Eesti Keeleressursside Keskus

Kadri Vider (kadri.vider@ut.ee)

Eesti Keeleressursside Keskus
www.keeleressursid.ee

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

Eesti Keeleressursside Keskus (EKRK)

- Eesti keeleressursside keskus on humanitaaride teadustaristu – erinevates uurimisasutustes paiknevate, veebist ligipääsetavate andmehoidlate võrgustik, mis võimaldab autentimise teel juurdepääsu mitmel erineval tasemel kasutajatele.

- Lisaks olemasolevate ja uute, loodavate keeleressursside kogumisele ja arhiveerimisele käivitatakse süsteem olemasolevate keeleressursside tutvustamiseks ja potentsiaalsete kasutajate koolitamiseks.

EKRK missioon

- milleks?
  - Luua taristu, mis võimaldaks kõigile uurijatele **keeleressursside ja -tehnoloogiate kättesaadavuse**
  - Keelest sõltumatuid vahendeid on võimalus kasutada ja jagada
  - Keelest sõltuvaid vahendeid on võimalik üle kanda

- kuidas?
  - Ühendades eksisteerivad digitaalsed arhiivid ja tagades nende kättesaadavuse veebi kaudu
  - Pakkudes keeletehnoloogia vahendeid kui veebiteenust, mis kasutab arhiveeritud andmeid
  - Meta-andmeid kättesaadavaks tehes

