KULTTUURIHISTORIAN TUTKIMUS
Lähteistä menetelmiin ja tulkintaan

Toimittaneet Rami Mähkä, Marika Ahonen, Niko Heikkilä,
Sakari Ollitervo & Marika Räsänen

CULTURAL HISTORY — KULTTUURIHISTORIA 17

Kustantaja Kulttuurihistorian seura, Turku
ISBN 978-952-68776-8-6 (pehmeäkantinen)
ISBN 978-952-68776-9-3 (PDF)
ISSN 1458-1949

Kulttuurihistorian seura
c/o Kulttuurihistoria
20014 Turku
http://kulttuurihistoria.net

Ulkoasu: Henri Terho ja Heli Rantala
Taitto: Heli Rantala

Kannen kuva: Valtionarkiston (nyk. Kansallisarkisto) tutkijasali tarjosi historian-tutkijoille ajanmukaiset puitteet vuonna 1901. Kuvaaja: Atelier Nyblin A.B. Lähde: Museovirasto, Historian kuvakokoelmat, HK19650527:2. CC BY 4.0.

Painopaikka: BoD — Books on Demand, Norderstedt, Saksa 2022

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

DIGITOIDUT SANOMA- JA AIKAKAUSLEHDET LÄHDEAINEISTONA

Heidi Hakkarainen, Heidi Kurvinen, Petri Paju,
Hannu Salmi ja Satu Sorvali[1]

Digitaaliset lehtiaineistot kuuluvat nykyisin monen historiantutkijan aineistopohjaan joko pääasiallisena tai täydentävänä lähteenä. Tässä artikkelissa keskitymme Kansalliskirjaston digitoituun sanoma- ja aikakauslehtikokoelmaan, jonka käyttöliittymää on kehitetty pitkäjänteisesti 1990-luvun jälkipuoliskolta lähtien. Aineistoa on käytetty laajasti koko 2000-luvun ajan, ja tutkijat ovat pohtineet myös digitaalisten aineistojen luotettavuutta sekä niiden käyttöön liittyviä menetelmällisiä kysymyksiä.[2] Digitaalisten lehtiaineistojen huolellinen ja perinpohjainen käyttö vaatiikin tietoisuutta niiden erityispiirteistä. Artikkelimme tavoitteena on tarjota tähän menetelmällisiä välineitä. Avaamme sekä digitoitujen kokoelmien tutkimusmahdollisuuksia että aineistojen käyttöön liittyviä rajoituksia. Kuinka digitoitua lehtikokoelmaa kannattaa

[1] Artikkelin sisältö on suunniteltu kollektiivisesti, mutta alalukujen pohjatekstit on kirjoitettu yksin tai työpareina seuraavasti: Johdanto (Kurvinen ja Paju), Aineiston materiaalisuus ja lähdekritiikki (Kurvinen ja Salmi), Sanahakujen rajat ja mahdollisuudet (Hakkarainen ja Sorvali), Oman korpuksen rakentaminen (Sorvali), Lähi- ja etälukeminen (Paju ja Salmi), Lopuksi (Hakkarainen). Kaikki ovat osallistuneet tekstin kommentointiin ja hiomiseen.

[2] Esim. Paju, Rantala & Salmi 2019; Jensen 2021.

lähestyä lähdeaineistona? Miten monin eri tavoin lehtikokoelmia voi käyttää tutkimuksessa? Minkälaisia käyttötapoja lehtiaineistojen hyödyntämiseen on kehitetty?

Tarkastelumme keskittyy tekstisisältöjen analyysiin, mutta tietokoneavusteinen tutkimus voi yhtä hyvin kohdentua meta- eli kuvailutietoihin, lehtikuviin tai tekstin yhteydessä käytettyihin taulukoihin. Artikkelissa esitetyt ajatukset ja ohjeet pätevät yleisellä tasolla niin eri maiden sanoma- ja aikakauslehtiä sisältäviin verkkopalveluihin kuin laajemminkin digitaalisiin aineistoihin. On kuitenkin hyvä muistaa, että kansalliskirjastojen ja muiden tahojen tarjoamat käyttöliittymät ovat aina yksilöllisiä. Digitaalisissa kokoelmissa on samankaltaisia toimintoja, koska palveluja on kehitetty toisilta mallintaen ja rinnan, mutta niissä piilee myös eroavaisuuksia. Tästä johtuen tutkijan on tunnettava käyttämänsä aineistokokonaisuuden tausta, sen vahvuudet ja heikkoudet sekä tarjotun käyttöliittymän ominaisuudet.[3] Vain siten hän pystyy käyttämään aineistoaan tietoisesti ja tekemään sen pohjalta kestäviä tulkintoja.

Aineiston materiaalisuus ja lähdekritiikki

Tietokoneen näytöltä katsottaessa digitaaliset lähteet voivat tuntua aineettomilta, mutta niillä on materiaalinen luonne, joka on olennaista huomioida tulkintoja tehtäessä. Alkuperäisten sivujen sijasta digitoidut sanoma- ja aikakauslehtikokoelmat pohjautuvat pääsääntöisesti mikrofilmeihin, joiden digitointi on ollut nopeampaa ja taloudellisempaa kuin painettujen, usein yhteennidottujen lehtivuosikertojen skannaaminen.

Sanomalehtien mikrofilmaus alkoi toisen maailmansodan aikana tai pian sen jälkeen. Keskeistä oli vähentää painettujen lehtien käyttöä ja siten niiden kulumista. Suomessa mikrokuvaus alkoi vuonna 1951, ja mikrofilmi oli Kansalliskirjaston primäärinen tallennusalusta 2000-luvun alkuun asti.[4] Tosin digitaalisen kulttuurin vallattua alaa ja Internetin yleistyessä 1990-luvulla vahvistui käsitys siitä, että tulevaisuudessa järkevintä olisi tallentaa lehdet nimenomaan digitaalisesti.

[3] Ks. esim. Jensen 2021; Bødker 2018, 1115.
[4] Beals & Bell 2020, 6; Oiva, Nivala & Salmi 2018.

Digitointityö ymmärrettiin saumattomana jatkeena mikrokuvaukselle: analogisista tallenteista siirryttiin binäärisiin digitallenteisiin. Juuri tästä syystä useimmat kokoelmat – Suomen lisäksi muun muassa Australian, Iso-Britannian ja Yhdysvaltojen digitoidut lehdet – näyttävät ”mikrofilmimäisiltä”, mustavalkoisilta ja hieman epätarkoilta. Näin ollen kuvan muuntaminen tekstiksi on erityisen haasteellista. Hyvä esimerkki tästä on Singaporen kansalliskirjaston 1800-luvun aineisto, jonka pohjana on brittiläisen mikrofilmiyrityksen tarjoama materiaali. Heikkolaatuisen tekstin ihmissilmä tulkitsee helposti, mutta koneelle sen tunnistaminen on haasteellista.[5]

