KATSAUS

Kansalliskirjaston historialliset sanoma- ja aikakauslehdet avoimena digitaalisena datana - datapaketteja, rajapintoja, käyttäjiä ja tutkimusongelmia

Kimmo Kettunen

Kansalliskirjasto

kimmo.kettunen@helsinki.fi

https://orcid.org/0000-0003-2747-1382

Tuula Pääkkönen

Kansalliskirjasto

tuula.paakkonen@helsinki.fi

https://orcid.org/0000-0003-3958-9732

Tässä artikkelissa luodaan katsaus Kansalliskirjaston digitoitujen lehtiaineistojen avoimen datan tutkimuskäyttöön. Lehtiaineistoista julkaistiin vuonna 2017 vuodet 1771–1910 kattava datapaketti, ja sen tutkimuskäytöstä on kertynyt tähän mennessä hiukan yli vuoden kokemus. Sivuamme katsauksessa myös aineiston verkkokäyttöä tutkimuksessa. Esittelemme lisäksi myös ohjelmistorajapintoja, joiden kautta aineistoihin pääsee käsiksi.

Asiasanat: historialliset sanomalehdet; digitoidut tekstiaineistot; avoin tiede

Kansalliskirjaston Mikkelin toimipiste on digitoinut Suomessa julkaistuja sanoma- ja aikakauslehtiä sekä pienpainatteita vuodesta 1998 alkaen. Aineisto on käytettävissä Kansalliskirjaston digi.kansalliskirjasto.fi (Digi) haku- ja esitysjärjestelmässä[1]. Lehtiaineiston määrä on kasvanut palvelussa vuosien aikana tasaisesti, samoin aineiston käyttäjien määrä. Alkuun lehtiaineistoa

1 http://digi.kansalliskirjasto.fi

Pysyvä osoite: https://doi.org/10.23978/inf.77412

Kuva 1: Digi.kansalliskirjasto.fi:n etusivu

oli verkkopalvelussa vapaasti saatavilla vuosilta 1771–1910. Vuoden 2017 helmikuun alussa aineisto avattiin saataville vuoden 1920 loppuun saakka. Vuoden 2018 alusta avoimuuden aikajana sai yhdeksän lisävuotta vuoden 1929 loppuun. Koko aineisto vuosilta 1771–1910 on saatavilla myös avoimen datan jakelupakettina verkkopalvelun sivuilta (Pääkkönen & Kervinen, 2016). Aineistossa on periaatteessa kaikki julkaistut lehdet, mutta joistain lehdistä saattaa puuttua yksittäisiä numeroita, joita ei ole kirjaston kokoelmassa. Myös lehden huono kunto on saattanut estää digitoinnin. Puuttuvat niteet ilmenevät Digin hakujärjestelmän lehtikohtaisessa näkymässä päivämäärän oranssina merkintänä.

Kansalliskirjasto ilmoittaa verkkosivuillaan digitoitujen sivujen kokonaismääräksi vuoden 2018 toukokuun lopulla 13 427 754. Luku sisältää sanoma- ja aikakauslehdet sekä pienpainatteet. Vapaasti käytettävissä lehdissä vuosilta 1771–1929 on yhteensä noin 4.07 miljoonaa sivua sanomalehtiä ja noin 3.37 miljoonaa sivua aikakauslehtiä. Sanomalehtiä palvelussa on kirjoitushetkellä 892 nimikettä, aikakauslehtiä 3763. Digitoitu kokoelma on karttunut viime vuosina vuosittain 1–2 miljoonalla sivulla. Kuva 1 esittää kokoelman käyttöliittymän aloitussivua.

Kansalliskirjaston digitoitu lehtikokoelma on osa maailmanlaajuista laajenevaa historiallisten lehtien digitoitua tarjontaa. Vuonna 2012 Europeana-projekti arvioi Euroopassa olevan digitoituja sanomalehtiä noin 129 miljoonaa sivua ja noin 24 000 nimikettä (Dunning, 2012). Samana vuonna arvioitiin, että vain noin 17 prosenttia Euroopan kirjastojen sanomalehdistä on digitoitu (Stroeker & Vogels, 2012). Varovainen arvio digitoitujen lehtinimikkeiden määrästä maailmanlaajuisesti vuonna 2015 oli yli 45 000. Suurin osa aineistosta on

digitoitu Euroopassa ja Yhdysvalloissa (The "State of the Art", 2015). Aineistojen tämänhetkinen määrä on varmasti paljon suurempi eri puolilla tehdyn digitoinnin ansiosta. Jo Pohjoismaissa aineistoja on avoinna noin 84 miljoonaa sivua (Pääkkönen, Rautiainen, Ryynänen, & Uusitalo, 2018).

Sanomalehtiaineisto avoimena datana

Avoimella datalla tarkoitetaan yleensä digitaalisessa muodossa olevia sisältöjä tai dataa, joita kuka tahansa voi käyttää vapaasti ja maksutta, sekä muokata ja jakaa uudelleen mihin tahansa käyttötarkoitukseen[2]. Aineiston käyttöä digitaalisten aineistojen esitysjärjestelmän kautta ei aina pidetä avoimen datan käyttönä, mutta se muodostaa kuitenkin tärkeän osan aineiston käytöstä. On myös huomattava, että tämä käyttö mahdollistaa aineistoihin tutustumisen ja saattaa johtaa laajempaan avoimen datan käyttämiseen. Sen vuoksi esittelemme myös tutkimuskäyttöä verkkoaineiston esitysjärjestelmän kautta. Samasta syystä esittelemme myös aineiston käyttöön liittyviä ohjelmistorajapintoja.

