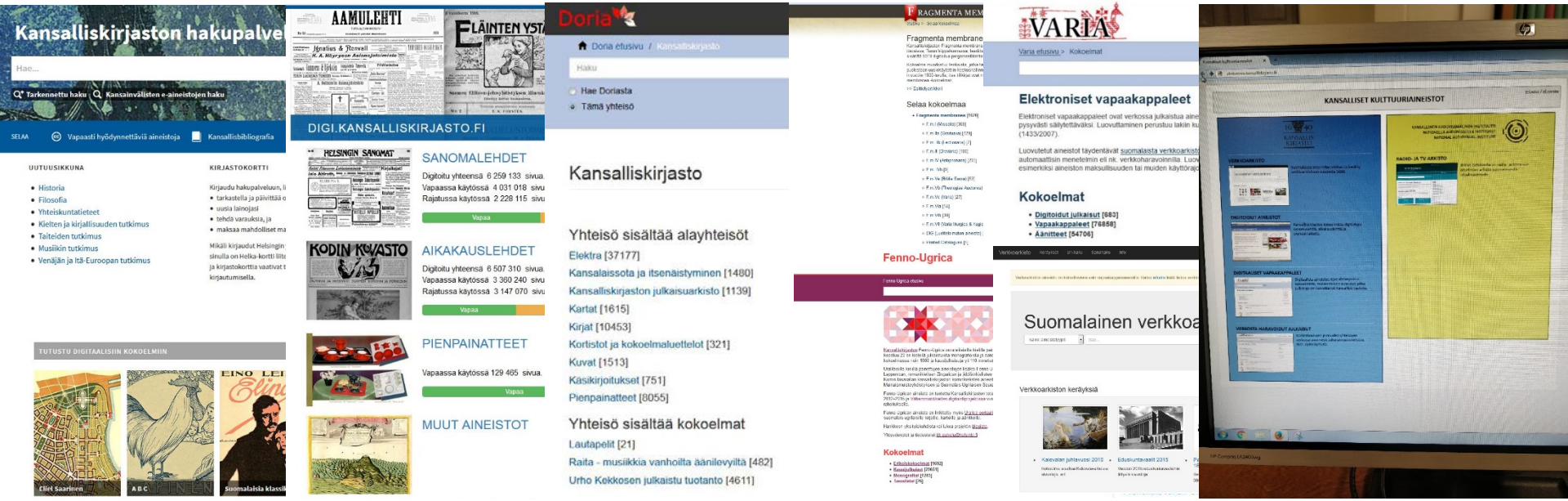
Kansalliskirjaston digitaaliset aineistot ja data

Jussi-Pekka Hakkarainen
Tietoasiantuntija
Erikoiskokoelmat
jussi-pekka.hakkarainen@helsinki.fi

Joitain viimeaikaisia kehityslinjoja

- Digitalia (2015 – 2019)
- Kansalliskirjaston DH-politiikka (2016)
- digi.kansalliskirjasto.fi:n kehittäminen (2016-)
- Avoin data / datakatalogi (2017)
- Verkkoarkistoinnin ja -haravoinnin hyödyntäminen tutkimuksessa (2018-)
- Massadigitoinnista kohti tutkijalähtöisempää digitointia (2019-)



Digitalia ja COMHIS

- Kansalliskirjaston ja Kaakkois-Suomen ammattikorkeakoulun yhteisen tiedonhallinnan tutkimus- ja kehittämiskeskus **Digitalian** (2015 – 2019) tavoitteena on syventää osaamista digitaalisten aineistojen hallinnassa ja säilyttämisessä.
  - **Digitalian I** vaihe jatkui 31.7.2017 saakka. Sanomalehtiaineiston suomenkieliselle osalle tehtiin sanatason laatuarviota, aineiston optisen luvun tasoa kehitettiin ja selvitettiin jälkikorjauksen mahdollisuuksia.
  - **Digitalia II** -projekti alkoi 1.8.2017. Sen tavoitteena on nimien tunnistamisen kehittäminen (NER, Stanford NER) ja nimihaun digi.kansalliskirjasto.fi:hin luominen, OCR:n parempi laatu, sanomalehtiaineiston kuvien luokittelun ja artikkelien eristämisen edistäminen.

- Suomen akatemian digitaalisten ihmistieteiden tutkimusrahoituksen **COMHIS-projekti** (2016-19). Kansalliskirjasto tukee aineistojen käytettävyyttä. Vuoden 2017 aikana Kansalliskirjasto avannut Digi.kansalliskirjasto.fi:n rajapinnat (OAI-PMH, OpenURL) metatietojen hakua varten uudelleen. Rajapinnat on myös kuvattu.

Digitaalisten ihmistieteiden politiikka

- Digitaalisilla ihmistieteillä tarkoitetaan 1) laskennallisten metodien hyödyntämistä humanistisessa ja yhteiskuntatieteellisessä tutkimuksessa sekä 2) digitaalisuuden ja digitaalisen kulttuurin tutkimusta.

- Kansalliskirjaston kontekstissa *digitaaliset ihmistieteet* (Digital Humanities) tarkoittaa tutkimukseen soveltuvien aineistojen, asiantuntijapalveluiden, infrastruktuurien kehittämistä ja tarjoamista yhteistyössä tiedeyhteisön ja muiden toimijoiden kanssa.

- DH-politiikka ja teesit

Avoin data ja datakatalogi

- DH-politiikan jalkauttamista Kansalliskirjaston omaan toimintaan.