Digitaalisten kokoelmien materiaaliset lähtökohdat vaikuttavat siihen, millaisia mahdollisuuksia tutkimukselle avautuu nykypäivänä. Massadigitointien käynnistyessä 1990-luvulla ei vielä ollut selvää, että tulevaisuudessa lehtiin tehtäisiin sana- tai fraasihakuja.[6] Nykyiset hakujärjestelmät perustuvat optiselle tekstintunnistukselle (OCR, *optical character recognition*), jossa OCR-ohjelmiston avulla kuvasta tunnistetaan teksti. Tällä tunnistamisella on materiaaliset ehtonsa, jotka eivät rajaudu vain mikrofilmien merkitykseen. Vaikka aineisto digitoitaisiin alkuperäisestä julkaisusta, materiaalisuus voi vaikuttaa tekstintunnistukseen: kirjakkeet ja niiden muotit eivät ole välttämättä olleet standardisoituja, huokoinen paperi on käsitellyt mustetta eri tavoin kuin hiottu paperi ja sivun mittasuhteet ovat vaikuttaneet siihen, miten isolla tai pienellä kirjasinkoolla teksti on painettu. Myös paperin väri ja laatu vaikuttavat siihen, millaisia merkityksiä eri lehtien teksteistä voidaan tulkita.

Digitaalisten kokoelmien muodostumisen ohella materiaalisuus liittyy kiinteästi tietokoneen ruudulta luettavien tekstien analysoimiseen. Paperilehteä selaava tutkija näkee lähteensä materiaalisen muodon konkreettisesti edessään, mutta digitaalista jäljennettä tarkastelevalla tutkijalla lehden fyysinen koko häipyy helposti näkyvistä. Tutkija löytää toki viitteitä aineistonsa materiaalisuudesta metatiedoista (Kuva 1), mutta kyse on viime kädessä numeroista, joiden kuvittelu materiaalisena objektina jää hänen mielikuvituksensa varaan.[7] Materiaalisuudella

[5] Salmi 2020, 47–49.
[6] Prescott 2018, 49–71.
[7] Esim. Huistra & Mellink 2016, 221–222; Bastiansen 2020, 39–40.

```
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xlink="http://www.w3.org/TR/xlink"
 xsi:noNamespaceSchemaLocation="//192.168.10.50/docworks/docWORKS/schema/alto-1-2.xsd">
  <Description>
    <MeasurementUnit>mm10</MeasurementUnit>
    <sourceImageInformation>
      <fileName>//192.168.10.35/dwin/sanomalehdet/hameen_sanomat/Mf75393/0010.tif</fileName>
    </sourceImageInformation>
    <OCRProcessing ID="OCRPROCESSING_1">
      <preProcessingStep>
        <processingSoftware>
          <softwareCreator>CCS Content Conversion Specialists GmbH, Germany</softwareCreator>
          <softwareName>CCS docWORKS</softwareName>
          <softwareVersion>6.2-1.8</softwareVersion>
        </processingSoftware>
      </preProcessingStep>
      <ocrProcessingStep>
        <processingSoftware>
          <softwareCreator>ABBYY (BIT Software), Russia</softwareCreator>
          <softwareName>Finereader</softwareName>
        </processingSoftware>
      </ocrProcessingStep>
    </OCRProcessing>
  </Description>
  <Styles>
    <TextStyle ID="TXT_0" FONTSIZE="28" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_1" FONTSIZE="24" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_2" FONTSIZE="29" FONTFAMILY="Times New Roman"/>
    <TextStyle ID="TXT_3" FONTSIZE="10" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_4" FONTSIZE="11" FONTFAMILY="Times New Roman"/>
    <TextStyle ID="TXT_5" FONTSIZE="9" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_6" FONTSIZE="11" FONTFAMILY="Times New Roman" FONTSTYLE="bold italics"/>
    <TextStyle ID="TXT_7" FONTSIZE="11" FONTFAMILY="Times New Roman" FONTSTYLE="italics"/>
    <TextStyle ID="TXT_8" FONTSIZE="8" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_9" FONTSIZE="11" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_10" FONTSIZE="8" FONTFAMILY="Times New Roman" FONTSTYLE="bold italics"/>
    <TextStyle ID="TXT_11" FONTSIZE="8" FONTFAMILY="Times New Roman" FONTSTYLE="bold"/>
    <TextStyle ID="TXT_12" FONTSIZE="35" FONTFAMILY="Times New Roman"/>
    <TextStyle ID="TXT_13" FONTSIZE="29" FONTFAMILY="Times New Roman" FONTSTYLE="italics"/>
    <TextStyle ID="TXT_14" FONTSIZE="13" FONTFAMILY="Times New Roman"/>
    <TextStyle ID="TXT_15" FONTSIZE="23" FONTFAMILY="Times New Roman"/>
    <TextStyle ID="TXT_16" FONTSIZE="17" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_17" FONTSIZE="23" FONTFAMILY="Times New Roman" FONTSTYLE="italics"/>
    <TextStyle ID="TXT_18" FONTSIZE="13" FONTFAMILY="Times New Roman" FONTSTYLE="bold"/>
    <TextStyle ID="TXT_19" FONTSIZE="13" FONTFAMILY="Fraktur"/>
    <TextStyle ID="TXT_20" FONTSIZE="12" FONTFAMILY="Fraktur"/>
```

Kuva 1. Lehtiaineiston metatiedot löytyvät ALTO XML -tiedostosta. ALTO (*Analyzed Layout and Text Object*) on metatietostandardi ja XML tiedostotyyppi. Tiedostossa on kuvattu muun muassa, millä ohjelmistolla optinen tekstintunnistus on tehty, mistä numerosta ja sivusta on kyse, missä hakemistossa sivun TIF-muotoinen kuvatiedosto sijaitsee, millaisia fontteja ja pistekokoja sivulla on käytetty ja miten tekstit ja kuvat sijoittuvat. Kuva on *Hämeen Sanomien* ensimmäinen sivu 3.7.1891, Kansalliskirjaston Digi-palvelu, https://digi.kansalliskirjasto.fi/sanomalehti/binding/610067/page-1.xml. Haettu 9.12.2020.

on kuitenkin merkitystä, sillä se auttaa ymmärtämään sitä laajempaa mediaympäristöä, jossa analyysin kohteena oleva yksittäinen artikkeli on julkaistu. Digitaalisten hakujen tuottamien tulosten tulvassa julkaisukonteksti nimittäin hämärtyy helposti, jos tutkija ei kiinnitä siihen tietoisesti huomiota. Kontekstualisointi on näin ollen yksi niistä historiantutkijan perustyökaluista, jotka on syytä muistaa myös digitaalisissa tutkimusympäristöissä.[8]

Tutkimusmielessä digitoidut lehdet ovat lähtökohtaisesti eri asia kuin painetut sanoma- ja aikakauslehdet, mistä johtuen niitä kan-

[8] Esim. Jensen 2021. Journalistisen kontekstin merkityksestä ks. myös esim. *Digital Journalism* -lehden numero 9 (2018), joka keskittyy digitaalisiin lehtiarkistoihin journalismin historian tutkimuksessa.