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

EKRK kui konsortsium

- konsortsiumleping allkirjastati 2.12.2011

TARTU ÜLIKOOL
1632

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT

Eesti Keele Instituut

Euroopa Liit
Euroopa Regionaalarengu Fond

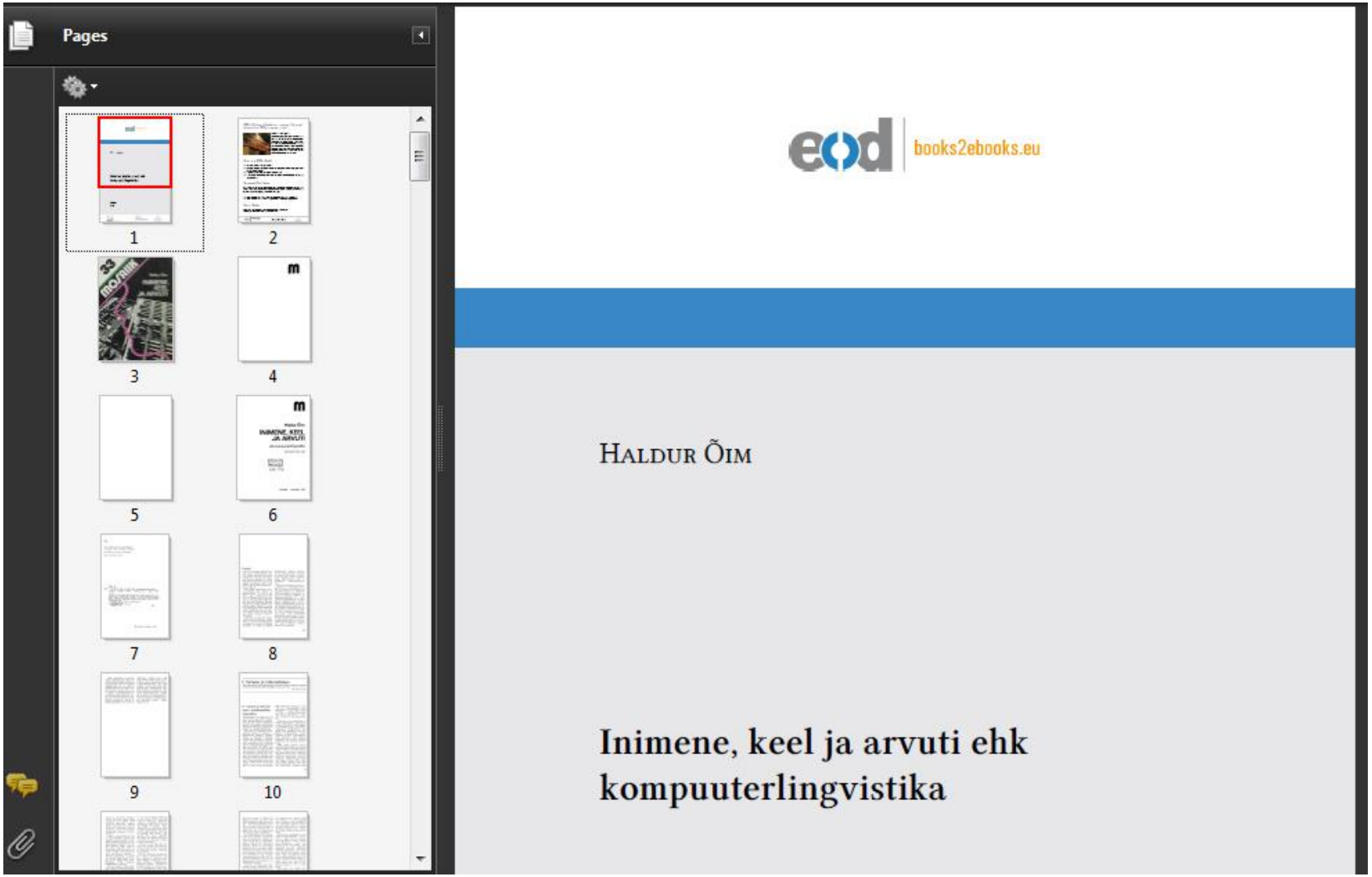
Eesti tuleviku heaks

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-andmestik
  - Kirjeldav struktureeritud info ressursi kohta, näiteks info teavikute kohta kataloogikaardil raamatukogus, digimaailmas sageli kasutajale peidetud märgenditega info digiobjekti kohta.
  - Sisaldab kokkulepitud elementide komplekti (näiteks DCMI, IMDI, IsoCAT), mis võimaldab ka otsingut filtreerida

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Digitaalsed andmed



Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Mis on keeleressurss?

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
automatiseeritud i

Erinevalt inimese ajust ei unusta arvuti vastava signaali saamisel kõik jälle välja laduda. Seepärast ongi ehk kõige kasulikum arvutile üldse mitte midagi ütelda.

C. Ford, «Mõtlemise õpetus»

2.I. 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 artarbijate 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

Euroopa Liit
Euroopa Regionaalarengu Fond

Eesti tuleviku heaks

Humanitaarteadlaste andmekogud

- Tänapäevaks on digitaliseeritud 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
- Humanitaarteadlased sageli ei tea
  - mis on keeleressursid (KR)
  - kas ja kuidas KR neid võiks aidata

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

EKRK Euroopa plaanis

- CLARIN – www.clarin.eu
  Common Language Resources and Technology Infrastructure
  - Tugeva kasutajatoega, teadlastele orienteeritud võrgustik
  - Koosneb eri tüüpi keskustest, mis ühendavad kasutajaid, ressursse ja tugiteenuseid
  - EKRK on Eesti CLARINi keskus
- META - www.cs.ut.ee/metanord/
  - Ühtne repositooriumite võrgustik Euroopa mitmekeelsuse toeks
  - Võrgustiku sõlmed vahendavad ressursse ühtses meta-andmete vormis ja ühtsetel tingimustel
- DASISH – www.dasish.eu
  - Sotsiaal- ja humanitaarteaduste andmeteenuste taristu
  - Võrgustikus CLARIN, DARIAH, ESS, CESSDA, SHARE

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

CLARIN LRT inventory: Estonian

VLO Home >> Faceted Browser Resources

Estonian search

COLLECTION : all > CLARIN LRT inventory

LANGUAGE

Estonian (29)
English (7)
Hungarian (6)
Bulgarian (5)
Czech (5)
Romanian; Moldavian; Moldovan (5)
Danish (4)
Dutch (4)
Finnish (4)
French (4)
more...

CONTINENT

GENRE

COUNTRY

Estonia (22)
Germany (2)
Latvia (2)
Bulgaria (1)
Canada (1)
Croatia (1)
Czech Republic (1)
Hungary (1)
Italy (1)
Lithuania (1)
more...

