Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen

Kimmo Kettunen, Kansalliskirjasto
Tuula Pääkkönen, Kansalliskirjasto
Mika Koistinen, Kansalliskirjasto

Informaatiotutkimuksen päivät 2016

3.11.2016

Teemat

- Aineistot yleisesti
- Aineistojen käyttö ja käyttäjät
- Aineistojen laatu ja laadun parantaminen

Aineistot

Erilaista vanhaa tekstiaineistoa on digitoitu usean vuosikymmenen ajan ja materiaalia on saatavilla eri kielillä huomattavia määriä. Digitoitujen vanhojen tekstien kokonaismäärien arvioiminen on vaikeaa, mutta muutaman esimerkin mainitseminen riittää kertomaan aineistojen laajuudesta:

- **Europeana-projekti** arvioi vuonna 2012, että digitoituja vanhoja sanomalehtiä oli Euroopan tasolla olemassa noin 24 000 nimikettä ja yli 128 miljoonaa sivua (Dunning, 2012).

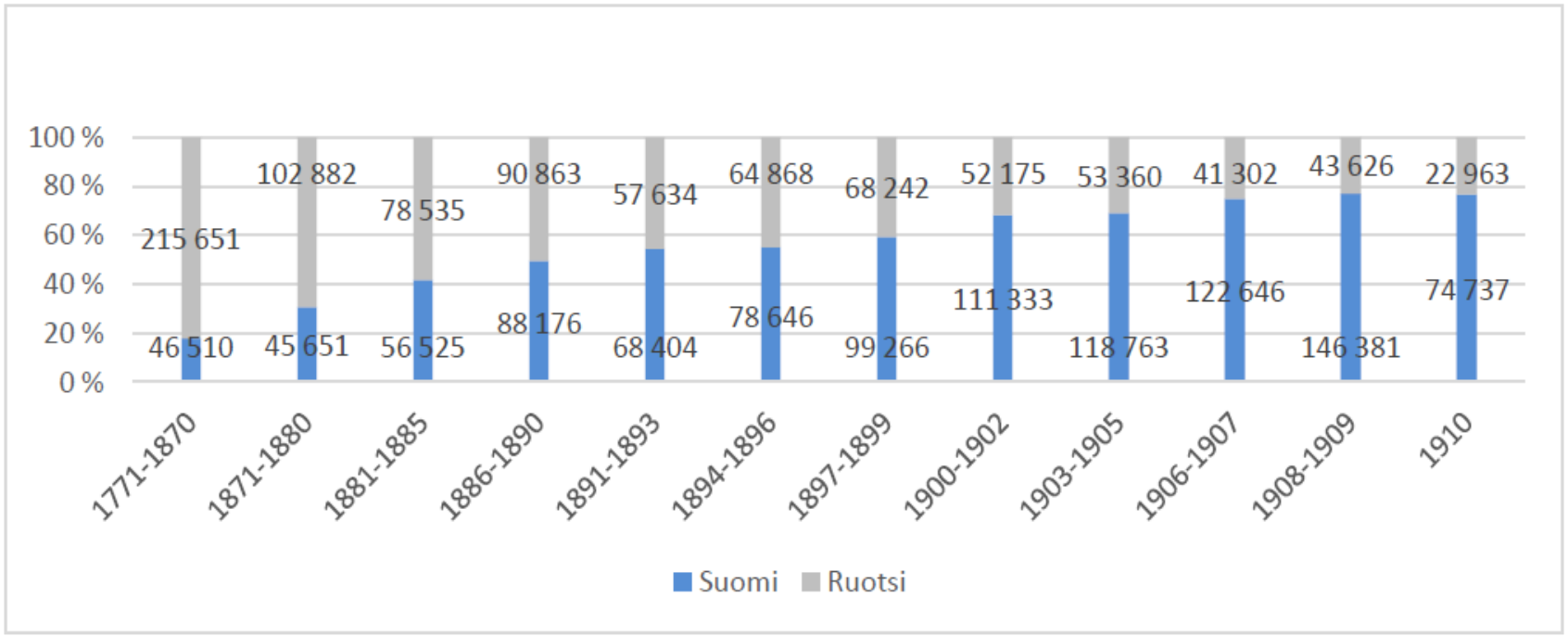
Aineistot

- British Libraryn 1800-luvun lehtikokoelmassa 19th Century British Library Newspapers Database on noin 3 miljoonaa sivua ja 70 nimikettä.
- Gutenberg-projekti ilmoittaa, että heillä on saatavilla yli 50 000 digitoitua kirjaa.
- Yhdysvaltalaisella Hathitrustilla on digitoituna 14 558 573 nidosta, joissa on 5 095 500 550 sivua.
- Kansallisarkisto ilmoittaa digitoitujen aineistojensa määräksi vuoden 2016 keväällä 42 244 873.

