Eesti Keeleressursside Keskus

www.keeleressursid.ee

Kadri Vider
EKRK tegevjuht
kadri.vider@ut.ee

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

ARCHIMEDES

Haridus- ja Teadusministeerium

Eesti Keeleressursside Keskus (EKRK)

- Riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006-2010)“ projekt 2008-2010 TÜ-s
- CLARINi projektipartner TÜ 2008-2011
- CLARIN ERICu riiklik keskus alates 2012
- Eesti teaduse infrastruktuuride teekaardi objekt
- Eestis 3 juhtiva keeletehnoloogiaüksuse (TÜ, TTÜ KübI, EKI) konsortsium

=> Riiklikult olulise humanitaarteaduste infrastruktuuri osa

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

EKRK Euroopa plaanis

- CLARIN Common Language Resources and Technology Infrastructure - www.clarin.eu
  - 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
Common Language Resources and Technology Infrastructure

- CLARIN = Common Language Resources and Technology Infrastructure - Ühine keeleressursside ja –tehnoloogia infrastruktuur
  - ESFRI projekt 2008 – 2011, 32 partnerit 22 riigist, Eestist osales Tartu Ülikool
  - European Research Infrastructure Consortium ehk ERIC alates 29.02.2012
- www.clarin.eu

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

CLARIN ERIC

- Põhieesmärgiks on muuta olemasolevad keeleressursid ja keeletehnoloogia vahendid kättesaadavaks, vastastikku toimivaks ning stabiilseteks teenusteks, mida kasutajad saavad vajaduse korral ka oma tarbeks kohandada
- Juhtriik Holland; asutajariigid Austria, Bulgaaria, Tšehhi Vabariik, Taani, Saksamaa, Poola, Eesti
- CLARIN-ERIC on loodud määramata ajaks