Kansalliskirjasto on strategiassaan vuosille 2016–2020 (Tuori, 2016) määritellyt tietovarantojen avoimuuden tärkeäksi tavoitteekseen. Kirjastossa on luotu sekä Avoin Kansalliskirjasto (Tuori, 2016) että Digitaalisen humanismin -politiikka, joilla pyritään lisäämään avoimuutta ja tutkimusyhteistyötä. Suunta on yleinen kautta maailman (Terras, 2015). Kun alkuperäinen digitoinnin jälkikäsittelyn lopputulos tarjotaan tutkijoille ja kehittäjille käytettäväksi, on helpompi löytää uusia tutkimus- ja käyttömahdollisuuksia, joilla voisi esimerkiksi parantaa aineiston laatutasoa tai käytettävyyttä. Tavoitteena on tarjota laajoja aineistoja, joista voidaan tehdä kvantitatiivista analyysiä, mikä mahdollistaa tutkimukselle oleellisen informaation löytämisen (Tolonen & Lahti, 2015). Aineiston saatavuus hyödyttää muun muassa digitaalisten ihmistieteiden tutkimusta.

Vuoden 2016 aikana suunniteltu ja toteutettu digitoitujen historiallisten sanoma- ja aikakauslehtien avoin aineistopaketti[3] koostuu ALTO XML-tiedostoista (ALTO: Technical Metadata for Layout and Text Objects). Suomessa julkaistuista sanomalehdistä vuosilta 1771–1910 ja aikakauslehdistä vuosilta 1816–1910 (Pääkkönen ym., 2016). Tiedostot on jaettu datapaketissa lehtien ilmestymisvuosien mukaan eri jaksoihin, ja yksittäisten ladattavien aineistopakettien koko on pyritty pitämään kohtuullisena. Kokonaisuus on kuitenkin laaja, noin 193 gigatavua. Paketissa yksi tiedosto vastaa julkaistun lehden sivutiedostoa. Sivukohtaiseen ALTO XML -tiedostoon on liitetty sivun

2 https://fi.okfn.org/about/visiojaarvot/

3 https://digi.kansalliskirjasto.fi/opendata/submit

Kuva 2: Avoimen datan sivutiedoston rakenne

```
<metadata>
      <title>Uusi Suometar</title>
      <identifier type="issn">1457-4721</identifier>
      <published format="edtf">1871-07-03</published>
      <issue>77</issue>
   <pageOrder>1</pageOrder>
   <pageLabel></pageLabel>
      <language>fin</language>
      <contentType>newspaper</contentType>
      <originalPublisher/>
     <latestPublisher>Uuden Suomettaren Oy</latestPublisher>
   <publishingPlace country="fi">Helsinki</publishingPlace>
      <copyright>Copyrights expired</copyright>
      <license>ESIVERSIO - EI EDELLEENJAKELUUN</license>
   <pageIdentifier>1830866</pageIdentifier>
   <bindingIdentifier>427528</bindingIdentifier>
      <imageURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/427528/image/1</imageURL>
      <pdfURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/427528/pdf</pdfURL>
      <browseURL>http://digi.kansalliskirjasto.fi/sanomalehti/binding/427528/#?page=1</browseURL>
      <pingURL/>
      <ocrVersion/>
      <lastModified>2007-11-01T00:11:00</lastModified>
   </metadata>
```

Kuva 3: Jakelupakettien metadata

metatiedot sekä pelkkä sivun tekstisisältö ilman XML-koodausta.[4] Kuvassa 2 on esitetty sivutiedoston rakenne.

Kuvassa 3 on esitetty jakelupakettien metadatan sisältö.

4 Aineiston vuodet 1911-1920 ovat saatavissa ALTO XML-muodossa Kielipankista: `https://korp.csc.fi/download/Digilib/1875_1920/`

Metadatassa kuvataan julkaisutietojen lisäksi kielitieto sekä aineiston sijainti kirjaston esitysjärjestelmässä. Aineiston metadata on pyritty saamaan mahdollisimman virheettömäksi, mutta jakelupaketin koostamisessa havaittiin virheitä muun muassa kielimerkinnöissä. Havaitut virheet pyritään korjaamaan, mutta korjausten näkyminen aineistossa voi kestää.

Tässä artikkelissa luodaan pääasiassa katsaus Kansalliskirjaston digitoitujen lehtiaineistojen avoimen datan tutkimuskäyttöön. Aineiston datapaketin tutkimuskäytöstä on kertynyt tähän mennessä hiukan yli vuoden kokemus, joten saadut kokemukset ovat alustavia. Aineiston rakennetta ja tiedostojakoa on kuvattu tarkemmin artikkelissa Pääkkönen ja kumppanit (2016) sekä vuoden 2016 Informaatiotutkimuksen päivien esityksessä[5], joten emme käsittele niitä tässä katsauksessa. Sivuamme katsauksessa myös aineiston verkkokäyttöä tutkimuksessa. Esittelemme lisäksi myös ohjelmistorajapintoja, joiden kautta aineistoihin pääsee käsiksi.

Avoin data ja tutkimuskäyttö

Aineiston tutkimuskäyttö jakaantuu karkeasti ottaen muutaman suuremman projektin ja yksittäisten tutkijoiden välille[6]. Havaintojemme mukaan erityisesti erilliset aineistopaketit ovat synnyttäneet uusia mahdollisuuksia tutkijoille. Esimerkiksi viimeisimmässä Digital Humanities in the Nordic Countries 2018 -konferenssissa[7] vanhoja digitoituja sanomalehtiä käytettiin useassa eri tutkimuksessa monessa eri yliopistossa. Suuri osa aineiston käyttäjistä oli historioitsijoita, jotka tutkivat eri teemoja. Kuvassa 4 esitetään aineiston latausikkuna, ja kuvassa 5 esitetään datapaketin lataukset 6. huhtikuuta 2018 mennessä.