Aineistot

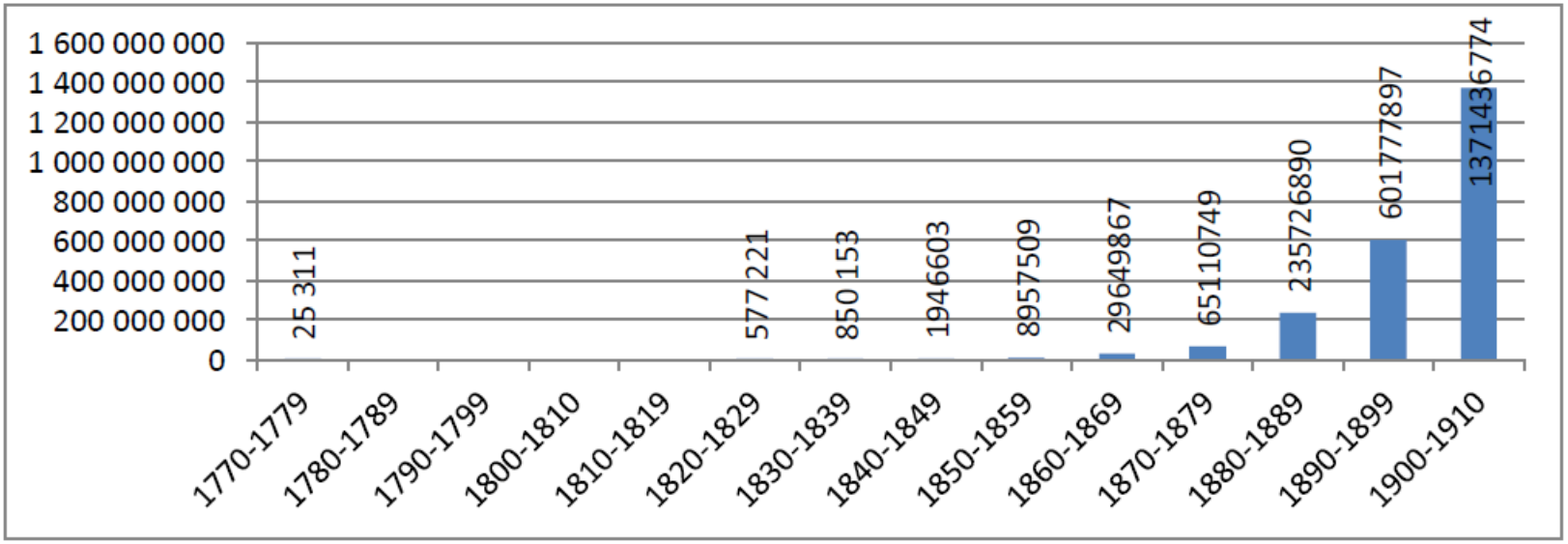
- **Kansalliskirjasto** ilmoittaa verkkosivuillaan (digi.kansalliskirjasto.fi) digitoitujen sivujen kokonaismääräksi vuoden 2016 lokakuussa 10 442 560 . Luku sisältää sanoma- ja aikauslehdet sekä pienpainatteet.
- Koko lehtiaineistosta vuodet 1771–1910 ovat vapaasti käytettävissä, ja niissä on yhteensä noin 3 miljoonaa sivua. Sanomalehtiä on 445 nimikettä, aikakauslehtiä 3141.

