Tuula Pääkkönen & Jukka Kervinen

Historiallisten digitoitujen sanoma- ja aikakauslehtien avaaminen avoimena datana tutkijoille

Tuula Pääkkönen, orcid.org/0000-0003-3958-9732, Kansalliskirjasto, tuula.paakkonen@helsinki.fi;
Jukka Kervinen, Kansalliskirjasto, jukka.kervinen@helsinki.fi

Kansalliskirjasto on uudessa strategiassaan (Tuori, 2016) vuosille 2016-2020 määritellyt tietovarantojen avoimuuden tärkeäksi tavoitteekseen. Tähän liittyen kirjastossa on luotu sekä Avoin Kansalliskirjasto (Tuori, 2016) ja Digitaalisen humanismin politiikat, joilla pyritään systemaattisesti lisäämään avoimuutta ja tutkimusyhteistyötä (Hormia-Poutanen, 2016). Tähän liittyen Kansalliskirjaston digitoiduista aineistoista (digi.kansalliskirjasto.fi) on nyt vuoden 2016 alussa valmisteltu ladattava aineistopaketti, joka tullaan avaamaan saman verkkopalvelun kautta ladattavaksi kaikille. Ajatuksena on, että kun alkuperäinen digitoinnin jälkikäsittelyn lopputulos tarjotaan tutkijoille ja muille kehittäjille käyttöön, löytyisi uusia tutkimus- ja käyttömahdollisuuksia, joilla voisi mm. parantaa aineiston laatutasoa. Tavoitteena on myös tarjota laajoja aineistoja, joihin voidaan tehdä kvantitatiivista analyysiä, mahdollistaen tutkimukselle oleellisen informaation löytämisen (Tolonen & Lahti, 2015).

Tähän mennessä Kansalliskirjaston digitoituja aineistoja on saanut tutkimuskäyttöön Helsingin yliopiston Kielipankki-palvelun kautta, johon FinClarin-konsortio on luonut mm. suomenkielisestä aineistosta N-gram korpuksen (FinClarinAineistoFNC1, ei pvm.), sekä Korppalvelun (Borin, Forsberg, & Roxendal, Johan, 2012) suomalaisen version kautta. Kansalliskirjaston avoimuuden ilmentyminä on toimineet mm. aineiston toimitus The European Library (TEL)-palveluun (The European Library, 2016) ja vuoden 2016 alussa tehty rajapinta, jota kautta on haettavissa Finna.fi-palvelun osallistujaorganisaatioiden tarjoamien sisältöjen kuvailutietoja (Kansalliskirjasto, 2016a), minkä lisäksi suomalaisen sanasto- ja ontologiapalvelu Fintolla on myös avattu oma rajapinta (Kansalliskirjasto, 2016b). Kansalliskirjaston pitkän tähtäimen tavoitteena onkin päästä, erityisesti metatietovarannoissa, viiden tähden laatutasoon, tarjoten data omalla julkaisualustallaan eri muodoissa (Kansalliskirjasto, 2015). Tämän lisäksi myös varsinaisia sisältöjä Kansalliskirjastosta pyritään tarjoamaan nykyisen käyttöliittymän lisäksi myös aineistopaketteina tai myös rajapinnan kautta, jos kyselyjä alkaa tulla enemmänkin.

Vuoden 2016 aikana suunniteltu ja toteutettu aineistopaketti koostuu ALTO (Technical Me-

tadata for Layout and Text Objects) (Library of Congress, 2016) XML-tiedostoista, Suomessa julkaistuista sanomalehdistä 1771-1910 ja aikakauslehdistä 1816-1910 (Pääkkönen, Kervinen, Nivala, Kettunen, & Mäkelä, 2016) Tiedostot on jaettu ilmestymisvuosien mukaan eri jaksoihin, yrittäen pitää ladattavat aineistopakettien koko kohtuullisena. Yksi ALTO-tiedosto vastaa yhtä sivutiedostoa, ja Kansalliskirjaston valmistelemaan XML-tedostoon on lisäksi lisätty sivun metatiedot ja tekstisisältö.