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

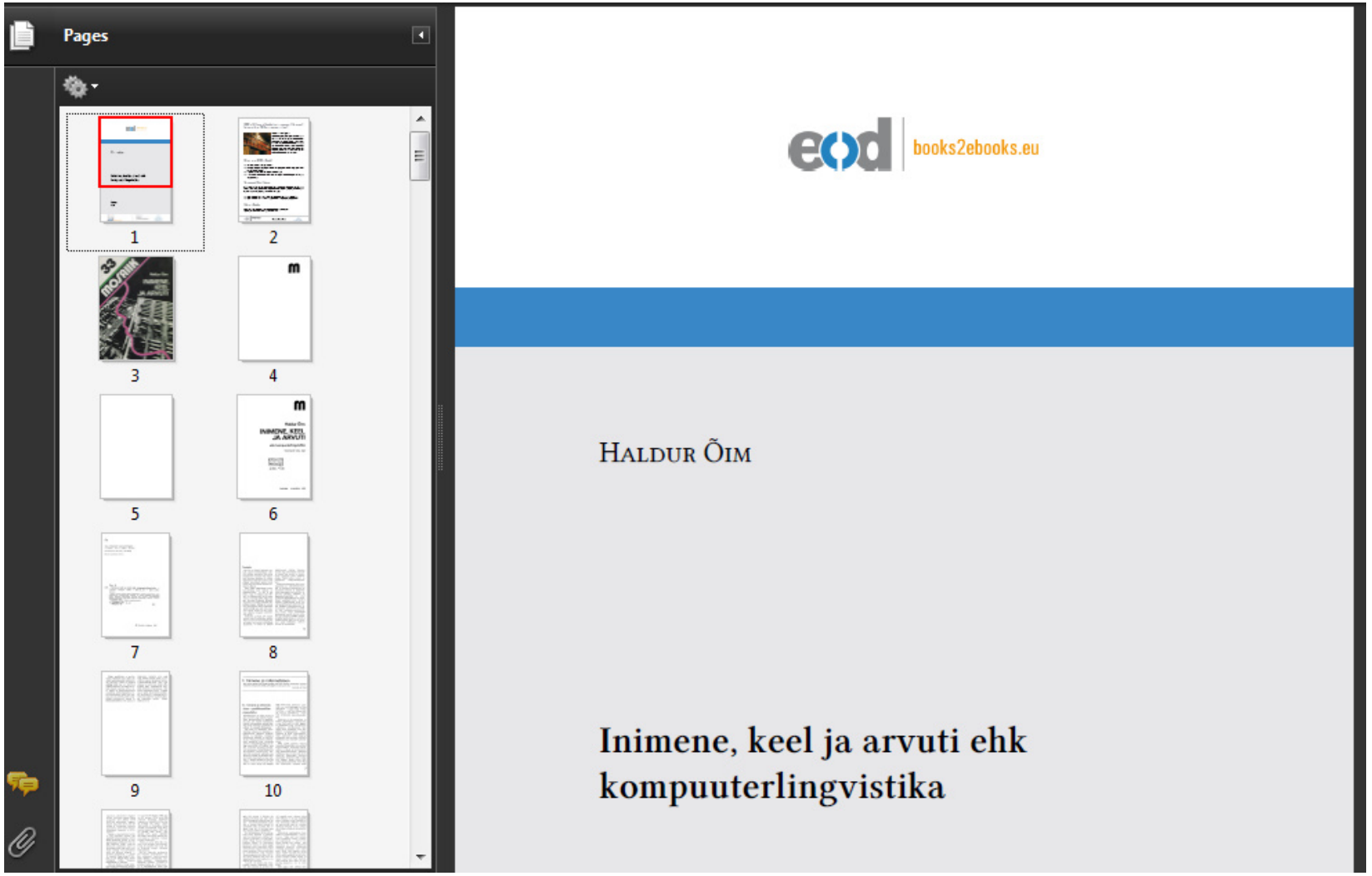
Humanitaarteadlaste andmekogud

- 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