Suomea ja ruotsia pääasiassa



Kuva 1. Suomen ja ruotsin osuus sanomalehtiaineistossa sivuina

Suomenkielisen aineiston sanamääräinen jakautuminen ajallisesti

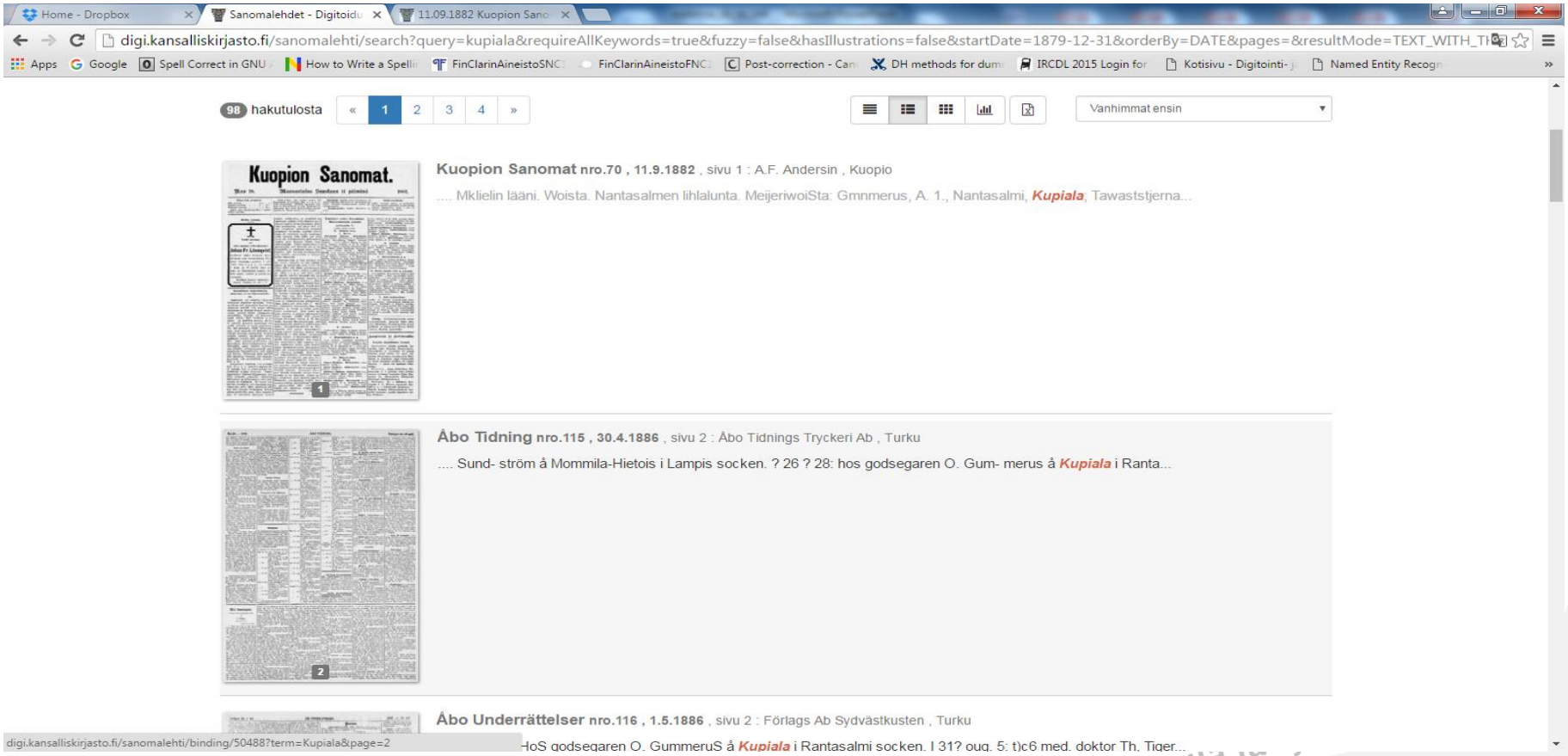


Kuva 2. Lehtiaineiston sanamäärät vuosikymmenittäin; vuosina 1780–1819 julkaistiin vain ruotsinkielisiä lehtiä

