

Kansalliskirjaston digitaaliset aineistot ja data

Jussi-Pekka Hakkarainen

Tietoasiantuntija

Erikoiskokoelmat

jussi-pekka.hakkarainen@helsinki.fi



KANSALLISKIRJASTO

Joitain viimeaikaisia kehityslinjoja

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

The image is a collage of various digital library services and search results. It includes:

- Kansalliskirjaston hakupalve**: A search interface with a search bar and filters for 'Tarkennettu haku' and 'Kansainvälisten e-aineistojen haku'.
- Digi.kansalliskirjasto.fi**: A search results page for 'SANOMALEHDET' (Digitally 8 259 133 pages, Available 3 300 240 pages, Indexed 2 228 116 pages) and 'AIKAKAUSLEHDET' (Digitally 6 507 310 pages, Available 3 300 240 pages, Indexed 3 147 070 pages).
- Doria**: A search interface for 'Kansalliskirjasto' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Kansalliskirjasto**: A search results page for 'Yhteisö sisältää alayhteisöt' (Elektra [37177], Kansalaisoita ja itsenäistyminen [1480], Kansalliskirjaston julkaisuarkisto [1139], Kartat [1615], Kirjat [10453], Kortit ja kokoelmaluettelot [321], Kuvat [1513], Kasikirjoitukset [751], Pienpainatteen [8055]).
- Fragmenta MEMO**: A search interface for 'Fragmenta MEMO' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- VARIA**: A search interface for 'VARIA' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Elektroniset vapaakappaleet**: A search results page for 'Elektroniset vapaakappaleet' (Elektroniset vapaakappaleet ovat verkossa julkaisua alne pyydyt säilytettäväksi. Luovuttaminen perustuu lakiin lu [1453207]).
- Kokoelmat**: A search results page for 'Kokoelmat' (Ditoidut julkaisut [843], Vapaakappaleet [7868], Aanitit [54706]).
- Fenno-Ugrica**: A search interface for 'Fenno-Ugrica' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Suomalainen verkko**: A search interface for 'Suomalainen verkko' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Kansalliset kulttuuriaineistot**: A search interface for 'Kansalliset kulttuuriaineistot' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Yutu**: A search interface for 'Yutu' with a search bar and filters for 'Hae Doriasta' and 'Tämä yhteisö'.
- Yhteisö sisältää alayhteisöt**: A search results page for 'Yhteisö sisältää alayhteisöt' (Elektra [37177], Kansalaisoita ja itsenäistyminen [1480], Kansalliskirjaston julkaisuarkisto [1139], Kartat [1615], Kirjat [10453], Kortit ja kokoelmaluettelot [321], Kuvat [1513], Kasikirjoitukset [751], Pienpainatteen [8055]).
- Yhteisö sisältää kokoelmat**: A search results page for 'Yhteisö sisältää kokoelmat' (Lautapelit [21], Raita - musiikkia vanhoilta äänilevyiltä [482], Urho Kekkosen julkaistu tuotanto [4611]).

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 laaturarviota, 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](https://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.
- Poliitikassa 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)

tietomallimäärittelykset

kokoteksti- ja ääniteaineistot

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

Pelkkä verkkosivusto ei ole tietoaineisto!

Avoim 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-ugrialaisten 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-лэн историја : жендодом курс : 3-од да 4-од классјаслы учебник
Julkaistu: Сыктывкар : Коми Государственной Издательство, 1938



Nimeke: Советской Социалистической Республикајас Союзлэн Конституција : Основной Закон
Julkaistu: Сыктывкар : Коми Государственной Издательство, 1938



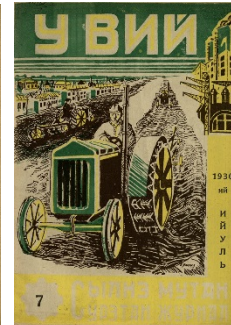
Nimeke: СССР-са народной овмөс развитиялэн коймөд пятилетный план : ВКП (б) XVIII-бд съезд вылын доклад да заключительной кив : март 14-17 луньяс, 1939 eo
Tekijä: Молотов, В.
Julkaistu: Сыктывкар : Коми государственной издательство, 1939