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, //

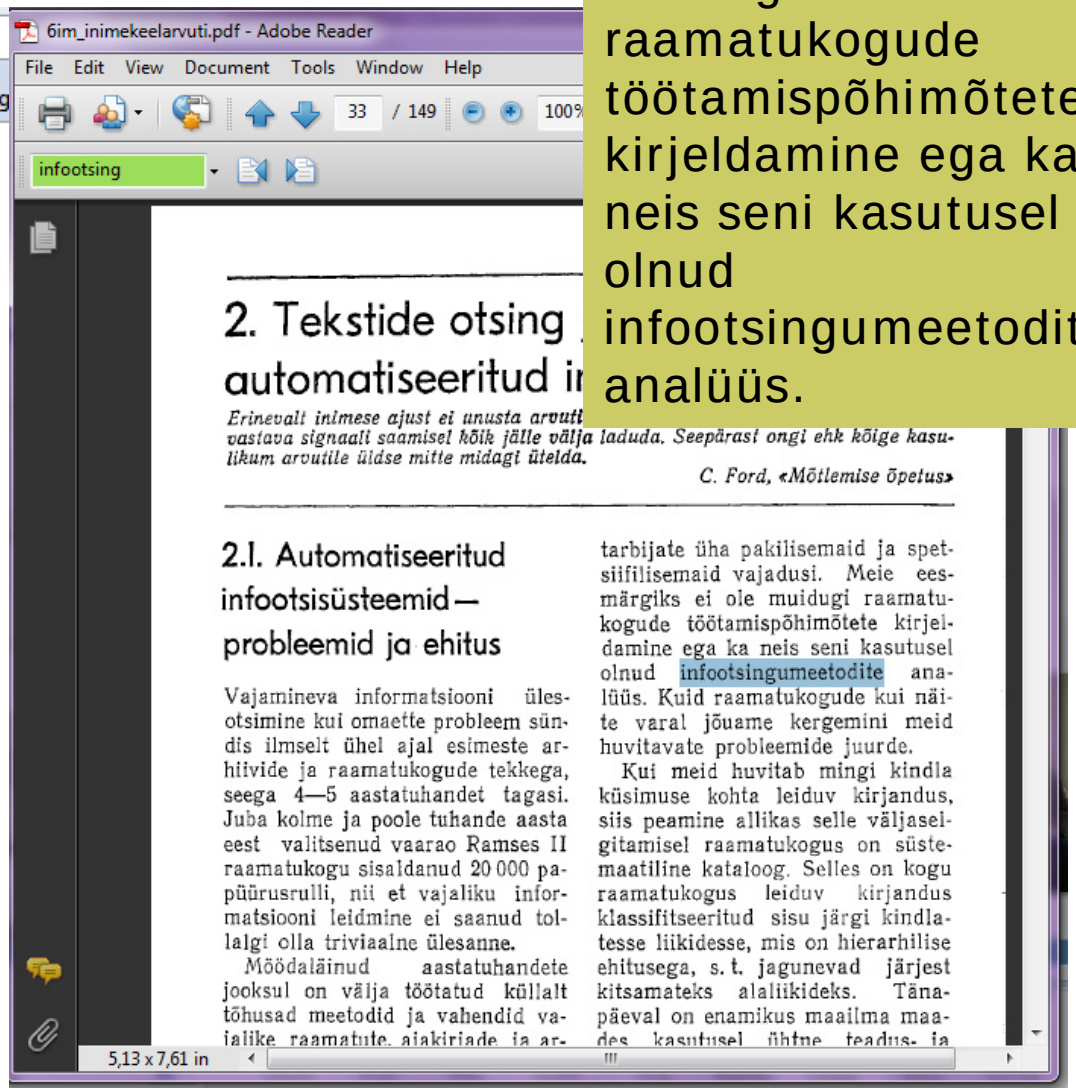
Kuid
kuid+0 //_J_ //
kuu+id //_S_ pl p, //

