

Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüüp

Peep Küngas, Dr. ing.
Arvutiteaduse instituut
Tartu Ülikool

Inimesed

- Martin Luts
- Aivi Kaljuvee
- Margus Tremuth
- Siim Orasmaa
- Aleksandr Tkatšenko
- Enn Petersoo
- Mark Baranin
- Janar Nagel
- Mariana Kukhtyn

Rahastamine

- EKKTT rahastus alates 2009
 - Projekti vanuseks varsti 2 aastat
- Eelarve 2009 + 2010
 - $602\,240 + 575\,000 = 1\,177\,240$ krooni

Eesmärk

- Edendada tehnoloogiat süntaksi- ja semantikapõhiste infootsingute toetamiseks nii veebis kui desktop rakenduses
 - Projekti vahetulemused on sisendiks eesti keele morfoloogilisele ja semantilisele analüüsile

Taustast

- Paremate infootsingute tulemuste tagamiseks on **vaja kombineerida ressursse**
 - süvaveebis, veebis, dokumendihoidlates, lokaalsetes arvutites ja mujal
 - süvaveebi moodustavad online-andmebaasid ja –andmeteenused, mille sisu ei ole otsingumootorite poolt indekseeritav
 - sisu erinevates keeltes
- **Keeruline on siduda** eesti- ja muukeelsete dokumentide **sisu**
 - sõnastikest ei piisa
 - Wordnet ei sisalda valdkonnaspetsiifilist infot IS kontekstis
- Nimega üksuste tuvastamist (i.k. *named entity recognition* - **NER**) **pole vaadeldud kui tervikut** laiema hulga rakenduste kontekstis
- Avalikus sektoris on **loomisel** mitmed **valdkonnaontoloogiad**, mille abil kirjeldatakse semantiliselt **riigi IS andmeteenused** (eestikeelse süvaveebi oluline osa)

Nutika otsimootori prototüüp

- Kombineerib struktureerimata dokumentide sisu süvaveebi andmeallikatega
 - Rakendab **NERi struktureerimata info töötlemiseks**
 - Sobivate andmeteenuste valimiseks rakendatakse **automaatse kompositsiooni algoritme** ja valmislahendusi
 - Kasutab **ontoloogiaid** erinevate **ressursside kombineerimiseks**
 - **Kasutajaliidese** dünaamiline **valimine vastavalt** tuvastatud ja kombineeritud **info semantikale**

Tehnoloogiad



NER
(tekstilised
ressursid)



Kompositsioon
(süvaveebi
ressursid, nt X-
tee)



Ontoloogiad
(üldontoloogia,
valdkondlikud
ontoloogiad)



Vahevara
(mõistete
süsteem +
API)

Rakendused / visualiseerimine (otsisüsteem)

Kasutuslugu (1)

- Kasutaja teeb päringu
- Sõnad kasutaja päringus viiakse algvormi
- Kasutaja päringust **tuvastatakse nimega üksused** ning ülejäänud sõnad seotakse lingvistilise ontoloogia elementidega
- Teostatakse süvaveebi ressursside (veebiteenused) otsing
 - mille sisendiks saab anda nimega üksused ja ajaväljendid
 - mille väljundiks on andmeelemendid, mille semantika on kirjeldatud lingvistilise ontoloogia elementidega konkreetse rakenduse kontekstis

Kasutuslugu (2)