nattaa ajatella alkuperäisten lehtien uudelleenjulkaisuina.[9] Tällöin on helpompi muistaa, että niiden käyttö eroaa merkittävästi paperilehtien lukemisesta. Tutkimukselle onkin eduksi, jos tutkija tutustuu digitaalisen aineiston rinnalla lehtien alkuperäisiin versioihin. Se auttaa kuvittelemaan, miten aineistoa on alun perin luettu ja miltä lehti on tuntunut lukijoiden käsissä. Digitaalisten uudelleenjulkaisujen avulla päästään kuitenkin riittävän lähelle alkuperäisiä lehtiä, jotta mennyttä voidaan lähestyä niitä analysoimalla alkuperäisten sanoma- ja aikakauslehtien tavoin. Samalla digitaalinen kokoelma tarjoaa mahdollisuuksia, joita aiemmilla lehtipinojen tarkastelijoilla tai mikrofilmattujen lehtien käyttäjillä ei ollut. Digitoidut lehtiarkistot ovat mahdollistaneet myös kokonaan uudenlaisten tutkimuskysymysten esittämisen.[10]

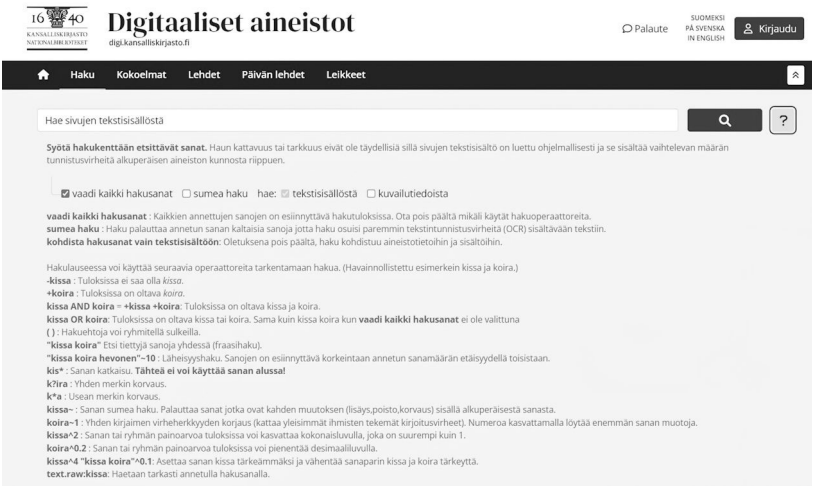
Digitoitujen lehtiaineistojen muuntunut materiaalisuus edellyttää tutkijalta uudenlaista lähdekriittisyyttä, ja tutkimuksessa on alettu puhua erityisestä digitaalisesta lähdekritiikistä tai digitaalisten arkistojen lukutaidosta. Yksinomaan se, mitä jää digitaalisten kokoelmien ulkopuolelle, vaatii tutkijalta tietoista pohdiskelua.[11] Esimerkiksi suomalaiset naistenlehdet ovat pääsääntöisesti digitoimatta, ja suomenkielisistä sanomalehdistä ainoastaan *Etelä-Suomen Sanomat*, *Länsi-Savo* ja *Maaseudun Tulevaisuus* löytyvät artikkelin kirjoitushetkellä koko julkaisuhistoriansa ajalta digitoituina. Lehtikokoelman koostumus jättääkin väistämättä tiettyjä lehtiä kokoelman pohjalta tehtyjen analyysien ulkopuolelle ja siten vinouttaa käsitystämme menneisyyden julkisesta keskustelusta. Vaikka Kansalliskirjaston lehtikokoelmaa pidetään muuten perustellusti varsin kattavana tekijänoikeusvapaan aineiston osalta,[12] se ei ole täydellinen tai virheetön. Kokoelmasta esimerkiksi puuttuu sano-

[9] Esim. Bastiansen 2020, 39–40; Cordell 2017, 188–225.

[10] Esim. Bødker 2018, 1115–1116.

[11] Esim. Elo 2016; Jensen 2021.

[12] Tekijänoikeuslain mukaan digitaalisten tekstien oikeudet kuuluvat alkuperäisille kirjoittajille ja heidän perikunnilleen 70 vuotta kuoleman jälkeen. Kansalliskirjaston ja Kopioston sopimuksen mukaan aineisto on vuoteen 1939 avointa kaikille käyttäjille. Sopimuksen mukaan yliopistojen opiskelijat ja tutkijat voivat käyttää tekijänoikeussuojattua aineistoa vuodesta 1940 eteenpäin Haka-tunnistautumisen kautta. Tämä sopimus on voimassa 31.12.2022 asti. Ks. lähemmin https://www.kansalliskirjasto.fi/fi/projektit/tutkain-2020-2022. Haettu 9.12.2020.



Kuva 2. Kansalliskirjaston Digi-palvelusta avautuu listaus hakuohjeista painamalla ylhäällä oikealla nähtävää kysymysmerkkiä. Kuva: Kansalliskirjaston Digi-palvelu, https://digi.kansalliskirjasto.fi/search?query. Haettu 9.12.2020.

malehtien numeroita sieltä täältä, ja julkaisujen metatiedoissa saattaa esiintyä ongelmia, joita toki korjataan jatkuvasti.

Tutkijan on myös syytä tiedostaa omien taitojensa merkitys digitaalisista kokoelmista saataville hakutuloksille. Aikaisemmin paperisten lehtien käyttö tai mikrofilmattuun lehteen tutustuminen alkoi tyypillisesti niiden selailulla, mutta digitaaliseen lehteen tai kokoelmaan sukelletaan herkästi suoraan yksinkertaisen sanahaun avulla. Myös käyttöliittymät ohjaavat epäsuorasti tämänkaltaiseen aineistojen käyttöön, sillä esimerkiksi Kansalliskirjaston Digitaaliset aineistot -sivusto näyttää käyttäjälleen ensimmäiseksi juuri sanahakukentän. Sanahaku on kieltämättä yksinkertaisin, ja sellaisena oleellinen sekä tehokas hakumuoto ja selkeä esimerkki erosta entiseen, paperilehtien ja mikrofilmien aikaan. Se on myös digitaalisten aineistojen käyttöön tottumattomalle tutkijalle helppo tapa aloittaa tutustuminen aineistoihin. Yksinkertaisten sanahakujen kautta aineisto jää kuitenkin ohuesti tutkituksi ja kokonaisuuksien hahmottaminen käy mahdottomaksi.[13] Tutkijan kan-

[13] Esim. Broersma & Harbers 2018, 1151.

nattaakin perehtyä myös käyttöliittymän edistyneempiin haku- ja selailumahdollisuuksiin (ks. Kuva 2), kuten osoitamme seuraavaksi.[14]