Aineistot

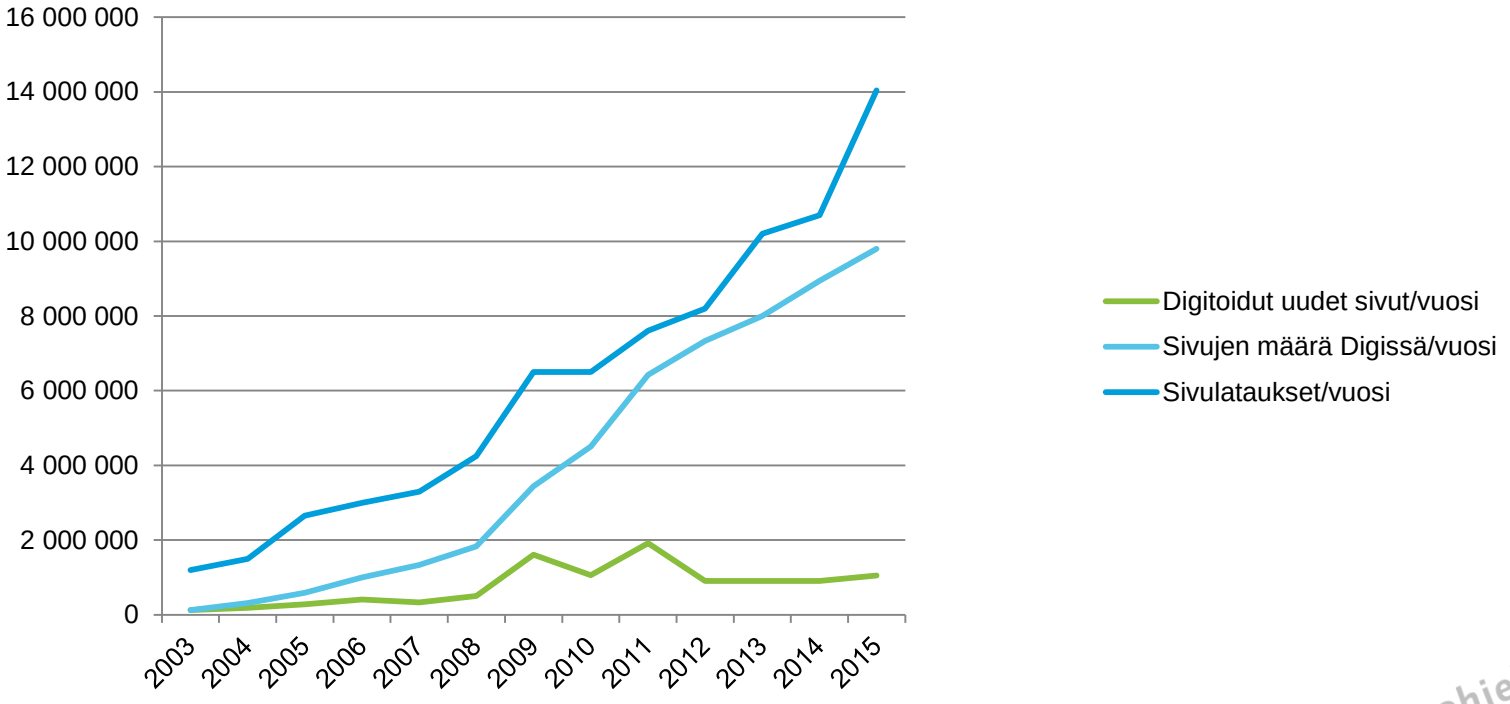
- Digi, kansalliskirjaston digitaaliset kokoelmat

Aineistot: Digi



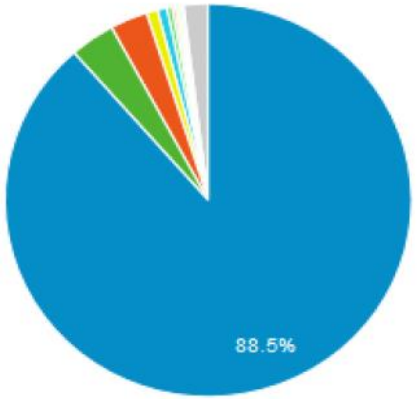
Aineistot: Digi

Digin tilastoja



Google-tilastot 1.1.2014-8.5.2015, top 10 maata

1.	Finland	88.51%
2.	Sweden	3.59%
3.	United States	2.98%
4.	Russia	0.89%
5.	Germany	0.74%
6.	United Kingdom	0.42%
7.	Åland Islands	0.25%
8.	Spain	0.24%
9.	Norway	0.22%
10.	Canada	0.20%

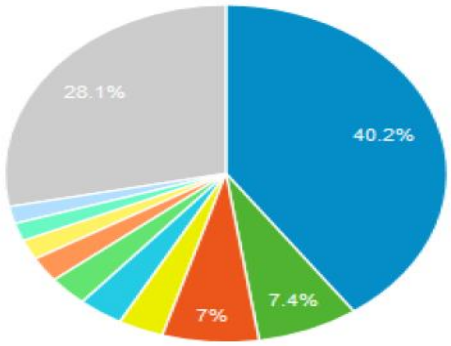


1.1.2014-8.5.2015

Käyttäjätilasto Suomen osalta

In Finland

1.	Helsinki	40.19%
2.	Tampere	7.42%
3.	Turku	7.03%
4.	Oulu	3.31%
5.	Espoo	3.30%
6.	Jyvaskyla sub-region	2.90%
7.	Mikkeli Province	2.40%
8.	Vantaa	1.92%
9.	Joensuu	1.73%
10.	Lahti	1.65%



Hölttä, 2016:

https://tampub.uta.fi/handle/10024/98714

Tutkinut Kansalliskirjaston ja Kansallisarkiston digitoitujen aineistojen käyttöä, kevät 2015, 112