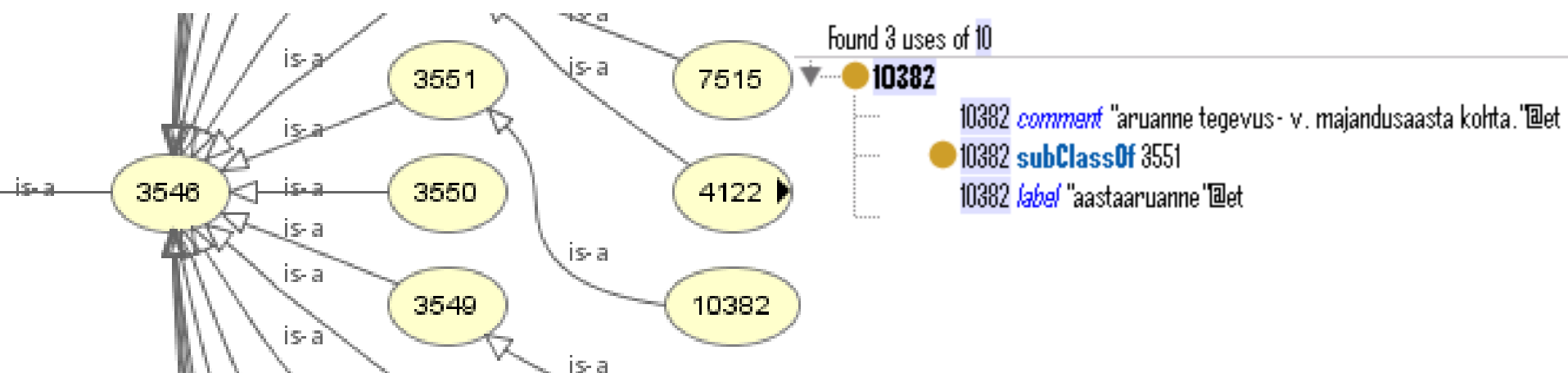
- Teenuseid otsitakse nii kompleks-, kui atomaarsete teenuste seast
 - Kompleksteenuse koosneb mitmest atomaarsest teenusest
- Valitud teenuste visualiseerimiseks valitakse sobiva semantikaga visuaalsed komponendid
- Kasutajale kuvatakse veebi kui süvaveebi otsingu tulemusi

Kasutaja teeb otsingu



- Demo aadressil <http://xml-services.ioc.ee/>
- Päringust leitakse nimega üksus Organisatsioon="Elion" (NER lahendus)
