Euroopan unioni
Euroopan aluekehitysrahasto

Vipuvoimaa EU:lta 2014–2020

Historiallisten digitoitujen sanoma- ja aikakauslehtien avaaminen avoimena datana tutkijoille

Tuula Pääkkönen, Jukka Kervinen
Tietojärjestelmäasiantuntija
Informaatiotutkimuksen päivät 3.-4.11.2016

Kansalliskirjasto, Helsingin yliopiston erillislaitos

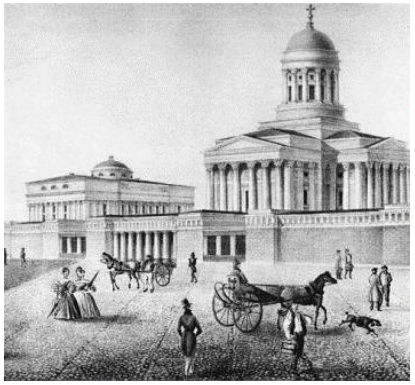
#natlibfi

Digitointi- ja Konservointi-keskus

2015

1990

1828 Keisarillinen Aleksanterin yo



1707 Vapaakappalelaki

Turun akatemia

Akatemiatalo. Akademihuset. The Academy. C.C. Gjörwell. 1815.

1640

Lähde: http://www.kansalliskirjasto.fi/yleistieto/kirjastotietoutta/historia.html

Aineistot digi.kansalliskirjasto.fi :ssä

Sanomalehdet


SUOMI.

Mikkelissä, Keskiviikkona 28 p. Marraskuuta.

N:o 0. 1888.

SUOMI,

KOULUJEN VIRSIKIRJA,

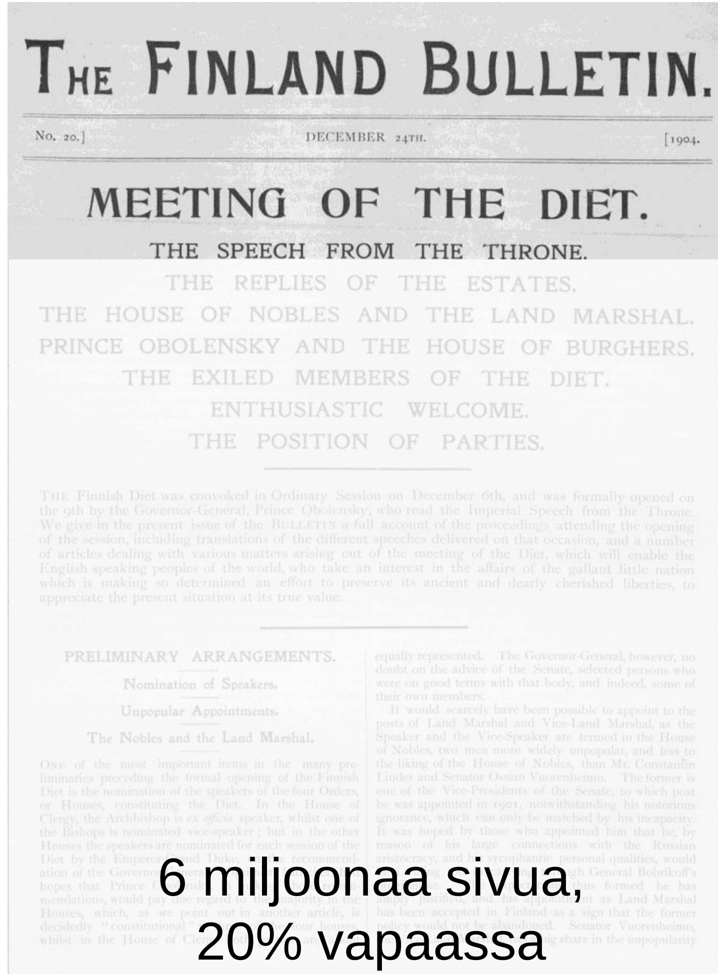
A. Westerholm

Lääkäri.

Mikkelin postikonttori

Yli 4 miljoonaa sivua, ~50% vapaassa verkkokäytössä

Aikakauslehdet


THE FINLAND BULLETIN.

No. 20.] DECEMBER 24TH. [1904.

MEETING OF THE DIET.

THE SPEECH FROM THE THRONE.

THE REPLIES OF THE ESTATES.

THE HOUSE OF NOBLES AND THE LAND MARSHAL.

PRINCE OBOLENSKY AND THE HOUSE OF BURGHERS.

THE EXILED MEMBERS OF THE DIET.

ENTHUSIASTIC WELCOME.

THE POSITION OF PARTIES.

THE Finnish Diet was convoked in Ordinary Session on December 6th, and was formally opened on the 9th by the Governor-General, Prince Obolensky, who read the Imperial Speech from the Throne. We give in the present issue of the BULLETIN a full account of the proceedings attending the opening of the session, including translations of the different speeches delivered on that occasion, and a number of articles dealing with various matters arising out of the meeting of the Diet, which will enable the English speaking peoples of the world, who take an interest in the affairs of the gallant little nation which is making so determined an effort to preserve its ancient and dearly cherished liberties, to appreciate the present situation at its true value.

PRELIMINARY ARRANGEMENTS.