Aineistoa käyttävät projektit

COMHIS

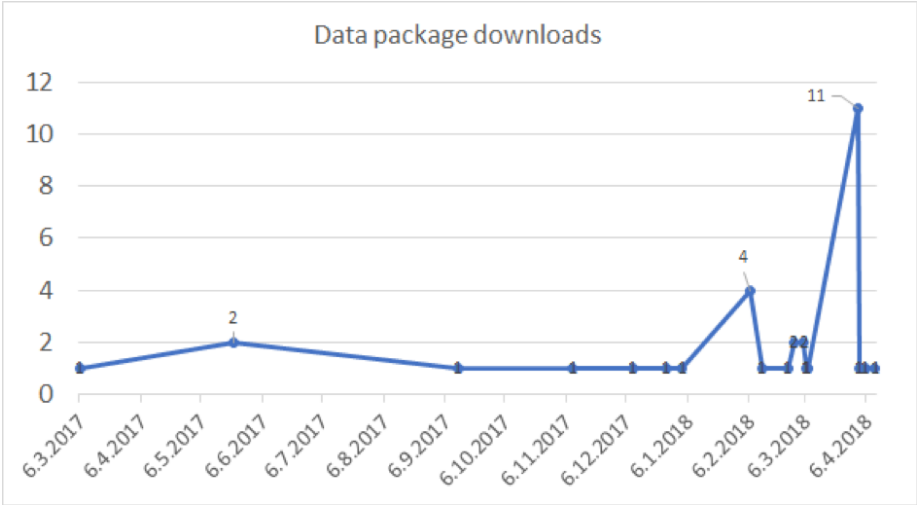
Tähän mennessä aineistopaketin keskeinen käyttäjä on ollut tutkimuskonsortio *Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910*

5 https://journal.fi/inf/article/view/59442

6 Tietomme aineiston tähänastisesta käytöstä ei kata kaikkea käyttöä. Tutkijat ilmoittavat aineistomme käytöstä julkaisuissaan tai raporteissaan, mutta käyttötietojen keskitetty kerääminen ei ole toistaiseksi ollut mahdollista. Jonkin verran olemme onnistuneet keräämään tietoa aineistojen käytöstä sivulle digi.kansalliskirjasto.fi.

7 https://www.helsinki.fi/en/helsinki-centre-for-digital-humanities/dhn-2018

Kuva 4: Avoimen datan lataaminen: https://digi.kansalliskirjasto.fi/opendata/submit



Kuva 5: Datapaketin lataukset 6.4. 2018 mennessä

(COMHIS)[8], joka on Suomen Akatemian rahoittama Helsingin yliopiston, Turun yliopiston ja Kansalliskirjaston yhteinen projekti. Projektissa on tutkittu muun muassa uutisten leviämistä eri puolella Suomea 1800-luvulla saman uutisen toistoina tai hiukan muokattuina versioina (Ginter ym., 2018; Vesanto, Nivala ym., 2017; Vesanto ym., 2017).

Menetelmällisesti projekti on ollut kiintoisa: uutisten toistoja on kyetty löytämään OCR-laadun ongelmista huolimatta käyttämällä bioinformatiikkaan kehitettyä proteiinien ja nukleotidien sekvenssien samankaltaisuuden sekvensointiohjelmaa NCBI BLASTia tekstien samankaltaisuuden toteamiseen (Vesanto, Nivala ym., 2017). BLAST on osoittautunut tehokkaaksi välineeksi tekstien samankaltaisuuksien etsimisessä: parhaimmillaan se löytää yhteneviä tekstejä, jotka eroavat toisistaan merkkitasolla 60-prosenttisesti. BLAST vaatii tosin myös hyvät it-resurssit: parhaiten sen ajaminen isoilla tekstimäärillä onnistuu Tieteen tietotekniikan keskuksen CSC:n supertietokoneympäristössä. Toistuvia tekstejä on erotettu kolmea erilaista tyyppiä: mainokset, pitkän aikavälin jälkeen uudelleen julkaistavat tekstit (esimerkiksi isänmaallinen ylioppilaslaulu "Ännu på tidens mörka vågor") sekä todelliset viraalit uutiset, jotka ovat levinneet ensimmäisen julkaisemisensa jälkeen laajasti ja nopeasti ympäri Suomea eri lehdissä. (Vesanto ym., 2017). Esimerkiksi tieto Henrik Gabriel Porthanin patsaan julkistamisesta levisi elokuussa 1864 Suomalaisen Kirjallisuuden Seuran Suomenmaalle lähettämän kirjeen uudellenjulkaisuina 18 lehdessä nopeasti kautta koko Suomen (Salmi, 2018).[9]

COMHISin toisessa osiossa Helsingin yliopiston tutkijat ovat tutkineet paljon sanomalehtiaineiston metadataa. Tutkimuksessa on tähän mennessä saatu tuloksia muun muassa lehtien julkaisuformaateista (sivukoko ja -määrä), julkaisutiheydestä, lehtien levikistä, nimikemääristä, suomen ja ruotsinkielisten julkaisujen määrästä sekä painopaikoista. Tämän tiedon perusteella aineistosta muodostetaan alikorpuksia, joiden avulla tutkitaan tarkemmin julkisen keskustelun muokkautumista ja muuttumista lehdistössä. (Tolonen, Marjanen, Roivainen, & Lahti, 2017).

Kansalliskirjaston COMHISin osuudessa on tutkittu muun muassa tekstin optisen luvun laadun parantamista sekä erisnimien eristämistä aineistosta (Koistinen, Kettunen, & Pääkkönen, 2017; Ruokolainen & Kettunen, 2018). Tässä projektin osuudessa pyritään erityisesti parantamaan aineiston käytettävyyttä.

---

8 https://www.utu.fi/fi/yksikot/hum/yksikot/kulttuurihistoria/tutkimus/Sivut/comhis.aspx

9 Lehtiaineiston toistojen hakutietokanta on osoitteessa http://comhis.fi/clusters

Lisäksi on syytä mainita Suomen Akatemian rahoittama projekti *Supporting Evolving Search Tasks in Digital Environments via Fuzzy String Matching* (Kumpulainen & Keskustalo, 2017). Siinä tutkitaan informaatiotutkimuksen näkökulmasta erityisesti historioitsijoiden tarpeita digitoitujen aineistojen sekä myös digitoitujen vanhojen sanomalehtiaineistojen käytössä. Projektin tarkoitus on tuottaa menetelmiä ja välineitä, jotka tukevat historioitsijoita tutkimustehtävissä, joissa aineisto muodostuu erityisesti historiallisista digitaalisista lehtiaineistoista (Emt.).