- Politiikassa mainittuja tavoitteita:
  - Avataan tekijänoikeuksista vapaita digitoituja teksti- ja datamassoja konelukuisessa muodossa.
  - Tarjotaan / avataan rajapintoja ja testataan niitä yhteistyössä tutkijoiden kanssa (Finna,Digi, Doria, metatietovarannot).
  - Avataan verkkoarkisto ja muut e-vapaakappaleet soveltuvasti tutkijakäyttöön.
  - Edistetään verkkoaineistojen keräämistä, arkistointia ja tutkimuskäyttöä koskevaa julkista keskustelua.
  - Tarvittaessa hankitaan ulkopuolisten tuottamaa dataa tutkimusaineistoksi.

- Kansallisbibliografian avaaminen avoimena datana, joulukuu 2017.

Avoin data ja datakatalogi

tietoaineistot

- bibliografinen metatieto
- auktoriteettitiedot (ml. sanastot ja ontologiat)
- tietomallimääritykset
- kokoteksti- ja ääniteaineistot

Tietoaineiston tulee olla **saatavilla** koneellista analyysia, jälkikäsittelyä ym. varten esimerkiksi *rajapintapalveluna* ja/tai *ladattavina tiedostoina*.

Pelkkä verkkosivusto ei ole tietoaineisto!

Avoin data ja datakatalogi

- Kansalliskirjaston itse tuottamaa dataa ja metadataa voivat kaikki hyödyntää vapaasti. Tietovarantoja ja rajapintoja, jotka on julkaistu avoimena datana CC0-lisenssillä. Lisätiedot aineistoista ja niiden käyttömahdollisuuksista on koottu datakatalogiin.

- Kansallisbibliografia Fennican data on hyödynnettävissä useilla eri tavoilla. Fennican tietoja pääsee selaamaan avoimen datan päälle rakennetun käyttöliittymän avulla.

- Oman datakatalogin ja tietoaineistojen dokumentaation avulla tavoitellaan parempaa näkyvyyttä Kansalliskirjaston datalle sekä mahdollistetaan Kansalliskirjaston tietoaineistojen saavutettavuuden kansallisissa ja kansainvälisissä datakatalogeissa.

Suku- ja Vähemmistökielten digitointiprojektit

- Toteutettu Koneen Säätiön rahoituksella vuosiuna 2012-2017, osana **Koneen Säätiön Kieliohjelmaa**.
- Tarkoituksena on pienten suomalais-ugrilaisten kielten, suomen sekä Suomen vähemmistökielten **dokumentointi** ja niiden **aseman vahvistaminen**.
- Tavoitteena on saattaa kieliaineistoja **tiedeyhteisön** ja **muun yhteiskunnan** avoimeen käyttöön.
- Projektissa digitoitiin 20 uralilaisella kielellä painettuja aineistoja ja aineiston jalostamisvälineitä **kielentutkimukselle** ja **kansalaistieteelle**.
- Digitoituja monografianimekkeitä on kokoelmassa noin 1500 ja kausijulkaisuja yli 110 nimekettä.

Kielten dokumentointia

- Keskeisiä komponentteja kielen dokumentaatiossa ovat:
  - 1) aineistojen **tallentaminen**, mukaan lukien siihen liittyvän metadatan tuottaminen
  - 2) kieliaineistojen **siirrettävyys**
  - 3) arkistointi ja arkistoidun aineiston **avoin saavutettavuus**
  - 4) **lisäarvon tuottaminen** mm. annotoimalla, transkriboimalla ja linkittämällä
  - 5) aineiston **mobilisointi**, eli sen hyödynnettävyys kolmansissa järjestelmissä.

[Peter Austin, 12.6.2015; Lyle Campbell & Bryn Hauk, 17.8.2015; Michael Rießler 20.8.2015]

- Kielen dokumentointi tarkoittaa yhä enemmän sitä, mitä Kansalliskirjasto tehtäviensä puolesta jo tekee. Vrt. myös **Avoin tiede ja tutkimus**.

1) Aineistojen tallentaminen

Fenno-Ugrica

Fenno-Ugrica etusivu > Monografiat > Kirjat

Hae Hakuohjeet

Tämä kokoelma Hae Fenno-Ugricasta

Kirjat

Uusimmat viitteet

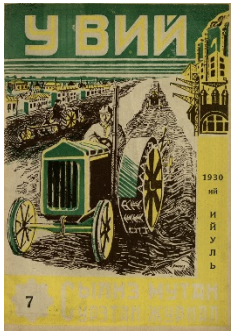
Nimeke: СССР-лӧн історія : жеӈдӧдӧм курс : 3-ӧд da 4-ӧд классъяслы учебӈик
Julkaistu: Сыктывкар : Коми Государственнӧй Издательство, 1938

Nimeke: Сӧветскӧй Социалистическӧй Республикаяс Союзлӧн Конституция : Основнӧй Закон
Julkaistu: Сыктывкар : Коми Государственнӧй Издательство, 1938

Nimeke: СССР-са народнӧй овмӧс развитиелӧн коймӧд пятилетньӧй план : ВКП (б) XVIII-ӧд съезд вылын доклад да заключительнӧй кыв : март 14-17 луньяс, 1939 во
Tekijä: Молотов, В.
Julkaistu: Сыктывкар : Коми государственнӧй издательство, 1939

Nimeke: Гурт комсомоллы политикалы дышетскон книга : нырисетйез книга
Tekijä: Барков, В. Н.
Julkaistu: Ижкар : Удкнига, 1929

Nimeke: Вашкала дуннелэн историез : средней школалэн 5-6 классъёсыслы учебник
Julkaistu: Ижевск : Удмуртгосиздат, 1943



