

# 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



Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



# Teemat

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

Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# 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).

Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# 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.

Kestävää kasvua ja työtä -ohjelma



# 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.

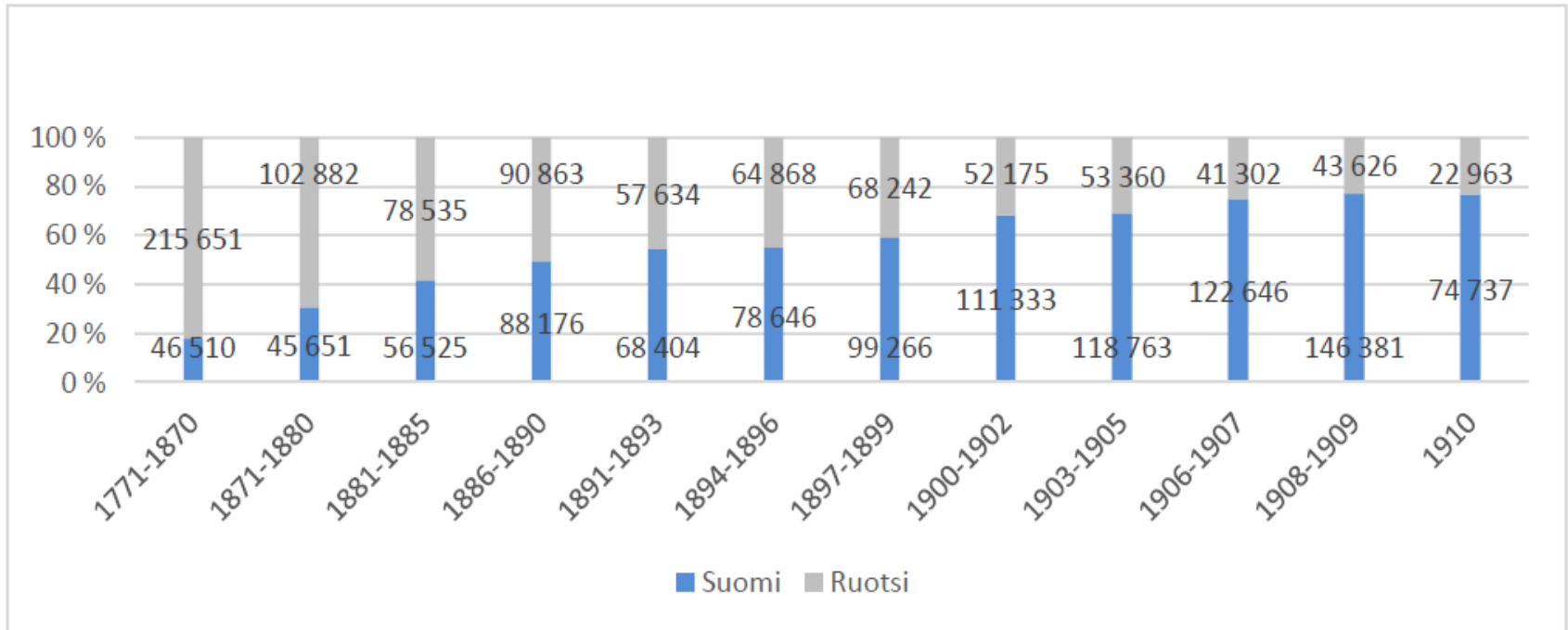
Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