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 õpetada.

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

Euroopa Liit
Euroopa Regionaalarengu Fond

Eesti tuleviku heaks

CLARINi missioon

(on ka 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
  - Kõik see tugineb tugevatele keskustele, mis suudavad pakkuda vajalikke teenuseid ja millel on garanteeritud riiklik toetus
  - Meta-andmeid kättesaadavaks tehes – sellega tegeleb intensiivselt META-NETi algatus META-SHARE

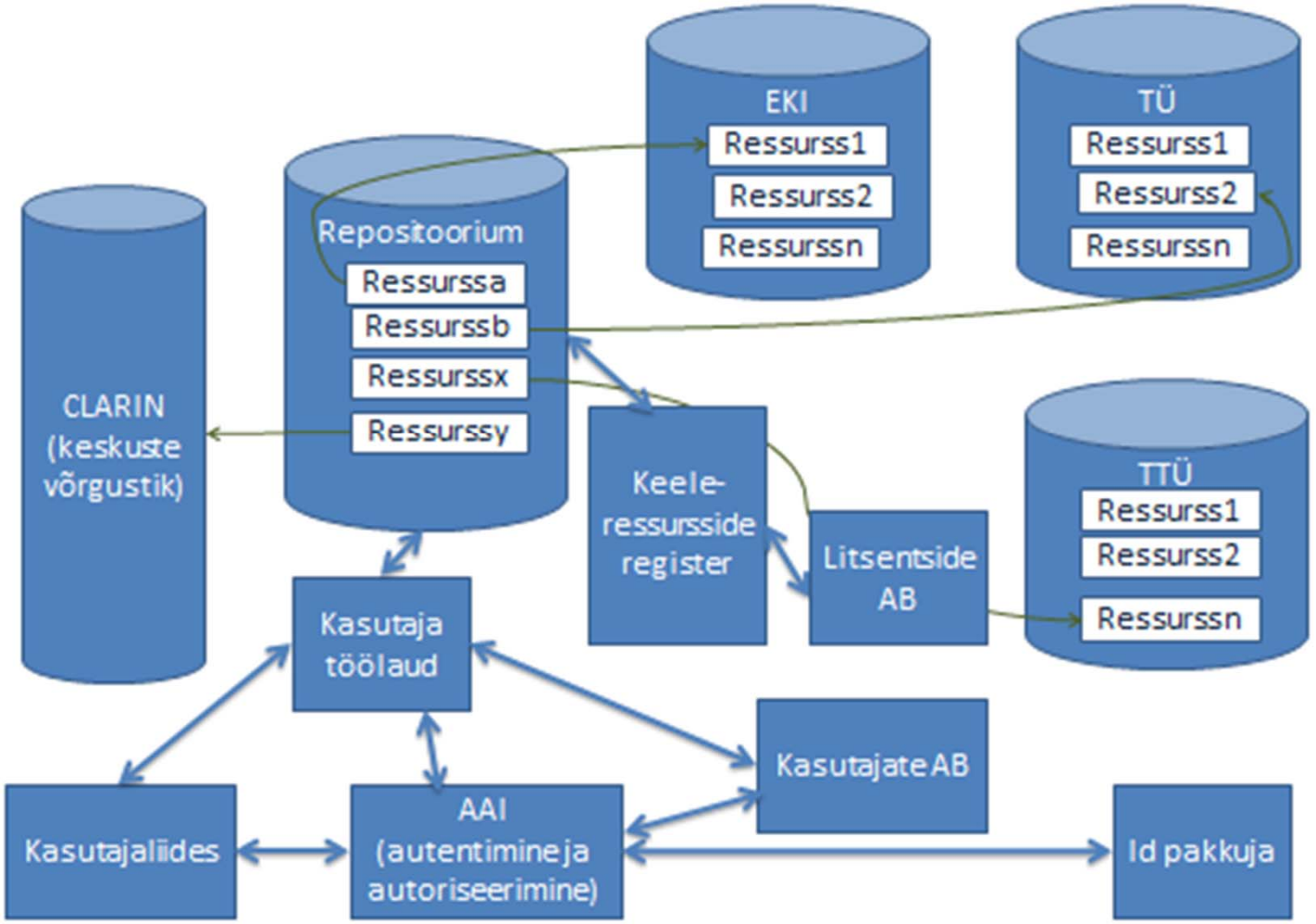
Euroopa Liit
Euroopa Regionaalarengu Fond
Eesti tuleviku heaks

