DIGITAALINEN TIETO HALTUUN

RATKAISUJA DIGITAALISTEN AINEISTOJEN HALLINTAAN JA KÄYTTÖÖN

Miia Kosonen (toim.)

XAMK
Kaakkois-Suomen ammattikorkeakoulu

Miia Kosonen (toim.)

DIGITAALINEN TIETO HALTUUN

RATKAISUJA DIGITAALISTEN AINEISTOJEN HALLINTAAN JA KÄYTTÖÖN

XAMK KEHITTÄÄ 11

KAAKKOIS-SUOMEN AMMATTIKORKEAKOULU
MIKKELI 2017

XAMK KEHITTÄÄ 11

DIGITAALINEN TIETO HALTUUN

Kannen kuva: Mainostoimisto Ilme Oy
Taitto- ja paino: Grano Oy
ISBN: 978-952-344-019-7
ISBN: 978-952-344-020-3 (PDF)
ISSN: 2489-2467 (nid.)
ISSN 2489-3102 (verkkojulkaisu)
julkaisut@xamk.fi

DIGITAALINEN TALOUS KASVAA

Kaakkois-Suomen ammattikorkeakoulu on parhaillaan päivittämässä strategiaansa. Nopeasta toimintaympäristön muutoksesta kertoo, että vuonna 2015 hyväksytty strategia on jo nyt uudistamisen tarpeessa. Yksi keskeisimpiä muutoksen ajureita on digitalisaatio. Se muuttaa korkeakouluja niin toimintatavoiltaan kuin rakenteiltaan. Luonnollisesti vaikutus ulottuu myös tutkimus- ja kehittämistoiminnan painopisteisiin.

Kaakkois-Suomen ammattikorkeakoulu profiloituu vahvana tutkimus- ja kehitystoiminnan korkeakouluna. Digitaalinen talous on yksi sen neljästä tki-painoalasta vastaisuudessakin, vaikka tutkimustoiminnan kärkiä on syytä entisestään terävöittää. Myös uuden liiketoiminnan etsimiseen ja hallittujen riskien ottamiseen uusien tutkimusteemojen löytämisessä tulee investoida.

Ammattikorkeakoulun soveltavan tutkimuksen odotetaan synnyttävän toiminta-alueelle myös uutta yritystoimintaa, uutta työtä ja elinvoimaa.

Digitaalinen talous painoalana yhdistää digitaalista osaamista ja liiketaloutta. Tämä on perusteltua, koska digitalisaation ytimessä on liiketoimintamallien muuttuminen. Huolimatta esimerkiksi verkkokaupan ja -palvelujen nopeasta kasvusta, on viimeaikaisten selvitysten (HS 6.6.2017) perusteella noussut huoli siitä, että huomattava osa pienistä ja keskisuurista suomalaisista yrityksistä on jäämässä ”digipudonneiksi” tai ”digieksyneiksi”. Tarvitaan siis sekä digitaalisen liiketoiminnan osaajien koulutusta että yhdessä yritysten kanssa tehtävää soveltavaa tutkimusta.

Mikkelissä tutkimus on keskittynyt digitaaliseen tiedonhallintaan, sähköisten aineistojen pitkäaikaissäilytykseen ja niiden käytettävyyteen. Kotkassa pääpaino on kyberturvallisuudessa ja Kouvolassa pelillisyydessä. Kiinnostus sähköiseen arkistointiin heräsi Mikkelissä jo viitisentoista vuotta sitten. Sen ansiosta perustettiin aikanaan ammattikorkeakoulun tytäryhtiö Disec Oy, joka tarjoaa kuvantamisen tietojärjestelmäpalveluita ja tiedon hallinnan ratkaisuja erityisesti terveydenhuollon sektorille.

Kuten tästä julkaista käy selville, nyt työn alla on muun muassa Kansalaisarkisto. Innovatiivisia sovellutuksia digitaalisen tiedon uusista käyttötavoista on vireillä myös esimerkiksi digitaalisessa rakennussuunnittelussa.

Digitaalisten aineistojen tutkimus- ja kehitystoiminnassa tarvitaan tieteenalarajat ylittävää yhteistyötä, mistä Xamkin, Kansalliskirjaston digitointi- ja konservointikeskuksen ja Helsingin yliopiston yhdessä perustama Digitalia on oiva esimerkki. Haluamme antaa oman panoksemme Mikkelissä kasvussa olevalle digitaaliselle arkistoklusterille.

OKM-sopimus 2017–2020 edellyttää ammattikorkeakoululta tutkimustoiminnan rakenteiden – yhtenä näistä Digitalia – edelleen vahvistamista ja panostamista yrittäjyyteen, innovaatiotoimintaan ja tutkimustulosten kaupallistamiseen.

Mikkelissä 12.6.2017
Kalevi Niemi
Kehitysjohtaja, Kaakkois-Suomen ammattikorkeakoulu

KIRJOITTAJAT

Anssi Jääskeläinen, TkT, TKI-asiantuntija
Kaakkois-Suomen ammattikorkeakoulu

Eero Kausalainen, eläkkeellä oleva historian harrastaja

Kimmo Kettunen, FT, tutkimuskoordinaattori
Kansalliskirjasto

Liisa Uosukainen, DI, IT-asiantuntija
Kaakkois-Suomen ammattikorkeakoulu

Miia Kosonen, KTT, TKI-asiantuntija
Kaakkois-Suomen ammattikorkeakoulu

Mika Koistinen, DI, tietojärjestelmäasiantuntija
Kansalliskirjasto

Mikko Tolonen, FT, professori
Helsingin yliopisto

Noora Talsi, YTT, tutkimusjohtaja
Kaakkois-Suomen ammattikorkeakoulu

Teemu Ruokolainen, TkT, tutkijatohtori
Kansalliskirjasto

Tenho Kokkonen, FM, tutkija
Päivälehden arkisto

Tuula Pääkkönen, FM, tietojärjestelmäasiantuntija
Kansalliskirjasto

SISÄLTÖ

JOHDANTO – MIKSI DIGITALIA?