Sanahakujen rajat ja mahdollisuudet

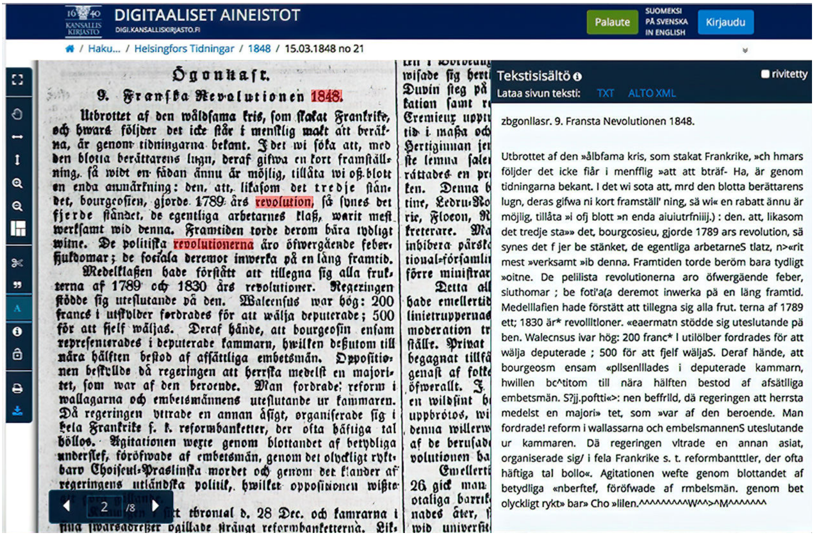
Tietokoneavusteinen analyysi edellyttää, että tutkittavat dokumentit on skannattu tai valokuvattu digitaaliseen muotoon kuvatiedostoiksi, jolloin niiden sisältämät tekstit voidaan ajaa optisen tekstintunnistuksen läpi ja muuntaa koneellisesti luettaviksi raakateksteiksi. Arkistot ja kirjastot tarjoavat digitoitujen sanoma- ja aikakauslehtiaineistojen lukemiseen usein käyttöliittymän, jossa voi tutkia OCR-tekstejä. OCR-tekstin laatuun vaikuttaa kuitenkin se, millä kirjasintyypillä teksti on painettu. Esimerkiksi vanha fraktuura on usein haastavaa optiselle tekstintunnistamiselle. Myös eri kielet ja kielten sekoittuminen aineistossa, historialliset variaatiot, ladonta, typografia ja lähteen digitoinnin laatu vaikuttavat siihen, kuinka hyvin tekstintunnistaminen toimii (ks. Kuva 3). Hyvälaatuinenkin OCR-teksti sisältää aina jonkin verran virheitä, ja pahimmassa tapauksessa niin kutsuttua hälyä tai kohinaa (*noise*) on niin paljon, että se vaarantaa tulosten luotettavuuden.[15]

OCR-tekstin puutteista on hyvä olla tietoinen, koska tekstin louhimiseen ja koneoppimiseen liittyvät digitaaliset työkalut hyödyntävät sitä. Sanahakuja on mahdollista tehdä paitsi yksittäiseen tekstitiedostoon, myös sellaisiin tietokantoihin, joissa haetaan sanoja koko aineistosta. Usein tämä indikoidaan hakuna aineistosta tai tekstisisällöistä (*full text search*). Näin haku voidaan kohdistaa aineiston tekstisisältöihin sen kuvailutietojen sijasta.

Sanahaun käyttöliittymä ohjaa hakutoimintoa; millä ehdoilla ja rajauksilla hakuja voi tehdä, kuinka suuri osa digitoidusta aineistosta on haun piirissä ja miten haun tulokset esitetään. Monissa hakukäyttöliittymissä voi tehdä edistyneempiä ja tarkempia hakuja käyttäen esimerkiksi Boolen operaattoreita (AND/OR/NOT), asiasanoja (*key words*) sekä ennalta määrättyjä hakukenttiä, joiden avulla voi etsiä esimerkiksi

[14] Kannattaa muistaa myös aikoinaan sanomalehtiaineistosta ihmisvoimin tehdyt, nykyään digitoidut hakemistot. Ks. Kokko 2017.

[15] Ks. Jarlbrink & Snickars 2017.



Kuva 3. Kansalliskirjaston käyttöliittymässä kohta "Näytä tekstinä (OCR)" avaa näkymän, jossa kuva tekstistä ja OCR-tunnistettu tekstisisältö näkyvät rinnakkain. Kuvassa *Helsingfors Tidningar* -lehden 15.3.1848 ilmestyneen numeron sivu. Kirjoitus helmikuun vallankumouksen puhkeamisesta Pariisissa vuonna 1848 on haettu hakusanoilla "1848" AND "revolution" AND "Paris". Kuva: Kansalliskirjaston Digipalvelu https://digi.kansalliskirjasto.fi/sanomalehti/binding/495290?term=1848&term=revolution&term=revolutionerna&term=Paris&term=PariS&page=2&ocr=true. Haettu 9.12.2020.

teoksen otsikon tai sen tekijän perusteella (*field-restricted search*), lainausmerkkien ympäröimiä fraaseja eli määrättyjä sanojen yhdistelmiä (*phrase search*), sekä sanojen variaatioita (*wild card*) ja niiden välisiä etäisyyksiä (*proximity*). Lisäksi hakusanan katkaiseminen katkaisumerkillä * antaa useissa hakukoneissa suuremman määrän hakutuloksia, sillä haku sisältää tällöin kaikki sanan taivutusmuodot.

Tekstintunnistuksen laatu määrää sen, kuinka luotettava sanahaun tulos on. Kirjoitusmuodoltaan samankaltaiset sanat saattavat sekoittua. Koneellinen haku ei myöskään löydä kaikkia niitä sanoja, jotka esiintyvät aineistossa. Tämän vuoksi puhutaan vääristä positiivisista ja negatiivisista tuloksista (*false positives and false negatives*).[16] Lisäksi sanahakuja on kritisoitu kontekstin puuttumisesta ja lehtikuvien merkityk-

[16] Eri tavoista tehdä hakuja ks. lisää Huistra & Mellink 2016, 220–229.

sen kasvun sivuuttamisesta. Kuvien ja tekstien yhdistelmien tutkiminen onkin haastavaa, sillä se edellyttää, että kuvia voi erikseen hakea käyttöliittymässä.[17]

Sanahaut ovat tutkijan määrittelemiä ja sisältävät hänen ennakkokäsityksiään tutkittavasta aineistosta sekä siitä, mitä hän haluaa etsiä ja löytää.[18] Näin ollen on hyvä pohtia ja myös kyseenalaistaa omien hakusanojensa käyttöä. Lisäksi hakusanojen käyttöön liittyvien valintojen avaaminen tutkimusprosessista kirjoitettaessa lisää tutkimuksen toistettavuutta ja luotettavuutta. Tässä yhteydessä on hyvä huomata sanojen merkitysten muuttuminen. Esimerkiksi 1800-luvun alun suomenkielisessä lehdistössä ”vallankumouksen” sijaan puhuttiin usein ”metelistä” tai ”kapinasta” (vrt. ruotsi, *upprörelse*, saksa, *Aufruhr*), kun taas sana ”vallankumous” vakiintui suomen kieleen vasta vuoden 1848 jälkeen.[19] Perehtyminen laajemmin tutkittavan ajankohdan kieleen ja keskusteluihin auttaa löytämään hakusanoja, jotka kuvaavat oman aikansa ihmisten kielenkäyttöä. Myös synonyymit ja sanat, joilla on samankaltainen merkitys, saattavat muuttaa hakutulosta. Synonyymien käyttö hakusanoina ja erilaisten vaihtoehtoisten hakuratkaisujen kokeileminen onkin erittäin suositeltavaa.[20]