Kansainväliset projektit

Kansalliskirjaston aineistoa käytetään myös kansainvälisessä projektissa *Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840–1914* (Beals ym., 2017).[10] Projektissa on yhdeksän osapuolta kuudesta maasta: Hollannista, Isosta-Britanniasta, Meksikosta, Saksasta, Suomesta ja Yhdysvalloista. Projektissa pyritään avaamaan 1800-luvun ja 1900-luvun alun uutisvirtoja maiden ja mantereiden välillä rajoista ja kielistä riippumatta. Projekti alkoi vuonna 2017 ja päättyy vuonna 2019.

Toinen kansainvälinen hanke, joka tulee käyttämään sanomalehtien aineistopakettia on toukokuussa 2018 aloittanut NewsEye-projekti, joka saa rahoituksensa EU:n Horizon-ohjelmasta. Hankkeessa on mukana tietojenkäsittelyn ja digitaalisten ihmistieteiden tutkijoita ja kirjastoja Itävallasta (Innsbruck ja Wien), Ranskasta (La Rochelle, Pariisi), Saksasta (Rostock), sekä Suomesta (Helsingin yliopisto ja Kansalliskirjasto). Kolmevuotisessa projektissa kehitetään tutkijoille menetelmiä ja välineitä laajojen digitoitujen lehtiaineistojen käyttöön.

Yksittäiset historioitsijat

Monet yksittäiset historiantutkijat käyttävät Kansalliskirjaston lehtiaineistoa suoraan Digin verkkopalvelussa.Yksittäisistä tutkijoista ja heidän tutkimuksistaan voidaan mainita ajallisuuden käsitettä sosialistisesta lehdistöstä tutkiva Risto Turunen (Turunen, 2018), niin sanottua puolukkaryntäystä tutkiva Matti La Mela (2016, 2018) sekä Heikki Kokko, joka on tutkinut väitöskirjassaan ihmiskäsityksen muutosta suomenkielisen kansanosan kulttuurissa 1800-luvun puolivälissä (Kokko, 2016). Lisäksi Kokko on aloittamassa projektia sanomalehtien lukijakirjeistä 1800-luvulla. Onni Pekonen on väitöskirjassaan (Pekonen, 2014) tarkastellut suomalaisen parlamentaarisen elämän kehitystä 1800-luvun lopun säätyvaltiopäivillä ja 1900-luvun alun yksikamarisessa eduskunnassa.

10 http://osf.io/wa94s

Paketin vaikutukset

Avoimen datan aineistopakettien vaikutuksesta tutkimukseen ja sen suuntautumiseen on toistaiseksi liian varhaista sanoa mitään kovin laajamittaista. Tutkimuksia on käynnistynyt ja uusia seuraa, ja vasta pitemmällä ajalla näkyy, mitä vaikutuksia aineistoilla on ollut ja miten aineistot ovat esimerkiksi suunnanneet tutkimusta tai muokanneet tietämystä. Selvää kuitenkin on, että laajojen aineistojen saatavuus digitaalisessa muodossa muuttaa tapaa tehdä tutkimusta: vähintäänkin tutkimuksen tekemisen tehokkuus ja työskentelyolot paranevat, kun ei tarvitse käyttää fyysistä alkuperäisaineistoa (Gooding, 2017). Luultavimmin myös tutkimuksen sisällöt ja tavat tehdä tutkimusta muuttuvat.

Vaikutus aineistoon

Aineistopaketin ensimmäinen versio *Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910* -projektin tutkimusryhmän käyttöön toi Kansalliskirjastolle näkyviä hyötyjä. Aineiston läpikäynti auttoi löytämään metadatan ja sisältöjen puutteita ja epätarkkuuksia, mikä auttaa aineiston laadun parantamisessa. Uuteen verkkopalvelun versioon, joka julkaistiin syksyllä 2016, metadatan epätarkkuuksia korjattiin ja korjauksia jatketaan resurssien salliessa. Aineistosta on valmiina myös ajanjakson 1911–1918 datapaketti, ja se julkaistaan myöhemmin vuonna 2018.

Palautetta aineistopaketeista

Yksi aineistopakettien haasteista on niiden suuri koko – voisi olla hyvä jakaa aineistopaketteja pienemmiksi kokonaisuuksiksi, esimerkiksi hyödyntäen aineiston lähdekieltä. Esimerkiksi DHN2018-konferenssissa käyttäjät olivat pohjoismaisia, joten heillä oli toiveita pääosin ruotsinkielisistä aineistoista. Halutun kielisen aineiston löytäminen vaatii nyt aineistopakettien purkamista ja selailua kielen perusteella. Toisaalta jos käyttäjä haluaa hyödyntää kaikkea aineistoa, nykyisen tyyppiset aineistopaketit voivat olla soveltuvimmat. Käyttötapauksia on kuitenkin erilaisia, joten joudumme ehkä harkitsemaan sekä kielikohtaisia että vuosiväleittäin jaettuja paketteja. Tämä jakelumalli vaatii säilytystilaa ja muita resursseja enemmän, mutta tutkijan käyttökokemus voi jaotellussa mallissa olla parempi.

Aineistopaketteja tulisi myös tehdä lisää sitä mukaa kuin digitointi etenee vuosittain. Jos aineistojen käyttöoikeudet mahdollistavat jakelun, tutkijoille voidaan antaa aineistopaketteja tutkimuksen raaka-aineistoksi.

Rajapintoja ja käyttömahdollisuuksia

Kansalliskirjaston digitoituja lehtiaineistoja voi käyttää myös muilla tavoin kuin käyttämällä Digin esitysjärjestelmää tai lataamalla itselleen aineistopaketteja. Esittelemme lyhyesti rajapinnat, jotka mahdollistavat esimerkiksi aineiston täsmäkäytön. Lisäksi kuvaamme Kansalliskirjaston datakatalogia.