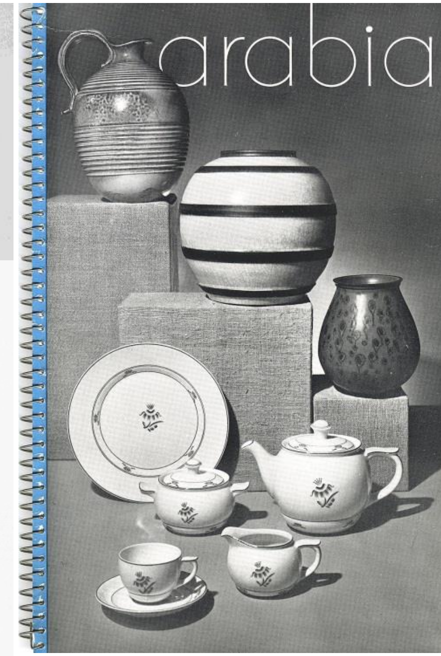
Nomination of Speakers.

Unpopular Appointments.

The Nobles and the Land Marshal.

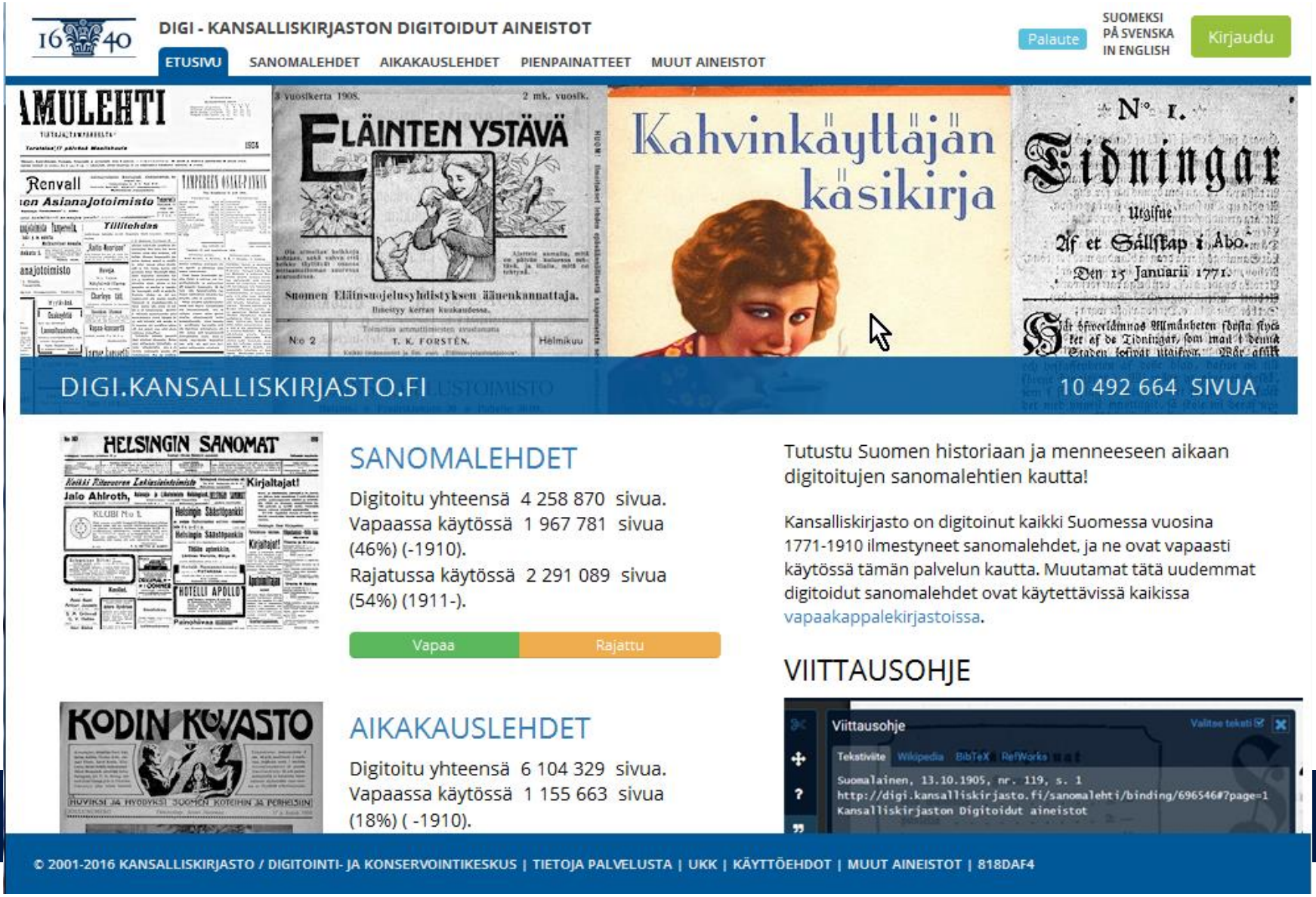
6 miljoonaa sivua, 20% vapaassa verkkokäytössä

Pienpainatteet



130.000 sivua, 100%

digi.kansalliskirjasto.fi



Avoin Kansalliskirjasto

Konteksti - Kehitysprojektit

	Projektin kutsumanimi	Päätarkoitus	Kotisivu
Digitalia Digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskus	Digitalia	Tekstinlouhinta, menetelmäkehitys (tekstinkorjaus (OCR), NER)	http://www.digitalia.fi/
SUOMEN AKATEMI FINLANDS AKADEM ACADEMY OF FINL	Comhis	Uutisten siirtyminen lehdestä toiseen , TY, NatlibFi, HY	https://wiki.helsinki.fi/display/Comhis/
Vipuvoimaa EU:lta 2014–2020 Euroopan unioni Euroopan aluekehitysrahasto	Aviisi	Tekijänoikeusten alaista aineistoja laajempaan käyttöön piloteilla	http://blogs.helsinki.fi/digiaviisi/