Noora Talsi, YTT, *tutkimusjohtaja, Kaakkois-Suomen ammattikorkeakoulu*

Digitaalisuus on yksi aikamme megatrendeistä ja tunnistettu myös Suomessa. Digitaalisuus mielletään helposti lähinnä palveluiden sähköistämiseksi ja osaprosessien digitalisoinniksi. Digitaalisuus ei kuitenkaan ole vain analogisten aineistojen digitointia tai palvelun muuttamista sähköiseksi, vaan digitaalinen maailma muuttaa käytännön tapoja toimia. Se pakottaa ajattelemaan uudella tavalla ja tekemään toisin. Digitaalisuus vaatii niin uusia teknologioita kuin uusia taitoja käyttää niitä. Digitaalisuuden hyödyntäminen edellyttää ennen kaikkea laadukasta tietojen hallintaa.

Digitaalisen yhteiskunnan tärkein tuotantotekijä on data. Datan määrä kasvaa maailmassa koko ajan ja dataa on myös aiempaa helpommin saatavilla. Data on kuitenkin käsittelemättömänä ja tulkitsemattomana arvotonta. Data saa arvonsa vasta kun siitä jalostetaan merkityksellistä tietoa, jonka alkuperä ja ominaisuudet tunnetaan. Datasta ei siis ole pulaa, mutta datan tehokas hyödyntäminen vaatii jatkuvaa tutkimusta ja kehittämistyötä, jotta entistä parempia menetelmiä olisi saatavilla datan tehokkaaseen käsittelyyn ja hyödyntämiseen.

Samalla kun julkinen keskustelu nostaa digitalisaatiosta ja globalisaatiosta esiin uhkia, kuten työpaikkojen vähenemisen ja työn siirtymisen halvemman työvoiman maihin, on huomattava, että digitaalinen yhteiskunta on yltäkylläisyyden yhteiskunta. Siinä missä teollisessa yhteiskunnassa kamppailtiin resurssien riittävyydestä, digitaalisissa tuotteissa ja palveluissa resurssit ovat rajattomat. Teollista yhteiskuntaa leimasi niukkuus: autoja, rautateitä tai pesukoneita ei riitä kaikille ja kaikkialle. Digitaalinen tuote, palvelu tai sovellus on puolestaan rajattomasti ja välittömästi kaikkien saatavilla.

Globalisaatio ja vapaa liikkuvuus kiinnittyvät digitalisaatioon yhtä saumattomasti kuin nationalismi ja protektionismi teolliseen yhteiskuntaan. Niinpä Suomessakin on mahdollista kehittää sellaisia digitaalisia tuotteita, palveluja tai sovelluksia, jotka voidaan ottaa nopeasti käyttöön missä päin maailmaa tahansa. Kun työvoiman määrän ja pääoman merkitys vähenee, hyvien ideoiden, korkean osaamisen ja ennen kaikkea tehokkaiden tiedonjakelun alustojen merkitys kasvaa. Digitaalisessa yhteiskunnassa toimiminen edellyttää datan ymmärrystä, sen tehokasta hyödyntämistä ja jakamista. (Pohjola, 2015)

DIGITALIA SYNTYI MIKKELISSÄ OLEVAN OSAAMISEN YMPÄRILLE

Mikkelissä on vahva alueellinen digitaalisuuden osaamiskeskittymä, johon kuuluu niin julkisen kuin yksityisen ja kolmannen sektorin toimijoita. Menestyminen nopeasti muuttuvassa, globaalissa maailmassa edellyttää osaamista ja näkemystä digitaalisuuden hyödyntämisessä. Menestys rakentuu organisaatioiden kyvylle kehittää, soveltaa ja kaupallistaa sitä uutta tietoa ja osaamista, jota digitaalinen tiedonhallinta voi tarjota.

Laajemmin katsottuna kyse on tietojohtamisesta: erilaisista menetelmistä ja prosesseista, joilla tiedosta luodaan arvoa.

Digitaalisen tiedonhallinnan osaamisen syventämiseksi perustettiin Digitalia. Kaakkois-Suomen ammattikorkeakoulu, Helsingin yliopisto ja Kansalliskirjasto perustivat kesällä 2015 yhteistyössä Mikkeliin digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskuksen, Digitalian. Digitalian perustamistyötä varten saatiin rahoitus hankkeelle ”Digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskus – Digitalia” (2015-2017). Rahoituksen myönsi reiluksi kahdeksi vuodeksi Etelä-Savon maakuntaliitto Euroopan unionin aluekehitysrahastosta.

Kaakkois-Suomen ammattikorkeakoulu ja Kansalliskirjaston Digitointi- ja konservointikeskus ovat vuosituhannen alusta tehneet rinnakkaista tutkimus- ja kehittämistoimintaa Mikkelissä. Kaakkois-Suomen ammattikorkeakoulu on tehnyt sähköisen säilyttämisen ja arkistoinnin tutkimus- ja kehitystyötä jo 15 vuotta erilaisissa kehittämishankkeissa. Kehittämistä on tehty niin 3D-mallintamisen, elävän kuvan ja äänen digitoinnin kuin pitkäaikaissäilytyksen saralla (Palonen, 2015). Samoin Kansalliskirjaston Digitointi- ja konservointikeskus on tehnyt vuosikymmenet töitä lehtiaineistojen digitoinnissa ja digitoitujen sanomalehtiaineistojen käytettävyyden parantamisessa.