- Eesti Wordneti baasil loodud lingvistilisest ontoloogiast leitakse terminit "aruanne" sisaldava mõiste (3551) terminit "aastaruanne" sisaldav alamklass (10382)

Lingvistiline ontoloogia



Found 5 uses of 3551

10382

- 10382 *subClassOf* 3551

3551

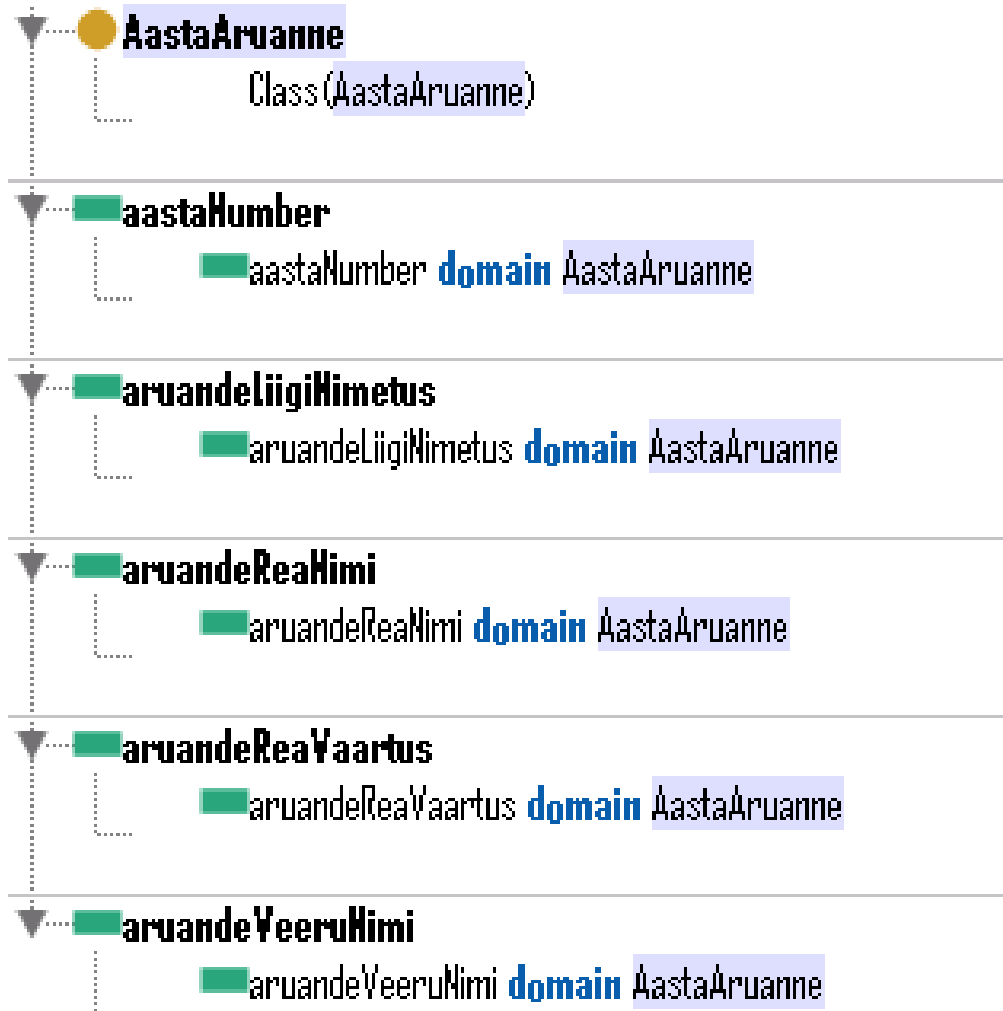
- 3551 *subClassOf* 3546
- 3551 *label* "aruanne"@et
- 3551 *label* "seletus"@et
- 3551 *comment* "mingi nähtuse v. sündmuse põhjendamine lähema kirjeldamise ja lisaandmete esitamisega."@et

