeste arhiivide ja raamatukogude tekkega, seega 4—5 aastatuhandet tagasi. Juba kolme ja poole tuhande aasta eest valitsenud vaarao Ramses II raamatukogu sisaldanud 20 000 papüürusrulli, nii et vajaliku informatsiooni leidmine ei saanud tollalgi olla triviaalne ülesanne.

Möödaläinud aastatuhandete jooksul on välja töötatud küllalt tõhusad meetodid ja vahendid vajalike raamatute, ajakirjade ja ar-

tarbijate üha pakilisemaid ja spetsiifilisemaid vajadusi. Meie eesmärgiks ei ole muidugi raamatukogude töötamispõhimõtete kirjeldamine ega ka neis seni kasutusel olnud infootsingumeetodite analüüs. Kuid raamatukogude kui näite varal jõuame kergemini meid huvitavate probleemide juurde.

Kui meid huvitab mingi kindla küsimuse kohta leiduv kirjandus, siis peamine allikas selle väljaselgitamisel raamatukogus on süstemaatiline kataloog. Selles on kogu raamatukogus leiduv kirjandus klassifitseeritud sisu järgi kindlatesse liikidesse, mis on hierarhilise ehitusega, s. t. jagunevad järjest kitsamateks alaliikideks. Tänapäeval on enamikus maailma maades kasutusel ühtne teadus- ja

5,13 x 7,61 in

Andmekogud, arhiivid ja repositooriumid

- Digiteeritud tohutu maht humanitaarteaduste uurimisandmeid, enamik nendest on keelepõhised
- Paljud sellised arhiivid kasutavad erinevaid standardeid, sõltuvalt uurimise eesmärgist on andmed erineva detailsuse või struktuuriga
- Ka andmetele ligipääs on korraldatud eri viisidel

Digitaalse keeleandmestiku koondamise, arhiveerimise ja ligipääsuga tegeleb ka **Eesti Keeleressursside Keskus (EKRK)**

- teadustaristu, mis võimaldaks kõigile uurijatele keeleressursside ja -tehnoloogiate kättesaadavuse
- ühendades eksisteerivad digitaalsed keelearhiivid ja tagades nende kättesaadavuse veebi kaudu
- keelest sõltumatuid vahendeid on võimalus kasutada ja jagada, keelest sõltuvaid vahendeid on võimalik üle kanda

EKRK

Open Language Archives Community = OLAC http://www.language-archives.org/

Infot 41-s avatud keelearhiivis ja -repositooriumis leiduva keelelise materjali ja andmestiku kohta üle maailma

Andmestiku otsikategooriaid:

üksikud keeled, keeleperekonnad, regioonid ja riigid,

Andmestiku tüüp: lähtetekstid (korpused), leksikonid, keelekirjeldused;

ligipääs: on-line või off-line

Uurimisvaldkond: keeledokumendid, süntaks, üldkeeleteadus, tüpoloogia, korpuslingvistika, fonoloogia, semantika, morfoloogia, foneetika…

Diskursuse tüüp: laul, dialoog, narratiiv, kõne

DCMI* tüüp: tekst, heli, pilt, liikuv pilt, andmestik, kollektsioon, tarkvara

Formaat: variatsioonid teemal text/html/application/pdf/ image/audio…

*Dublin Core Metadata Initiative

= põhimõtteliselt avatud mudel, andmete vabalt jagamise vastutus on andmete jagajatel

EKRK

OLAC otsing


Participating Archives • OLAC
OLAC Language Resource Catalog
Search for language resources
go
Navigating the Catalog
Catalog Home
Search Strategies
Advanced Search
New: Records recently added or modified
Quick Links
Browse by Language
Browse by Country
Browse by Linguistic Field
Browse by Linguistic Type
Browse by Language Family
Contacts
Email Us
More information
OLAC Homepage
OLAC FAQ
Archive
Sort: by frequency alphabetically
The Language Archive's IMDI portal (87642)
SIL Language and Culture Archives (22377)
Alaska Native Language Archive (13172)
Graduate Institute of Applied Linguistics Library (8176)
Ethnologue: Languages of the World (7413)
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (7380)
WALS Online RefDB (7328)
The Rosetta Project: A Long Now Foundation Library of Human Language (6571)
California Language Archive (6418)
A Digital Archive of Research Papers in Computational Linguistics (3280)
WALS Online (2678)
The LINGUIST List Language Resources (2273)
CHILDES Data repository (275)
TALKBANK Data repository (204)
POLLEX-Online (65)
The Natural Language Software Registry (62)
TST-Centrale (61)
African Language Materials Archive (53)
Multimodal Learning and teaching Corpora Exchange (37)
Language resources at the Text Laboratory (33)
Speech and Language Data Repository (SLDR, formerly CRDO-Aix) (33)
U Bielefeld Language Archive (13)
Language Commons Language Corpora (10)
The Typological Database Project (8)
Cornell Language Acquisition Laboratory (CLAL) / Virtual Center for Language Acquisition (VCLA) (6)

CLARIN VLO (Virtual Language Observatory)

CLARIN (Common Language Resources and Technology Infrastructure) – www.clarin.eu

Virtual Language Observatory (http://www.clarin.eu/vlo/)