SUBJECT

ORGANISATION

University of Tartu (8)
Institute of Cybernetics at Tallinn University of Technology (2)
Department of Linguistics and Nordic Studies, University of Oslo (1)
Joint Research Centre of the EU (1)
Netherlands Institute of Advanced Study (NIAS), Collegium Budapest (1)
Tilde (1)
Tilde as coordinator, Eurotermbank consortium (1)

RESOURCETYPE

Written Corpus (10)
Spoken Corpus (5)
Lexicon / Knowledge Source (4)
Treebank (3)
Terminological Resource (2)
Aligned Corpus (1)
Grammar (1)
Web Service (1)

Showing 1 to 10 of 29 << < 1 2 3 > >>

name	description
Arborest	149 sentences, VISL tagset
BABEL Estonian Database	The database consists of three sets: - Many Talker Set: 30 males, 30 females; each to read 50...
Constraint Grammar of Estonian	general written, Constraint Grammar
Corpus of Old Written Estonian	Corpus of texts written fully or partly in Estonian, from 13.-19. century; 1,5 million words
Corpus of Present-day Written Estonian	written general; 95 mio words; TEI/SGML
Corpus of Spoken Estonian	spoken general; 1 mio words; local tagset
Corpus of Written Estonian	4.4 mio words; TEI/SGML
Corpus query for Estonian corpora	Web application for querying the automatically morphologically disambiguated Mixed corpus of...
Corpus with Disambiguated Word Senses	100000 words, word senses based on TEKsaurus (Estonian Wordnet)
Database of Estonian Multi-word Verbs	17 500 entries

META-SHARE repositooriumide võrgustik

- META-NETi võrgustiku (*A Network of Excellence forging the Multilingual Europe Technology Alliance*, http://www.meta-net.eu/) hajustaristu, mille metadata registri (www.meta-net.eu/meta-share) ja repositooriumi andmetele pääseb vabalt ligi võrgustikku kuuluvate sõlmede (node) kaudu.
- Tartu Ülikool osales META-NETi projektis META-NORD, mille üks ülesandeid panna püsti oma võrgustikusõlm metaandmete registri ja repositooriumiga: metashare.ut.ee

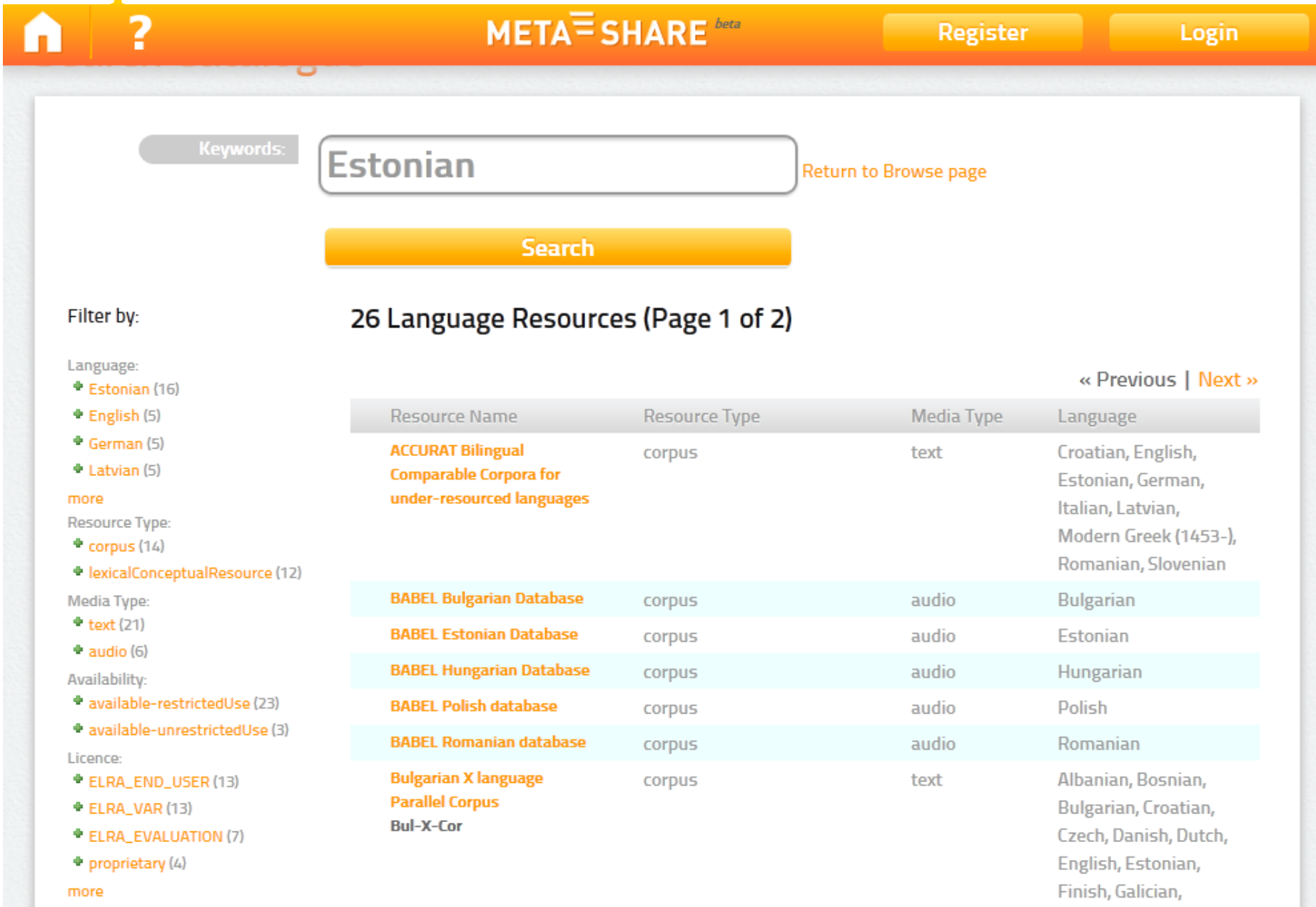
Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