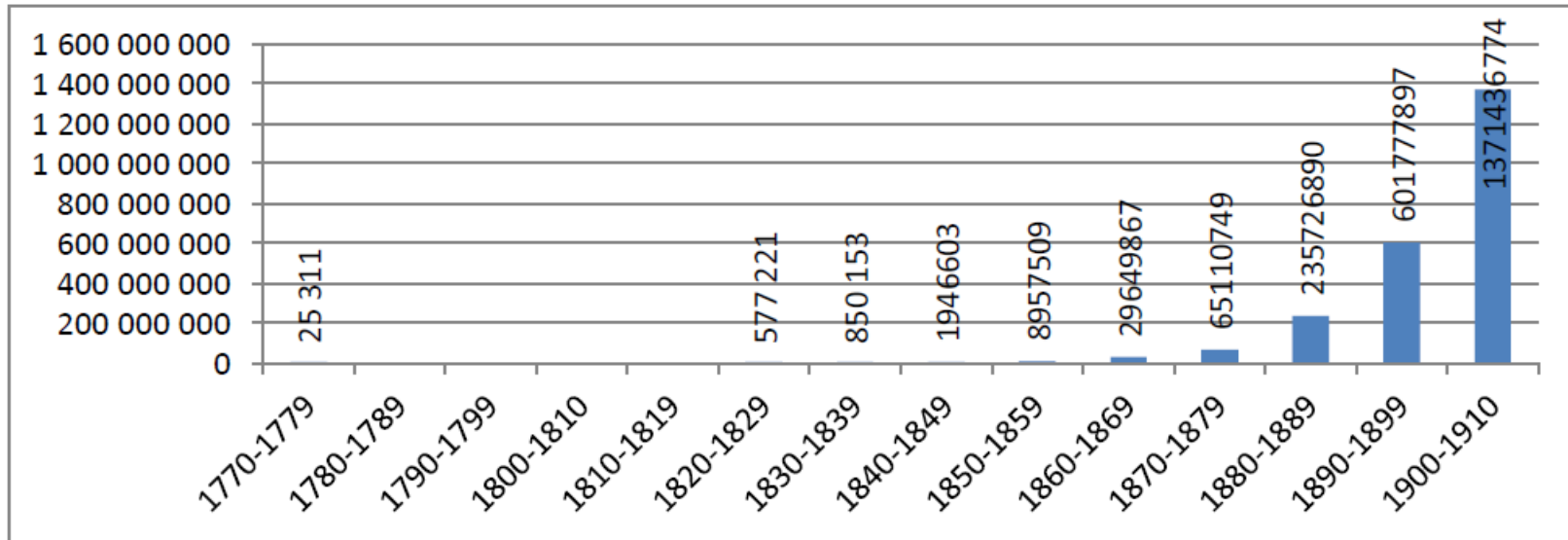
Euroopan unioni  
Euroopan aluekehitysrahasto

# 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

Kestävää kasvua ja työtä

# Aineistot

- Digi, kansalliskirjaston digitaaliset kokoelmat

The screenshot shows the website 'digi.kansalliskirjasto.fi'. The main navigation bar includes 'ETUSIVU', 'SANOMALEHDET', 'AIKAKAUSLEHDET', 'PIENPAINATTEET', and 'MUUT AINEISTOT'. A banner at the top right says 'SUOMEKSI PÅ SVENSKA IN ENGLISH' with 'Kirjautu' and 'Palaute' buttons. A survey notice reads: '06.09.2016: Vastaathan digin käytön kyselyyn. Kiitos! Please respond to survey about digi.' Below this are four featured collections: 'Suur-Savo', 'KAUPPALEHTI SUOMEN LIIKKEISTEN AAMUKANNATTAJA', '1898 NOKIA 1936-37 PÄÄLLYSKENKIEN KUVITETTU HINNASTO', and 'Mikke'. A blue bar at the bottom of the featured items displays 'DIGI.KANSALLISKIRJASTO.FI' and '10 287 538 Sivua'. Below the featured items are three sections: 'SANOMALEHDET' (with a 'Helsingin Sanomat' thumbnail), 'AIKAKAUSLEHDET' (with a 'Kodin Kuvasto' thumbnail), and 'VIITTAUSOHJE' (with a 'Viittausohje' thumbnail). Each section includes a description of digitized and available pages, and a progress bar for 'Vapaa' (green) and 'Rajattu' (orange) pages.

**DIGI.KANSALLISKIRJASTO.FI** 10 287 538 Sivua

**SANOMALEHDET**  
Digitoitu yhteensä 4 104 136 sivua.  
Vapaassa käytössä 1 963 455 sivua (47%) (-1910).  
Rajatussa käytössä 2 140 681 sivua (53%) (1911-).

**AIKAKAUSLEHDET**  
Digitoitu yhteensä 6 053 937 sivua.  
Vapaassa käytössä 1 147 791 sivua (18%) (-1910).  
Rajatussa käytössä 4 906 146 sivua (82%) (1911-).