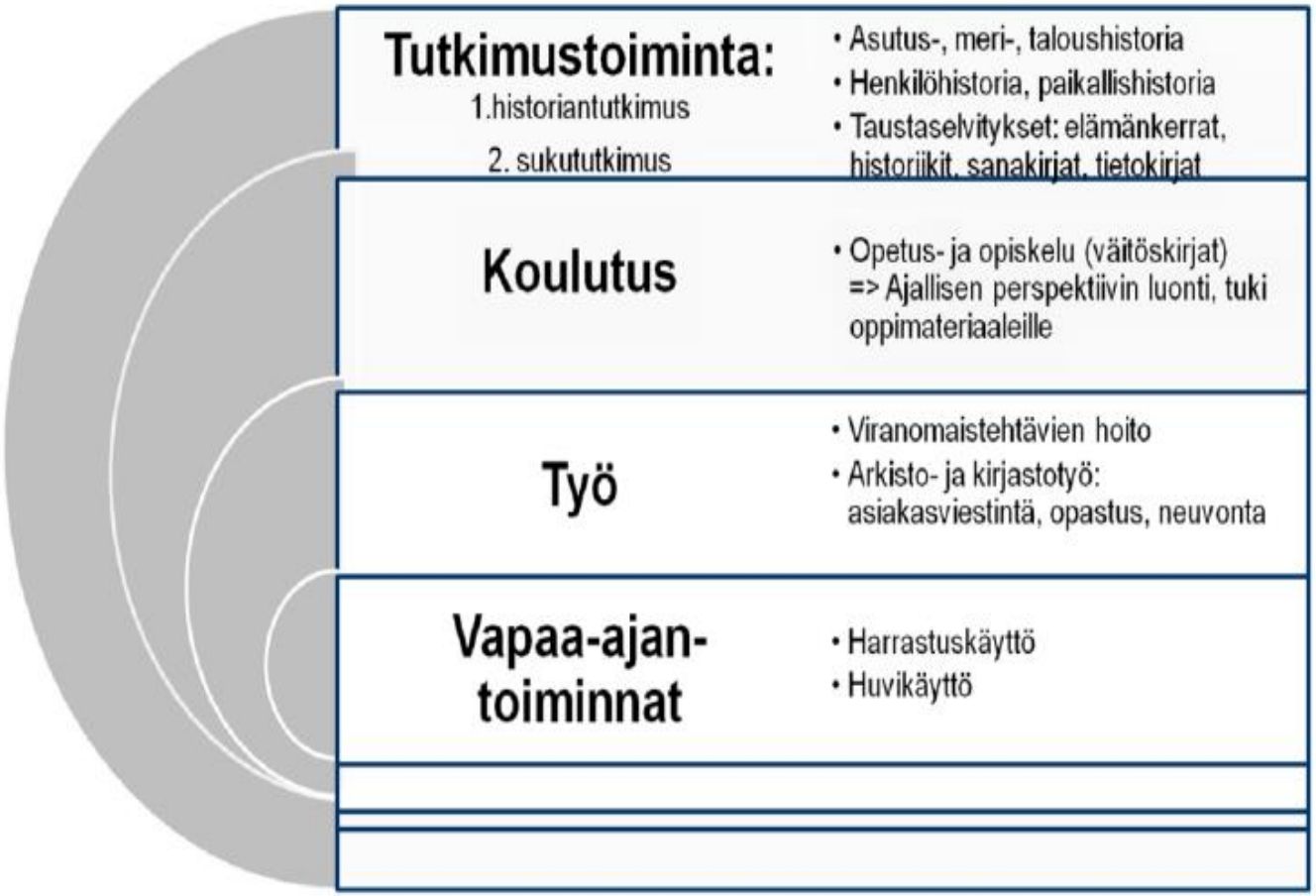
Taulukko 4. Digin aineistolataukset v. 2011–2014 (Pääkkönen 2015.)

DIGIn käyttötilastot v. 2011–2014	Uniikit (vierailijat) kävijät	%:a v. 2011 – 2014	Käyntien (vierailujen) määrä	Käyntiä/ kävijä	%:a v. 2011 –2014	Avattujen kuvatiedostojen (sivujen) latausmäärät	Sivuja/ vierailu	%:a v.2011 –2014
Vuosi 2011	141 878	25	256 131	1,81	25	7 605 671	29,69	22
Vuosi 2012	150 908	26	275 092	1,82	27	8 178 474	29,73	24
Vuosi 2013	162 766	28	290 985	1,79	28	10 218 309	35,12	29
Vuosi 2014	122 477	21	211 934	1,73	20	8 854 402	41,78	25
Yhteensä	**578 029**	**100**	**1 034 142**		**100**	**34 856 856**		**100**
Keskiarvot v. 2011–2014	144 507		258 536	1,79		8 714 214	33,71	