META-SHARE andmestik

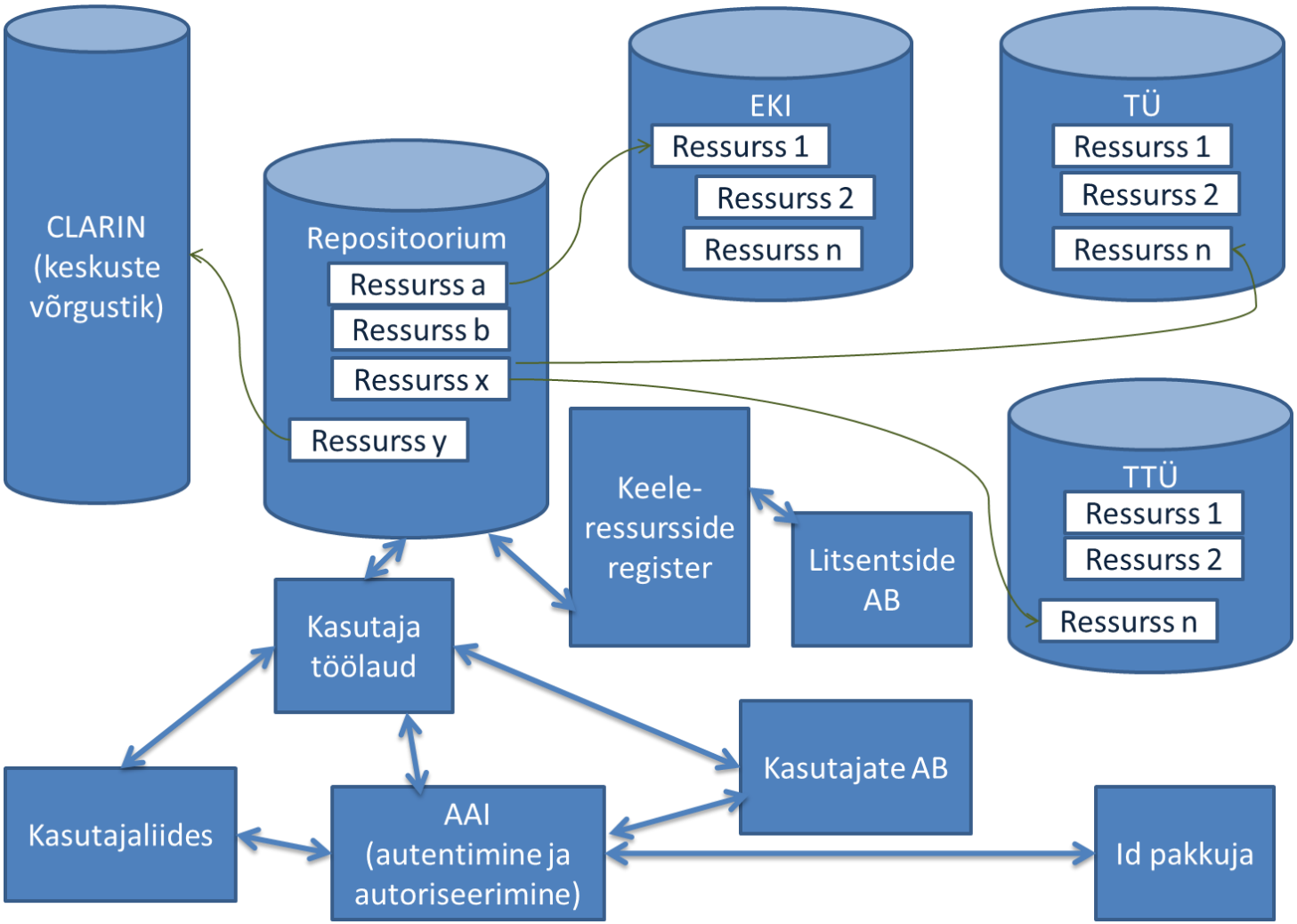
- Registri andmeteks on repositooriumis säilitatava keeleressursi meta-andmed:
  - Esindatud keeled, kas üks või mitmekeelne ressurss, paralleelne vm struktuur mitmekeelsetel
  - Ressursi tüüp, meediumi ja esitusviisi tüüp (tekst/audio…, kirjalik/suuline keel…), formaat
  - Kättesaadavus
  - Kasutuslitsentsi tüüp ja kasutajaskonna piirangud (k.a inim- või masinkasutatavus)
  - Võimalikud KT rakendusvaldkonnad