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



Nimeke: Вашкала дуннелэн историез : средней школапэн 5-6 классъёсызлы учебник
Julkaistu: Ижевск : Удмуртгосиздат, 1943



2) Kielaineistojen siirrettävyys

Suomeksi In English По-русски

Fenno-Ugrica etusivu > Monografiat > Kirjat > Näytä viite

Hae Hakuohjeet

Tämä kokoelma Hae Fenno-Ugricasta

G+ Jaa 0 Tweet

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](#)



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

[Avaa tiedosto](#)

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

Datapaketti, sis.
AltoXML, CSV,
TIFF, TXT

Käyttökopiona
PDF

Selaa kokoelmaa

- [Nimekkeet](#)
- [Tekijät](#)
- [Julkaisuajat](#)
- [Asiasanat](#)
- [Uusimmat](#)
- [Selaa kielen mukaan](#)
- [Julkaisutyyppi](#)
- [Sivukartta](#)

Omat tiedot

- [Kirjauudus](#)
- [Profiili](#)
- [Tallennukset](#)

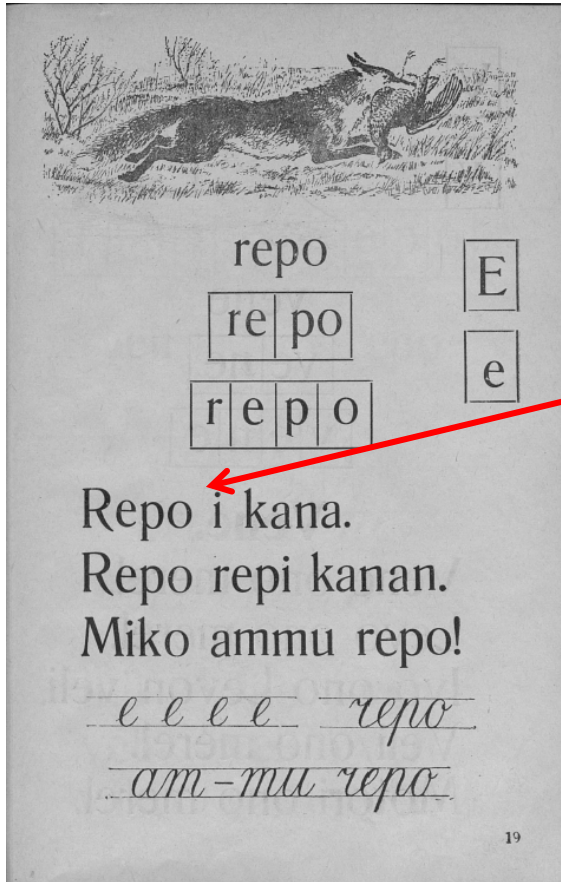
Toiminnot

- [Muokkaa tietuetta](#)
- [Eksportoi tietue](#)
- [Eksportoi metadata](#)