2) Kieliaineistojen siirrettävyys

Saam bukvar

Cerniakov, Saxkre

Julkaisun pysyvä osoite on http://urn.fi/URN:NBN:fi-fe2016051212331

Nimi: sme_4-2_1933_DATA.zip
Koko: 338.5Mt
Formaatti: Unknown
Kuvaus: ZIP package containing ...
Lataukset 36

Avaa tiedosto

Datapaketti, sis. AltoXML, CSV, TIFF, TXT

Nimi: sme_4-2_1933.pdf
Koko: 32.30Mt
Formaatti: PDF
Lataukset

Avaa tiedosto

Käyttökopiona PDF

Nimeke: Saam bukvar
Muu nimeke: Букварь на саамском языке
Tekijä: Cerniakov, Saxkre
Julkaistu: Moskva ; Leningrat : Ucpedgiz, 1933.
Научно- исслед. ассоц. Ин-та народов севера ЦИК СССР

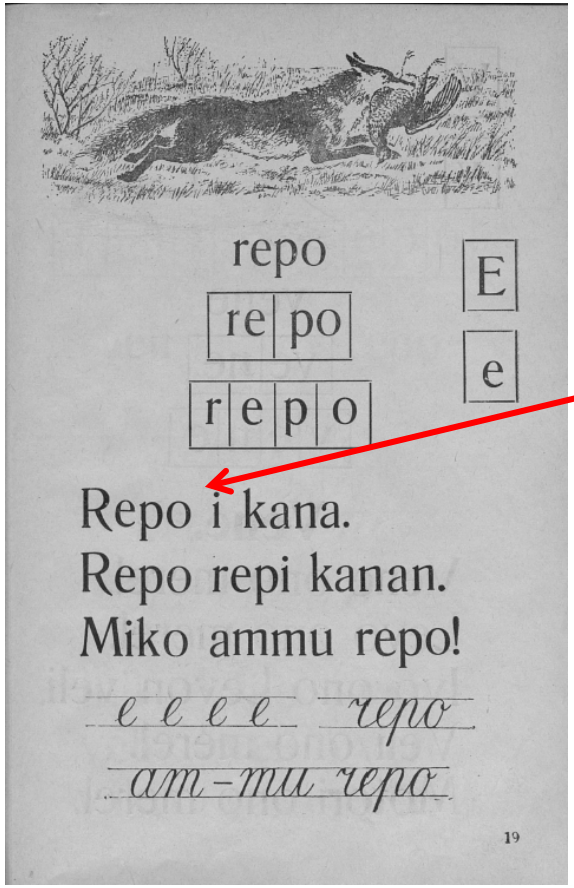
Selaa kokoelmaa
- Nimekkeet
- Tekijät
- Julkaisuajat
- Asiasanat
- Uusimmat
- Selaa kielen mukaan
- Julkaisutyyppi
- Sivukartta