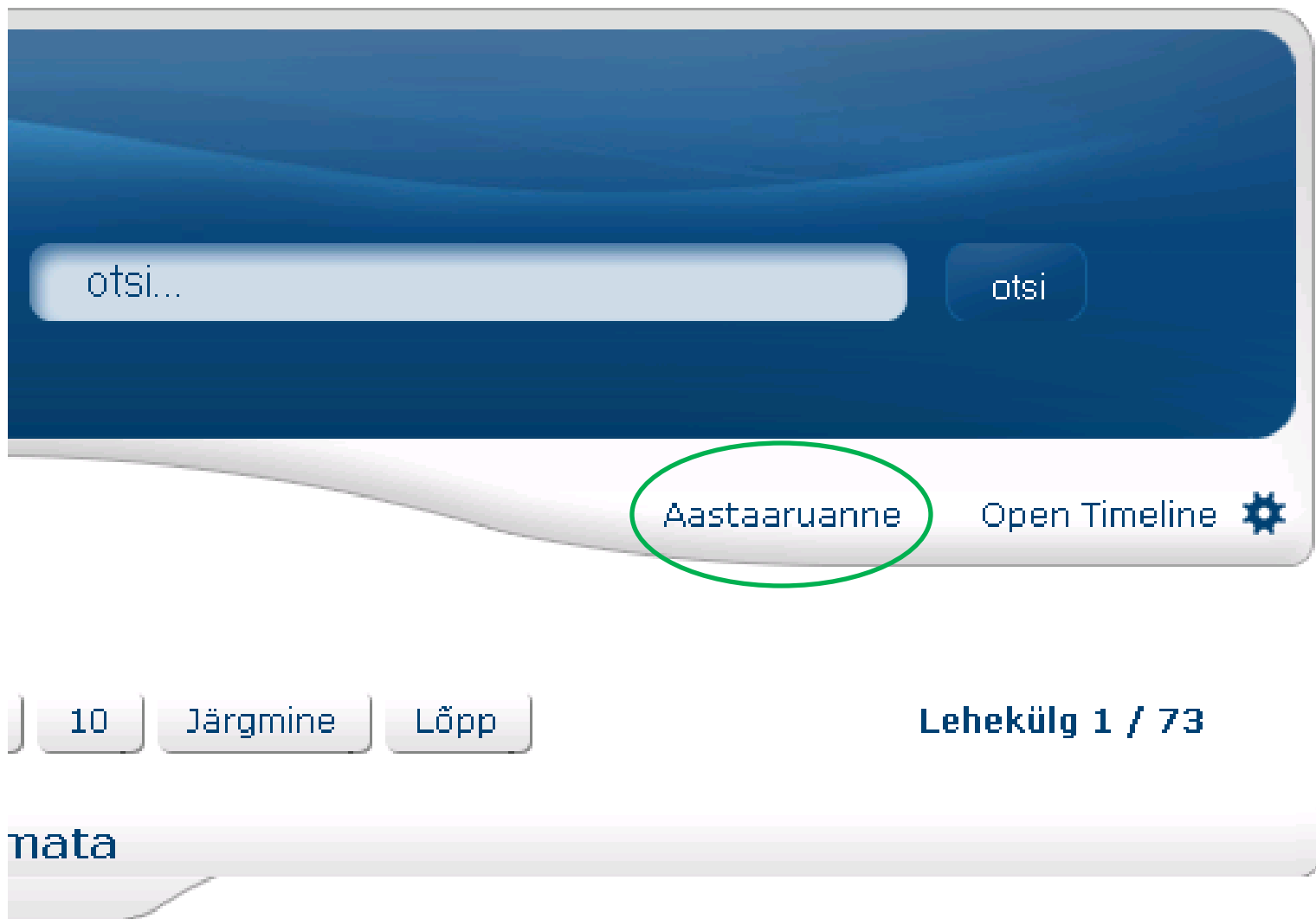
Lingvistiline vs rakendusontoloogia

- Rakendusontoloogiat kasutame süvaveebiressursside semantiliseks kirjeldamiseks (WSDL, XSD, SA-WSDL)
- Rakendusontoloogia element **ro:AastaAruanne** on seotud vastava lingvistilise ontoloogia elemendiga **ling:10382** (owl:equivalentClass):

```
<owl:Class rdf:about="#AastaAruanne">  
  <equivalentClass rdf:resource="&ling;1038"/>  
</owl:Class>
```
- Rakendusontoloogias defineeritakse klasside detailid süvaveebi spetsiifikast lähtudes (nt andmetüüpomadused, mis on vajalikud veebiteenuste semantiliseks kirjeldamiseks)
- **Andmetüüpomaduste järgi leitakse sobivad infot andvad süvaveebiressursid**

Rakendusontoloogia fragment





Leitud süvaveebiressursside komplekt koosneb 3st Äriregistri päringust: **findBusinesses**, **getListOfAnnualReports**, **getAnnualReportData**

Ettevõtte nimetus	Registri kood
<u>Elioni Spordiklubi</u>	80013034
<u>Elion Ettevõtte Aktsiaselts</u>	10283074
<u>Dandelion Teeninduse OÜ</u>	10504960
<u>Elioni Tallinna Ametiühing</u>	80106474
<u>OÜ HELION GRUPP</u>	10680674
<u>ELION KINNISVARA OÜ</u>	10957638
<u>Gelion Lar Osaühing</u>	10963188
<u>Elion Ametiühing</u>	80211395
<u>Osaühing HelionPartner</u>	11119201
<u>OSAÜHING Meliond</u>	11157992
<u>Adelion Kids OÜ</u>	11445432
<u>Elionora Galu⁰⁰kevit⁰⁰</u>	11814289
<u>Ofelion OÜ</u>	11863566
<u>Intelions OÜ</u>	11938152
<u>Elionaara OÜ</u>	11968118
<u>OÜ Keelion Translations</u>	11004662
<u>OÜ Seelion</u>	11004320
<u>osaühing Helion Ehitus</u>	11141471
<u>APHELION OSAÜHING</u>	11477716
<u>Triskelion OÜ</u>	11537752
<u>Exelion OÜ</u>	10967068
<u>Adelion & Partnerid Õigusbüroo OÜ</u>	11013218
<u>Exelion plus OÜ</u>	11702254
<u>Osaühing HelionBaltic</u>	10974223