**VIITTAUSOHJE**



# Aineistot: Digi

The screenshot shows a web browser window with the URL `digikansalliskirjasto.fi/sanomalehti/search?query=kupiala&requireAllKeywords=true&fuzzy=false&hasIllustrations=false&startDate=1879-12-31&orderBy=DATE&pages=&resultMode=TEXT_WITH_T...`. The search results are displayed in a grid. The first result is for **Kuopion Sanomat** nro.70, 11.9.1882, page 1, by A.F. Andersin, Kuopio. The second result is for **Åbo Tidning** nro.115, 30.4.1886, page 2, by Åbo Tidnings Tryckeri Ab, Turku. The third result is for **Åbo Underrättelser** nro.116, 1.5.1886, page 2, by Förlags Ab Sydvästkusten, Turku. Each result includes a thumbnail of the newspaper page and a snippet of text. The word **Kupiala** is highlighted in red in the snippets. At the bottom of the screenshot, there is a URL: `digikansalliskirjasto.fi/sanomalehti/binding/50488?term=Kupiala&page=2`.



# Aineistot: Digi

Home - Dropbox x Sanomalehdet - Digitoidu x 11.09.1882 Kuopion Sanom x 11.09.1882 Kuopion Sanom x

digikansalliskirjasto.fi/sanomalehti/binding/38333?page=1&term=Kupiala

Apps Google Spell Correct in GNU How to Write a Spell FinClarInAineistoSNC FinClarInAineistoFNC Post-correction - Can DH methods for dum IRCDL 2015 Login for Kotisivu - Digitointi-j Named Entity Recogn

16 40 DIGI - KANSALLISKIRJASTON DIGITOIDUT AINEISTOT

ETUSIVU SANOMALEHDET AIKAKAUSLEHDET PIENPAINATTEET MUUT AINEISTOT

SUOMEKSI Kirjautu PÄ SVENSKA IN ENGLISH

HAKU LEIKKEET LEHDET PÄIVÄN LEHDET ARTIKKELIT

Nimekkeet / Kuopion Sanomat / 11.09.1882 Kuopion Sanomat no 70

# Kuopion Sanomat.

N:o 70.

Maanantaina Syyskuun 11 päivänä

1882.

## Ulkous-hinta Kuopiossa:

Koko vuodelta . . . . . 4 m. — p.  
Puolelta vuodelta . . . . . 2 " 50 "  
Neljännes vuodelta . . . . . 1 " 50 "  
Johon tulee kotiinkantopalkkaa 1 markan jälkeen vuodelta.

**Jako-aika:** Lehti jaetaan kunkin arki Maanantaina ja Torstaina kello 2 j. p. p. uudessa kirjapainossa. Yhtäisiä numeroita myydään B. Weurlander'in ja Wiivi Nordbergin kirjakaupoissa kuin myös anniskelu-yhtiön myymäläpuodissa B. Pirisen talossa 10 p. kappale.

**Ilmoituksia** jätetään uuteen kirjapainoon 10 pennin maksusta peit-riviltä eli sitä vastaavasta alalta. Tilauksia vastaan-otetaan Tohtori Nylanderin talossa.  
**Toimituspaikka:** Tohtori Nylander'in talossa.

## Waihto-ilmoituksia

kaikkiin maamme suomen- ja ruotin-kielisiin sanomalehtiin pantawiksi vastaanotetaan uudessa kirjapainossa, naan eli alle 50 pennin pienintäkään ilmoitusta.

## Kuolon sanomia.



Tietää annetaan  
että

rakas puolisoni, tulkivahtimestari

Fr. Lönnqvist.

maisen vaikutuksen, ja etupäässä ne mahtavat ryhmät, joihin Kuopion puolen woinäytteet olivat järjestetyt. Yleisesti myönnettiin, että nämä voit eivät ainoastaan paljoutensa puolesta ansainneet huomiota; myöskin niiden laatu oli ylipäänsä warsin kiitettävä, jonka todistaa sekin seikka, että niin juuri osa korkeimmistakin palkinnoista jaettiin juuri Kuopion läänin näyttelleanijaille. Tässä mainittakoon siunnessä, että Kuopion woi jo ulkomaillaakin on woitannut lauean käyttämisen sekä suuriakin leinakuoneita

## Tähdellinen luettelo Savonlinnan Meijeriinäyttelyssä jaetuista palkinnoista. \*)

(Jatkoa viime n:roon).