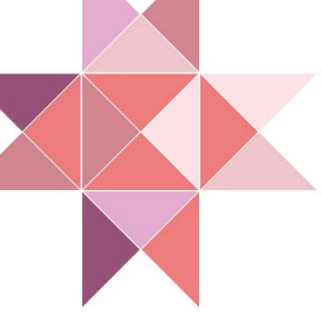
Hallinnointi

- 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>

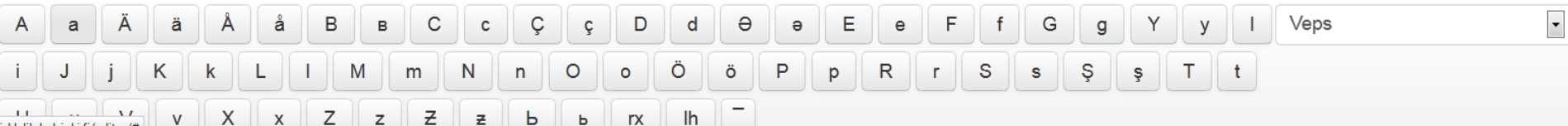
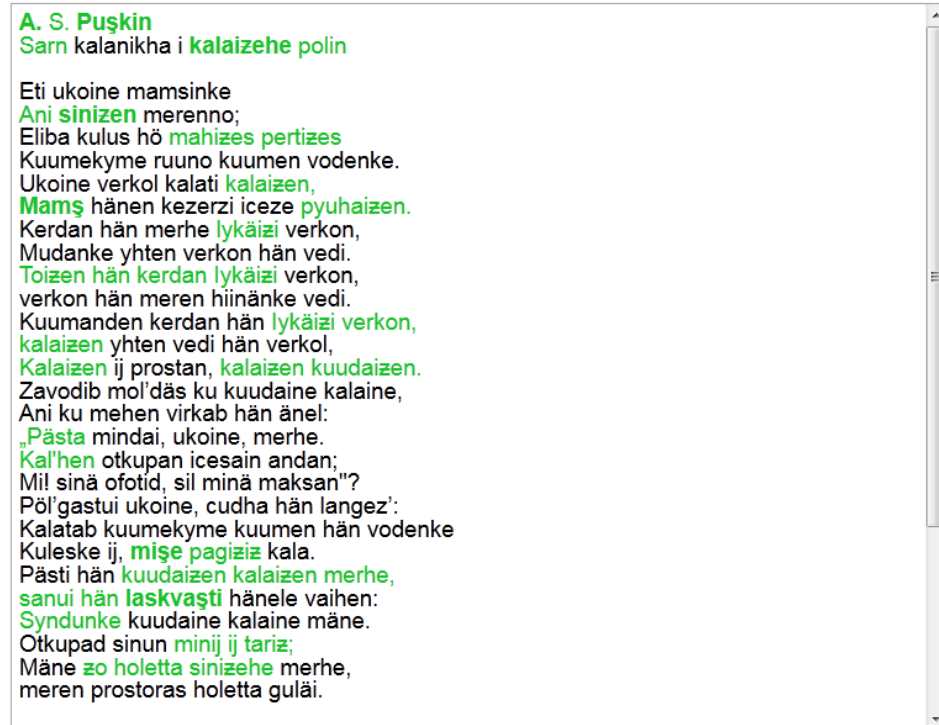
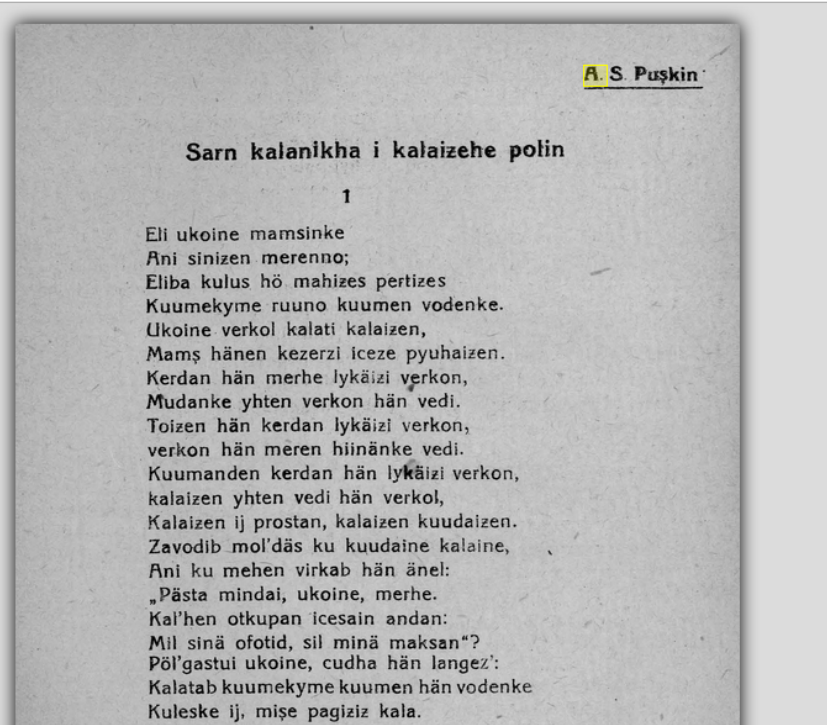


3) Avoin saavutettavuus

- Fenno-Ugrican aineistot ovat avoimesti saavutettavissa.
- Aineistot PAS-kelpoisia, eli säilytettävyyšnäkökulma 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 digitoitiprojektissa 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:
bröd
- 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.