Oman korpuksen rakentaminen

Sanoma- ja aikakauslehtiä hyödyntävän tutkijan voi olla hyödyllistä rakentaa ja koota kokotekstiaineistosta oma korpus eli kokoelma tutkimusaihetta käsitteleviä tekstejä, ellei sellaista ole jo olemassa.[21] Pelkästä tekstikokoelmasta korpus erottuu siten, että se on rakennettu tiettyyn tarpeeseen ennalta määritellyn kriteeristön mukaisesti.[22] Kun tutkimus kohdistuu vain pieneen osaan kokotekstiaineistoa, ja hakutoiminnossa on vaikeaa tai mahdotonta tehdä sopivia rajauksia, korpuksen raken-

[17] Maurantonio 2014, 88–102.
[18] Nicholson 2013, 59–73.
[19] Rantala 2019.
[20] Esim. Hoekstra & Koolen 2018.
[21] Piotrowski 2019, 12.
[22] Ibid, 11–12.

taminen on järkevää. Esimerkiksi Kansalliskirjaston digitoima aineisto on segmentoimatonta, eli hakua ei voi kohdistaa vain tiettyyn tekstityyppiin, kuten pääkirjoituksiin tai vaikkapa mainoksiin. Yhdessä tutkimuksessa voidaan käyttää useampia korpuksia tai korpusten alikorpuksia.[23]

Ensin tulee selvittää, onko tutkimukseen sopiva korpus jo olemassa. Niitä löytyy esimerkiksi Kielipankin ylläpitämästä META-SHARE -palvelusta.[24] Jos sopivaa korpusta ei ole olemassa, se voidaan luoda automaattisesti, manuaalisesti tai näiden yhdistelmällä.[25] Manuaalisessa työstössä tiedot kerätään lehdistä yksitellen käsin. Jos tutkimusaineisto on laaja, on suositeltavaa, ellei jopa välttämätöntä luoda korpus ainakin osittain automaattisesti. Tähän on olemassa valmiita sovelluksia ja useimmiten verkkopalveluissa on jokin tällainen toiminto olemassa.[26] Hybridimallissa korpus luodaan automaattisesti halutulla aineistorajauksella ja saatavissa olevilla aineiston metatiedoilla ja myöhemmin sitä rikastetaan uusilla metatiedoilla. Hyvä esimerkki tästä on 1800-luvun sanomalehtien paikalliskirjeistä koostuva Translocalis-tietokanta, joka on rakennettu Kansalliskirjaston aineistosta.[27]

Korpuksen suunnittelu on tehtävä huolella, jotta sen avulla voidaan saada vastauksia tutkimuskysymyksiin ja näin välttää turhaa työtä. Mallintaminen lähtee liikkeelle aineiston rajaamisesta eli pohdinnasta, mitä kaikkea korpukseen sisällytetään. Datan määrä ei ole ratkaiseva tai laatua lisäävä tekijä, vaan se, mitä tutkimuskysymyksiä on tarkoitus ratkaista. Korpuksen suunnittelussa korostuu lähdekriittisyys: sen tulisi olla tasapainoinen, tarpeeksi edustava ja mielekäs kokonaisuus. Jos korpuksesta puuttuu jotain oleellista, se puuttuu myös tutkimuksen analyysistä.[28] Liian laajaa korpusta taas ryhdytään helposti pitämään kopiona siitä tekstistä, josta se on alun perin koostettu.[29] Korpuksen

[23] Ks. esim. Piotrowski 2019, 13.

[24] https://metashare.csc.fi/. Haettu 2.11.2020.

[25] Ks. lähemmin Tolonen & Lahti 2015.

[26] Esim. Kansalliskirjaston Digi-palvelussa voi viedä hakutulokset tai Leikekirjan sisällön Exceliin.

[27] Translocalis-tietokanta, Suomen Akatemian Kokemuksen historian huippuyksikkö, Tampereen yliopisto, https://research.uta.fi/hex-fi/translocalis/. Haettu 3.12.2020.

[28] Piotrowski 2019, 9, 13.

[29] Piotrowski 2019, 16. Ks. myös Hoekstra & Koolen 2018, 87.

kokoa kannattaa rajoittaa siitäkin syystä, että sen koostaminen vie aikaa. Lisäksi on pohdittava tarkkaan, mitä metatietoja aineistosta tallennetaan. Jos metatiedot kerätään manuaalisesti, ei ole ajankäytöllisesti järkevää, että aineiston joutuu käymään läpi useaan kertaan, eikä se ole aina edes mahdollista. Aineistosta tallennettavia kuvailutietoja tulee pohtia myös tietosuojan näkökulmasta, joten on hyvä tutustua tutkimuseettisen neuvottelukunnan ohjeisiin.[30]

Korpuksen formaatin suhteen on huomioitava, että muunto myöhemmin toiseen formaattiin voi olla vaikeaa tai jopa mahdotonta. Microsoftin Excel lienee yksi käytetyimmistä formaateista, josta saa koostettua erilaisia taulukkoja ja graafeja ja jossa on erilaisia laskenta- ja suodatustoimintoja. Yksi vaihtoehto on koota korpus leikekirjatoiminnolla, jos kokotekstin tallennusalusta tukee sitä. Sellainen tarjotaan esimerkiksi Kansalliskirjaston Digi-palvelussa tunnistautuneelle käyttäjälle. Tällöin kirjoitukset kerätään lehdistä valitsemalla halutun tekstin alue ja tallentamalla leikkeestä otsikko, aihealue, tyyppi ja avainsanat. Leikekirjaa voi tarkastella Kansalliskirjaston Digi-palvelussa ja sen sisällön voi ladata itselleen Excel-muodossa. Taulukkoon tulostuvat edellä mainittujen metatietojen lisäksi lehden nimeke, ISSN-koodi, päivämäärä, lehden numero, URL-osoite, muistiinpanot, luontipäivämäärä ja OCR-teksti.

Lopuksi on vielä selvitettävä, minne korpus tallennetaan ja keiden pitää päästä siihen käsiksi. Jos korpus sisältää arkaluonteista tietoa, se tulee suojata asiattomalta käytöltä. Jos se sisältää tietoa tunnistettuun tai tunnistettavissa olevaan elävään henkilöön, se on henkilörekisteri ja siihen pätee tietosuoja-asetus.[31]

Tieteen avoimuus palvelee kaikkia tiedon uudelleenkäytettävyyden ja resurssien järkevän käytön vuoksi. Tutkimuksen läpinäkyvyydelle ja toistettavuudelle sekä tulosten luotettavuuden arvioinnille onkin tärkeää kirjoittaa auki, miten ja millä perusteella korpus on kerätty ja miten sen avulla on päädytty tiettyihin tutkimustuloksiin.[32]

[30] Ks. luvut 3.5.–3.7, https://tenk.fi/sites/tenk.fi/files/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2019.pdf. Haettu 2.11.2020.