### II. Mikkelin lääni.

#### A. Woista.

**Nantafaimen kirkkotalon Meijeriwoista:** Lnen palkinto: Summerus, A. J., Nantafalmi, Kupiala; Tawastijerna, Nathalie, Sääminki, Aholaks. — 2nen palkinto: Scheele, W., Sultawa, Partala; Lindfors, C. P. H., Tutola (2 kapp.); Ungern, Carolina, Nantafalmi, Putkifalo; Etman, Oskar, Nantafalmi, Waaherjalo. — 3mas palkinto: yhteensä 7. Maatiais-

2nen palkinto: Kuisi, Pekka, Kurkijoki, Etjenwaara. — 3mas palkinto: annettiin 2:lle näyttelleanijalle. **Moswientiwosta:** 3mas palkinto: annettiin 1:lle näyttelleanijalle.

**Sortavalan kirkkotalon Meijeriwoista:** 3mas palkinto: jaettiin 1. **Moswientiwosta:** 3mas palkinto: jaettiin 1.

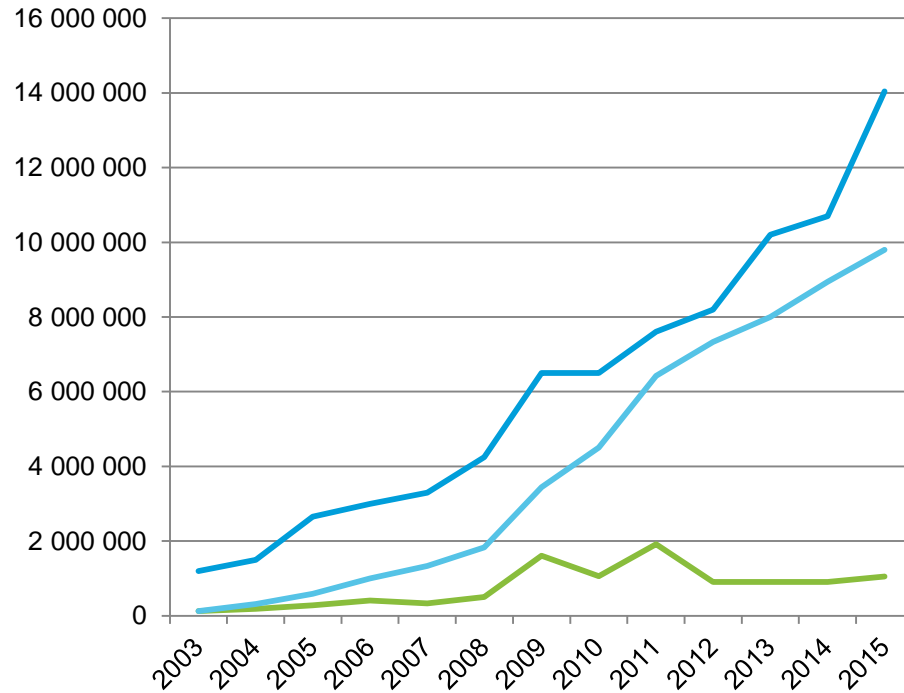
**Salmen kirkkotalon Meijeriwoista:** 2nen palkinto: Pelkonen, Jmpilahti. — 3mas palkinto: jaettiin 1. **Maatiaiswoista:** Lnen palkinto: Hoffrén, G. P., Jmpilahti, Laitiois.

#### B. Juustosta.

2nen palkinto: Graufelt, Arthur, Petäjäjärwi, Sakkola; Kurkijoen maanviljelysopisto. — 3mas palkinto: Kurkijoen maanviljelysopisto: Graufelt Arthur, Petäjäjärwi, Sakkola.