Digitaalinen ketju

1011001010101
1011001010101
1011001010101

Aineistojen vastaanotto/palautus

Mikrokuvaus/ digitoinnin valmistelu/ konservointi

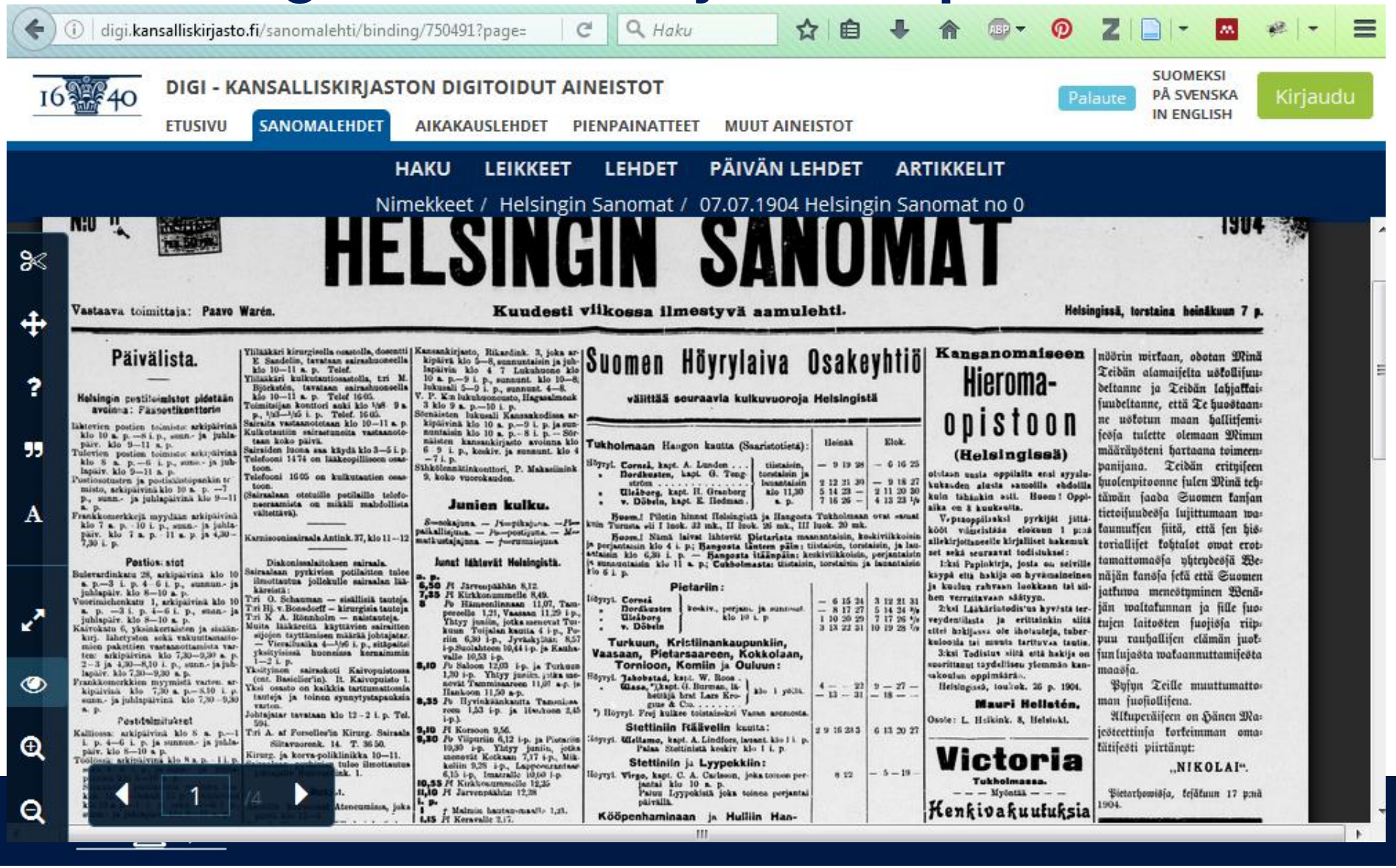
Skannaus

Jälkikäsittely: rakenteellinen analyysi

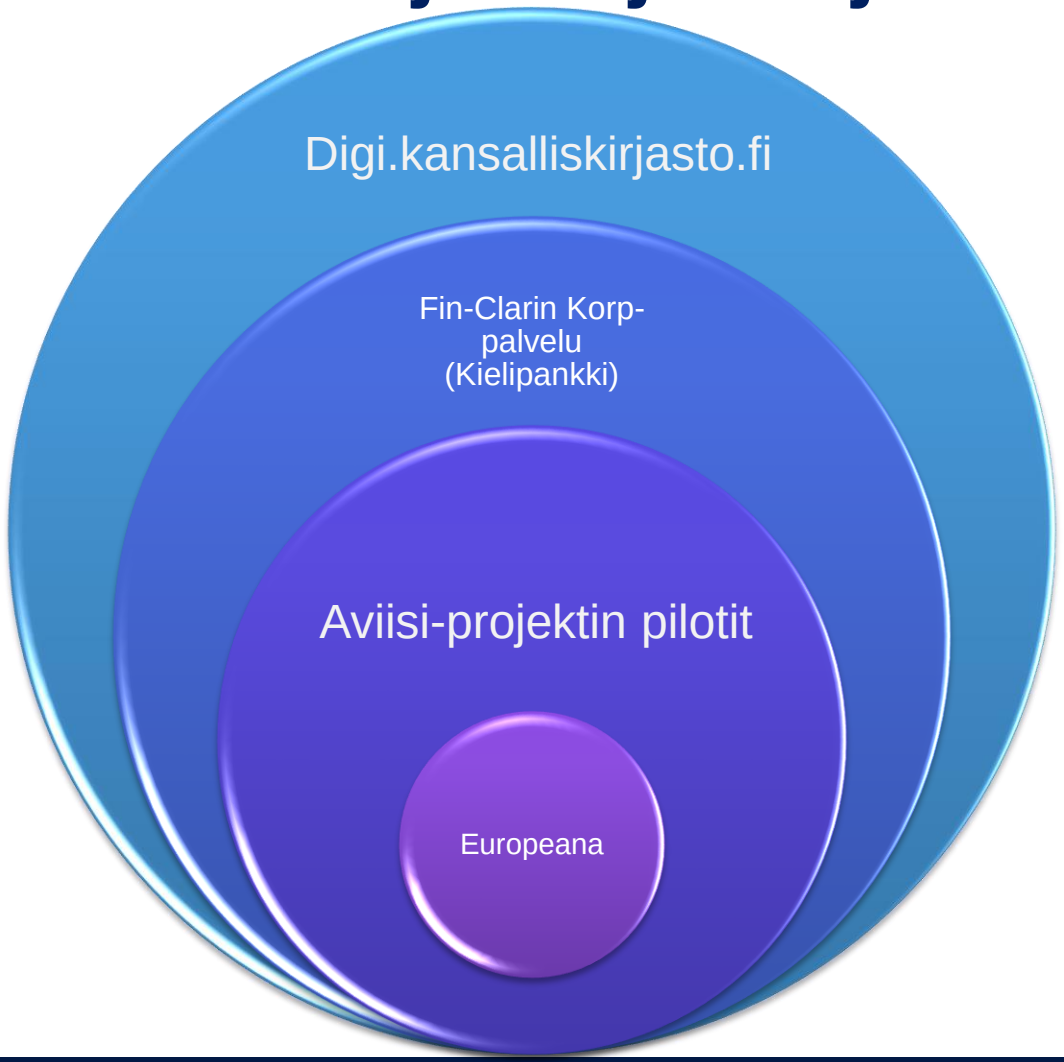
Käyttöönsaattaminen ja säilytys

ALTO XML
METS XML

Sivu digi.kansalliskirjasto.fi -palvelussa



Aineistojen sijainteja



Digi: yli 10 miljoonaa sivua, ~30% saatavilla yleisessä verkkopalvelussa. 70% vapaakappalekirjastossa.