Hölttä, 2016

Historiallinen sanomalehtikirjasto oli toiseksi suosituin verkkopalvelu lähes kaikissa vastaajaryhmissä (kaikki vastausryhmät: akateemiset – historia, ja akateemiset – perinne, kulttuuri ja museologia) Tutkijan ääni -kyselyn tuloksissa. Sukututkijoille ja historian harrastajille historiallinen sanomalehtikirjasto oli neljänneksi käytetyin verkkopalvelu ja Doria-julkaisuarkisto kuudenneksi tärkein. Akateemisille historiantutkijoille kaksi tärkeintä palvelua olivat KA:n Digitaaliarkisto ja KK:n historiallinen sanomalehtikirjasto. (Hupaniittu 2012, 26–27.) Merkittävä huomio kyselyn tuloksissa oli hyvin käytännönläheinen. Työekonomia sanelee rajoituksia akateemisten tutkijoiden tekemälle tutkimukselle: laajoja sanomalehtiaineistoja on mahdollista läpikäydä ainoastaan digitointien ansiosta. (Hupaniittu 2012, 39.)

Hölttä, 2016



Kuva 4. Käytön tyypit ja osa-alueet

Aineiston laatu: tekstit

- Digitoitujen historiallisten aineistojen tunnettu ongelma ovat optisen merkintunnistuksen (OCR) tuomat virheet
- Ongelmat syntyvät monista lähteistä: alkuperäiset painetut aineistot ovat heikkolaatuisia, kuluneita jne. Käytetyt painofontit (erityisesti fraktuura) ovat hankalia tunnistettavia ohjelmille
- Tuloksena on enemmän ja vähemmän virheitä digitoiduissa aineistoissa. Pääsääntönä se, että vanhempi aineisto on vaikeampaa digitoitavaa
- Virheet heikentävät aineiston käytettävyyttä: hakukoneet toimivat heikommin, luettavuus hankalampaa, kaikenlainen jälkikäsittely hankaloituu
- Esimerkki: British Libraryn Century Newspaper Project: sanojen oikeellisuusaste n. 78 % (http://www.dlib.org/dlib/july09/munoz/07munoz.html)

Laatu

- Digin aineistojen skannaus on aloitettu 2000-luvun alkuvuosina
- OCR-ohjelmistojen laatu ei vielä paras mahdollinen
- Lehtien painolaatu ja kunto vaihtelevaa
- Kirjasinlaji fraktuura tuottaa ongelmia
- → OCR:n lopputulos on kirjavaa

- Aamulehti 17.9. 1905

Kuitenkin jää asema nyt aiwan
tohjekſi kuin ſilloin. Japani on tun-
nuſtettu nyt ſuurwallaksi ſen ſijaan,
että Simonoſekiſſa ſitä kohdeltiin
kuin poikamuliikkaa, jonka oli pakko
hellittää towerinſa kaulukſesta. Ja-
pani on nyt ſaanut lujan jalanſijan
Aaſian mantereella Korean ja Port
Arthurin herrana eikä ſen walta jää
ſuinkaan tuntumattomaksi Mandſhu-
riaſſa. Tuhanſittain japanilaiſta lii-
kemieſtä ja teolliſuuden harjoittajia
on ſinne ſodan aikana aſettunut löy
täen ſieltä edulliſemmat waikutus-
piirit kuin liian taajaan aſutuilla
ſaarilla. Mantſuria on ylen rikas
luonnon antimiſta ja ankara innoi-
tus ſen eteläiſeſſä niemen kärjeſſä
muodoſtuu japanilaisten käſiſſä hei
dän mannermaa-waltanſa keskipis-
teekſi, josta pelwon ja kunnioituksen
tunne kaikkialle lewiää.

Virhe-esimerkkejä sanoista, jotka esiintyvät 100 kertaa (IFLA, 2014)

Freq.	OCR form	Correct form	Edit distance	Translation
100	ytsimiclisesti	yksimielisesti	3	unanimously
100	yslämällisesti	ystäwällisesti	2	kindly
100	todistuskappaleilla	todistuskappaleilla	0	with the pieces of evidence
100	peltikattovernissaa	peltikattovernissaa	0	tin roof varnish
100	mastaaminen	wastaaminen	1	answering
100	lyfymylfeen	kysymykseen	4	into the question
100	knstannnksella	kustannuksella	2	at the expense of
100	glasgomista	glasgowista	1	from glasgow
100	annisleluosaleyhtiön	anniskeluosakeyhtiön	2	of the licensed limited liability company
100	amioliitoista	awioliitoista	1	of marriages

Table 5. Examples of word forms that appear one hundred times in the Digi corpus. The correct form is also shown as well as the edit distance between the original and the correct form.