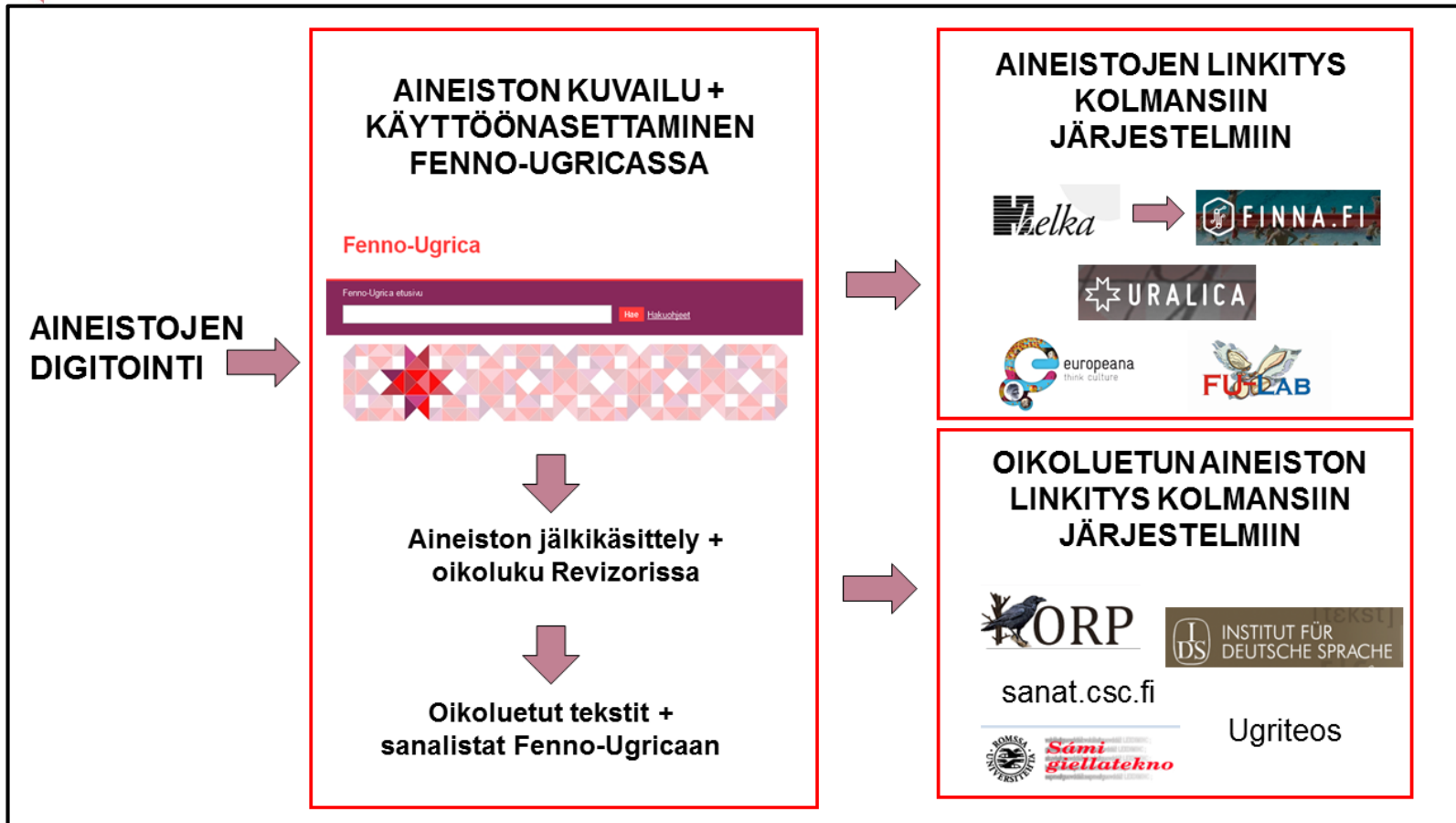
- Tampereen yliopiston Vanha kirjasuomen kurssi oli menestys.