Molempien työ on hyödyttänyt laajasti alueen muita toimijoita, erityisesti muistiorganisaatioita. Toimijat ovat tehneet myös yhteistyötä aikaisemmissa hankkeissa. Nyt oli kuitenkin aika syventää ja vahvistaa yhteistyötä yhteisellä hankkeella ja yhteisillä päämäärillä, jotka tukevat alueen tutkimus- ja kehittämistoimintaa ja mahdollistavat sen nostamisen kansallisesti ja kansainvälisesti tunnustetuksi kärkialaksi.

Digitalia syntyikin halusta tehdä uudenlaista yhteistyötä digitaalisia aineistoja käsittelevien tahojen kanssa. Rinnakkaista yhteistyötä haluttiin syventää konkreettiseksi yhdessä tekemiseksi Mikkelin digitaalisen tiedonhallinnan asiantuntijaorganisaatioiden välillä. Digitalia on uudenlainen osaamiskeskittymä ja yhteistyötapa, jossa yhdistyy Kansalliskirjaston, yliopistomaailman ja ammattikorkeakoulukentän osaaminen.

Kullakin toteuttajalla on myös erittäin laajat asiantuntijaverkostot kansallisesti ja kansainvälisesti. Näitä on hyödynnetty Digitaliassa aktiivisesti ja rakennettu samalla uusia kumppanuuksia. Kaikkien Digitalian toimijoiden verkostot tukevat ja vahvistavat toisiaan. Digitalia on myös osoitus digitaaliseen tiedonhallintaan liittyvän osaamisen vahvuudesta Mikkelissä.

Näin Mikkeli on ottanut merkittäviä askeleita kohti digitaalisten aineistojen hallinnan ja käytön johtavaa asemaa Suomessa. Digitalian myötä on syntynyt positiivista pioneerihenkeä ja yhteinen visio, jossa tutkimus, kaupunki ja yritykset toimivat yhdessä liittyen erilaisten tietovirtojen hyötykäyttöön.

DIGITALIA PARANTAA TIEDON SAATAVUUTTA JA KÄYTETTÄVYYTTÄ

Ratkaisua lähdettiin hankkeessa hakemaan kahteen tärkeään kysymykseen: Miten datasta saadaan tietoa? Miten saadaan tieto käyttöön? Tässä artikkelikokoelmassa vastataan näihin kysymyksiin useista eri näkökulmista ja avataan tarkemmin kahden

vuoden aikana tehtyä työtä. Artikkeleita ovat kirjoittaneet paitsi Digitalian omat asiantuntijat myös keskeiset yhteistyökumppanit. Digitalian toiminta on verkostomaista ja tiedon ja tutkimuksen hyödyntäjät ovat keskeisessä roolissa.

Artikkelit esittelevät Digitaliassa kehitettyjä ratkaisuja tiedon merkityksen, käytettävyyden parantamisen sekä uusien sovellusten näkökulmista. Tiedon merkitystä ja arvoa pohtivat erityisesti Mikko Tolonen ja Miia Kosonen artikkeleissaan. Tiedon käytettävyyden parantamisesta kirjoittavat Tuula Pääkkönen sekä Kimmo Kettunen, Mika Koistinen, Teemu Ruokolainen ja Tuula Pääkkönen. Erilaisia tiedon hyödynnettävyyttä parantavia sovelluksia ja niihin liittyvää kehittämistyötä esitellään Anssi Jääskeläisen, Liisa Uosukaisen ja Eero Kausalaisen sekä Anssi Jääskeläisen ja Tenho Kokkosen artikkeleissa.

Kaiken kaikkiaan Digitaliassa tehtävä tutkimus- ja kehitystyö vastaa yhteiskunnan digitaaliseen murrokseen. Lähtökohtana on, että digitaalista tietoa on helposti saatavilla ja hyödynnettävissä.

Suurten tietomassojen eli niin sanotun big datan hyödyntäminen on yhä useammin osa niin tutkijoiden, kansalaisten kuin yritystenkin arkea. Nykypäivän tietojohtaminen ei enää pyri yksinomaan tiedon ”omistamiseen” tai hallintaan organisaatioissa, vaan yhä useammin arvo syntyy niiden ulkopuolisen tiedon hyödyntämisestä ja yhdistelystä uusilla tavoilla. Esimerkiksi asiakastiedon louhiminen eri sosiaalisen median kanavista avaa yrityksille uusia mahdollisuuksia kulutustrendien ja yleisen mielipideilmaston ymmärtämiseen. Samalla yritykset saavat suoraa tietoa kuluttajilta päätöksentekonsa tueksi.

Hyödyntämisen edellytyksenä on, että tieto on avoimesti saatavilla. Tiedon avaamiselle on tyypillistä, että avaamisen hyötyjä ei voida ennalta tarkkaan määritellä tai ennustaa. Avaaminen tukee kuitenkin ketterää kehitystä ja datan saatavuutta. Tämä mahdollistaa palvelujen ja tuotteiden nopean ja edullisen toteutuksen. Kun tietoa, menetelmiä ja työkaluja on olemassa, ketterät ja erikoistuneet pk-yritykset voivat kehittää tiedosta julkisia palveluita ja luoda uutta liiketoimintaa.