Fin-Clarin: suomen- ja ruotsinkielisten sanomalehtien korpus

Aviisi: Länsi-Savo ja Maaseudun Tulevaisuus vuodet 1913–2016 (n. 1M sivua)

Europeana: Valikoima kansallisia ja alueellisia sanomalehtiä (n. 100.000 sivua)

Miksi aineistopaketti?

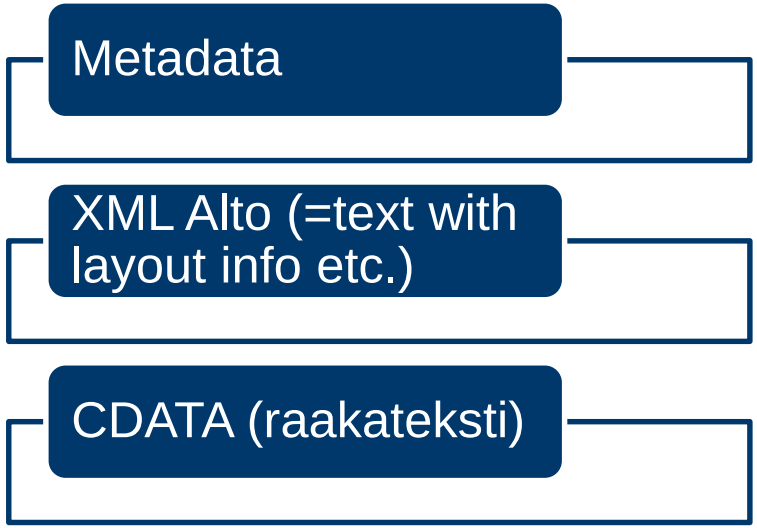
- Tapa tarjota koko aineisto kätevästi tutkijoille, esimerkiksi digitaaliseen humanismin käyttöön
- 1. protopaketit käytössä COMHIS-projektissa, jossa tutkitaan mm. uutisten siirtymistä lehdestä toiseen
- Lisäksi käyttöä Digital Humanities Hackathonissa (DH-kurssin loppuprojektissa)