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

http://metashare.ut.ee/



Keskuse komponendid



Keskuse pakutavad teenused

- Keeleressursside arhiveerimine ja haldamine
- Keeleressursside kogumine ja hindamine
- Ligipääs ja kasutajate koolitamine

Avatud nii keeleressursside pakkujatele, arendajatele kui ka keeleressursside kasutajatele, kes nõustuvad kasutuskorra ja litsentsitingimustega, kuid eelistatud on teadus-arendusasutuste kasutajad ja partnerid CLARINi liikmete seas.

Arhiveerimine

- ressursside metaandmete (ehk tekstilise kirjelduse) säilitamine registris
- ressursside koopiate säilitamine repositooriumis

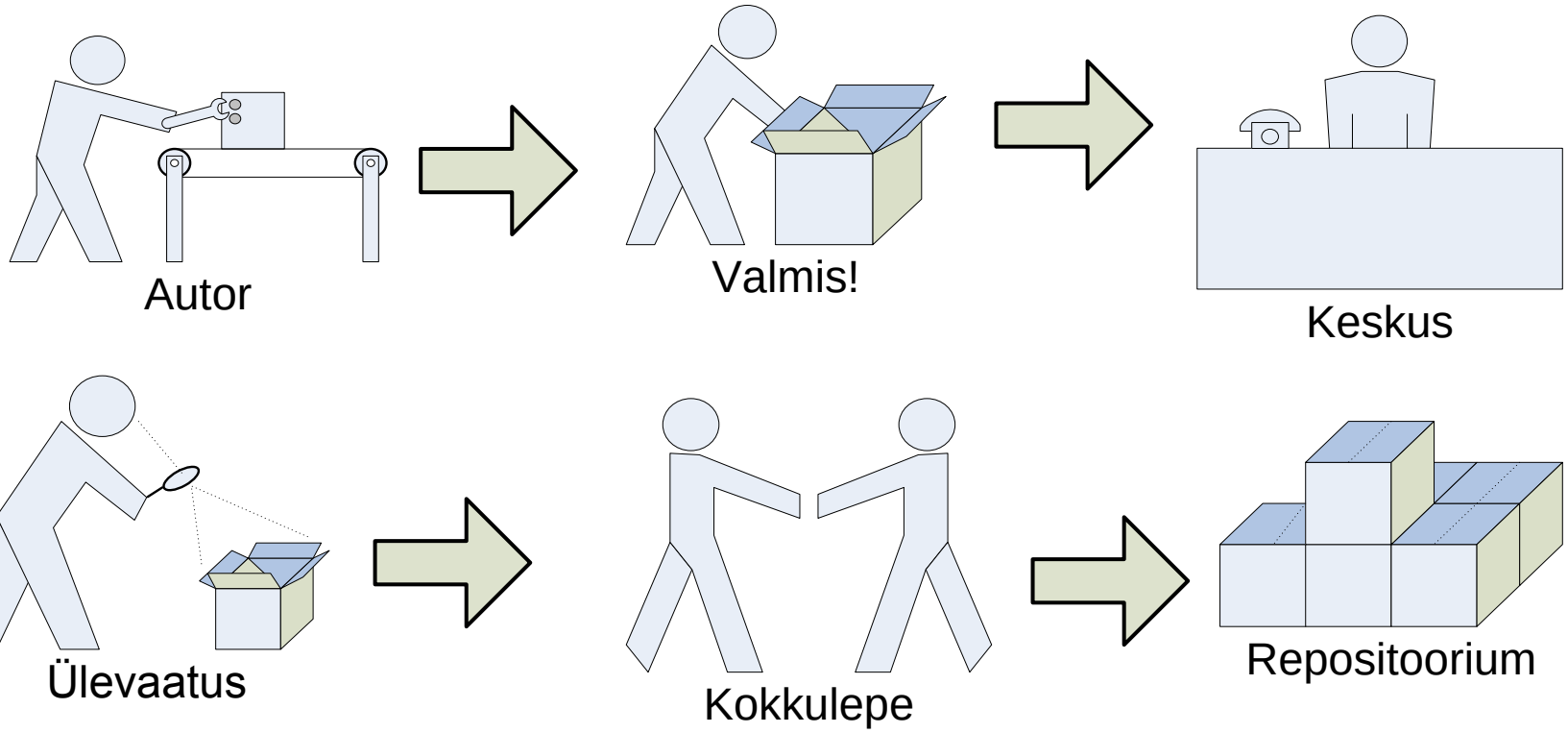
Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Ressursside kogumine