[31] Ks. tarkemmin https://tietosuoja.fi/tieteellinen-tutkimus. Haettu 2.11.2020.

[32] Hoekstra & Koolen 2018, 80, 92–93; Koolen, van Gorp & van Ossenbruggen 2019, 370; Piotrowski 2019, 16; Tolonen & Lahti 2018, 253–255.

Lähi- ja etälukeminen

Digitaalisten sanoma- ja aikakauslehtien käytössä lukemisen tekniikat ja tavat ovat keskeinen osa metodologiaa. Historiantutkimuksessa tarkkaa, merkitysten tulkintaan keskittyvää lukemista kutsutaan lähilukemiseksi (*close reading*): tavoitteena on ymmärtää, miten teksti välittää ja tuottaa merkityksiä.[33] Tekstiä tulee katsoa ”läheltä”, jotta laajemmat yhteydet voisi tavoittaa. Digitoituja aineistoja voi lähestyä lähilukemisen keinoin joko lukemalla suoraan skannattuja numeroita, sivu ja artikkeli kerrallaan, tai suodattamalla aineistoa hakutoimintojen, kuten sana- ja fraasihakujen avulla.

Lähilukemisen perusajatuksena on, että tutkija luo kokonaiskuvan lukemalla tarkasti ison määrän tekstejä, ja tämä laajempi kehys auttaa ymmärtämään ja kontekstualisoimaan yksittäisen tekstin tulkintoja ja rakenteita. Sen sijaan käsite etälukeminen (*distant reading*) on viime vuosina vakiintunut tarkoittamaan laajan digitaalisen tekstiaineiston säännönmukaisuuksien hahmottamista tietokoneohjelmien avustamana tai laskennallisesti, siis ikään kuin lähilukemisen vastakohtaa. Etäluennan sijasta on käytetty myös termiä kaukoluenta. Molemmat termit viittaavat lähestymistapaan, jonka alle voidaan laskea monenlaisia tekstikokoelmien haltuunoton menetelmiä. Olennaista on ajatus tietokoneen käyttämisestä lukemisen välineenä, jolloin tekstien täytyy olla koneluettavassa (*machine-readable*) muodossa. Tämä mahdollistaa tekstinlouhinnan, hahmottamisen suuressa mittakaavassa.

Etälukemisen käsitteen loi kirjallisuushistorian tutkija Franco Moretti vuonna 2000. Alkujaan Moretti ajatteli nimetä uuden tapansa hahmottaa valtavia tekstiaineistoja ”sarjalukemiseksi”, sillä kysymys oli pitkälti tekstiaineistosta lasketuista numerosarjoista ja niiden visualisoinneista sekä näiden perusteella tehdyistä tulkinnoista eli etälukemisesta.[34] Morettin provokatiivinen ajatus oli, että lähilukemisessa

[33] Lähilukemisesta ks. Federico 2016; Salmi 2020, 30–32.

[34] Moretti ammensi vaikutteita kulttuurihistorialliseen tutkimukseen suuresti vaikuttaneelta Annales-koulukunnalta ja nimenomaan sen määrällisiä tutkimustapoja painottaneelta suuntaukselta. Turun yliopiston kulttuurihistoriassa on tähän asti voimakkaasti painottunut annalistien esimerkki nimenomaan historiantutkimuksen laadullisten tutkimusmenetelmien kehittäjinä.

tutkija oli liian lähellä tekstiä, eikä kyennyt näkemään kokonaisuutta, tarpeeksi kaukaa.[35]

Etäluenta liittyy kiinteästi niin sanottuun historiantutkimuksen digitaaliseen käänteeseen. Tietoverkkojen sisältämiä valtavia digitaalisia aineistokokonaisuuksia oli 2000-luvun alussa mahdotonta enää käydä läpi tarkan lähiluvun keinoin. Tutkimusaineistona saattoi olla esimerkiksi tuhat digitoitua kirjaa tai vaikkapa 200 vuotta parlamentin keskusteluaineistoja, satojatuhansia tiiviitä sivuja. Tutkijan haasteeksi nousi se, kuinka materiaalista voisi saada edes jonkinlaisen yleiskäsityksen. Miten kohdistaa katse helpommin ja nopeammin oleellisiin kohtiin? Aineisto oli pakko käsitellä tietokoneavusteisesti.

Uusiin kysymyksiin ryhdyttiin kehittämään vastaamisen tapoja, joita nykyään kutsutaan nimellä digitaalinen historiantutkimus (*digital history*). Englannin kielessä digitaalinen historia viittasi alun perin 1990-luvulla tietoverkoissa esitettyyn ja sinne muokattuun historia-aineistoon, mutta suomeksi käsite otettiin laajemmin käyttöön 2010-luvulla puhuttaessa laajasti tietokoneavusteisesta historiantutkimuksesta.[36] Käytetään myös termiä laskennallinen historia (*computational history*), kun informaatioteknologian menetelmiä sovelletaan historiantutkimuksen aineistojen prosessointiin ja analyysiin.

Helpoimmin avautuva etäluennan tapa on muodostaa tekstistä sanapilvi (*word cloud*),[37] jossa tyypillisesti lasketaan ohjelmallisesti, nostetaan esiin ja ryhmitellään tutkitun tekstikokonaisuuden yleisimmät sanat. Mitä useammin sana esiintyy aineistossa, sitä suurempana se näkyy pilvessä (Kuva 4). Sanapilvi näyttää tekstin sisällön uudesta näkökulmasta ja auttaa siten laajemman kokonaisuuden hahmottamisessa. Monet etäluennan työkalut, kuten tilastolliseen laskentaan perustuva aihemallinnus (*topic modelling*) ovat huomattavasti monimutkaisempia.[38] Tämä on sekä vahvuus että heikkous, sillä tutkijan saattaa olla vaikea arvioida, millaiset laskennalliset ratkaisut ohjaavat aineiston käsittelyä.

[35] Moretti 2000, 57; Moretti 2005, 1; Paju 2020, passim.

[36] Tietokoneita tosin oli hyödynnetty historiantutkimuksen apuna jo 1960-luvun lopulta lähtien myös Suomessa. Ks. lähemmin Paju 2020, 21–44.

[37] Sanapilviä on mahdollista tehdä verkkosivujen kautta toimivissa ohjelmissa, kuten EdWordle, Voyant Tools ja Wordart.

[38] Aiheenmallinnuksesta ks. esim. Graham, Weingart & Milligan 2012.

Kuva 4. Sanapilvi on ensisijaisesti esimerkki etälukemisen lähtökohdasta ja tulosten visualisoinnin luomista tulkintahaasteista. Kuvan sanapilvi Aleksis Kiven romaanista *Seitsemän veljestä* (1870) näyttää ainakin keskeisten veljesten nimet. Lehtiteksteissä on tyypillisesti paljon erilaisia, tavanomaisia sanoja, joten sanapilvi sopii parhaiten luetteloiden hahmottamiseen. (Wordart.com, teksti Project Gutenberg.)