DIGITALIAN TYÖ HYÖDYTTÄÄ MONIA KOHDERYHMIÄ

Digitalian kohderyhmänä ja tulosten hyödyntäjinä ovat olleet ennen kaikkea muistiorganisaatiot, tutkijat ja kansalaiset. Esimerkiksi Digitaliassa kehitetty sähköpostien automaattinen arkistointi parantaa huomattavasti yksityisarkistojen ja yritysarkistojen toimintaa. Yritysarkistoilla on merkittävä rooli suomalaisen elinkeinoelämän tallentajina ja yhä useammin yrityksissä syntyvä tieto on digitaalisessa muodossa. Siinä missä vuosikymmenet sitten yrityksen perustajan tai toimitusjohtajan kirjeet voitiin arkistoida paperisina tuleville sukupolville, nykyisin yksityisarkistot kamppailevat kovalevyjen ja sähköpostitulvan kanssa. Tämäkin tieto on saatava tallennettua tuleville sukupolville tehokkaasti ja toimintavarmasti. Sähköpostien arkistoratkaisusta kertovat tarkemmin Anssi Jääskeläinen ja Tenho Kokkonen artikkelissa *Sähköpostien ihanuus ja kurjuus*.

Digitaliassa tehdään myös digitaalisen humanismin tutkimusta ja tutkimusaineistojen käytön parantaminen on keskeinen osa työtä. Tutkijat ovat tärkeä kohderyhmä.

Digitaalinen humanismi yhdistää humanistisen tutkimuksen ja modernit informaatio- ja kommunikaatioteknologiat. Digitaalisen tiedonhallinnan kehittymisen anti humanistiselle ja yhteiskuntatieteelliselle tutkimukselle onkin laajojen laadullisten aineistojen hyödyntäminen.

Ihmisvoimille liian raskaiden big data -aineistojen koneellinen käsittely mahdollistaa esimerkiksi aivan uudenlaisen ymmärryksen suomalaisuudesta yhteiskunnallisten keskustelujen seuraamisen kautta. Digitaalinen humanismin tutkimus avaa erityisesti Mikkelissä sijaitsevat Kansalliskirjaston aineistot uudenlaiseen käyttöön. Digitaalisesta humanismista kirjoittaa enemmän Mikko Tolonen tässä kokoelmassa. Kansalliskirjaston digitoitujen aineistojen käytön kehittämisestä ja digi.kansalliskirjasto.fi:n palveluista kirjoittavat Kimmo Kettunen, Mika Koistinen, Teemu Pääkkönen ja Tuula Pääkkönen.

Digitaliassa tehtävä työ hyödyttää myös kansalaisia. Kansalliskirjastossa tehty työ digitoidun aineiston käytön parantamiseksi palvelee niin kansalaisia, yrityksiä kuin tutkijoitakin. Kaakkois-Suomen ammattikorkeakoulussa kehitetty Kansalaisarkisto mahdollistaa henkilökohtaisen digitaalisen tiedon kokoamisen, esittämisen ja luotettavan pitkäaikaissäilyttämisen. Kansalaisarkisto tukee Digitalian profiloitumista digitaalisten aineistojen saatavuuden ja käytettävyyden kehittäjänä. Kansalaisarkiston kehityksestä ja pilotoinnista kertovat Liisa Uosukainen sekä Kansalaisarkiston pilottiasiakas Eero Kausalainen artikkelissaan *Kansalaisarkisto – Sukuyhteisön aarteet talteen digitaaliseen arkistoon*.

Digitalisaation edetessä yhä suurempi osa suomalaisia koskevasta tiedosta on digitaalisessa muodossa. Siksi digitaalista tietoa jalostavat, hyödyntävät ja säilyttävät ratkaisut nousevat entistä tärkeämpään rooliin. Digitalian tekemällä työllä on siis huomattava yhteiskunnallinen tilaus.

LÄHTEET

Palonen, O. 2015. ”Jotakin ne siellä Mikkelissä taas tekevät”. Digitaalisen tiedonhallinnan tutkimuskeskus sai rahoituksen. Faili 3/2015, 17–18.

Pohjola, M. 2015. Digitaalisuus ja tuottavuus finanssialalla. Helsinki.

DIGITAALINEN HUMANISMI JA DIGITALIA[1]

Mikko Tolonen, *FT, professori, Helsingin yliopisto*

DIGITAALISESTA HUMANISMISTA JA METODEISTA

Digitaalinen humanismi (tai digitaaliset ihmistieteet, kuten Suomen Akatemia on tämän määritellyt) koskee laskennallisten menetelmien yhdistämistä humanistiseen tutkimukseen (Digitaaliset ihmistieteet DIGIHUM (2016–2019), 2015). Kyseessä on vielä muotoutuva tutkimusperinne, jossa korostuvat informaatio- ja kommunikaatioteknologiat.

Usein määrittelyyn lisätään vielä huomio, että digitaalinen humanismi tutkii myös digitaalisuutta. Määritelmästä riippuen digitaalisella humanismilla voidaan siis tarkoittaa monenlaisia asioita digitaalisen kulttuurin tutkimuksesta kieliteknologiaan. Digitaalinen humanismi on keskeinen muutostekijä kaikissa muistiorganisaatioissa (kirjastot, arkistot, museot).

Yksi hyödyllinen reitti pois määritelmän tietystä epämääräisyydestä on yleiseurooppalaisen Dariah-EUn (Digital Research Infrastructure for the Arts and Humanities) muotoilu, jonka mukaan kyseessä on erityisesti humanistisen tutkimuksen muutos ja uudistuminen pikemmin kuin uuden, ”digitaalisen” tieteenalan syntyminen (Dariah Desir hankkeen dokumentti, 2016). Näin ei tarvitse korostaa termiä ”Digital Humanities”, vaan voidaan miettiä tutkimuksen luonnollista kehitystä esimerkiksi historiantutkimuksessa tai sosiologiassa. Tämä tukee myös Helsingin yliopistossa viime vuosina tapahtunutta kehitystä, jossa pelko humanistisen tutkimuksen jakautumisesta kahtia on ohjannut joitain toimintoja digitaalisen humanismin kehittämisessä (Sinnemäki & Tolonen, 2015).