Aineistopakettien synty

Aineistopaketti sanomalehdistä

- Sanomalehtien ja aikakauslehtien sivut -1910 asti

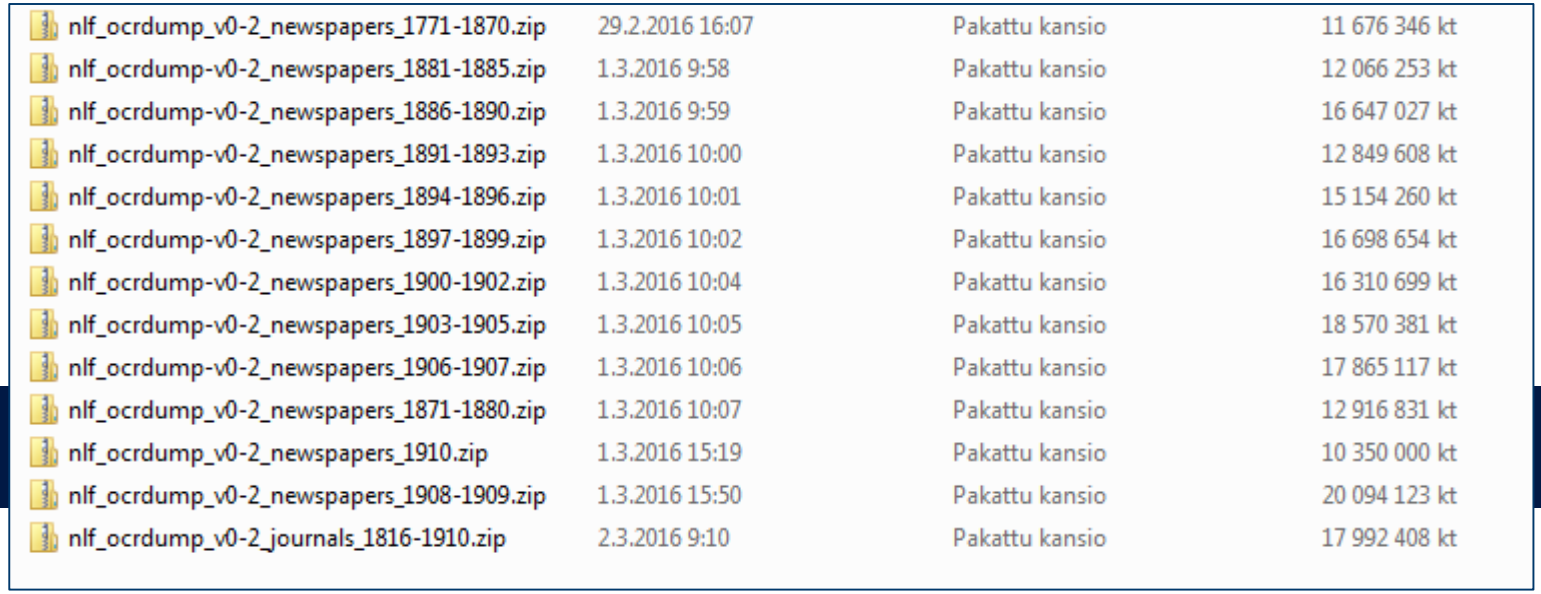
- 1 XML tiedosto per sivu



Lähde: Kimmo Kettunen, Informaatiotutkimuksen päivät 2016

Perusprosessi

- Tehtiin työkalu, joka poimii arkistopaketeista ALTO-tiedostot ja purkaa ne kansiorakenteeseen
  - SAN/by_year/1771-1870/fin/1775/1457-4683_1775-09-01_0_001.xml …
  - Vuosiväli/kieli/vuosi/ISSN_ILMESTYMISPVM_NRO_SIVUNRO



nlf_ocrdump_v0-2_newspapers_1771-1870.zip	29.2.2016 16:07	Pakattu kansio	11 676 346 kt
nlf_ocrdump-v0-2_newspapers_1881-1885.zip	1.3.2016 9:58	Pakattu kansio	12 066 253 kt
nlf_ocrdump-v0-2_newspapers_1886-1890.zip	1.3.2016 9:59	Pakattu kansio	16 647 027 kt
nlf_ocrdump-v0-2_newspapers_1891-1893.zip	1.3.2016 10:00	Pakattu kansio	12 849 608 kt
nlf_ocrdump-v0-2_newspapers_1894-1896.zip	1.3.2016 10:01	Pakattu kansio	15 154 260 kt
nlf_ocrdump-v0-2_newspapers_1897-1899.zip	1.3.2016 10:02	Pakattu kansio	16 698 654 kt
nlf_ocrdump-v0-2_newspapers_1900-1902.zip	1.3.2016 10:04	Pakattu kansio	16 310 699 kt
nlf_ocrdump-v0-2_newspapers_1903-1905.zip	1.3.2016 10:05	Pakattu kansio	18 570 381 kt
nlf_ocrdump-v0-2_newspapers_1906-1907.zip	1.3.2016 10:06	Pakattu kansio	17 865 117 kt
nlf_ocrdump_v0-2_newspapers_1871-1880.zip	1.3.2016 10:07	Pakattu kansio	12 916 831 kt
nlf_ocrdump_v0-2_newspapers_1910.zip	1.3.2016 15:19	Pakattu kansio	10 350 000 kt
nlf_ocrdump_v0-2_newspapers_1908-1909.zip	1.3.2016 15:50	Pakattu kansio	20 094 123 kt
nlf_ocrdump_v0-2_journals_1816-1910.zip	2.3.2016 9:10	Pakattu kansio	17 992 408 kt

Aineistopaketin XML - ylätaso

pageOCRData

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <pageOCRData xmlns="kk-ocr" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instan
3      <metadata> [22 lines]
26     <content version="1" lastModified="2009-09-16T00:09:00">
27         <altoXML> [4583 lines]
4611       <text conversion="digi-importer">
4612           <![CDATA[NtH I (25)
4613   Lehti ilmestyy joka arki maanantaina, kpsH
4614   vastaanotetaan lehden konttorissa. Hinta 25 p piel
4615   Pitempiaikaisista ilmotuksista Kuuri hinnanalennus. -4
```

Sivukohtainen metadata