Digi.kansalliskirjasto.fi -palvelussa on nyt käytettävissä OAI-PMH- ja OpenURL-rajapinnat, joiden avulla voi selata metadataa tai päästä tiettyyn aineiston osaan[11]. OAI-PMH on kirjastoalalla käytetty yleinen haravointirajapinta, jonka avulla erilaiset hakujärjestelmät kuten Europeana voivat näyttää aineistoja usean kirjaston kautta. Europeana on käyttänyt Digin aineistoja osin OAI-PMH-rajapintaa hyödyntäen, ja osin aiemmin luotujen aineistopakettien kautta, joihin valittiin otos suomalaisista sanomalehdistä.

OAI-PMH

OAI-PMH[12] on rajapinta, jota tutkijat eivät ole hyödyntäneet Digissä - sen pääkäyttäjä ovat olleet Europeana ja kirjaston sisäiset järjestelmät. Sen kautta on mahdollista hakea tietoja iteratiivisesti: ensin haetaan lista tiedoista, joita on aineistotyyppikohtaisesti (sanomalehti tai aikakauslehti). Sen jälkeen tietueen tunnisteen avulla saadaan perusmetatiedot jokaiselle lehden numerolle. Aineistomäärän takia rajapinta palauttaa kerralla sata nidettä, jonka jälkeen haluamaansa kyselyä voi jatkaa jatkokoodilla (*resumption token*) eteenpäin. Parhaiten rajapinta soveltuu hakemaan uusia tietueita annetun päivämäärän jälkeen kaikkien tietueiden läpikäymisen sijaan. Kuvassa 6 on OAI-PMH-rajapinta.

OpenURL

OpenURL[13] on rajapinta, jonka avulla käyttäjä voi navigoida aineistoon helposti tietäessään tietyn metadatan lehden numerosta.[14] Lehden numeron avulla on mahdollista päästä muihinkin lehden alikomponentteihin, kuten sivutekstiin tai ALTO XML-muotoiseen tiedostoon, jossa on talletettuna lehden rakenne

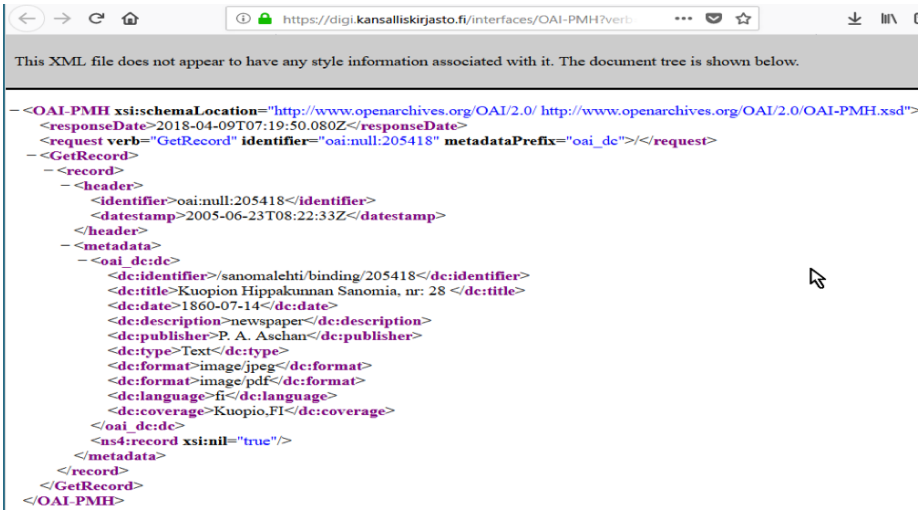
---

11 https://wiki.helsinki.fi/display/Comhis/Interfaces+of+digi.kansalliskirjasto.fi

12 https://www.openarchives.org/pmh/

13 https://www.niso.org/publications/z3988-2004-r2010-openurl-framework-context-sensitive-services

14 OpenURL-linkillä http://digi.kansalliskirjasto.fi/openurl/query.html?genre=journal&date=1888-01-03&issn=0355-6913&spage=2 käyttäjä ohjataan Aamulehden (ISSN: 0355-6913) päivämäärän 3.1.1888 lehteen, sivulle 2. Linkki ohjaa Digiin suoraan lehden sivukuvaan huolimatta siitä, mikä on lehden nidetunnus (id), joka on Digissä sisäisesti käytetty tunniste jokaiselle lehden numerolle.



Kuva 6: OAI-PMH-rajapinta

```
http  -b  https://digi.kansalliskirjasto.fi/sanomalehti/binding/379687/page-1.txt  |
head -n 30

Tillusktnta: Tampereella j<ilopaifo'»sll . . . . 4.s« 3 . c» 1 " Tampereen poltilonttoorl^a . . b^. 4' 84. I,i? Muissa mcamme
postilaitoksissa 5 88. 4. >« !>< Venäjällä nstistt lähetettynä 4 r. ?öt. «r bal. . ?. 1?.2o . Ytsityiset rumerot matsa»at 10 pe,

Ashteä kauptngiVfa tilctct. jakaa ja myy: Kosken länsipuoleUll : I. F. Olenin lir!°l«.p°i«sa. «°up. Paperit.".. Äug Helen l» I.
H°8»»n Punpulit. Suutari Partanen viaat N. Gcanbcrg ja H. Malinen Kuninkaan!, »«relli, G. Lindberg Sllenlulsen tai. Läntisen
pllllllnl. w°rrella.
```