Lokakuussa 2016 pyrimme selvittämään suomalaisten humanististen alojen tutkijoiden digitaalisten aineistojen käyttöä. Kysely lähetettiin kahdeksaan yliopistoon. Kyselytutkimuksen idea oli katsoa tutkimuksen koko elinkaarta. Inés Matresin toteuttamaan kyselyyn saatiin 239 vastausta ja siitä käy selväksi, että humanististen alojen tutkijoiden digitaalisten metodien käyttö vaihtelee suuresti (Matres & Tolonen, 2016).

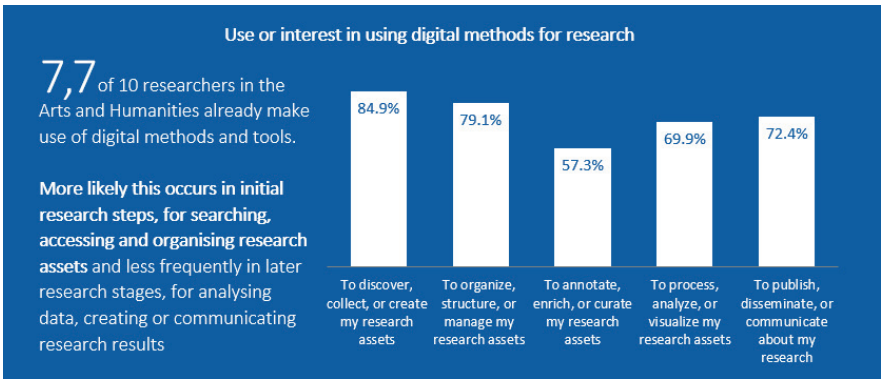
Mielenkiintoa uusien metodien käyttöön ja erilaisten työkalujen tehokkaampaan soveltamiseen riittää. Kuitenkin tulokset painottuvat tiedonhakuun. Vaikuttaa siltä, että metodien käytön erottelu alla olevan kaavion mukaisesti on monille tutkijoille vierasta. Erilaiset analysointi- ja visualisointimenetelmät tai laskennallisten mahdollisuuksien käyttö on useille tutkijoille vielä tuntematon alue.

Tämä johtuu osittain siitä, että humanistinen tutkimus ei ole perinteisesti järjestynyt luonnontieteiden tapaan tutkimusryhmien ympärille. Tietyt tutkimusmenetelmät

[1] Tämä teksti perustuu eri tapahtumissa pitämiini esitelmiin sekä Digital humanities -opetukseen Helsingin yliopistossa yhdessä Eetu Mäkelän kanssa. Osa tekstiä julkaistaan myöhemmin Gaudeamuksen historiantutkimuksen teorioita käsittelevässä kirjan artikkelissa, jonka olen kirjoittanut yhdessä Leo Lahden kanssa. Lisäksi Digitalian toimintaan liittyen lainaan sitä varten laadittuja dokumentteja.

ovat yhdistyneet samalla hyvin homogeeniseen tutkijaidentiteettiin eri aloilla, minkä seurauksena erilaisia laskennallisia menetelmiä sekä tilastollista lähestymistä on myös vieroksuttu humanistien parissa. Kyselytutkimuksen tulokset korostavat, että meidän tulee ajatella Dariahin mallin mukaisesti digitaalista humanismia infrastruktuurina, joka tukee erilaisista taustoista tulevia tutkijoita heidän oman tutkimustraditionsa kehittämisessä ja mahdollistaa yhteistyön tiederajojen yli – ei siis uutena syntyvänä alanaan.

Taulukko 1. *Digitaalisten metodien käyttö humanistisessa tutkimuksessa*



HUMANISTISEN INFRASTRUKTUURIN TARVE

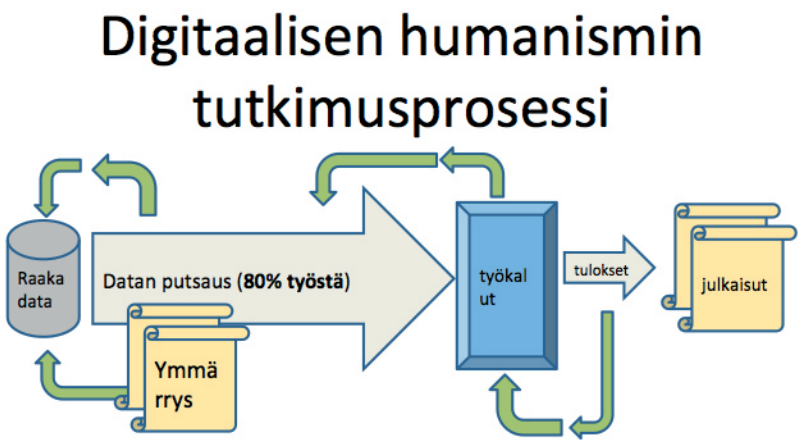
Tarvitaan infrastruktuuri, joka tukee humanistien toimintaa ylittäen tieteenalojen rajoja. Heldig-tutkimuskeskus Helsingin yliopistossa edistää tällaista kehitystä (www.heldig.fi). Myös Digitalialla on hyvä mahdollisuus kehittyä keskeiseksi kansalliseksi toimijaksi tällä alalla. Jos mietitään tätä tehtävää, voidaan esittää lyhyt lista digitaalisen humanismin tavoitteista:

1. Humanistiset tutkimuskysymykset keskiössä
2. Monitieteinen yhteistyö (vrt. laskennallinen biologia)
3. Monipuoliset aineistot (ei vain teksti, materiaalinen kulttuuri)
4. Menetelmien osalta pääosin soveltavaa tutkimusta (vrt. bioinformatiikka)
5. Avoin tiede ja aineistojenhallinta (raakadata, tutkimusdata, julkaisut)
6. Yhteistyö tutkimuksen ja aineistojenhaltijoiden välillä äärimmäisen tärkeää (eri tilanne kuin luonnontieteissä).