```
<?xml version="1.0" encoding="UTF-8"?>
<pageOCRData xmlns="kk-ocr" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instan
    <metadata>
        <title>Suur-Savo</title>
        <identifier type="issn">1458-8528</identifier>
        <published format="edtf">1906-03-02</published>
        <issue>1</issue>
        <pageOrder>1</pageOrder>
        <pageLabel></pageLabel>
        <language>fin</language>
        <contentType>newspaper</contentType>
        <originalPublisher/>
        <latestPublisher>Suur-Savon Kustannus-Osakeyhtiö</latestPublisher>
        <publishingPlace country="fi">Mikkeli</publishingPlace>
        <copyright>Copyrights expired</copyright>
        <license>ESIVERSIO - EI EDELLEENJAKELUUN</license>
        <pageIdentifier>3544459</pageIdentifier>
        <bindingIdentifier>703916</bindingIdentifier>
        <imageURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/703916/
        <pdfURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/703916/pd
        <browseURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/703916
        <pingURL/>
        <ocrVersion/>
```

Aineistopaketin ALTO-XML

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <pageOCRData xmlns="kk-ocr" xmlns:xsi="http://www.w3.org/2001/XM
3      <metadata> [22 lines]
26     <content version="1" lastModified="2009-09-16T00:09:00">
27        <altoXML>
28        <alto xsi:noNamespaceSchemaLocation="http://schema.kansa
29     <Description>
30        <MeasurementUnit>mm10</MeasurementUnit>
31        <sourceImageInformation>
32           <fileName>//192.168.10.35/dwin/sanomalehdet/suur_sav
33        </sourceImageInformation>
34        <OCRProcessing ID="OCRPROCESSING_1">
```

ALTO XML:n saa myös digistä

ALTO XML -esimerkki

Tämän XML-dokumentin mukana ei näytä olevan mitään tyyli- tai muotoilutietoa. Dokumentin hierarkia-puu on alla.

```
−<alto xsi:noNamespaceSchemaLocation="http://schema.kansalliskirjasto.fi/alto/alto-1-2.xsd">
  +<Description></Description>
  −<Styles>
      <TextStyle ID="TXT_0" FONTSIZE="10" FONTFAMILY="Times New Roman" FONTSTYLE="bold italics"/>
      <TextStyle ID="TXT_1" FONTSIZE="8" FONTFAMILY="Times New Roman"/>
      <TextStyle ID="TXT_2" FONTSIZE="10" FONTFAMILY="Times New Roman"/>
      <TextStyle ID="TXT_3" FONTSIZE="10" FONTFAMILY="Times New Roman" FONTSTYLE="italics"/>
      <TextStyle ID="TXT_4" FONTSIZE="13" FONTFAMILY="Courier New"/>
      <ParagraphStyle ID="PAR_CENTER" ALIGN="Center"/>
      <ParagraphStyle ID="PAR_LEFT" ALIGN="Left"/>
      <ParagraphStyle ID="PAR_BLOCK" ALIGN="Block"/>
      <ParagraphStyle ID="PAR_RIGHT" ALIGN="Right"/>
    </Styles>
  −<Layout>
    −<Page ID="P4" PHYSICAL_IMG_NR="4" HEIGHT="2206" WIDTH="1479" PRINTED_IMG_NR="4" PC="0.951">
      −<TopMargin ID="P4_TM00001" HPOS="0" VPOS="0" WIDTH="1479" HEIGHT="183">
        −<TextBlock ID="P4_TB00001" HPOS="544" VPOS="147" WIDTH="376" HEIGHT="32" STYLEREFS="TXT_0 PAR_CENTER">
          −<TextLine ID="P4_TL00001" HPOS="544" VPOS="149" WIDTH="376" HEIGHT="30">
            <String ID="P4_ST00001" HPOS="544" VPOS="149" WIDTH="105" HEIGHT="27" CONTENT="Otavan" WC="0.99" CC="310300"/>
            <SP ID="P4_SP00001" HPOS="649" VPOS="176" WIDTH="20"/>
            <String ID="P4_ST00002" HPOS="669" VPOS="149" WIDTH="251" HEIGHT="30" CONTENT="JoulukirjaUisuus" WC="0.78" CC="0174250005600000"/>
          </TextLine>
        </TextBlock>
        −<TextBlock ID="P4_TB00002" HPOS="185" VPOS="156" WIDTH="14" HEIGHT="19" STYLEREFS="TXT_1 PAR_LEFT">
          −<TextLine ID="P4_TL00002" HPOS="185" VPOS="157" WIDTH="14" HEIGHT="18">
            <String ID="P4_ST00003" HPOS="185" VPOS="157" WIDTH="14" HEIGHT="18" CONTENT="4" WC="1.00" CC="9"/>
          </TextLine>
        </TextBlock>
      </TopMargin>
      <LeftMargin ID="P4_LM00001" HPOS="0" VPOS="183" WIDTH="25" HEIGHT="1754"/>
      <RightMargin ID="P4_RM00001" HPOS="1475" VPOS="183" WIDTH="4" HEIGHT="1754"/>
      <BottomMargin ID="P4_BM00001" HPOS="0" VPOS="1937" WIDTH="1479" HEIGHT="269"/>
```

Raakateksti

pageOCRData

```
1    <?xml version="1.0" encoding="UTF-8"?>
2    <pageOCRData xmlns="kk-ocr" xmlns:xsi="http://www.w3.org/2001/XMLSchema-i
3        <metadata> [22 lines]
26       <content version="1" lastModified="2009-09-16T00:09:00">
27           <altoXML>  [4583 lines]
4611         <text conversion="digi-importer">
4612         <![CDATA[NtH I (25)
4613 Lehti ilmestyy joka arki maanantaina, kpsH
4614 vastaanotetaan lehden konttorissa. Hinta 25 p piel
4615 Pitempiaikaisista ilmotuksista Kuuri hinnanalennus. -4
4616 maksua. Vaihtoilmotukset välitetään maammekaikJ
4617 puur-Sduo
4618 I Mikkelissä perjantaina maaliskuun 2 p:nä
4619 liviikkona ja perjantaina k:lo 5 i. p. ? Ilmotuksia Toimittajat: Tilaushi
4620 lukita riviltä tekstin edellä ja 20 p. jälkeen tekstin. Akseli Koponen, v
4621 f- Vieraskieliset ilmotukset suomennetaan ilman eri Tavataan varmimmin kl
4622 Toimisto ja konttori 0. B Blomfeltin talossa. Telef. n:o 209. konttorissa
4623 1906
4624 koko vuosik. 3: 75, V? vuosik 2: 17, 1 kuukuusi 41 p.;
4625 losikerta 1: 80, 1 kuukausi 35 p.; kotiinkanto 75 p. ?
4626 ;sityisnumerot 10 p. ? Lehteä saadaan tilata lehden
4627 ?euduilla postitoimistoista ja asiamiehiltä ympäri lääniä.
4628 uur-Sauo
```

ALTO XML vai teksti?