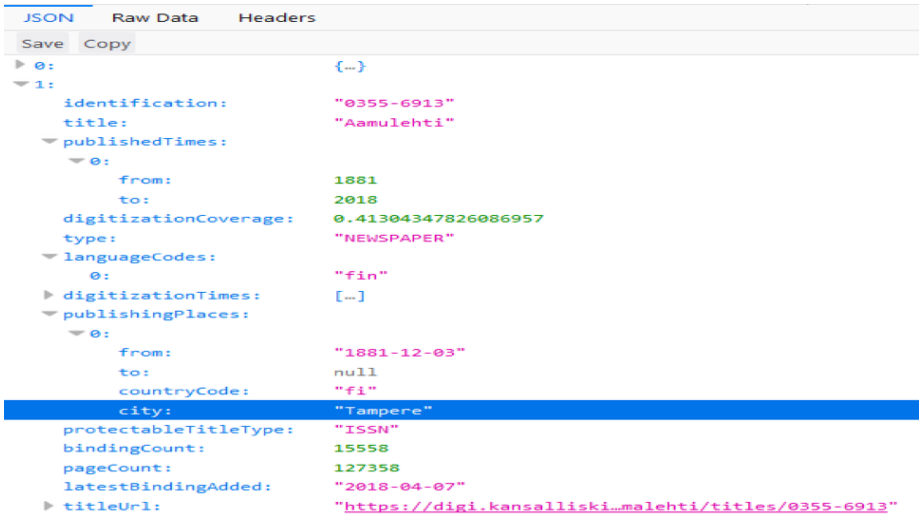
Kuva 7: Sivutekstin ja ALTO:n hakeminen OpenURL:illa

ja sivuteksti siinä muodossa missä jälkikäsittelyn tekstintunnistus on saanut tekstin tunnistettua. Esimerkki sivutekstin ja ALTO:n hakemiseksi suoraan on kuvassa 7.

JSON

Kaikkien lehtien metatiedot on mahdollista saada digi.kansalliskirjasto.fi -palvelun 'Lehdet'-näkymästä[15] sanomalehdistä tai aikakauslehdistä JSON-muodossa[16]. Tiedot sisältävät mm. julkaisupaikan, ilmestymisvuodet ja digitointikattavuuden suhteessa lehden sivujen laskettuun kokonaismäärään. Kansalliskirjasto käyttää Lehdet-näkymän JSON-rajapintaa kansalliskirjas-

15 https://digi.kansalliskirjasto.fi/api/newspaper/titles?language=fi
16 https://www.json.org/



Kuva 8: Metatiedot JSON-muodossa

to.finna.fi -palvelun lehtilistauksien näyttämiseen. SFX-linkitysjärjestelmän lehtitiedot on päivitetty vuonna 2018 tämän rajapinnan avulla. Kuvassa 8 näytetään metatietoja JSON-muodossa.

Kansalliskirjaston datakatalogi

Kansalliskirjasto avasi vuoden 2017 lopussa datakatalogin, jossa on tietoja rajapinnoista sekä tietoaineistoja eri aineistotyypeistä ja kokoelmista[17]. Eri aineistoista on lyhyet esittelykortit, joissa kerrotaan aineiston määrä, käyttöehdot ja kuinka aineiston saa käyttöönsä. Kokotekstiaineistoja on 17 kappaletta, metadataa kuudessa palvelussa ja erilaisia sanastoja kymmenen. Tavoitteena on, että datakatalogi toimii yhtenä paikkana, josta on helposti löydettävissä kaikki digitaalisena tarjolla olevat aineistopaketit. Suunnitelmissa on myös lisätä aineistoja datakatalogiin sitä mukaa kuin soveltuvia aineistojoukkoja tulee tarjolle. Toisaalta käytettävät aineistot voivat myös muuttua jos aineiston käyttöoikeudet muuttuvat, joten datakatalogissa on aina ajantasaisin tieto.

Lopuksi

Digitoidut vanhat sanomalehdet eivät ole ongelmaton tutkimusaineisto. Niiden keskeisiä ongelmia ovat muun muassa tekstintunnistuksen laatu, tekijänoikeuk-

17 http://data.nationallibrary.fi/

sien aiheuttamat ongelmat aineistojen saatavuudessa, kokoelmien kattavuus, hakujärjestelmän ominaisuudet jne.

Optisen luvun aiheuttamat laatuongelmat ovat ikävä realiteetti, jonka kanssa joudutaan elämään toistaiseksi. Esimerkiksi Euroopan alueen digitoitua lehtiaineistoa eri kirjastoista yhteen koonnut Europeana-projekti on arvioinut lehtisisältönsä sanatason laatua (Pletschacher, Clausner, & Antonacopoulos, 2015). Yli puolella kielistä noin 80 % tekstien sanoista oli aineistoissa oikein, mutta suomen, vanhan saksan, latvian, venäjän, ukrainan ja jiddishin aineistoissa vain vajaat 70 % sanoista oli oikein. Pienillä kielillä ja mutkikkaammilla kirjoitusjärjestelmillä julkaistu sisältö saattaa siis olla sisällöltään huonolaatuisempaa kuin valtakielten aineisto.

Jotkut digitaalisen humanismin tutkijat ovat suhtautuneet hyvinkin kriittisesti lehtien laatuongelmiin. Esimerkiksi Jarlbrink ja Snickars (2017) kertovat omista kokemuksistaan Aftonbladetin digitoidun version käytössä. Heidän mielestään optisen luvun aineistoon tuottamat laatuongelmat tekevät tutkimuksen monelta osin mahdottomaksi. He peräävät digitoitujen lehtiaineistojen aineistojen laadulle parempaa kontrollia.

Muut tutkijat ovat olleet myös kriittisiä, mutta aineiston laadun vaikutus tutkimukseen riippuu paljon myös tutkimuksen aiheesta. Cordell (2017) kuvaa omia kokemuksiaan, ja korostaa sitä, miten optisen luvun luoma uusi versio alkuperäisestä dokumentista on monivaiheinen prosessi, jota pitää tarkastella myös syntyhistoriansa kautta. Digitoitujen tekstien virheiden tutkijalle aiheuttamista käytännön ongelmista antavat esimerkkejä muun muassa Traub ja kumppanit (2015) sekä Hitchcock (2013). Traub ja kumppanit haastattelivat historiantutkijoita ja kysyivät heidän tutkimusongelmiaan. Sen perusteella he arvioivat, miten hyvin lehtiaineiston hakujärjestelmät kykenivät auttamaan tutkimuksessa. Joissain tutkimuskysymyksissä järjestelmät toimivat, joissain eivät. Tutkijat ovat yleisesti ottaen tietoisia optisen luvun aiheuttamista aineiston virheistä, mutta eivät kykene itse arvioimaan, miten virheet vaikuttavat heidän tutkimuksen suorittamiseensa. Tässä saattaisi auttaa tarkempi tieto digitointiprosessista ja tilastotiedot tekstin virheiden määrästä. (Emt.) Aineiston käyttö hakujärjestelmän kautta rajaa tutkimuskysymyksiä, mutta toisaalta datapaketin itsenäinen käyttö vaatii kohtalaisen edistyneitä tekstitietojen käsittelyn välineitä ja taitoja joko yksittäisestä käyttäjältä tai laajemmalta tutkimusryhmältä.