- hõlmab andmeid CLARIN Language Resource Inventory (kasutab IMDI arhiivi, OLACi, ELRA ja CLARINi andmeid) või Language Tool inventory (kombineerib CLARINi vahendid ja DFKI NLP Software Registry data) kasutades profileeritud otsingut
- Võimaldab sirvida CLARINi kogutud **meta-andmete** kataloogi

= meta-andmed on avatud, keeleandmestiku avatus andmestiku jagajate vastutusel ja korraldada

EKRK

Estonian search

COLLECTION

childes (152)
Endangered Languages (55)
CLARIN LRT inventory (29)
European Language Resources Association (14)
WALS RefDB (12)
CHILDES Data repository (3)
Ethnologue: Languages of the World (2)
talkbank (2)
A Digital Archive of Research Papers in Computational Linguistics (1)
ODIN - The Online Database of Interlinear Text (1)
more...

LANGUAGE

Estonian (210)
English (17)
Bulgarian (11)
Russian (11)
German (8)
Hungarian (8)
Swedish (8)
Danish (7)
French (7)
Latvian (7)
more...

CONTINENT

Europe (159)
Asia (45)
Unspecified (3)
North-America (2)

GENRE

discourse (203)
primary_text (19)
language_description (1)
unspecified (1)

COUNTRY

Estonia (182)
Russian Federation (46)
United States (3)
Germany (2)
Latvia (2)
Bulgaria (1)
Canada (1)
Croatia (1)
Czech Republic (1)

SUBJECT

unspecified (176)
unknown (6)
syntax (4)
typology (4)
general_linguistics (3)
language_acquisition (3)
discourse_analysis (1)
personal narrative (1)
phonetics (1)

Showing 81 to 90 of 277 << < 5 6 7 8 9 10 11 12 1

name	description
BABEL Hungarian Database	Desktop/Microphone
BABEL Polish database	Desktop/Microphone
BABEL Romanian database	Desktop/Microphone
Basic Course in Estonian	
Constraint Grammar of Estonian	general written, Constraint Gram
Conversation MOVIN Corpus	Political Talk Shows
Corpus of Old Written Estonian	Corpus of texts written fully or pa in Estonian, from 13.-19. century million words
Corpus of Present-day Written Estonian	written general; 95 mio words; TEI/SGML
Corpus of Spoken Estonian	spoken general; 1 mio words; loc tagset
Corpus of Written Estonian	4.4 mio words; TEI/SGML

META-SHARE repositooriumide võrgustik

META-NETi võrgustiku (*A Network of Excellence forging the Multilingual Europe Technology Alliance*, http://www.meta-net.eu/) hajustaristu, mille meta-andmete registrile ja repositooriumi andmetele (www.meta-net.eu/meta-share) pääseb vabalt ligi võrgustikku kuuluvate sõlmede (node) kaudu.
EKRK võrgustikusõlm metashare.ut.ee Eesti keeleressursside metaandmete registri ja repositooriumiga.

Registri andmeteks on repositooriumis säilitatava keeleressursi meta-andmed:
Esindatud keeled, üks või mitmekeelne ressurss; ressursi tüüp, meediumi ja esitusviisi tüüp (tekst/audio…, kirjalik/suuline keel…), formaat
NB! Meta-andmed ka kättesaadavuse kohta:
Kasutuslitsentsi tüüp ja kasutajaskonna piirangud (k.a inim- või masinkasutatavus)

EKRK

Keywords: Estonian

Return to Browse page

Search

Filter by:

Language:
- Estonian (16)
- English (5)
- German (5)
- Latvian (5)

more

Resource Type:
- corpus (14)
- lexicalConceptualResource (12)

Media Type:
- text (21)
- audio (6)

Availability:
- available-restrictedUse (23)
- available-unrestrictedUse (3)

Licence:
- ELRA_END_USER (13)
- ELRA_VAR (13)
- ELRA_EVALUATION (7)
- proprietary (4)

more

26 Language Resources (Page 1 of 2)

« Previous | Next »

Resource Name	Resource Type	Media Type	Language
ACCURAT Bilingual Comparable Corpora for under-resourced languages	corpus	text	Croatian, English, Estonian, German, Italian, Latvian, Modern Greek (1453-), Romanian, Slovenian
BABEL Bulgarian Database	corpus	audio	Bulgarian
BABEL Estonian Database	corpus	audio	Estonian
BABEL Hungarian Database	corpus	audio	Hungarian
BABEL Polish database	corpus	audio	Polish
BABEL Romanian database	corpus	audio	Romanian
Bulgarian X language Parallel Corpus Bul-X-Cor	corpus	text	Albanian, Bosnian, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finish, Galician,

Ligipääs keeleandmestikule EKRK-s

- Ressurssidel 3 tüüpi kasutuslitsentse
  - Vaba kasutus kõigile (näiteks Creative Commons)
  - Kasutamiseks teadustöö eesmärkidel (ACA)
  - Kasutamiseks eritingimustel (mitte-kommerts või isikuandmetega seotud)

- Kasutajate võimalused sõltuvalt kuuluvusest
  - Laialdasimad konsortsiumipartneritel
  - CLARINi partnerid jt teaduskasutajad
  - Avalikkus

EKRK

Eesti Keeleressursside Keskus

Tänan tähelepanu eest!