Omat tiedot
- Kirjaudu ulos
- Profiili
- Tallennukset

Toiminnot
- Muokkaa tietuetta
- Eksportoi tietue
- Eksportoi metadata

Hallinnointi
- Käyttöoikeudet
  - Käyttäjät
  - Ryhmät

2) Kieliaineistojen siirrettävyys



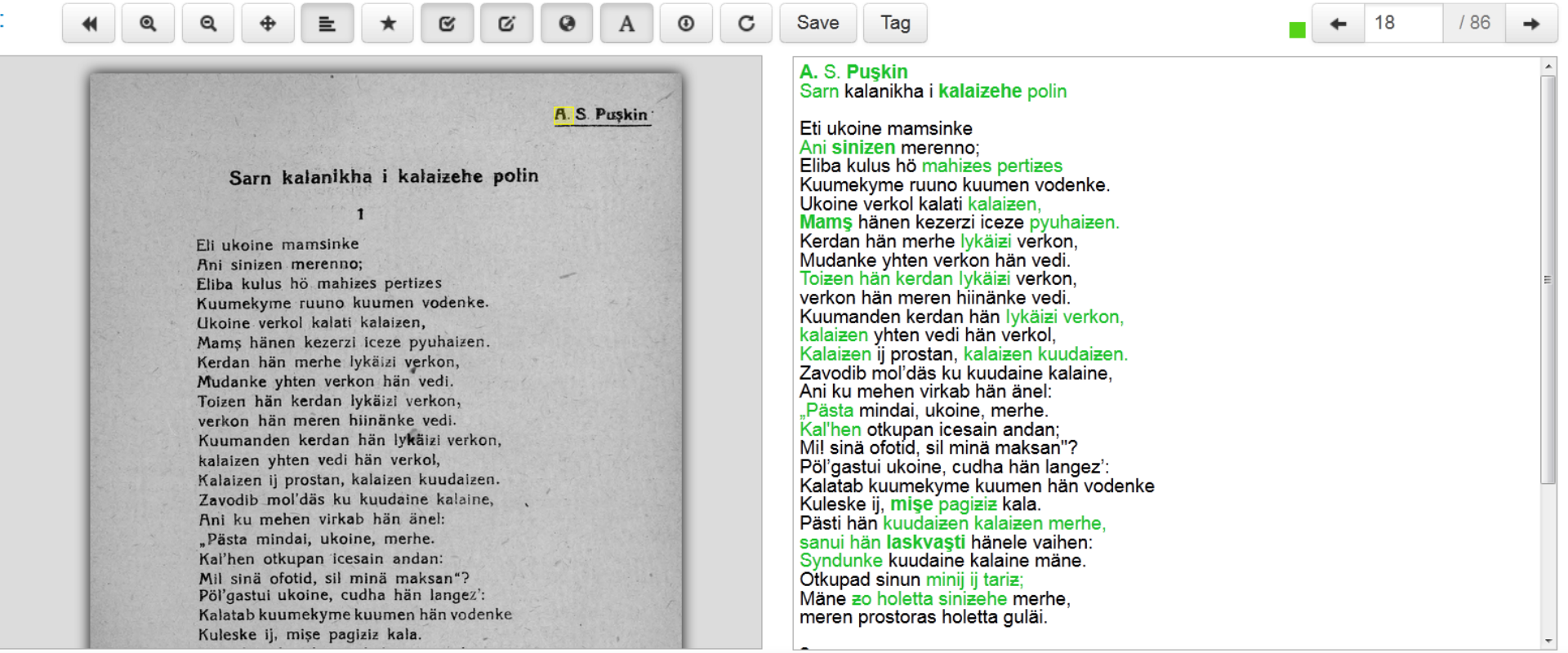

```
- <TextLine HPOS="283" VPOS="1461" WIDTH="798" HEIGHT="158">
    <String HPOS="283" VPOS="1461" WIDTH="326" HEIGHT="158" CONTENT="Repo"/>
    <SP HPOS="610" VPOS="1473" WIDTH="62"/>
    <String HPOS="673" VPOS="1473" WIDTH="24" HEIGHT="106" CONTENT="i"/>
    <SP HPOS="698" VPOS="1461" WIDTH="66"/>
    <String HPOS="765" VPOS="1461" WIDTH="316" HEIGHT="122" CONTENT="kana."/>
  </TextLine>
- <TextLine HPOS="281" VPOS="1651" WIDTH="1084" HEIGHT="160">
    <String HPOS="281" VPOS="1651" WIDTH="328" HEIGHT="160" CONTENT="Repo"/>
    <SP HPOS="610" VPOS="1693" WIDTH="62"/>
    <String HPOS="673" VPOS="1663" WIDTH="230" HEIGHT="146" CONTENT="repi"/>
    <SP HPOS="904" VPOS="1651" WIDTH="60"/>
    <String HPOS="965" VPOS="1651" WIDTH="400" HEIGHT="120" CONTENT="kanan."/>
  </TextLine>
- <TextLine HPOS="279" VPOS="1843" WIDTH="1128" HEIGHT="154">
    <String HPOS="279" VPOS="1845" WIDTH="320" HEIGHT="120" CONTENT="Miko"/>
    <SP HPOS="600" VPOS="1885" WIDTH="66"/>
    <String HPOS="667" VPOS="1881" WIDTH="366" HEIGHT="84" CONTENT="ammu"/>
    <SP HPOS="1034" VPOS="1881" WIDTH="66"/>
    <String HPOS="1101" VPOS="1843" WIDTH="306" HEIGHT="154" CONTENT="repo!"/>
  </TextLine>
```

3) Avoin saavutettavuus

- Fenno-Ugrican aineistot ovat avoimesti saavutettavissa.
- Aineistot PAS-kelpoisia, eli säilytettävyysnäkökulma on huomioitu.
- Teoksiin liittyvät tekijänoikeudelliset kysymykset on ratkaistu selvittämällä ja/tai sopimalla. Pyrkimys täyteen avoimuuteen tulisi olla aina päämääränä, sillä näin mahdollistetaan:
  - aineistojen esittäminen
  - aineistojen mobilisointi, eli hyödyntäminen kolmansissa järjestelmissä.

4) Lisäarvon tuottaminen

Sukukielten digitointiprojektissa on tuotettu myös vapaan lähdekoodin tekstieditori, **Revizor**, jonka avulla teosten konetunnistettua tekstiä voidaan korjata ja muokata.



4) Lisäarvon tuottaminen

- Kieliaineistojen korjaamisessa pyrittiin hyödyntämään erityisesti **kohdennettua joukkoistamista**.

- Joukkoistamalla aineistojen oikolukua voidaan yhdistää ammattilaisten osaamista optimoimaan monimutkaistenkin työtehtävien ihmislähtöistä suorittamista valistuneiden kansalaistieteilijöiden voimin.

- Pyrittiin **vastavuoroisuuteen**, jolloin kansalaistieteilijät ja heidän yhteisönsä hyötyvät hankkeen tuloksista.

4) Lisäarvon tuottaminen

- Perinteinen talkoistaminen:
  bród

- Kohdennettu joukkoistaminen:
  brød

…tai jopa enemmän
brød / bröd / bread /
Bröt / leipä / chleb etc.

4) Lisäarvon tuottaminen

- Joukkoistaminen onnistui vain osittain:
  - Sopivia kansalaistieteilijöitä oli vaikea löytää.
    - Inkeroinen ja vepsä onnistuivat.
    - Yhteistyö kansalaisjärjestöjen oli hankalaa.
    - Yliopistot ja kirjastot eivät kiinnostuneet tästä työstä.
- Prosessit olivat liian pitkiä.
- Tehtävät eivät olleet vaikeita, mutta osaaminen oli este.
- Vastavuoroisuutta ei aidosti syntynyt.
- Oikolukua lopulta jouduttiin ostamaan.

- Tampereen yliopiston Vanha kirjasuomen kurssi oli menestys.