- Käyttötarkoitus


```
<TextBlock ID="P1_TB00001" HPOS="91" VPOS="4262" WIDTH="907" HEIGHT="91" STYLEREFS="TXT_0 PAR_LEFT">
<TextLine ID="P1_TL00001" HPOS="128" VPOS="4264" WIDTH="870" HEIGHT="40">
<String ID="P1_ST00001" HPOS="128" VPOS="4270" WIDTH="24" HEIGHT="34" CONTENT="")" WC="0.63" CC="61"/>
<SP ID="P1_SP00001" HPOS="152" VPOS="4304" WIDTH="22"/>
<String ID="P1_ST00002" HPOS="174" VPOS="4269" WIDTH="43" HEIGHT="35" CONTENT="Ks." WC="0.95" CC="110"/>
<SP ID="P1_SP00002" HPOS="217" VPOS="4304" WIDTH="28"/>
<String ID="P1_ST00003" HPOS="245" VPOS="4269" WIDTH="181" HEIGHT="29" CONTENT="Wirolaisten" WC="0.90" CC=
<SP ID="P1_SP00003" HPOS="426" VPOS="4304" WIDTH="30"/>
<String ID="P1_ST00004" HPOS="456" VPOS="4266" WIDTH="156" HEIGHT="35" CONTENT="kansallista" WC="0.85" CC="
<SP ID="P1_SP00004" HPOS="612" VPOS="4304" WIDTH="25"/>
<String ID="P1_ST00005" HPOS="637" VPOS="4266" WIDTH="266" HEIGHT="35" CONTENT="sankariiunoclmaa" WC="0.8
<SP ID="P1_SP00005" HPOS="903" VPOS="4304" WIDTH="25"/>
<String ID="P1_ST00006" HPOS="928" VPOS="4264" WIDTH="70" HEIGHT="32" CONTENT="Kal-" WC="0.89" CC="11221"/>
</TextLine>
```


1_1871-07-03_77_001_rawb.txt – Muistio

Tiedosto Muokkaa Muotoile Näytä Ohje

') Ks. Wirolaisten kansallista sankariiunoclmaa ?Kal-
ewi pocg".
Mutta lähtekäämme näistä manhoista kirkoista,
jotka kaikki kyllä ansaitsemat huomiota, ihmettelemään
Tuomiopäätä eli boomia, tuota kortcaaatclis
pesää. Waikka huoneiden rakennustapa
tässä ylipäänsä osoittaa mähäls »nyöhempää aikaa,
tuin itse »vanhassa kaupungissa, on doomi
kuitenkin »vanhin historiallinen paikka. Tähän yhdistyy
muistoja kaikista niistä eri ajoista jolloin
eri kansat omat kaupunkia ja maata hallinneet:
maanasuwat Virolaiset, Tanskalaiset, Saksalaiset,
Ruotsalaiset ja Venäläiset. Tuomiopää on korkea
talktimuorenhuippu, joka jyrkillä seinillä kohoaa
kaupungin ja scuduu yli. Wuorihuipun

Käyttöesimerkki: raakateksti tiedostosta

```
python pick_textfromxml.py -i ..\\data\\1457-
4721_1871-07-03_77_001.xml
```

Tai useammasta hakemiston tiedostosta:

```
for %F in (..\data\*.xml) do python pick_textfromxml.py -i
%F -o %~nF_rawb.txt    (windows)
```

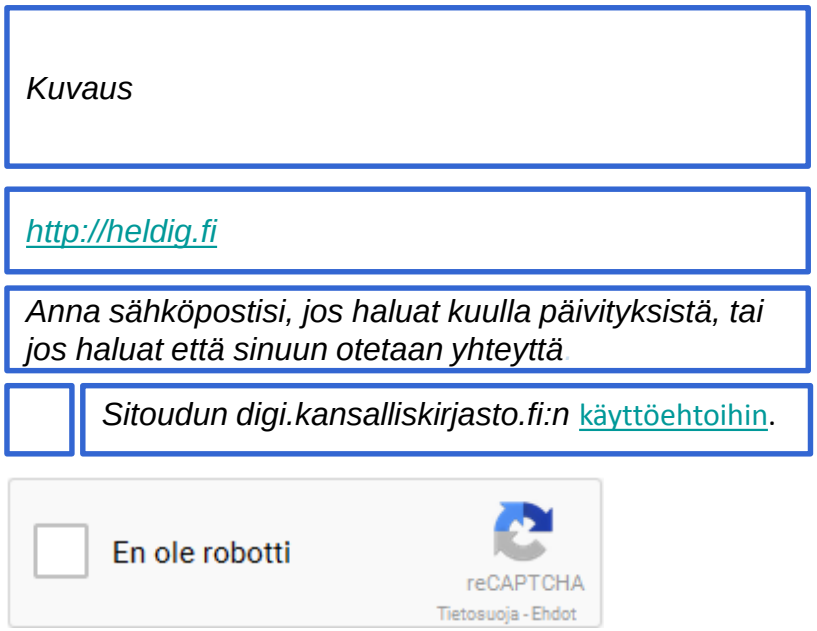
Ym. beta-kehitysversio, lataa itsellesi

- https://github.com/TuulaP/writings/tree/master/src

Digi.kansalliskirjasto.fi / Avoin data

BETA

- Käyttötarkoitus



- Saatavilla olevat aineistot

- Kysely käyttötarkoituksesta (vapaamuotoinen)