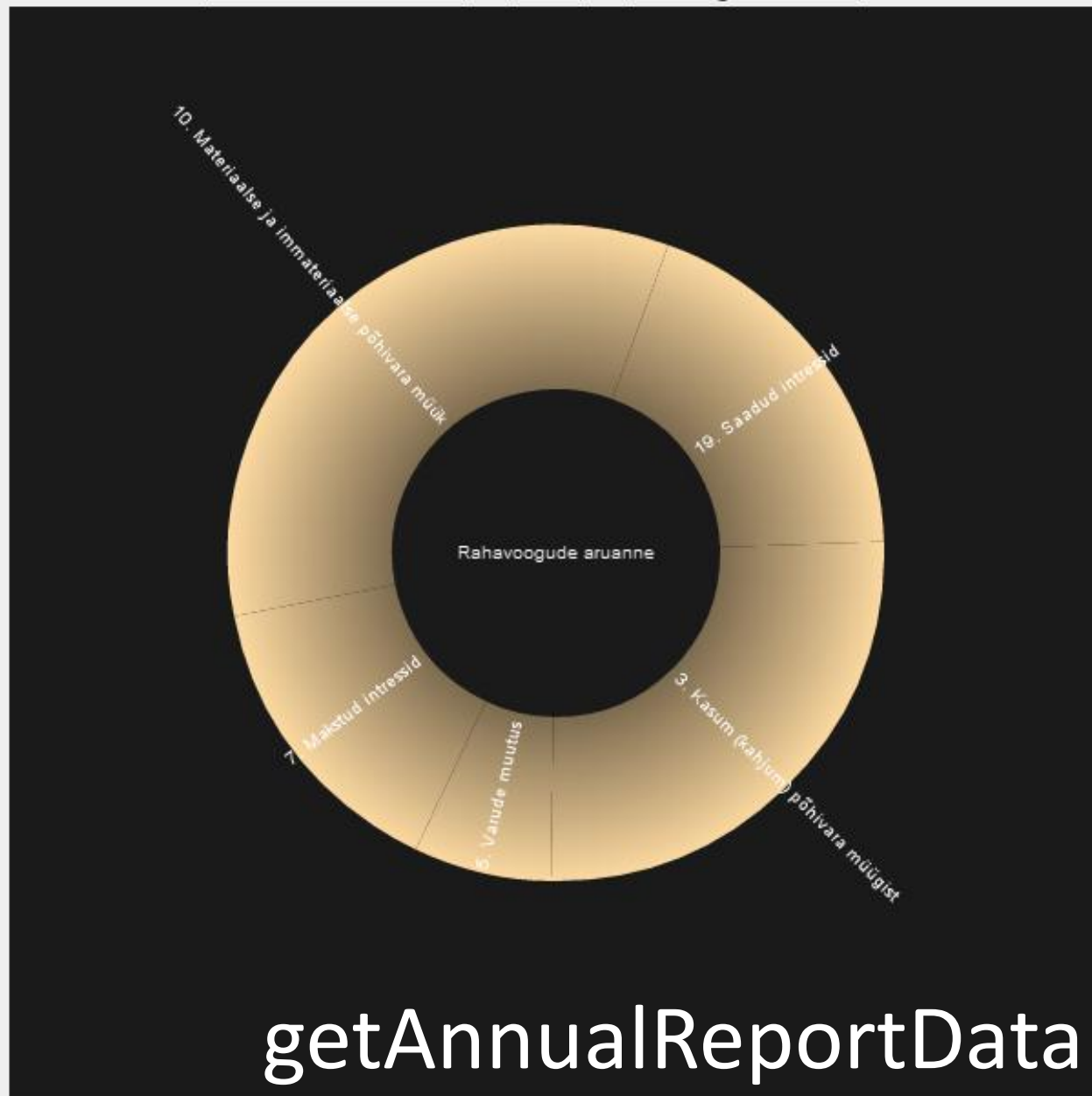
[Kasumiaruanne skeem 1](#) | [Bilans](#) | [Rahavoogude aruanne](#)

getListOfAnnualReports

[Kasumiaruanne skeem 1](#)

[Bilanss](#)