# Digin tilastoja



- Digitoidut uudet sivut/vuosi
- Sivujen määrä Digissä/vuosi
- Sivulataukset/vuosi







Kestävää kasvua ja työtä -ohjelma

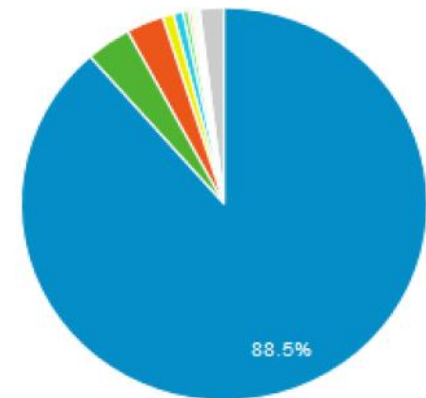
Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# 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%

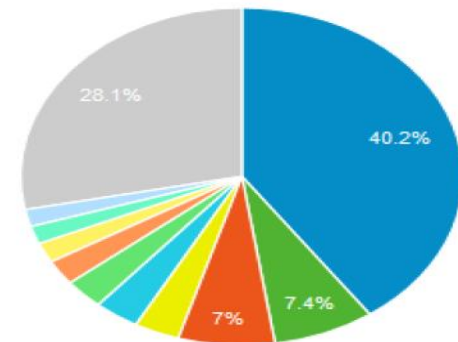


1.1.2014-8.5.2015

# 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.	■ Mikkelin alue	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

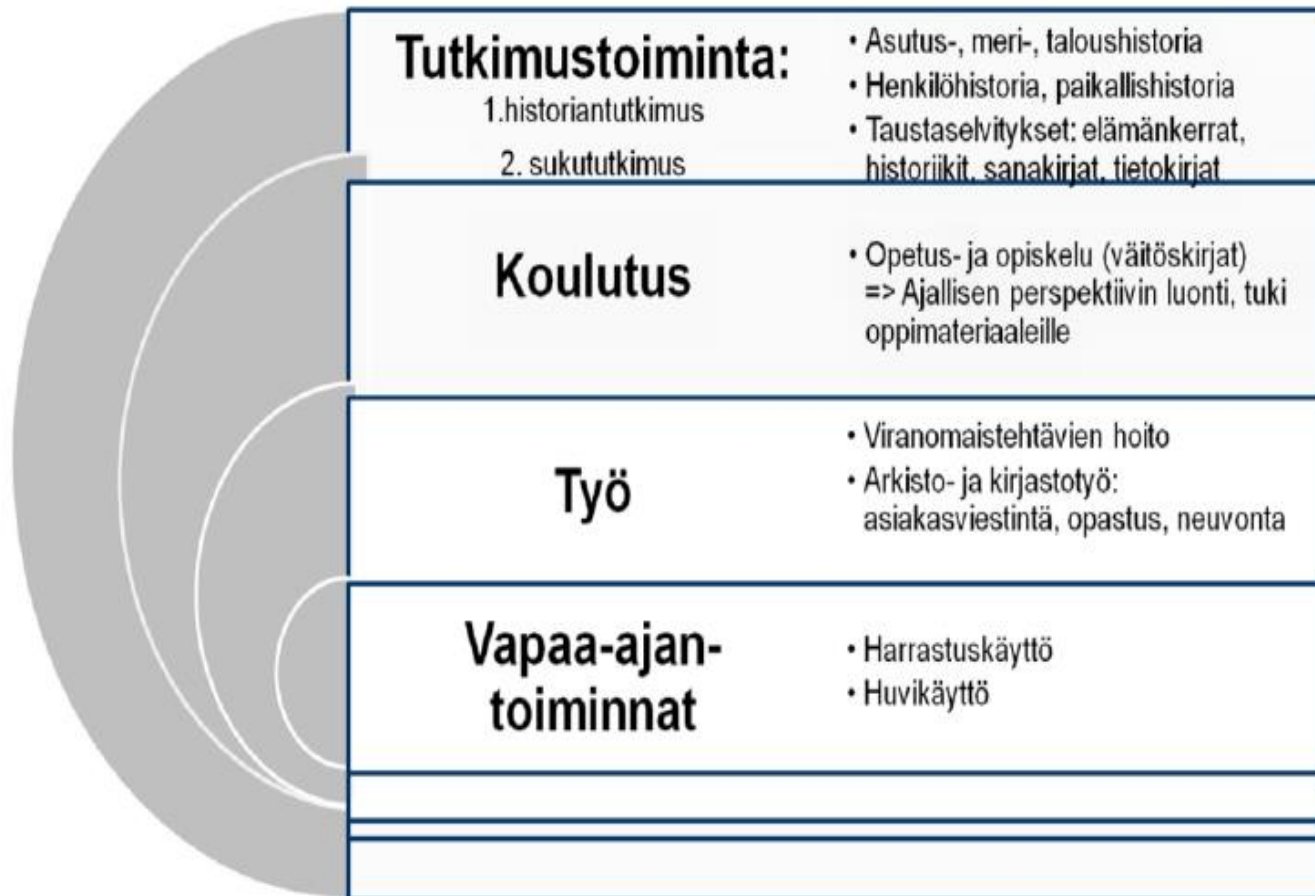
Taulukko 4. Digin aineistolataukset v. 2011–2014 (Pääkkönen 2015.)