4) Lisäarvon tuottaminen

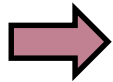
- Sanalistat julkaistiin .csv-formaatissa Fenno-Ugricassa, mistä ne ovat sellaisinaan saatavilla.
- Ei aivan korpuslingvistiikkaan kelpaavaa laatua, mutta hyvää kuitenkin.
 - Ei analyysiä.
 - Henkilöstö osaamiskapeikossa?
- Pitäisi käsittää pikemminkin **raakadatan** kuin valmiina tuotteena (rajallinen 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 features word 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

UIT The arctic university of Norway > [Giellatekno](#) >



[Ruoktu/Home](#) [Divvun](#) [Dicts](#) [oahpa.no](#) [Samiske språk - Saami Languages](#) [Andre språk - Other Languages](#)

- ▶ Saamelaiskielet
- ▶ Uralilaiset kielet
 - Uralilaiset ja muut kielet
- ▶ Ersä
- ▶ Suomi
- ▶ Länsimari
- ▶ Inkeri
- ▶ Hanti
- ▶ Komi
- ▶ Kveeni
- ▶ Liivi
- ▶ Itämari
- ▶ Mokša
- ▶ Nenetsi
- ▶ Nganasani
- ▶ Livvi
- ▶ Udmurtti
- ▶ Vepsä
- ▶ Muut kielet
 - Uralilaiset ja muut kielet
 - ▶ Burjaatti
 - ▶ Kornj
 - ▶ Fääri
 - ▶ Grönlanti
 - ▶ Iñupiaq
 - ▶ Pohjoishaida
 - ▶ Ojibwe
 - ▶ Plains Cree
 - ▶ Venäjä

Tervetuloa saamen kieliteknologian sivuille

[Davvisámegiili](#) [Norsk](#) [English](#) [Suomeksi](#) [Русский](#)

Note

FinUgRevita tutkii meidän kieliohjelmia. Auta heitä täyttämällä oheinen nettikysely:

SAAMI SURVEY

Resurssit

- **Muut uralilaiset kielet:** [ersä](#), [hanti](#), [inkeroinen](#), [komi](#), [kveeni](#), [liivi](#), [livvi](#), [mokša](#), [nenetsi](#), [nganasani](#), [niittymari](#), [suomi](#), [udmurtti](#), [vepsä](#), [võro](#), [vatja](#), [vuorimari](#).
- **Muut kielet:** [Burjaatti](#), [evenki](#), [fääri](#), [grönlanti](#), [iñupiaq](#), [korni](#), [norja](#), [ojibwe](#), [pohjoishaida](#), [plains cree](#), [venäjä](#).
- **Ordbøker**

Voit lähettää kommentteja ja kysymyksiä osoitteeseen: giellatekno@uit.no

Netidigisõnad [Etusivu](#) [Klikkaa-tekstissä](#) [Meistä](#)

Sanakirjoja

IŽORAN KEEL

Inkeroinen → Suomi (= Vaihda)

Suomi → Inkeroinen

hakusana

Etsi sanaa tai sanamuotoa analysoitavaksi tai määriteltäväksi.

- + LATVIEŠU VALODA
- + LIVŌ KEĻ
- + VADŌĀ ČEELI
- + VÕRO KIIL'

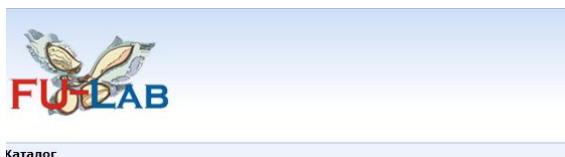
5) Aineistojen mobilisointi



Загрузки Орфосервис Словари Конвертеры ЦИЯТ Лаборатория Бюро перевода Небгаин

Загрузки

Здесь Вы можете загрузить бесплатные программы от команды FU-Lab.