- Keeleressursside keskuse partnerid
  - Koondada olemasolevad töörühmade ressursid
  - Hoida kättesaadaval ressursside vanemaid versioone
  - Ressursside ja tarkvara kasutamine üle veebi
- EKKTT programmi raames loodud ressursid
  - Loodud ressursid säilitatakse ühtses kohas
  - Võimaldatakse ressurssidele juurdepääs vastavalt kasutustingimustele
  - Võimalus ressursse edasi arendada
- Kõik teised soovijad

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

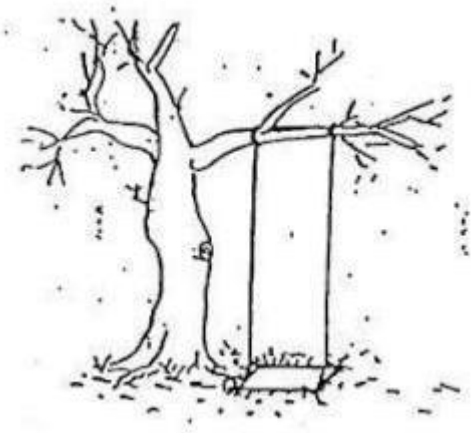
Kuidas ressurss meile jõuab? ☺



Euroopa Liit
Euroopa
Regionaalarengu Fond
Eesti tuleviku heaks

Ligipääs

- 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



Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Ligipääs

- Luuakse avalik veebipõhine ligipääs
  - www.keeleressursid.ee
- Luuakse kasutajagrupid, määratakse kasutusõigused.
- Eelisolukorras ligipääsu võimaldamisel arhiveerimisteenusele on CLARIN-ERIC liikmed.
- **Liidestus CLARINiga**
  - Regulaarne andmevahetus
  - SSO* autentimine, ligipääs rahvusvahelisse võrgustikku

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

SSO = Single-Sign-On

Koostöö mäluasutustega (1)

- Digiteerimise ja kopeerimise alane koostöö (ka juriidilisest aspektist)
  - tekstilise materjali osas huvitab meid ainult OCR-tud ehk tärktuvastatud materjal
  - helimaterjali analüüsiks ja massiliseks sisuotsinguks võimalik vastastikku kasulik koostöö KübI kõnetehnoloogia spetsialistidega
- Arhiveerimise ja pikaajalise säilitamise alane koostöö
  - varukoopiate deponeerimine
  - PID-süsteemi sünkroonimine

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Koostöö mäluasutustega (2)

- Sisu kasutamise autoriõiguse ja litsentsimise teemad
- Sisule ligipääsu teema, sealhulgas koostöö kasutajate autentimise ühise süsteemi alal
- Sisu semantilise annoteerimise teemad otsisüsteemide jaoks – see on ka keeletehnoloogiline ülesanne

Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Tänan tähelepanu eest!