4) Lisäarvon tuottaminen

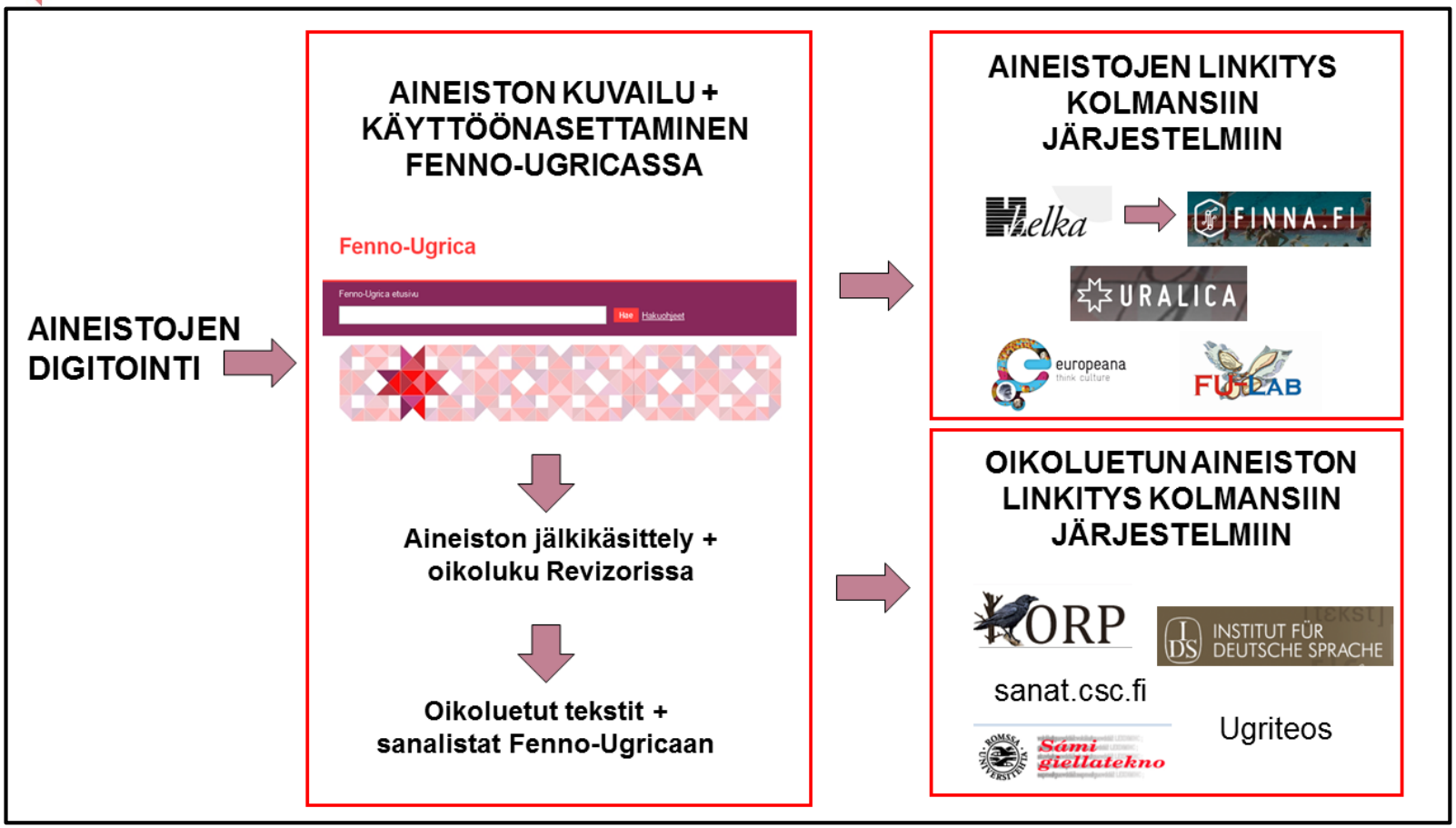
- Sanalistat julkaistiin .csv-formaatissa Fenno-Ugricassa, mistä ne ovat sellaisinaan saatavilla.

- Ei aivan korpuslingvistiikkaan kelpaavaa laatua, mutta hyvää kuitenkin.
  - Ei analyysiä.
  - Henkilöstö osaamiskapeikossa?

- Pitäisi käsittää pikemminkin **raakadatana** kuin valmiina tuotteena (rajallinen sanasto, vain otteita kielestä).