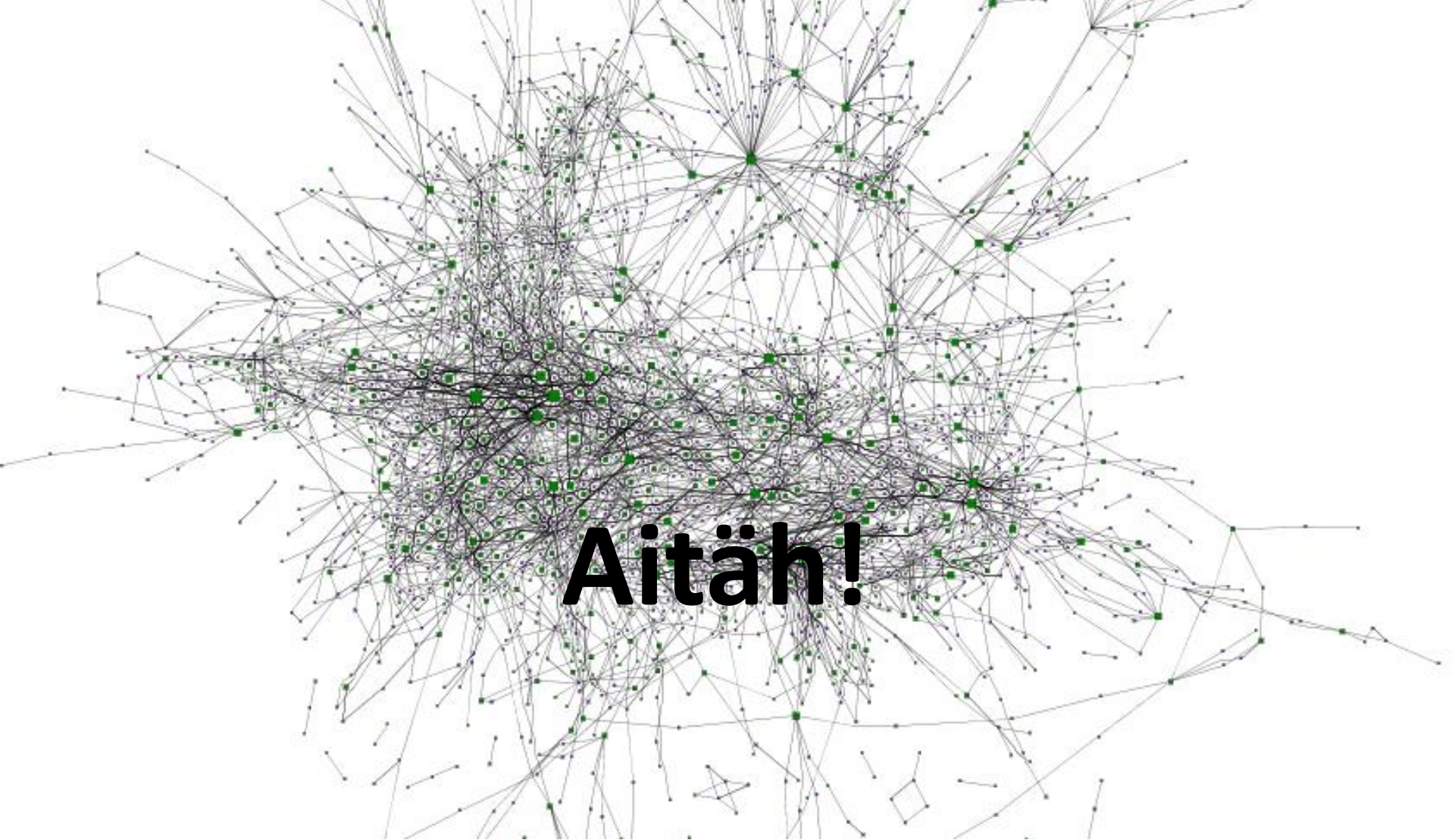
[Rahavoogude aruanne](#)



getAnnualReportData

Kokkuvõte

- Projekti raames on loodud NER realisatsioon, mida saab kasutada kas Java teegi või SOAP teenusena
- Projekt on esimeseks sammuks süvaveebi- ja veebiressursside kombineerimise lahenduse loomiseks kasutades eesti keeletehnoloogiaid
- Keeletehnoloogiate rakendamine semantilise veebi kontekstis on globaalselt üha enam aktuaalseks muutumas ja ka Eestil on siia panustada
- **Otsime kasutajaid infootsisüsteemi efektiivsuse mõõtmise jaoks**



Aitäh!

peep.kungas@ut.ee