Aineisto asettaa kuitenkin rajoituksensa tekstinlouhinnalle ja sen mahdollistamalle etälukemiselle. Esimerkiksi Kansalliskirjaston sanomalehtiaineistojen mittava etälukeminen on vaikeaa tai käytännössä hyödytöntä useimmilla menetelmillä, koska OCR-prosessointi on synnyttänyt aineistoon niin runsaasti satunnaista epätarkkuutta. Lukemattomat virheelliset merkit tuottavat etäluennan lähtökohdaksi uuden sekasotkun. Kun aineiston OCR-laatua saadaan tulevaisuudessa parannettua, etäluennalle avautuu kuitenkin uusia mahdollisuuksia. Tällä hetkellä Kansalliskirjaston palvelussa on koekäytössä Nimiapuri, joka perustuu nimentunnistukseen ja helpottaa henkilöiden paikantamista.[39]

OCR-virheistä huolimatta sanomalehtiaineiston etälukeminen on mahdollista kokoamalla aineistosta oma korpuksensa edellisessä alaluvussa kuvatulla tavalla. Kun tekstien OCR-laatua vielä korjataan manuaalisesti, tähän uuteen korpukseen voidaan soveltaa erilaisia etäluennan

[39] Nimiapuri, https://digi.kansalliskirjasto.fi/name-search. Haettu 9.12.2020.

työkaluja.[40] Muitakin mahdollisuuksia on toki kehitelty. Esimerkiksi sanomalehtien tekstintoistojen tunnistus laskennallisesti on auttanut luomaan kokonaiskuvan laajasta, miljoonien sivujen kokonaisuudesta. *Computational History and the Transformation of Public Discourse in Finland* -hankkeessa kehitettiin menetelmä, joka pystyi OCR-virheistä huolimatta löytämään aikavälillä 1771–1920 toistetut tekstit ja tekstikatkelmat.[41] Samoin lehdistöaineistoa voidaan tarkastella määrällisesti metatietojen avulla, kuten analysoimalla julkaisunimikkeiden sekä sivu- ja merkkimäärien muutoksia.

Kulttuurihistoriallisessa tutkimuksessa korostetaan sekä merkitysten tutkimista että tutkittavan ilmiön asettamista laajempiin merkitysyhteyksiin. Siksi lähi- ja etälukemista on usein välttämätöntä käyttää tutkimuksessa rinnakkain yhdistämällä erilaisia lukemisen tapoja ja tasoja. Frédéric Clavert on käyttänyt ilmaisua *une double lecture*, ”kaksinkertainen lukeminen”: kun tutkija katsoo lähdeaineistoa etäältä, hänen on samalla katsottava sitä myös läheltä.[42] Näkökulmien yhdistäminen, vuorottelu ja vertaaminen toisiinsa vievät tutkimusta tarkempiin ja yhä parempiin, vakuuttavampiin tuloksiin. Tekstinlouhinnan tai tekstiaineiston etälukemisessa saatujen tulosten tulkinta on usein prosessin vaikein osuus, jossa on eduksi palata valikoivaan lähilukemiseen. Tulkinnassa on kysyttävä, mistä esimerkiksi tiettyjen sanojen esiintyminen tyypillisesti lähekkäin, kuten aiheenmallinnuksen tapauksessa, oikeastaan kertoo ja mitä sanojen ryhmittymistä voi päätellä.

Etälukeminen soveltuu säännönmukaisuuksien hahmottamiseen laajoista kokonaisuuksista, uusien hypoteesien kehittelyyn ja mahdollisten aiempien aavistusten testaukseen. Näistä tutkija voi aiempaa systemaattisemmin ja rikkaammin tiedoin edetä ja palata lähilukemaan tekstejä sekä muita aineistoja. Samalla etälukeminen tuo tutkijan työkalupakkiin uusia mahdollisuuksia. Tämä ei koske vain ennalta asetettujen tutkimuskysymysten ratkaisemista vaan menetelmät laajentavat tutkijoiden mielikuvitusta siitä, millaisia kysymyksenasetteluja tulevai-

[40] Ks. esim. Hakkarainen & Iftikhar 2020, 259–278; Fridlund, Oiva & Paju 2020; Oiva et al. 2020. Ks. myös The Programming Historian, https://programminghistorian.org/. Haettu 9.12.2020.

[41] Rantala et al. 2019, 53–67; Salmi et al. 2021.

[42] Clavert 2012.

suudessa voitaisiin esittää, ja siten ulottaa historiantutkimuksen katsetta uusille alueille.

Lopuksi

Tämä artikkeli on keskittynyt Kansalliskirjaston digitoituun sanomalehtikirjastoon esimerkkinä siitä, millaisia menetelmällisiä lähestymistapoja sekä metodologisia haasteita digitoituihin lehdistöaineistoihin liittyy. Tutkimuskysymyksestä ja ongelmanasettelusta riippuen digitoidut sanoma- ja aikakauslehdet tarjoavat kiinnostavaa lähdemateriaalia erityyppisiin tutkimuksiin. Tutkijan on hyvä olla tietoinen paitsi digitoitujen aineistojen mahdollisuuksista myös niistä rajoituksista, joita aineistojen käyttöön liittyy. Samanaikaisesti digitaalinen kokoelma tarjoaa sellaisia tutkimuspolkuja, joita paperisten tai mikrofilmattujen lehtien käyttäjillä ei ollut.

Olemme tuoneet esiin digitoitujen lehtiaineistojen kytkennät paperilehtien sekä mikrofilmien materiaalisuuteen ja selvittäneet digitoimisprosessin eri haasteita. Optisen tekstintunnistuksen eli OCR-tekstin laatu on keskeinen aineistojen käyttöä määrittävä tekijä. Sana- ja fraasihaut, joita on mahdollista tehdä optisesti tekstintunnistettuun tekstiin, ovat käytetyin tapa hakea tietoa aineistosta. Sanahakuja on kuitenkin syytä käyttää kriittisesti ja läpinäkyvästi osana tutkimusprosessia. Oman korpuksen suodattaminen ja kokoaminen tekstimassasta tarjoaa puolestaan mahdollisuuden syventyä tarkemmin omalle tutkimukselle keskeiseen aineistoon ja luoda tulkinta siitä. Digitaalisten sanoma- ja aikakauslehtien käytössä eri lukemisen tavat ja tekniikat kulkevatkin usein rinnakkain tukien toisiaan. Perinteisen lähiluvun lisäksi suuria digitoituja aineistoja voidaan lähestyä tietokonepohjaisin menetelmin, joita kehitetään koko ajan. Uudenlaisia näkökulmia tutkittavaan aineistoon saadaan nimenomaan laskennallisia ja laadullisia etä- ja lähiluvun menetelmiä yhdistämällä. Kyse on ihmisen ja koneen yhteistyöstä, jossa ihmisen halu tietää, hänen tekemänsä valinnat ja tulkinnat ohjaavat tutkimusprosessia. Kysymys on lopulta luovuudesta, uusien kysymysten esittämisestä ja niiden ratkaisemiseen sopivien menetelmien kehittämisestä.

Tutkimuskirjallisuus

Bastiansen, Henrik Grue: Nettarkivenes digitale objekter som intellektuell utfordring: Mediehistorie 2.0 og behovet for en ny filologi. *Mediehistorisk Tidskrift* 1 (2020), 33–46.