Список словарей

- Большой коми-русский словарь (32 645)
- Большой нарийско-русский словарь (40 953) **обновлено**
- Большой русско-коми словарь (50 714) **обновлено**
- Большой удмуртско-русский словарь (44 628)
- Краткий нарийско-русский словарь неологизмов (1 869)
- Краткий русско-коми словарь компьютерных терминов (1 125)
- Краткий русско-коми словарь общественно-политических терминов (2 134)
- Краткий русско-коми топонимический словарь (47)
- Краткий русско-нарийский словарь делопроизводителя (926)
- Краткий русско-нарийский словарь неологизмов (2 001)
- Краткий русско-удмуртский словарь неологизмов (1 876)
- Краткий удмуртско-русский словарь неологизмов (2 435)
- Малый коми-русский словарь (2011) (9 615)
- Малый нарийско-русский словарь (8 104)
- Малый русско-коми словарь (2011) (12 485)
- Малый русско-нарийский словарь (1999) (14 909)
- Мансийско-русский словарь (3 907)
- Марий синоним мутер (1975) (1 513)
- Марийско-русский словарь неологизмов (5 233)
- Русско-нарийский словарь неологизмов (6 626)

новое - новое за последние 7 дней.
обновлено - обновлено за последние 14 дней.



Раскладки клавиатуры

Раскладки клавиатуры с символами, отражающими специфические буквы национальных алфавитов.



Электронные словари

Офф-лайн версии электронных словарей коми, марийского, удмуртского и алтайского языков.



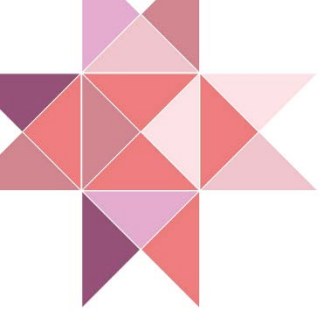
Модуль проверки правописания

Средства проверки орфографии для коми, марийского и удмуртского языков.



Учебные пособия

5) Aineistojen mobilisointi



 www.mari-language.com: ENGLISH | МАРИА | ПО-РУССКИ
Main page » Dictionary

Mari-English Dictionary

[Reset](#) | [Instructions](#) | [Morphological Analyzer](#)

<input type="text"/>	Search	• Mari → English	◦ Mari ← English	◦ Mari ⇄ English
Search:		• Whole Words	◦ All Matches	<input type="checkbox"/> Search Subentries
Stress, Palatalness:		• Extra Characters (*, ')	◦ Bold/Italics	◦ Do Not Annotate
Orthography:		• Cyrillic	◦ UPA	◦ IPA
Ordering:		• Alphabetical	◦ Reverse	
абвгдежзийклмнопрстуфхцшщъыьэюя				

WANCA Language Groups Languages Domains Login Sign up Instructions | Contact us SUKI



WANCA

(Beta version*)

Wanca (from Proto Uralic *wanca 'root') is a collection of links to web pages written in various Uralic languages. The pages have been found using the automated system developed in the SUKI project. At the moment, Wanca consists of 103 911 links to 1 399 sites containing pages written in 28 of the smaller Uralic languages.

You can browse the link collection through the languages and language groups or by the list of top-domains. See [instructions](#) for how to use Wanca. Native speakers and scholars are invited to create a personal account and help us verify the language labels given to the pages by our automatic language identifier. The verified links will be used for new crawls and to further improve our language identifier. After signing up you can apply for expert rights by sending us a message through our contact form.

In the SUKI project we are promoting the smaller Uralic languages and we do not intentionally collect links to pages written in Hungarian, Finnish or Estonian. Wanca is the result of the Language Programme of the Kone Foundation for small Finno-Ugric languages.

* The site is still under construction. When initially identifying the language of the links we found, we did not want to miss any potential texts written in small Uralic languages. Therefore, the site contains many links that turned out not to contain the target languages. We are re-identifying the links with a stricter filter, but with automatic language identification it is not possible to identify the language of all web pages correctly so we need the help of those who know these languages. We also wanted to get feedback on both the site and the links and, therefore, opened the site for public use already at this stage. A list of the recent updates made to Wanca can be found [here](#).