Keskuse funktsioonid

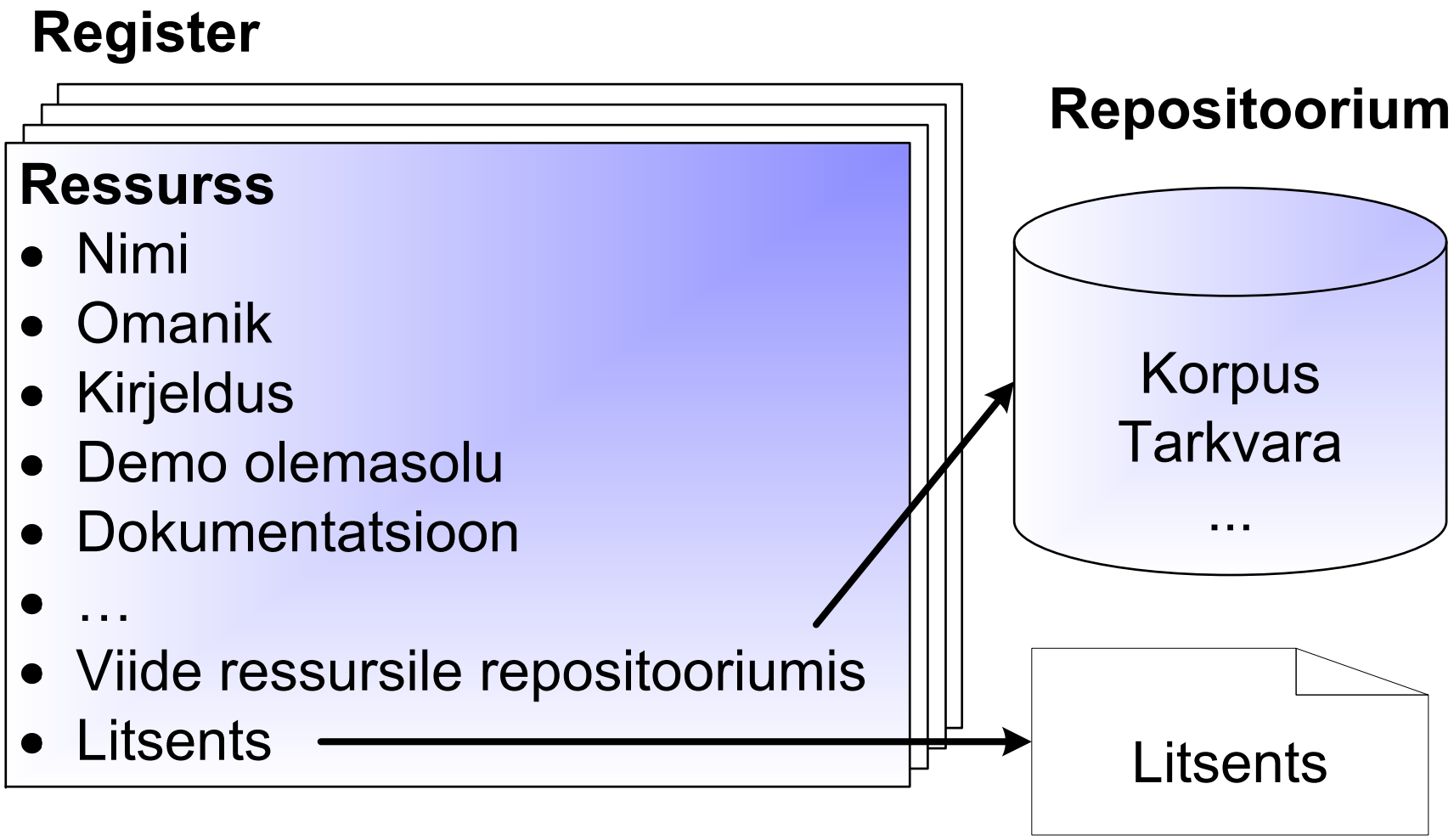
- Eesti keeleressursside keskus on infrastruktuur – erinevates uurimisasutustes paiknevate, veebist ligipääsetavate andmehoidlate võrgustik, mis võimaldab autentimise teel juurdepääsu mitmel erineval tasemel kasutajatele.

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

Keskuse komponendid



Register ja repositoorium (andmemudel)