Kerran esiintyvien sanojen virhe-esimerkkejä

Freq.	OCR form	Correct form	Edit distance	Translation
1	zzhdysvautki	yhdyspankki	5	union bank
1	zzznuirypäleitä	wiinirypäleitä	4	grapes
1	wiljelystartaltutsessa	wiljelystarkoituksessa	4	in a cultivation purpose
1	urheilutarloinksiin	urheilutarkoituksiin	3	for sports purposes
1	uratkakupoissa	urakkakupoissa (urakkakaupoissa)	1	in contract jobs
1	taitanuiubcsta	taitawuudesta	4	of dexterity
1	taitamattomundestani	taitamattomuudestani	1	from my ineptitude
1	taiötelelutanteren	taistelutanteren	4	of the battlefield
1	taioafliftiutpn	tavallisuuden	8	of the usual
1	taimokkaisuudcllllln	tarmokkaisuudellaan	6	with his/her vigor

Table 6. An illustrative selection of word forms that appear once in the Digi corpus. The correct form is shown with the edit distance to the original form. One of the original forms in a misspelling in the newspaper.

Joukkoistamalla korjattuja virheitä (n. 65 000 virhettä korjattu, eli noin 0,0001 %)

Vaikeat

Sana Digissä	**Korjaus**
Hmltleii!!.	Aamulehti.
CaMPMclla	Tampereella
KeMluMona	Keskiviikkona
ptthdiLtottua	puhdistettua
la11<ivi8t'in	Tallqvist'in

Helpot:

inyydään	myydään
Htuuwillalankoja	Puuwillalankoja
Wonaiä!!ä.	Wenäjällä,

Digin sanomalehtiaineiston laatuarvio

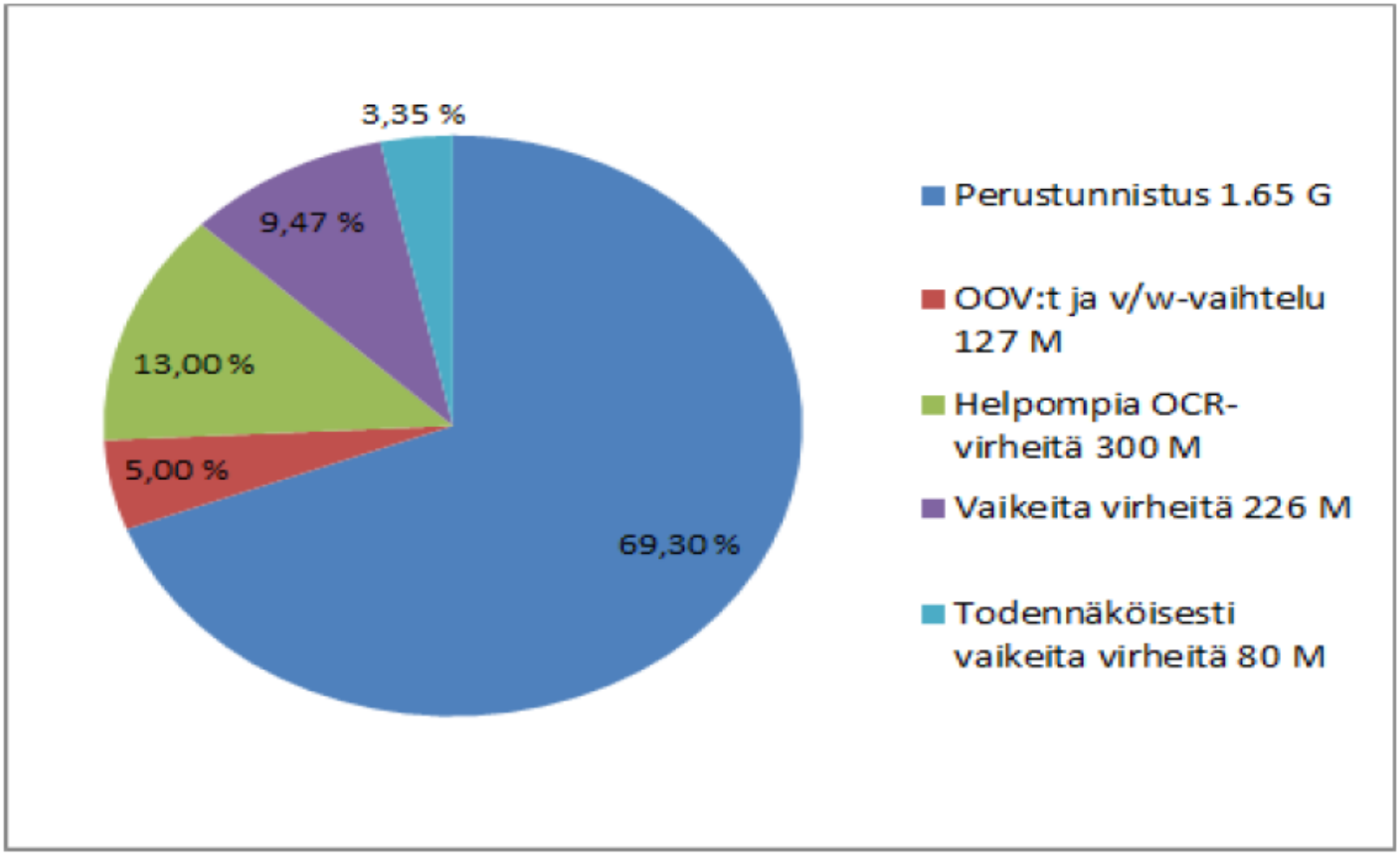
- Suomenkielisessä osuudessa vuosilta 1771-1910 on n. 2.4 G sanoja
- Käytetty kahta nykykielen automaattista morfologista analysaattoria, FINTWOLia ja Omorfia; muita keinoja automatisoituun analyysiin ei oikeastaan ole
- Ohjelmat tunnistavat kohtuullisen hyvin 1800-luvun loppupuolen aineistoa
- Arviossa käytettyä prosessia voidaan käyttää laadun arviointiin sen jälkeen, jos sanakantaan tehdään korjauksia (uudelleen OCR:ääminen ja jälkikorjaaminen)