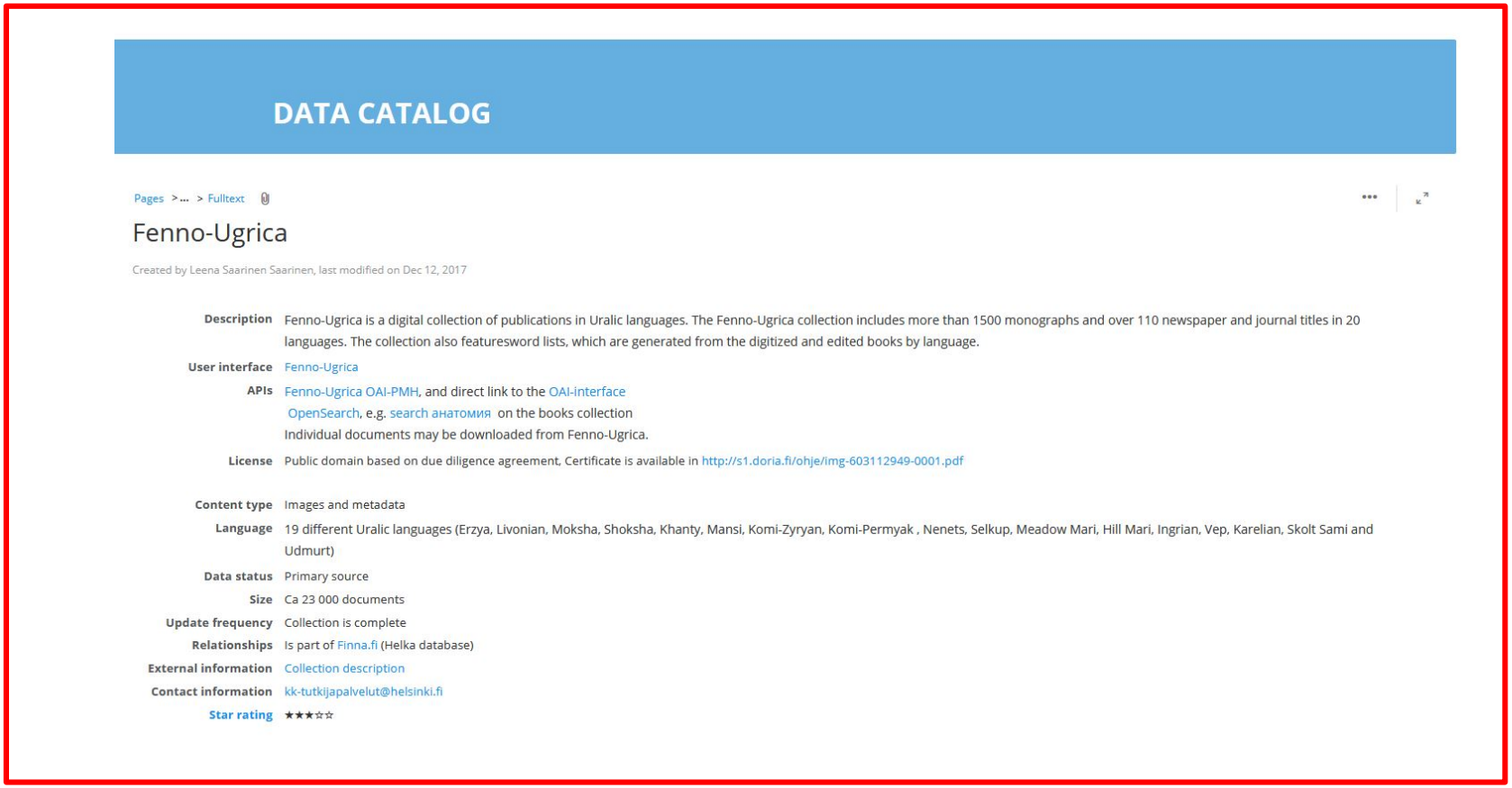
- Auttoi projektia parantamaan OCR-laatua.

5) Aineistojen mobilisointi



5) Aineistojen mobilisointi

- Saavutettavissa myös datakatalogista



DATA CATALOG

Pages > ... > Fulltext

Fenno-Ugrica

Created by Leena Saarinen Saarinen, last modified on Dec 12, 2017

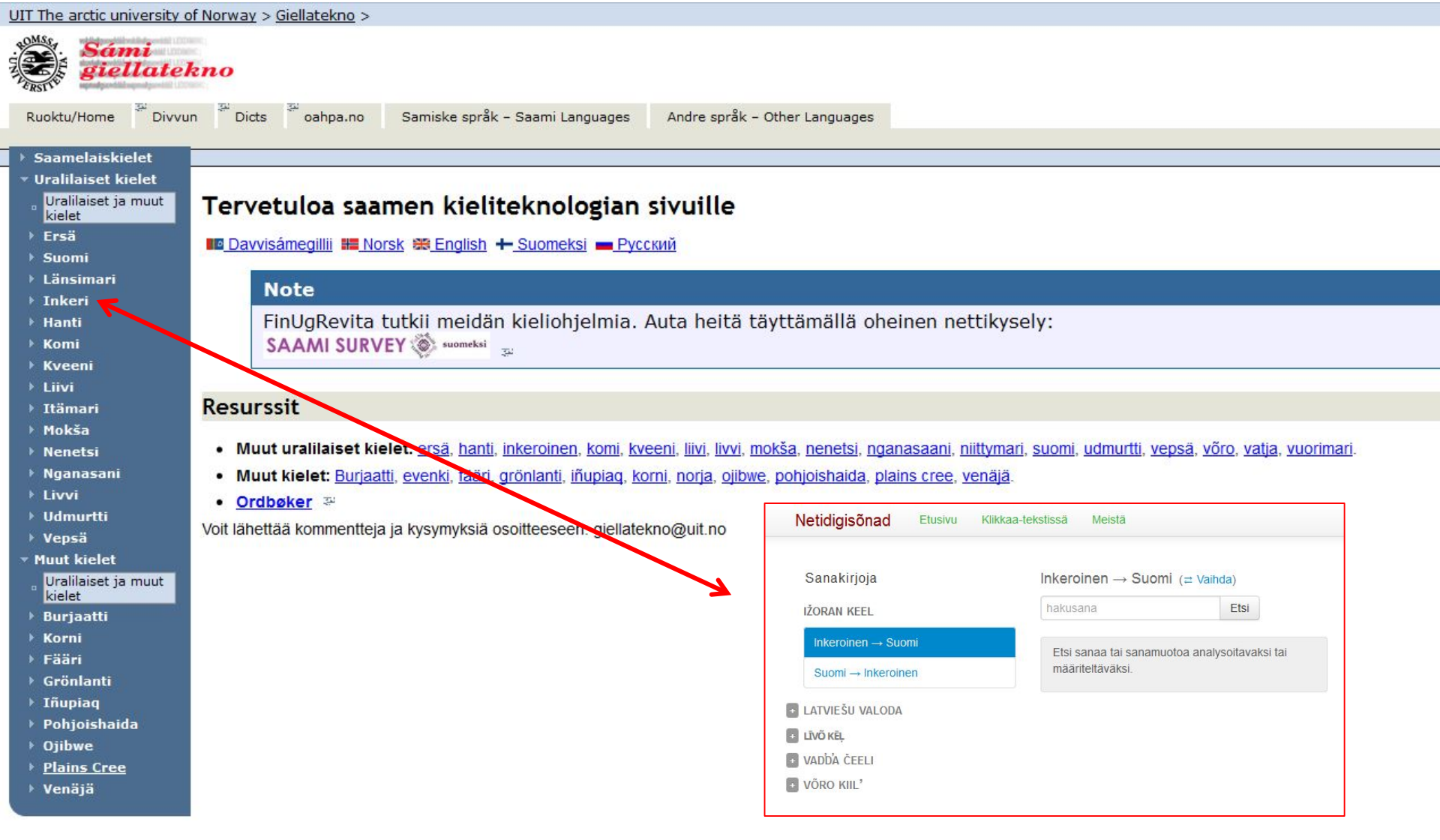
Description	Fenno-Ugrica is a digital collection of publications in Uralic languages. The Fenno-Ugrica collection includes more than 1500 monographs and over 110 newspaper and journal titles in 20 languages. The collection also featuresword lists, which are generated from the digitized and edited books by language.
User interface	Fenno-Ugrica
APIs	Fenno-Ugrica OAI-PMH, and direct link to the OAI-interface OpenSearch, e.g. search анатомия on the books collection Individual documents may be downloaded from Fenno-Ugrica.
License	Public domain based on due diligence agreement. Certificate is available in http://s1.doria.fi/ohje/img-603112949-0001.pdf
Content type	Images and metadata
Language	19 different Uralic languages (Erzya, Livonian, Moksha, Shoksha, Khanty, Mansi, Komi-Zyryan, Komi-Permyak , Nenets, Selkup, Meadow Mari, Hill Mari, Ingrian, Vep, Karelian, Skolt Sami and Udmurt)
Data status	Primary source
Size	Ca 23 000 documents
Update frequency	Collection is complete
Relationships	Is part of Finna.fi (Helka database)
External information	Collection description
Contact information	kk-tutkijapalvelut@helsinki.fi
Star rating	★★★☆☆