Kansallisarkiston ja Kansalliskirjaston digitoitujen aineistojen käyttäjäkyselyssä (Hölttä, 2016) on tutkittu digitoitujen aineistojen käyttöä. Kyselyn tuloksissa käyttäjien esiin nostama pääsyy digitaalisten aineistojen käyttämättömyyteen ei ollut aineiston laatu vaan se, että vapaa verkkokäyttö ei ole mahdollista joko siksi että aineistoa ei ole vielä digitoitu, tai että siinä on käyttörajoituksia (Hölttä, 2016). Tämä on selkeästi nähtävissä myös Digin käyttäjäpalautteista,

joista vuonna 2015 noin 25 % käsitteli joko aineiston ajallisen käytön takarajan pidentämistä, tai pyyntöjä saada oikeus tiettyyn tekijänoikeudellisesti rajattuun aineistoon verkon välityksellä. Aineistoa onkin avattu verkkopalvelussa Höltän tutkimuksen jälkeen tekijänoikeussopimuksien mahdollistamissa rajoissa 19 vuotta lisää vuosina 2017 ja 2018.

Painettu lehtiaineisto on, ainakin periaatteessa, tutkijan alkuperäinen aineisto, mutta sen käyttäminen vaatii yleensä matkustamista esimerkiksi vapaakappalekirjastoon tai Kansalliskirjastoon, joten aineiston saatavuudessa on rajoituksensa. Höltän (2016) tutkimuksen vastaajista 84% ilmoitti mieluummin käyttävänsä digitaalista aineistoa kuin alkuperäistä. Digitaalisen aineistojen käyttämisen perusteluissa mainittiin erityisesti käytön helppous, saavutettavuus ajasta, paikasta ja päätelaitteesta riippumatta, rajoittamaton käyttöaika ja löydettävyys. Aineistojen käyttäjinä on akateemisia tutkijoita, opettajia, toimittajia, sukututkijoita ja tiedon hankintaa vapaa-aikanaan eri syistä harrastavia. (Emt.) Samanlaisia vastauksia on saatu kansainvälisissä käyttäjätutkimuksissa erilaisista suurista digitoiduista lehtiaineistoista ja muusta digitoidusta historiallisesta aineistosta (Gooding, 2017; Hungenaert & Gillet, 2017; Sinn & Soares, 2014).

Olemme avanneet tässä katsauksessa Kansalliskirjaston digitoitujen historiallisten lehtien avoimena datana ja verkkodatana julkaistun materiaalin ominaisuuksia sekä luoneet katsauksen aineiston tähänastiseen tutkijakäyttöön. Lehtiaineisto on laaja, monipuolinen sekä myös kaksikielinen, joten se tarjoaa mahdollisuuksia monenlaiseen tutkimukseen. Luultavimmin olemme nähneet toistaiseksi vasta aineiston tutkimuskäytön alun: digitoidut sanomalehdet ovat alkamassa vakiinnuttaa asemaansa tärkeänä tutkimusinfrastruktuurin osana (Anderson, 2013). Ne ovat olemassa ja kehittyvät ja niiden käyttö tutkimuksessa lisääntyy.

Kiitokset

Artikkeli on osa Suomen Akatemian rahoittamaa tutkimushanketta *Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910,* jota rahoittaa Suomen Akatemian Digitaalisten ihmistieteiden ohjelma.

Kirjallisuutta

Anderson, S. (2013). What are research infrastructures? *International Journal of Humanities and Arts Computing*, *7*(1-2), 4–23. https://doi.org/10.3366/ijhac.2013.0078

Beals, M. H., Russell, I. G., Nyhan, J., Priani, E., Priewe, M., Salmi, H., … Hauswedell, T. (2017). Oceanic exchanges: tracing global information networks in historical newspaper repositories, 1840–1914. https://doi.org/10.17605/OSF.IO/WA94S

Cordell, R. (2017). "Q I-JTB THE RAVEN". Taking dirty OCR seriously. *Book History*, *20*(1), 188–225. https://doi.org/http://doi.org/10.1353/bh.2017.0006

Dunning, A. (2012). *European Newspaper Survey Report*. http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf

Ginter, F., Kanner, A., Lahti, L., Marjanen, J., Mäkelä, E., Nivala, A., … Vesanto, A. (2018). Metadata analysis and text reuse detection: reassessing public discourse in Finland through newspapers and journals 1771–1917. Teoksessa *DHN2018*. https://www.helsinki.fi/sites/default/files/atoms/files/dhn2018-book-of-abstracts.pdf

Gooding, P. (2017). *Historic Newspapers in the Digital Age. "Search All About It!"*. Routledge.

Hitchcock, T. (2013). Confronting the Digital. *Cultural and Social History*, *10*(1), 9–23. https://doi.org/10.2752/147800413X13515292098070

Hölttä, T. (2016). *Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat* (tohtorinväitöskirja). http://urn.fi/URN:NBN:fi:uta-201603171337

Hungenaert, J., & Gillet, F. (2017). *Studying user's digital practices and needs in Archives and Libraries. Final Report of the MADDLAIN project.* http://www.cegesoma.be/docs/images/stories/ceges/Recherche/MADDLAIN_grand_public_ENG_48p_.pdf

Jarlbrink, J., & Snickars, P. (2017). Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. *Journal of Documentation*, *73*(6), 1228–1243. https://doi.org/10.1108/JD-09-2016-0106

Koistinen, M., Kettunen, K., & Pääkkönen, T. (2017). Improving optical character recognition of Finnish historical newspapers with a combination of Fraktur & Antiqua Models and image preprocessing. Teoksessa *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden* (ss. 277–283). Linköping University Electronic Press, Linköpings universitet / National Library of Finland, The Centre for Preservation and Digitisation, Finland.