Tunnistus sanatyypeittäin ja juoksevina sanoina

Kokoelma	Sanamäärä	Omorfi 0.1:n tunnistamia	FINTWOLin tunnista-mia
Digin suomenkieliset sanat vuoteen 1850 sananmuodot	22.8 M	65.6 %	65.2 %
Digin suomenkieliset sanat 1851–1910 sananmuodot	2.385 G	69.3 %	---
Digin suomenkieliset sanat vuoteen 1850 sanatyypit	3.24 M	15.6 %	14.9 %
Digin suomenkieliset sanat 1851–1910 sanatyypit	177.3 M	3.8 %	3.5 %

Taulukko 1. Suomenkielisen sana-aineiston tunnistusprosentit

Tarkennettu arvio tunnistetuista ja tunnistamattomista sanoista



Kuva 5. Digin 1851–1910 sanaston arvioitu laatu

Tunnistusohjelmien muut versiot

Kokoelma	Sanamäärä	Omorfi 0.2	HisOmorfi
Digin suomenkieliset sanat vuoteen 1850 sananmuodot	22.8 M	66.3 %	70.8 %
Digin suomenkieliset sanat 1851–1910 sananmuodot	2.385 G	69.7 %	72.7 %
Digin suomenkieliset sanat vuoteen 1850 sanatyypit	3.24 M	16.0 %	19.4 %
Digin suomenkieliset sanat 1851–1910 sanatyypit	177.3 M	3.9 %	4.9 %

Taulukko 2. Tunnistustulokset Omorfi 0.2:lla ja HisOmorfilla[7]

Tunnistus ≠ sana oikein
Tunnistamattomuus ≠ sana väärin

- mli Num Roman Nom Sg → ilmeinen optisen luvun virhe, tunnistettu silti

- huu huu Part
  tain tai N Gen Sg → sana on jakautunut tavutuksen vuoksi väärin, ja molemmat erilliset osat on tunnistettu (*huutain* olisi käypää 1800-luvun suomea, mutta ei tulisi tunnistetuksi)

- Hei He Pron Nom Pl ?
  dan +? → tavutus on jakanut sanan, pitäisi olla *heidan*, tunnistamaton

- Samoinkuin +?? → kirjoitettu yhteen, jäänyt tunnistamatta
- ylöskannetaan +?? → kirjoitettu yhteen, jäänyt tunnistamatta

Laadun parantaminen

- Mahdollisuuksia on kaksi:
  1. uusi OCR-kierros
  2. Jälkikorjaus

  On tehty työtä Tesseractin opettamisessa fraktuura-fonttiin → on päästy hiukan paremmaksi, mutta ero pieni.

  Jälkikorjauksessa yhteistyötä FIN-CLARINin kanssa → 8-9 prosenttiyksikön parannus saatu

  Lopputulos: Todennäköisesti päästään 80+ prosenttiin sanojen tunnistettavuudessa

Kiitos kärsivällisyydestä