Tulossa

Digi.kansalliskirjasto.fi / Avoin data (2)

Saatavilla olevat aineistot

X Sanomalehdet 1771-1910 (ALTO,.zip)

Sanomalehdet (paketin koko)
1771-1870 (12Gb)
1871-1880 (13Gb)
1881-1885 (12Gb)
1886-1890 (16Gb)
1891-1893 (12Gb)
1894-1896 (15Gb)
1897-1899 (16Gb)
1900-1902 (16Gb)
1903-1905 (18Gb)
1906-1907 (17Gb)
1908-1909 (20Gb)
1910 (10Gb)

X Aikakauslehdet 1816-1910 asti (ALTO,.zip)

Lähetä tiedot

- Paketit ladattavissa verkkosivulta
- Latauksen jälkeen saat linkit joko sähköpostiin ja/tai sivulta
- Mahdollisuus lisätä uusia aineistoja samalle sivulle

Tulossa

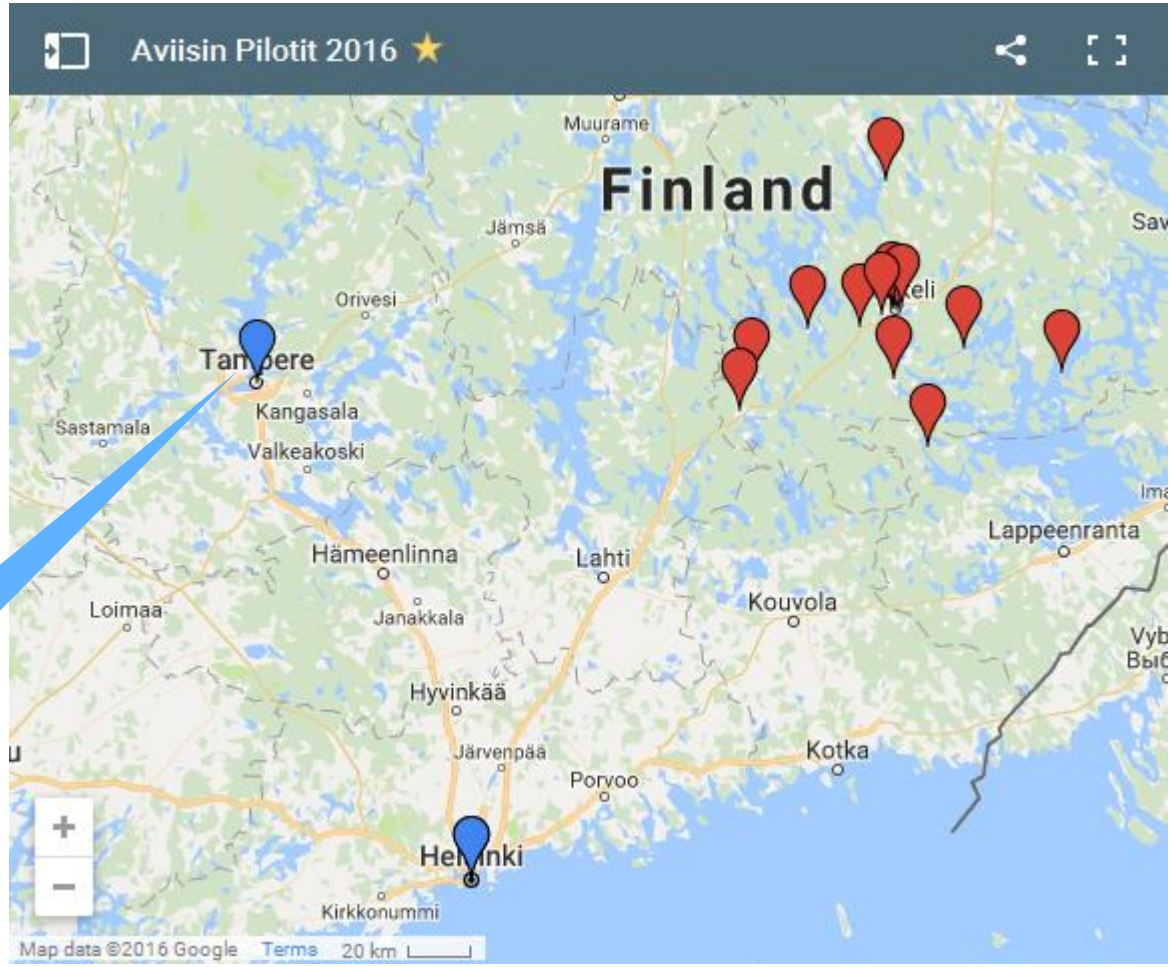
Mitä seuraavaksi?

- Kansalliskirjaston tutkijapalveluita ja aineistoja kehitetään edelleen.

- Ota yhteyttä jos sinulla ideoita tai tarvitset jotakin tiettyä aineistoa!
  - Digi.kansalliskirjasto.fi -> Palaute

- Digi.kansalliskirjasto.fi/avoindata sivu on tulossa…

Kiitos!



AVIISI-pilotti :
Tampereen yliopisto, Yhteiskunta-ja kulttuuritieteiden yksikkö, Kansanperinteen arkisto / http://www.uta.fi/yky/tutkimus/kansanperinne.html

Tuula.Paakkonen@helsinki.fi

Aiheesta lisää

- Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., & Mäkelä, E. (2016). Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. *D-Lib Magazine*, *22*(7/8). http://doi.org/10.1045/july2016-paakkonen