5) Aineistojen mobilisointi

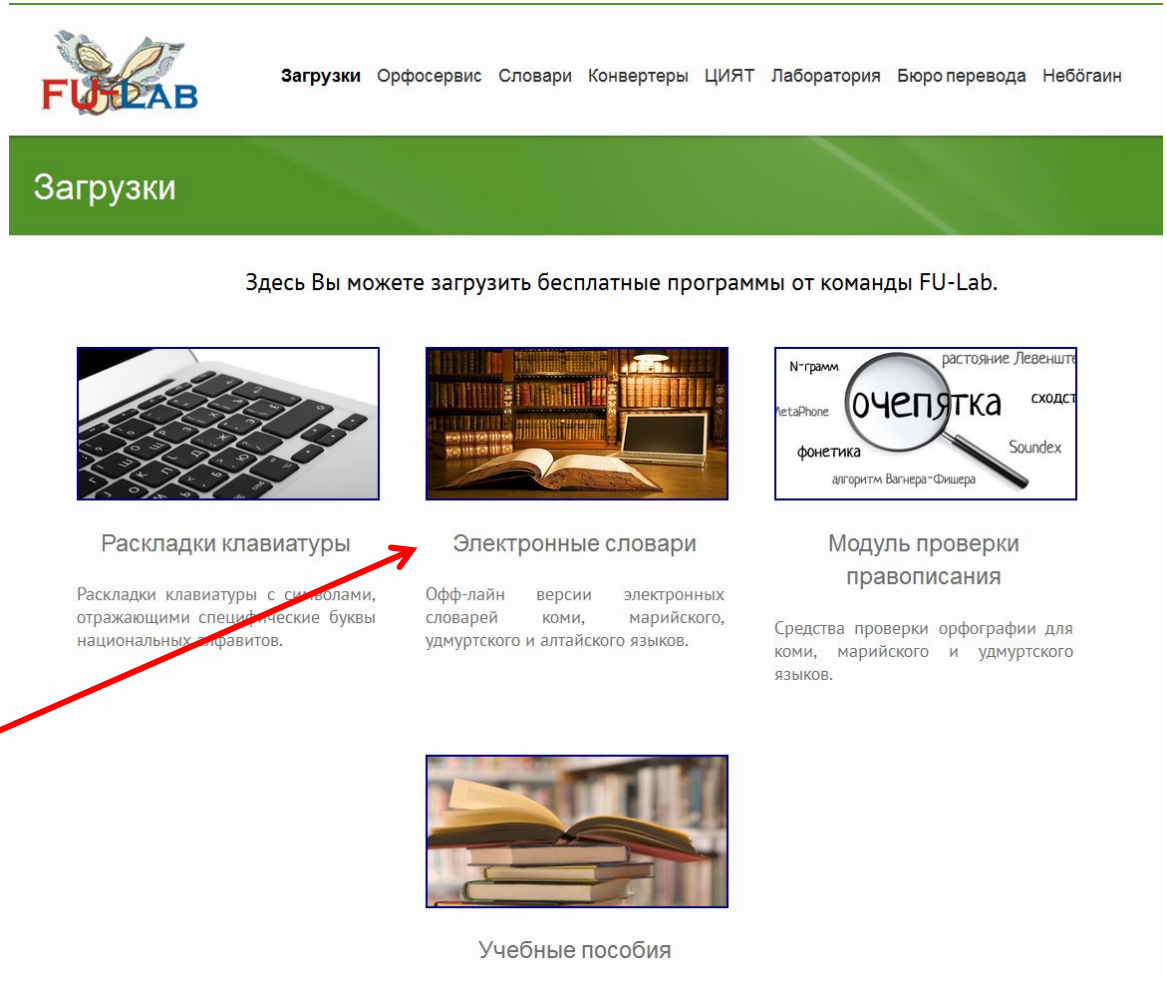
- Aineistojen jakaminen suoraan kieliteknologeille:
  - **Giellatekno** (Tromssa)
  - **FU-Lab** (Syktyvkar)
  - **Institut für Deutsche Sprache** (Mannheim)
  - **Venäjän Tiedeakatemia** (Moskova)