Kokko, H. (2016). *Kuviteltu minuus: ihmiskäsityksen murros suomenkielisen kansanosan kulttuurissa 1800-luvun puolivälissä* (tohtorinväitöskirja). Tampereen yliopisto. http://urn.fi/URN:ISBN:978-952-03-0282-5

Kumpulainen, S., & Keskustalo, H. (2017). Supporting evolving search tasks in digital environments via fuzzy string matching. Teoksessa *Heldig Summit*. http://static.seco.cs.aalto.fi/events/2017/heldig-summit/abstracts/kumpulainen-keskustalo.pdf

La Mela, M. (2016). *The politics of property in a European periphery : the ownership of books, berries, and patents in the Grand Duchy of Finland 1850-1910* (Thesis). European University Institute, Florence, Italy. https://doi.org/10.2870/604750

La Mela, M. (2018). Digitised newspapers and the geography of the nineteenth-century "lingonberry rush" in Finland. Teoksessa *DHN2018*. https://www.helsinki.fi/sites/default/files/atoms/files/dhn2018-book-of-abstracts.pdf

Pääkkönen, T., & Kervinen, J. (2016). Historiallisten digitoitujen sanoma- ja aikakauslehtien avaaminen avoimena datana tutkijoille. *Informaatiotutkimus*, *35*(3), 67–68. https://journal.fi/inf/article/view/59442

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). https://doi.org/10.1045/july2016-paakkonen

Pääkkönen, T., Rautiainen, J., Ryynänen, T., & Uusitalo, E. (2018). Open, extended, closed or hidden data of cultural heritage. Teoksessa *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*. Helsinki. http://ceur-ws.org/Vol-2084/

Pekonen, O. (2014). *Debating "the ABCs of parliamentary life" : the learning of parliamentary rules and practices in the late nineteenth-century Finnish Diet and the early Eduskunta* (tohtorinväitöskirja). University of Jyväskylä. http://urn.fi/URN:ISBN:978-951-39-5843-5

Pletschacher, S., Clausner, C., & Antonacopoulos, A. (2015). Europeana Newspapers OCR Workflow Evaluation. Teoksessa *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing* (ss. 39–46). New York, NY, USA: ACM. https://doi.org/10.1145/2809544.2809554

Ruokolainen, T., & Kettunen, K. (2018). À la recherche du nom perdu – searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection. Teoksessa *DAS2018, 13th IAPR International Workshop on Document Analysis Systems. Short papers booklet.* https://das2018.cvl.tuwien.ac.at/media/filer_public/85/fd/85fd4698-040f-45f4-8fcc-56d66533b82d/das2018_short_papers.pdf

Salmi, H. (2018). Viraalisuus – kulttuurihistoriallinen näkökulma. *Niin & näin*, (1), 71–79. http://netn.fi/artikkeli/viraalisuus-kulttuurihistoriallinen-nakokulma

Sinn, D., & Soares, N. (2014). Historians' use of digital archival collections: The web, historical scholarship, and archival research. *Journal of the Association for Information Science and Technology*, *65*(9), 1794–1809. https://doi.org/10.1002/asi.23091

Stroeker, N., & Vogels, R. (2012). *Survey report on digitisation in European cultural heritage institutions 2012. ENUMERATE Thematic Network*. https://www.egmus.eu/fileadmin/ENUMERATE/documents/ENUMERATE-Digitisation-Survey-2012.pdf

Terras, M. (2015). Opening Access to collections: the making and using of open digitised cultural content. *Online Information Review*, *39*(5), 733–752. https://doi.org/10.1108/OIR-06-2015-0193

Tolonen, M., & Lahti, L. (2015). Aatehistoria ja digitaalisten aineistojen mahdollisuudet. *Ennen ja nyt: historian tietosanomat*. http://www.ennenjanyt.net/2015/08/aatehistoria-ja-digitaalisten-aineistojen-mahdollisuudet/

Tolonen, M., Marjanen, J., Roivainen, H., & Lahti, L. (2017). Patterns of public discourse in Finland: combining meta-data from library catalogues and the Finnish historical newspaper library. Teoksessa *DHN 2017 Digital humaniora i Norden/Digital Humanities in the Nordic Countries*. http://dhn2017.eu/wp-content/uploads/2017/03/DHN2017_Book_of_Abstracts_20170313.pdf

Traub, M. C., Ossenbruggen, J. van, & Hardman, L. (2015). Impact analysis of OCR quality on research tasks in digital archives. Teoksessa S. Kapidakis, C. Mazurek, & M. Werla (toim.), *Research and Advanced Technology for Digital Libraries* (Vsk. 9316, ss. 252–263). Springer International Publishing. https://doi.org/10.1007/978-3-319-24592-8_19

Tuori, H.-K. (2016). Tehtävät ja strategia. *Kansalliskirjasto*. Text. https://www.kansalliskirjasto.fi/fi/tehtavat-ja-strategia

Turunen, R. (2018). Sculpting time: temporality in the language of Finnish socialism, 1895–1917. Teoksessa *DHN2018*. https://www.helsinki.fi/sites/default/files/atoms/files/dhn2018-book-of- abstracts.pdf

Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., & Ginter, F. (2017). Applying BLAST to text reuse detection in Finnish newspapers and journals, 1771(–1910. Teoksessa *Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23-24 May 2017* (ss. 54–58). Linköping. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf

Vesanto, A., Nivala, A., Salakoski, T., Salmi, H., & Ginter, F. (2017). A system for identifying and exploring text repetition in large historical document corpora. Teoksessa *Proceedings of the 21st Nordic Conference of Computational Linguistics. Gothenburg, Sweden, 23-24 May 2017* (ss. 330–333). Linköping. http://www.ep.liu.se/ecp/131/049/ecp17131049.pdf