Humanistiset tutkimuskysymykset pysyvät keskiössä, jos humanisteilla on keskeinen rooli ja mahdollisuus vaikuttaa siihen, millaisia työkaluja kehitetään ja mitä varten. Digitalialla on yhteys humanistiseen tutkimukseen, jota tulisi yhä kehittää.

Monitieteinen yhteistyö vaatii kärsivällisyyttä ja resursseja. Tärkeintä on yhteinen päämäärä ja kieli. Digitalian kaltaiset keskittymät voivat madaltaa tutkimuksen mahdollisuuksia päästä tätä tavoitetta kohti.

Aineistojen puolesta tekstiaineistot tulevat myös jatkossa näyttelemään keskeistä roolia humanistisessa tutkimuksessa, mutta samalla pitäisi pystyä laajentamaan näkökulmaa kohti materiaalisen kulttuurin tutkimusta sen eri muodoissa. Tässä Digitalialla olisi mahdollisuus kehittää visio, jossa lähestytään digitaalisia aineistoja ja digitaalista tietoa koko tiedontuottamisen ja -käytön elinkaaren kautta.



Kuva 1. *Digitaalisen humanismin tutkimusprosessi (kaavio: Eetu Mäkelä)*

DIGITAALISET AINEISTOT

Kuten yllä olevasta kaaviosta huomaamme, digitaalisen humanismin tutkimusprosessi tulee hyvin lähelle erilaisten digitaalisten aineistojen yleistä käytön kehittämistä. Tiedonlouhintaa varten datan on oltava koneluettavassa muodossa. Parhaassa tapauksessa tutkijat kehittävät aineistoja sekä niiden käyttöä yhdessä esimerkiksi muistiorganisaatioiden kanssa.

Myös tutkijoiden kynnys kehittää omia työkaluja on madaltunut. Tätä tulee ajatella ekosysteeminä, jossa voi löytyä yllättävääkin lisäarvoa yhteistyöstä, jota voidaan avustaa Digitalian suunnasta. Samalla tulee pitää huolta siitä, että skaalautuvuus toteutuu laajojen aineistojen käytössä. Tietojenkäsittelytieteen osalta tutkimus on soveltavaa.

Tärkein tehtävä tässä kokonaisuudessa voidaan muotoilla koskemaan avointa tiedettä sekä aineistojenhallintaa. Raakadataan, tutkimusdataan jsa avoimiin julkaisuihin (sekä datan julkaisuun) liittyy lukemattomia vielä ratkaisemattomia kysymyksiä, joissa Digitalia voi tulevaisuudessa ottaa keskeisen roolin. On kaikkien yhteinen tavoite edistää automaation laajempaa hyödyntämistä digitaalisten aineistojen käsittelyssä ja käyttöön saattamisessa, tunnistaa tiedonlouhintaan sopivat työkalut ja kehittää datan analysoinnin menetelmiä.

Laajojen historiallisten aineistojen käytössä on hyvä muistaa, että innovatiivinen tutkimus nousee usein ruohonjuuritasolta. Siksi aineistojen avoin jakelu on äärimmäisen tärkeä osa tutkimusyhteistyön kehittämistä. Kehitystä ei tapahdu, jos meillä on erikseen työkalujen tuottajat ja tutkijat. Aito yhteistyö aineistojen haltijoiden (esimer-

kiksi kirjastojen) ja tutkijoiden kanssa on välttämätöntä. Avoimuus sekä raaka- että tutkimusdatan kanssa on tie eteenpäin.

Esimerkkinä yhteistyöstä humanistien ja muistiorganisaatioiden välillä voidaan käyttää digitoituja lehtiaineistoja. Kansalliskirjaston digitointi- ja konservointikeskuksessa Mikkelissä tuotettavan ja ylläpidettävän digitoitujen sanomalehtien ja aikakauslehtien palvelun kehittämiseen sisältyy mm. aineistojen tekstintunnistuksen laadun parantaminen, joka on yhteinen myös tekstilouhintaa tekeville tutkimusryhmille.

Kyseessä on myös kansallisesti merkittävä tehtävä. Digitoitujen lehtien ja pienjulkaisujen verkkopalvelulla digi.kansalliskirjasto.fi on vuosittain noin 100 000 käyttäjää, ja siitä tehdään noin 18 miljoonaa sivulatausta. Palvelun käyttäjiä ovat yksityiset ihmiset ja yhteisöt. Laajan kansalaiskäytön ja tutkimuskäytön suosio on kasvanut jatkuvasti, yli 20% vuosittain. Digitoidut lehdet tarjoavat tutkimukselle ja kansalaisille laajan ja monipuolisen aineiston.

DIGITAALINEN HISTORIANTUTKIMUS

Digitalian piirissä tullaan jatkossa osallistumaan myös suoraan digitaaliseen historiantutkimukseen, koska Kansalliskirjasto on mukana Digitaalinen historiantutkimus ja julkisuuden muutos Suomessa 1640–1910 -hankkeessa (COMHIS). Hankkeessa tutkitaan julkista keskustelua historiallisissa sanomalehdissä ja suomalaista tiedontuotantoa osana eurooppalaista kehitystä.