Beals, M. H. & Bell, Emily: *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Loughborough 2020. DOI:10.6084/m9.figshare.11560059

Broersma, Marcel & Harbers, Frank: Exploring Machine Learning to Study the Long-Term Transformation of News. *Digital Journalism* 6:9 (2018), 1150–1164. DOI: 10.1080/21670811.2018.1513337

Bødker, Henrik: Journalism History and Digital Archives. *Digital Journalism* 6:9 (2018), 1113–1120. DOI: 10.1080/21670811.2018.1516114

Clavert, Frédéric: *Lecture des sources historiennes à l'ère numérique*, 14.11.2012, http://www.clavert.net/wordpress/?p=1061. Haettu 9.12.2020.

Cordell, Ryan: "Q i-jtb the Raven": Taking dirty OCR seriously. *Book History* 1 (2017), 188–225.

Elo, Kimmo: Digitaalisen historiantutkimuksen kenttää louhimassa. Teoksessa Kimmo Elo (toim.) *Digitaalinen humanismi ja historiatieteet*. Turun historiallinen yhdistys 2016, 11–35.

Federico, Annette: *Engagements with Close Reading*. Routledge 2016.

Fridlund, Mats, Oiva, Mila & Paju, Petri (toim): *Digital Histories: Emergent Approaches within the New Digital History*. Helsinki University Press 2020. DOI: https://doi.org/10.33134/HUP-5

Graham S., Weingart, S. & Milligan, I.: Getting started with topic modeling and MALLET. *The Programming Historian* 1 (2012), https://programming-historian.org/lessons/topic-modeling-and-mallet. Haettu 12.3.2021.

Hakkarainen, Heidi & Iftikhar, Zuhair: The Many Themes of Humanism: Topic Modelling Humanism Discourse in Early 19th-Century German-Language Press. Teoksessa Mats Fridlund, Mila Oiva & Petri Paju (toim.) *Digital Histories: Emergent Approaches within the New Digital History*. Helsinki University Press 2020. DOI: https://doi.org/10.33134/HUP-5

Hoekstra, Rik & Koolen, Marijn: Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52:2 (2018), 79–94. DOI:10.1080/01615440.2018.1484676

Huistra, Hieke & Mellink, Bram: Phrasing history: Selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49:4 (2016), 220–229. DOI: https://doi.org/10.1080/01615440.2016.1205964

Jarlbrink, Johan & Snickars, Pelle: Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive. *Journal of Documentation* 73:6 (2017), 1228–1243. DOI: https://doi.org/10.1108/JD-09-2016-0106

Jensen, Helle S.: Digital Archival Literacy for (all) Historians. *Media History* 27:2 (2021), 251–265. DOI: https://doi.org/10.1080/13688804.2020.1779047

Kokko, Heikki: Digitaalisten aineistojen sanomalehtihakemiston historiasta. *Scripta selecta – Kirjoituksia Kansalliskirjaston kokoelmista* 2017, http://blogs.helsinki.fi/scriptaselecta/2017/07/13/digitaalisten-aineistojen-artikkelihakemiston-historiasta/. Haettu 9.12.2020.

Koolen, Marijn, van Gorp, Jasmijn & van Ossenbruggen, Jacco: Toward a model for digital toolcriticism: Reflection as integrative practice. *Digital Scholarship in the Humanities* 34:2 (2019).

Maurantonio, Nicole: Archiving the Visual: The Promises and Pitfalls of Digital Newspapers. *Media History* 20:1 (2014), 88–102. DOI: https://doi.org/10.1080/13688804.2013.870749

Moretti, Franco: Conjectures on world literature. *New Left Review* 1:1 (2000), 54–68.

Moretti, Franco: *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso 2005.

Nicholson, Bob: The Digital Turn: Exploring the methodological perspectives of digital newspaper archives. *Media History* 19:1 (2013), 59–73. DOI: https://doi.org/10.1080/13688804.2012.752963

Oiva, Mila, Nivala, Asko & Salmi, Hannu: Digitized Newspapers at the National Library of Finland, Suomen kansalliskirjaston digitoidut sanomalehdet. *Oceanic Exchanges*, Helmikuu 20, 2018, https://oceanicexchanges.org/2018-02-20-data-reports-finland/. Haettu 8.11.2020.

Oiva, Mila, Nivala, Asko, Salmi, Hannu, Latva, Otto, Jalava, Marja, Keck, Jana, Martínez Domínguez, Laura & Parker, James: Spreading News in 1904: The Media Coverage of Nikolay Bobrikov's Shooting. *Media History* 26:4 (2020), 391–407. DOI: https://doi.org/10.1080/13688804.2019.1652090

Paju, Petri: The long road to 'digital history': History of computer-assisted research of the past in Finland since the 1960s. Teoksessa Mats Fridlund, Mila Oiva & Petri Paju (toim.) *Digital Histories: Emergent Approaches within the New Digital History*. Helsinki University Press 2020, 21–44. DOI: https://doi.org/10.33134/HUP-5

Paju, Petri, Rantala, Heli & Salmi, Hannu: Tietokannoista tulkintoihin: Digitaalisen historiantutkimuksen käytäntöjä. *Ennen ja nyt* 2 (2019).

Piotrowski, Michael: Historical Models and Serial Sources. *Journal of European Periodical Studies* 4:1 (2019), 8–18.

Prescott, Andrew: Searching for Dr. Johnson: The Digitization of the Burney Newspaper Collection. Teoksessa Siv Gøril Brandtzæg, Paul Goring & Christine Watson (toim.) *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century*. Brill 2018, 51–71. DOI: https://doi.org/10.1163/9789004362871

Rantala, Heli: Global 'revolution' in the early nineteenth-century Finnish press. *History of European Ideas* 45:5 (2019), 721–736. DOI: 10.1080/01916599.2018.1558908

Rantala, Heli, Nivala, Asko, Salmi, Hannu, Paju, Petri, Sippola, Reetta, Vesanto, Aleksi & Ginter, Filip: Tekstien uudelleenkäyttö suomalaisessa sanoma- ja aikakauslehdistössä 1771–1920. Digitaalisten ihmistieteiden näkökulma. *Historiallinen Aikakauskirja* 1 (2019), 53–67.

Salmi, Hannu, Paju, Petri, Rantala, Heli, Nivala, Asko, Vesanto, Aleksi & Ginter Filip: The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54:1 (2021), 14–28: DOI: https://doi.org/10.1080/01615440.2020.1803166

Salmi, Hannu: *What is Digital History*? Polity 2020.

Tolonen, Mikko & Lahti, Leo: Digitaaliset ihmistieteet ja historiantutkimus. Teoksessa Matti O. Hannikainen, Mirkka Danielsbacka & Tuomas Tepora (toim.) *Menneisyyden Rakentajat: Teoriat Historiantutkimuksessa*. Gaudeamus 2018.

Tolonen, Mikko & Lahti, Leo: Aatehistoria ja digitaalisten aineistojen mahdollisuudet. *Ennen ja nyt* 2 (2015).