- Hyödyntäminen eri projekteissa, mm.
  - **Uralica** (Kansalliskirjasto)
  - **SUKI-projekti** (Helsinki)
  - **Jazva-Komi** ja **Kiltinänsaamen annotointiprojekti** (Freiburg)
  - **Mari-Language.com** (Wien)
  - **Giellagas** (Oulu)
  - **Volgan alueen tutkimusyksikkö** (Turun yliopisto)
  - Yksittäiset tutkijat jne.

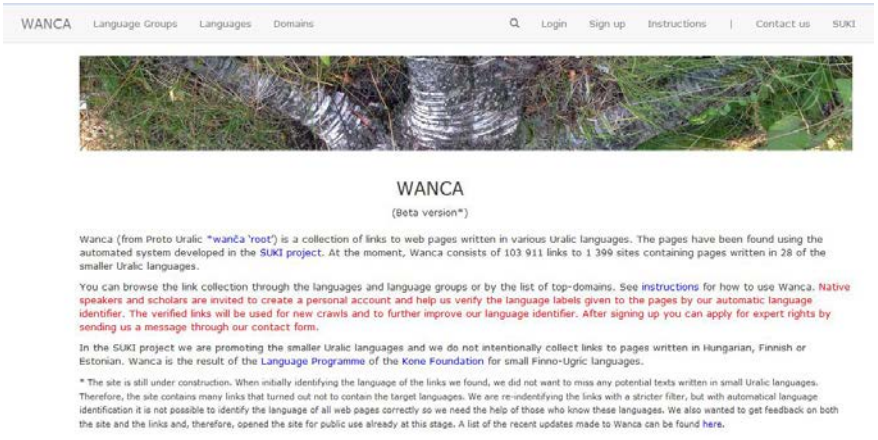
5) Aineistojen mobilisointi



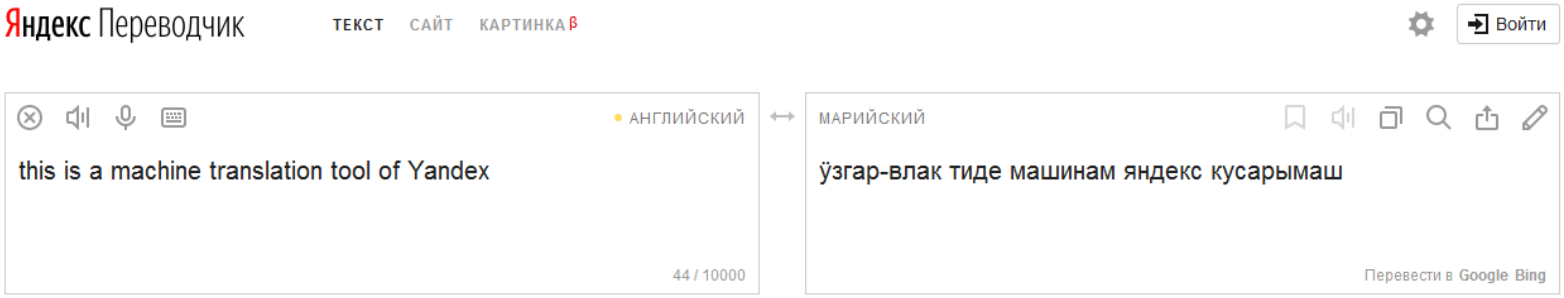
5) Aineistojen mobilisointi



5) Aineistojen mobilisointi



5) Aineistojen mobilisointi



http://lingvodoc.ispras.ru/dictionary/425/1/perspective/425/1/view

5) Aineistojen mobilisointi

- Taidekollektiivi Liiketilan MAA-performanssi.

Miten käyttäjät hyötyvät aineistoista?

- Odotetaan, että digitoiduilla aineistoilla on valtavat vaikutukset sekä yhteiskunnalle että tutkimukselle, mutta kukaan ei tiedä mikä se vaikutus on ja kuinka arvokas se on.

- Simon Tannerin mukaan numerot eivät ole ensinkään kovin tärkeitä, vaan **mahdollisuuksien luominen**:

Kun digitaalisia aineistoja ja niiden rikastamiseen tarkoitettuja välineitä käytetään, käynnistyy myös prosessi, jossa eri yhteisöjen ulottuville syntyy aikaisempaa laajempia mahdollisuuksia.

Lopuksi

- Kansalliskirjaston Digitaalisten Ihmistieteiden politiikka: www.kansalliskirjasto.fi/fi/aineistot/kirjaston-politiikat#digital-humanities--politiikka
- Kansalliskirjaston datakatalogi: http://data.nationallibrary.fi/
- Fenno-Ugrica: http://fennougrica.kansalliskirjasto.fi/
- Simon Tanner: Measuring the Impact of Digital Resources
- Lisätietoja: jussi-pekka.hakkarainen@helsinki.fi