Aineistopaketin ensimmäinen versio Kansalliskirjaston, Helsingin yliopiston ja Turun yliopiston yhteiseen COMHIS-projektin tutkimusryhmän käyttöön toi Kansalliskirjastolle näkyviä hyötyjä. Aineiston läpikäynti auttoi muutamien metadatan ja sisältöjen puutteiden ja epätarkkuuksien löytämisessä, jotka auttavat aineiston laadun parantamisessa. Uuteen verkkopalvelun versioon, joka julkaistiin syksyllä 2016, metadataepätarkkuuksia on jo korjattu ja korjauksia jatketaan sitä mukaa mitä muilta digitointitehtäviltä mahdollisuuksia löytyy.

Digitoituja sanoma- ja aikakauslehtiä voi tutustua osoitteessa . Verkosta löytyvä aineisto vuoteen 1910 asti on tekijänoikeuksista vapaata ja käytettävissä tutkimus, koulutus- ja opetuskäyttöön kuin myös muihin tarkoituksiin. Sanomalehtien digitointi on vuoden 2016 aikana edennyt eteenpäin vuoteen 1920 asti. Samanaikaisesti Kansalliskirjaston Aviisi-projektissa selvitellään kuinka tutkimus- ja opetuskäyttöön aineistoa saisi laajemmin käyttöön. Vain digitoitua aineistoa voi avata avoimena datana, joten samalla kun parannamme avoimuuden astetta, pyrimme myös varmistamaan, että aineistojen määrä kasvaisi jatkuvasti ja että tutkijoilla olisi myös digitoituihin aineistoihin pääsy monin eri keinoin.

Lähteet

Borin, L., Forsberg, M., & Roxendal, Johan. (2012). Korp – the corpus infrastructure of Språkbanken. Proceedings of LREC 2012. Istanbul: ELRA (ss. 474–478). Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf

FinClarinAineistoFNC1 | Kielipankki wiki. (ei pvm.). https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinAineistoFNC1 (20.8.2016)

Hormia-Poutanen, K. (2016). Kirjastot ja avoin tiede. Esitetty tilaisuudessa AMK-kirjastopäivät, Jyväskylä. http://amkkirjastopaivat.humak.fi/wp-content/uploads/sites/42/2016/03/AMK-kirjastop%C3%A4iv%C3%A4t_Hormia-Poutanen.pptx (8.10.2016)

Kansalliskirjasto. (2015). Kansalliskirjaston metatietovarantojen avaamisen suunnitelma 2015-2017. https://www.kiwi.fi/display/avoinkk/Kansalliskirjaston+metatietovarantojen+avaamisen+suunnitelma+2015-2017 (20.8.2016)

Kansalliskirjasto. (2016a). Finnan avoimen rajapinnan käyttöehdot - Finna. https://www.kiwi.fi/pages/viewpage.action?pageId=53839664 (20.8.2016)

Kansalliskirjasto. (2016b). Finton ja ontologioiden käyttöönotto - Finto - suomalainen sanasto- ja ontologiapalvelu. https://www.kiwi.fi/pages/viewpage.action?pageId=53839594 (20.8.2016)

Library of Congress. (2016). ALTO: Technical Metadata for Layout and Text Objects. https://www.loc.gov/standards/alto/ (20.8.2016)

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

The European Library. (2016). The European Library Open Dataset - The European Library. http://www.theeuropeanlibrary.org/tel4/access/data/opendata/details (20.8.2016)

Tolonen, M., & Lahti, L. (2015). Aatehistoria ja digitaalisten aineistojen madhollisuudet. Ennen ja nyt: historian tietosanomat, 2015(2). http://www.ennenjanyt.net/2015/08/aatehistoria-ja-digitaalisten-aineistojen-mahdollisuudet/

Tuori, H.-K. (2016). Tehtävät ja strategia. https://www.kansalliskirjasto.fi/fi/tehtavat-ja-strategia (8.10.2016)