Hankkeessa ovat mukana Kansalliskirjaston digitointi- ja konservointikeskuksen lisäksi Helsingin yliopiston humanistinen tiedekunta sekä Turun yliopiston kulttuurihistorian ja informaatioteknologian laitokset. Lähtökohtana on tutkia ja arvioida uudelleen suomalaisen julkisen keskustelun luonnetta ja laatua vuosina 1640–1910. Hankkeessa analysoidaan, miten kielirajat, eliittikulttuuri ja populaari keskustelu, tekstien uudelleen käyttö ja julkaisujen kanavat olivat vuorovaikutuksessa keskenään. COMHIS on uusi avaus muistiorganisaatioiden toiminnassa digitaalisen humanismin alalla; se edistää avointa tiedettä ja tiivistää tutkijoiden ja muistiorganisaatioiden avointa yhteistyötä.

Digitaalisuuden haasteet pitävät sisällään niin yhteiskunnan rakenteelliset muutokset kuin tutkimukseen ja opetukseen liittyvät muutokset. Kuten yllä jo todettiin, Helsingin yliopistoon on hiljattain perustettu Heldig, digitaalisen humanismin keskus. Kyseessä on hyvin kunnianhimoinen, kasvava hanke. Koska yhteistyö Digitalian kanssa on tiivistä, kyseessä on Mikkelin alueelle myös erinomainen mahdollisuus kehittyä ja erikoistua digitaalisen tiedon tutkimuksessa.

Esimerkiksi big datan käytön suhteen Suomi voi nousta kansainväliseen eturintamaan panostamalla humanistiseen tutkimukseen yhteistyössä julkisen sektorin kanssa. Humanistien tuhansien vuosien osaaminen ihmisten ajattelun ja käyttäytymisen historian mallintamisessa luo uusia tapoja ajatella systemaattisemmin digitalisoituvaa yhteiskuntaa ja sen perustavia rakenteita. Historialliset analyysit yhdistettynä laajeneviin tietoaineistoihin ovat väylä kohti huomista. Tähän tarvitaan digitaalisten aineistojen, niiden käsittelyn ja digitoimisprosessin tuntemusta, joita Digitalia tarjoaa.

BIG DATA JA ILMIÖKESKEISYYS

Käsi kädessä digitaalisuuden kanssa kulkee erilaisissa keskusteluissa big data. Sillä tarkoitetaan laajaa, järjestelemätöntä aineistoa, josta tutkimuksessa yritetään etsiä yleisiä trendejä usein tilastolliseen analyysiin luottaen. Karkeakin aineisto voi näin olla tutkimuksellisesti arvokasta, kun pyritään tunnistamaan laajoja tilastollisia säännönmukaisuuksia. Menetelmät voivat sietää huomattaviakin määriä satunnaisia virheitä ja epätarkkuuksia, joiden vaikutus laajempien trendien havaitsemiseen on usein rajattu. Laajojen aineistojen kvantitatiiviseen analyysiin perustuvassa tutkimuksessa datasta voidaan tehdä myös odottamattomia havaintoja. Nämä voivat suunnata tutkimusta tavalla, jota ei alunperin ole osattu ennakoida.

Molemmilla lähestymistavoilla – huolella laadittujen pienempien aineistojen ja laajempien karkeiden aineistojen tutkimuksella – on oma paikkansa. Lähestymistavasta riippumatta aineiston siistiminen käyttökelpoiseen muotoon on tutkimuksessa yksi eniten aikaa vievistä vaiheista.

Big data ymmärretään usein pelkistäen ja keskustelu aineiston koosta on vienyt paljon tilaa muulta keskustelulta. Aineiston kokoa huomattavasti mielenkiintoisempaa on muutos kohti ilmiöpohjaista lähestymistä tutkimukseen. Digitaalinen humanismi ja ns. big data -tutkimus onkin usein ilmiöpohjaista. Kohteena on tietty ilmiö (esimerkiksi kaupunki tai tiedontuotanto), jota tarkastellaan laajan, monimuotoisen ja usein digitaalisen aineiston kautta erilaisista näkökulmista, usein osana laajempaa tutkimuskokonaisuutta, jossa tietokoneavusteisiin menetelmiin yhdistyy esim. sosiolingvistinen tai sosiologinen tutkimus. Tähän liittyy ilmiön historiallinen analysointi, mutta tutkimuksen tarkoituksena on usein löytää myös näkökulmia, joilla on merkitystä esimerkiksi poliittisessa päätöksenteossa.

Syy ilmiökeskeisyydelle on usein käytäntö: eri tieteenalojen kohdatessa on järkevää tutkia nimenomaan ilmiöitä eri näkökulmista. Se pakottaa myös tutkijat ulos aikaisemmin hyvin tarkkaan varjelluista asiantuntijaroolin rajoista, joissa paradoksaalisesti tutkija itse on saanut määritellä oman mukavuusalueensa. Kun tarkoituksena on ymmärtää hankalasti hallittavia kokonaisuuksia, kasvaa tarve yhteistyölle eri asiantuntijoiden kanssa. Tämä on jotain, missä humanistit ovat olleet viimeisten vuosikymmenten eriytymisen aikana erityisen huonoja.

TUTKIMUKSEN MONIMUOTOISUUS

Sanomalehdet digitaalisen humanismin tutkimuskohteena ovat oiva esimerkki monimuotoisesta tutkimuksesta. Niitä voidaan käyttää niin kuin ennenkin yhtenä tutkimuksen lähderyhmänä muiden lähteiden joukossa. Samalla kuitenkin sanomalehtien materiaalisuus, fyysiset ominaisuudet, säännönmukaisuudet yms. muodostuvat mielenkiintoiseksi tutkimuskohteeksi, joka voi olennaisesti auttaa esimerkiksi laadullisissa tulkinnoissa julkisesta keskustelusta Suomessa.