DIGIn käyttötilastot v. 2011–2014	Unikit (vierailijat) kävijät	:%a v. 2011 – 2014	Käyntien (vierailujen) määrä	Käyntiä/kävijä	:%a v. 2011 –2014	Avattujen kuvatielosten (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
<b>Yhteensä</b>	<b>578 029</b>	<b>100</b>	<b>1 034 142</b>		<b>100</b>	<b>34 856 856</b>		<b>100</b>
<b>Keskiarvot v. 2011–2014</b>	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 digitoituissa aineistoissa. Pääsääntönä se, että vanhempi aineisto on vaikeampaa digitoitavaa
- Virheet heikentävät aineiston käytettävyyttä: hakukoneet toimivat heikommin, luettavuus hankalampaa, kaikenlainen jälkikäsitteily hankaloituu
- **Esimerkki:** British Libraryn Century Newspaper Project: sanojen oikeellisuusaste n. 78 %  
(<http://www.dlib.org/dlib/july09/munoz/07munoz.html>)

Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# 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 aivan toiseksi kuin silloin. Japani on tunnettu nyt suurella sen sijaan, että Simonofetsja sitä kohdettiin kuin poikamullista, jonka oli pakko heittää toverinsa kemuluksesta. Japani on nyt saanut lujan jalansijan Aasian mantereella Korean ja Port Arthurin herrana eikä sen valta jää suinkaan tontumattomaksi Mandchuriassa. Tuhansittain japanilaisia liisemiehiä ja teollisuuden harjoittajia on sinne sodan aikana asettunut löydäen sieltä edullisemmat vaihtuuspäivät kuin liian taajaan asutuilla saarilla. Maafanta on ylen rikas luonnon antimista ja ankara linnoitus sen eteläisessä niemessä järjessä muodostuu japanilaisten tässä heidän mannermaa-valtansa keskipisteeksi, josta pelvon ja kunnioituksen tunne laittavalle leviää.

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

Freq.	OCR form	Correct form	Edit distance	Translation
100	ytsimielisesti	yksimielisesti	3	unanimously
100	yslämällisesti	ystävällisesti	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	knstannuksella	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	taitanuiubesta	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	taimokkaisuudclllllln	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 is a misspelling in the newspaper.

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

## Vaikeat

### Sana Digissä

Hmitleii!!.

CaMPMcIla

KeMluMona

ptthdiLtottua

la11<ivi8t'in

### Korjaus

Aamulehti.

Tampereella

Keskiviikkona

puhdistettua

Tallqvist'in

## Helpot:

inyydään

Htuuwillalankoja

Wonaiä!!ä.

myydään

Puuwillalankoja

Wenäjällä,

Kestävää kasvua ja työtä -ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# Digin sanomalehtiaineiston laatuarvio

- Suomenkielisessä osuudessa vuosilta 1771-1910 on n. **2.4 G** sanoja
- Käytetty kahta nykykielen automaattista morfologista analysointia, 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)