 INSTITUT FÜR DEUTSCHE SPRACHE
Mitglied der Leibniz-Gemeinschaft 

Aktuelles | Organisationsstruktur | Forschung | Onlineangebote | Bibliothek | Service | Publikationen | Über uns

Direktion und zentrale Forschung | Grammatik | Lexik | Pragmatik | Öffentlichkeitsarbeit | FI - Technik | Verwaltung

Startseite :: Organisationsstruktur :: Direktion :: Korpuslinguistik | Korpuslinguistik :: Projekte :: KorAP

Direktion und zentrale Forschung

- :: Forschungsinfrastrukturen
- :: Korpuslinguistik
- :: Projekte
- :: Korpusausbau
- :: Analysemethodik
- :: KorAP
- :: Blog
- :: Recherchesystem
- :: DRuKoLA
- :: Abgeschlossene Projekte
- :: Bibliografie
- :: Sprache im öffentl. Raum
- :: Personal

KorAP - Korpusanalyseplattform der nächsten Generation

Ziele

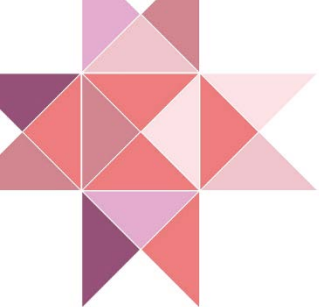
Systematisch zusammengestellte, elektronische Sammlungen von aufgezeichneten Kommunikationsakten, so genannte Korpora, sind mittlerweile die wichtigste empirische Grundlage der Sprachwissenschaft. Sie werden zur Bestätigung oder Widerlegung von Hypothesen verwendet und dienen auch als unmittelbarer Gegenstand explorativer Forschungsarbeiten. Gerade um große Korpora für Sprachwissenschaftler handhabbar zu machen, sind geeignete Werkzeuge unabdingbar, die in der Lage sind, sehr große Datenmengen verlustfrei zu verwalten und rechenintensive Funktionen für ihre methodisch valide Analyse anzubieten.

Mit dem Archiv für Gesprochenes Deutsch (AGD) und dem Deutschen Referenzkorpus (DeReKo) lagern am Institut für Deutsche Sprache die weltweit größten Sammlungen deutscher Sprachdaten. Um insbesondere auf Letzteres zugreifen zu können, wurde am IDS das Corpus Search, Management and Analysis System (COSMAS I und COSMAS II) geschaffen, das sich seit 1991 bzw. 2003 im Dauerbetrieb bewährt hat. Da auch COSMAS II jedoch bereits Anfang der Neunziger Jahre konzipiert wurde und der Arbeitsaufwand, derartige Software zu erweitern, mit steigender Lebensdauer und Komplexität überproportional steigt, wird es zunehmend schwieriger, die Software an die sich rasch wandelnden Bedarfe anzupassen. Indes haben sich sowohl die technischen als auch die wissenschaftlichen Rahmenbedingungen derart stark verändert, dass die Entwicklung eines neuartigen Analyse-Tools erstrebenswert ist.

Ziel des Projektes KorAP ist es somit, eine neuartige Korpusanalyseplattform zu entwickeln, die eine Grundlage für den methodisch validen Umgang mit very large corpora im Bereich der Sprachwissenschaft und insbesondere der empirisch germanistischen Forschung schafft.



KANSALLISKIRJASTO



5) Aineistojen mobilisointi

Яндекс Переводчик

ТЕКСТ САЙТ КАРТИНКА

Войти

АНГЛИЙСКИЙ

this is a machine translation tool of Yandex

44 / 10000

МАРИЙСКИЙ

Ўзгар-влак тиде машинам яндекс кусарымыш

Перевести в Google Bing

ISPRAS About Divisions Education Editions News

Institute for System Programming of the Russian Academy of Sciences



The Institute for System Programming (ISP) of the Russian Academy of Sciences was founded on January 25, 1994, on the base of the departments of System Programming and Numerical Software of the Institute for Cybernetics Problems of the RAS. ISP RAS belongs to the Division of Mathematical Sciences of the RAS.

Our business model has proved its vitality and efficiency for our industrial partners.

One of the reasons is that presence of high skilled developers in different areas of computer science and permanent manpower development allows us to successfully carry out complex industrial projects satisfying varying requirements of our partners.

The Institute employs more than 200 highly qualified researchers and software engineers, including 12 doctors of science and 45 philosophy doctors. Many employees of the Institute also work as professors in leading Russian universities.

[More](#)

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



KANSALLISKIRJASTO



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