Euroopa Liit
Euroopa
Regionaalarengu Fond
Eesti tuleviku heaks

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

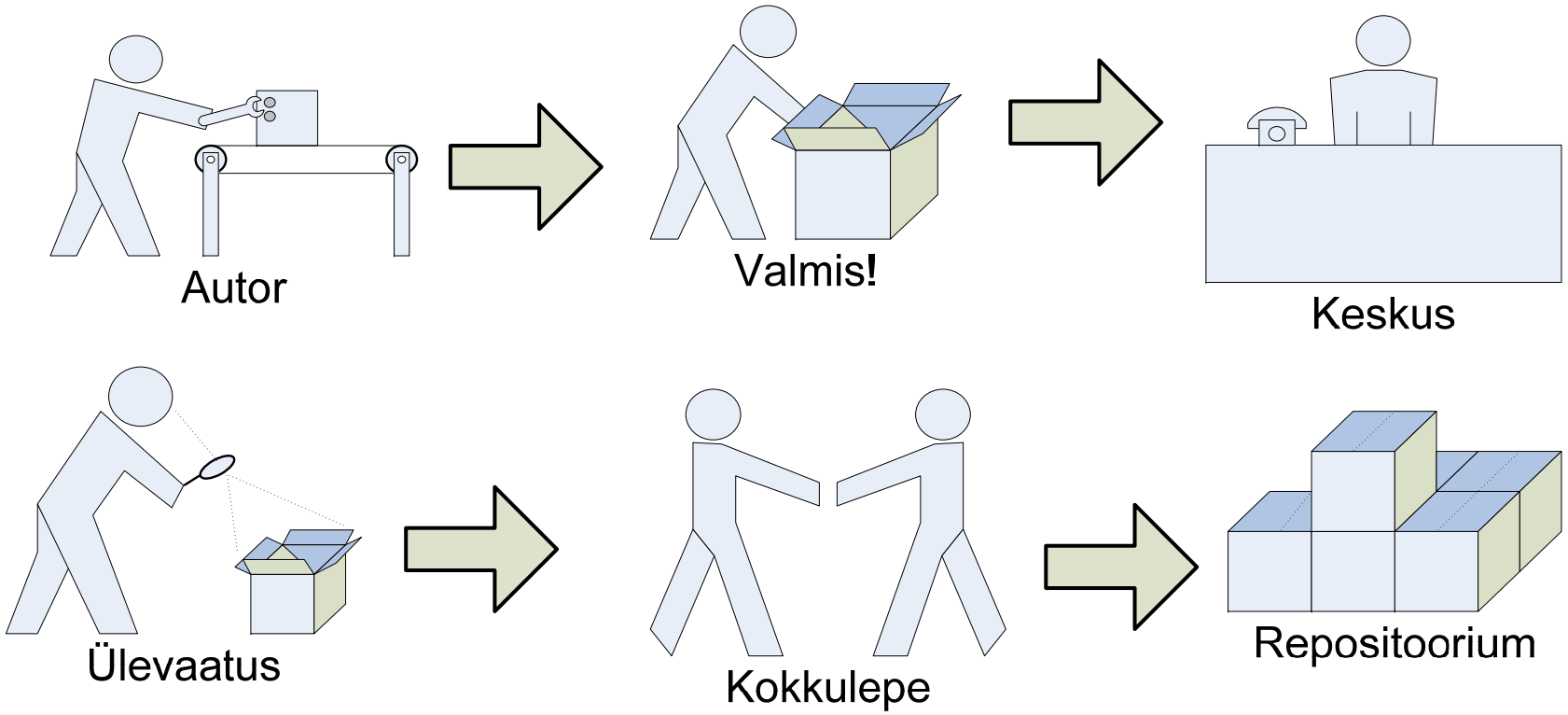
Ressursside kogumine

- 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

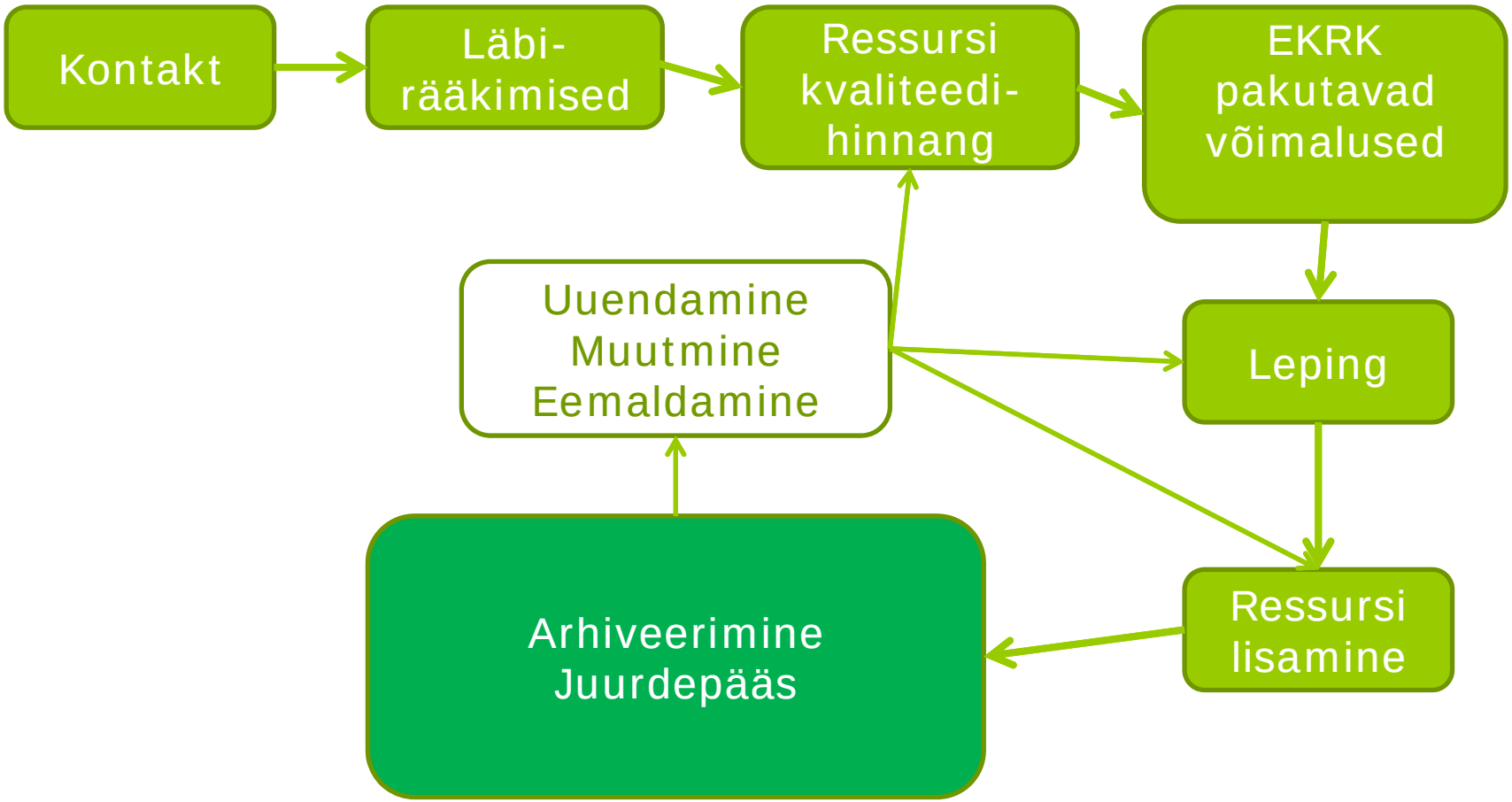
Eesti tuleviku heaks

Kuidas ressurss meile jõuab? ☺



Euroopa Liit
Euroopa
Regionaalarengu Fond
Eesti tuleviku heaks

Ressursi lisamine keskusse



Euroopa Liit
Euroopa
Regionaalarengu Fond
Eesti tuleviku heaks

Kasutusõigused ja ligipääs

KEELEVEEB

EUROPEAN LANGUAGE RESOURCES ASSOCIATION ELRA

MD

KEELEVARA

Eesti Keeleressursside Keskus

MD

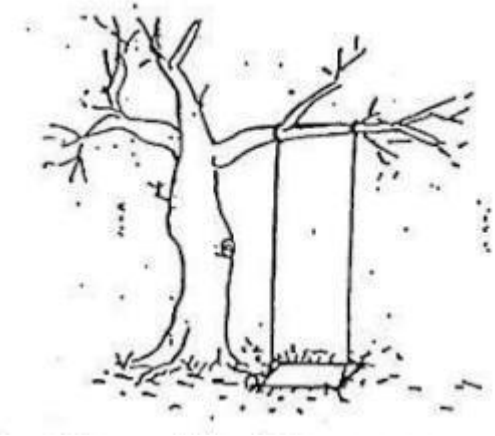
- Ressurssidele serveriruum – ka tarkvaralistele
- Võimalus ressursside arendamiseks keskuses

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

Kasutajate koolitamine

- arhiveerimis- ja töötlemisvõimaluste tutvustamine ressursside loojatele, arendajatele ja pakkujatele;
- kasutusvõimaluste tutvustamine ja koolitus tõhusamaks kasutamiseks ressursside kasutajatele, sealhulgas avalikkusele;
- litsentsimistingimuste tutvustamine

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

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!