Kestävää kasvua ja työtä ohjelma

Vipuvoimaa  
EU:lta  
2014–2020



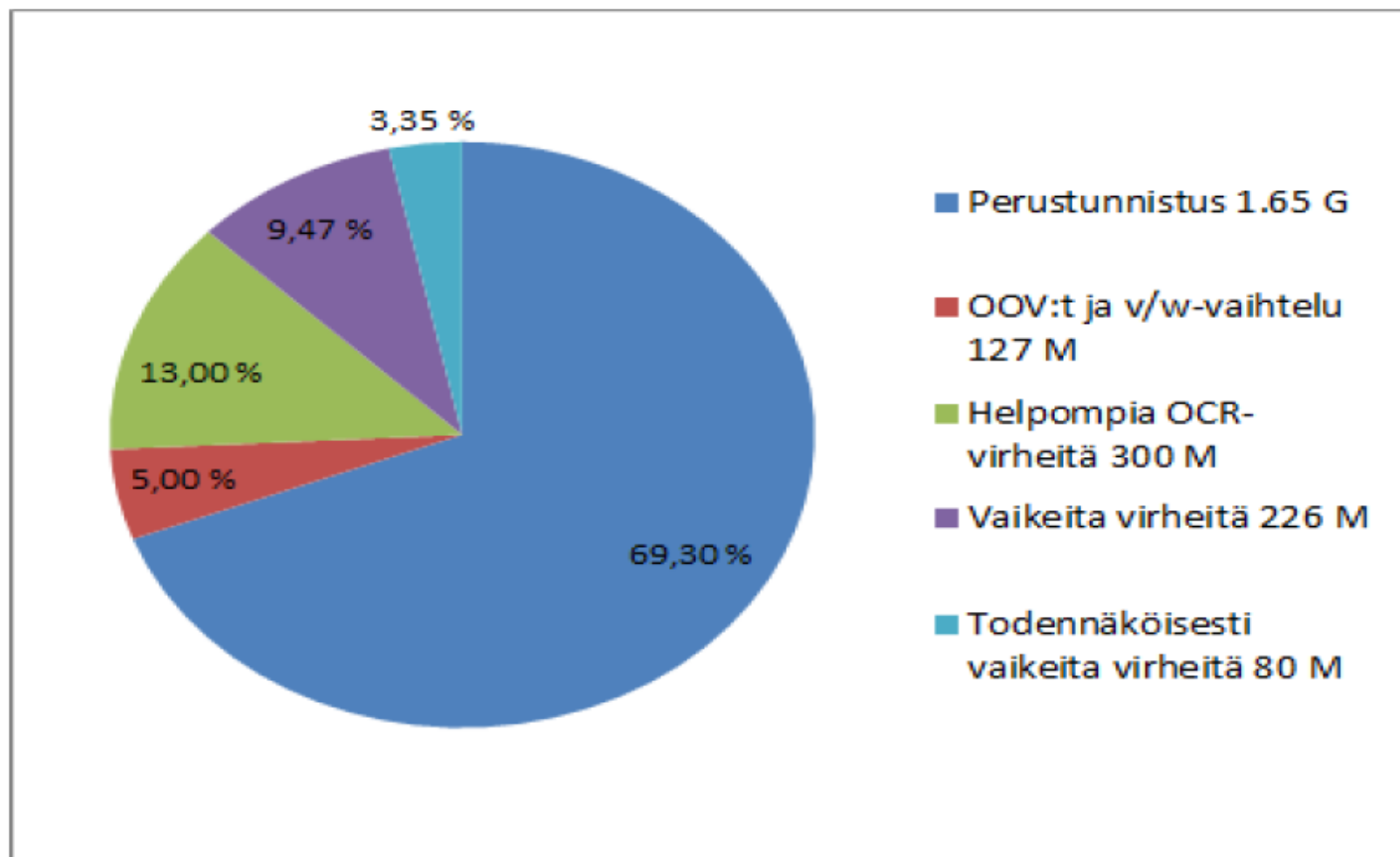
Euroopan unioni  
Euroopan aluekehitysrahasto

## Tunnistus sanatyypeittäin ja juoksevina sanoina

Kokoelma	Sanamäärä	Omorfi 0.1:n tunnistamia	FINTWOLin tunnistamia
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

<b>Kokoelma</b>	<b>Sanamäärä</b>	<b>Omorfi 0.2</b>	<b>HisOmorfi</b>
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<sup>7</sup>

# 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 ? → tavutus on jakanut sanan, pitäisi olla *heidan*, tunnistamaton  
dan +?
- Samoin kuin +?? → kirjoitettu yhteen, jäänyt tunnistamatta
- ylöskannetaan +?? → kirjoitettu yhteen, jäänyt tunnistamatta

Kestävää kasvua,

Vipuvoimaa  
EU:lta  
2014–2020



Euroopan unioni  
Euroopan aluekehitysrahasto

# 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

Kestävä kasvua ja työtä -ohjelma