Toinen erinomainen esimerkki ilmiöpohjaisesta lähestymisestä ovat erilaiset kaupunkitutkimuksen hankkeet. Tähän liittyy merkittäviä uusia mahdollisuuksia: tutkimuskenttä muuttuu ja kaupunki ilmiönä ymmärretään myös aivan eri tavalla kuin aikaisemmin. Voidaan tutkia, miten kaupunkien toimintaan liittyvää avointa ja suljettua

lähes reaaliaikaistakin dataa voidaan käsitellä ja hyödyntää kaupunkien kehityksen kontekstissa. Kaupunkeihin ja yhteiskuntaan liittyvää dataa on saatavilla yhä enemmän. Kyseessä on uudenlaisten tutkimusaineistojen ja -menetelmien käyttöönotto laadullisen tutkimuksen tueksi jo vakiintuneiden menetelmien rinnalle. Humanistien panos voi tuottaa tuoreita näkökulmia, kun tutkitaan tällaisten ilmiöiden kehittymistä ja yhteyksiä ihmisten käyttäytymiseen, hyvinvointiin, vuorovaikutukseen ja maailmankuvaan.

Suomen Akatemian digitaalisten ihmistieteiden määritelmässä mainitaan digitaalisuuden tutkimus. Kun mietitään digitaalisen tiedon luonnetta ja tarkastellaan asiaa tätä kautta, nähdään selvästi yhteys myös metodien käyttöön, kuten sen edellä määrittelin. Näin säilytetään keskusteluyhteys uusien menetelmien käytön ja digitaalisuuden tutkimuksen välillä. Digitaalinen humanistinen tutkimus voi olla myös sellaista, missä ei käytetä uusia metodeja, vaan keskitytään digitaalisen tiedon luonteeseen. Sen kautta tutkimus kuuluu digitaalisten ihmistieteiden piiriin.

Digitaalisissa ihmistieteissä korostuu tutkimuksen monitieteinen tutkimusote, yhteistyö yli tiederajojen sekä pragmaattisuus. Yksi pääväitteistäni onkin, että suurin muutos digitaalisen otteen yleistymisessä historiantutkimuksessa ei suinkaan ole teknologian lisääminen humanistien työkalupakkiin, vaan juuri ilmiöpohjaisen tutkimuksen mukaantulo. Se vaatii humanisteilta laajentumista oman tieteenalan ulkopuolelle. Ei ole enää relevanttia pohtia historiantutkimuksen metodeja ja teoriaa pelkästään suhteessa omaan oppialaan. Suurempi muutos on yrittää ymmärtää historiantutkimuksen aikaisempaa teoreettista pohjaa osana laajempaa humanistista ja sosiaalitieteiden perinnettä, jossa ihmisten toiminnan mallintaminen on kulkenut jo jonkin aikaa eri reittejä suhteessa historiantutkimukseen. Haasteena on sovittaa nämä paremmin yhteen tulevaisuudessa. Samalla myös tietokoneavusteisten menetelmien mahdollisuudet pitää pystyä hyödyntämään käytäntöön menettämättä humanistisen tutkimuksen hermeneuttista erityispiirrettä.

LOPUKSI

Yleisesti on ajateltu, että humanistit ovat kiinnostuneita nimenomaan laadullisista kysymyksistä. Kuitenkin jos tätä miettii esimerkiksi sen kontekstin tai aikajänteen kautta, mitä historiankirjoituksessa usein tavoitellaan (”valistus”, ”keskiaika”, ”antiikki”), niin erilaiset kvantitatiivinen-kvalitatiivinen -jaottelut eivät ole kovinkaan hedelmällisiä.

Yksi uusista suuntauksista onkin – aivan oikein – liikkua pois tästä vanhanaikaisesta jaottelusta, joka ei itse asiassa kerro mistään mitään, mutta on sen sijaan tehnyt monimutkaisista kysymyksistä turhan yksioikoisia. Aivan samalla tavalla luonnontieteissä ja muilla aloilla tarvittaisiin paljonkin humanistien osaamista liittyen vaikkapa erilaisiin elämään liittyviin sykleihin ja kulttuurin kerrostumiin.

Hyvän digitaalisen humanismin tutkimuksen tunnistaakin siitä, että se tietoisesti häivyttää vanhanaikaista laadullinen/tilastollinen -erottelua merkityksettömänä taka-alalle. Laadullisen tutkimuksen ja modernin tieteellisen laskennan yhdistäminen avaa polkuja kohti tutkimuksen uudistumista. Se vaatii kuitenkin pitkäjänteistä ja päämäärätietoista työtä (Tolonen & Lahti, 2018). Tämä on kenttä, jonka yhteyteen Digitalia voi rakentaa tulevaisuuttaan.

LÄHTEET

Dariah Desir hankkeen dokumentti, 2016. WWW-dokumentti. Saatavissa: http://cordis.europa.eu/project/rcn/207190_en.html [viitattu 27.4.2017]

Digitaaliset ihmistieteet DIGIHUM (2016-2019), 2015. WWW-dokumentti. Saatavissa: http://www.aka.fi/digihum [viitattu 27.4.2017]

Matres, I. & Tolonen, M. 2016. Finnish survey on digital research practice in the Arts and Humanities, WWW-dokumentti. Saatavissa: https://www.helsinki.fi/en/news/finnish-survey-on-digital-research-practice-in-the-arts-and-humanities [viitattu 27.4.2017]

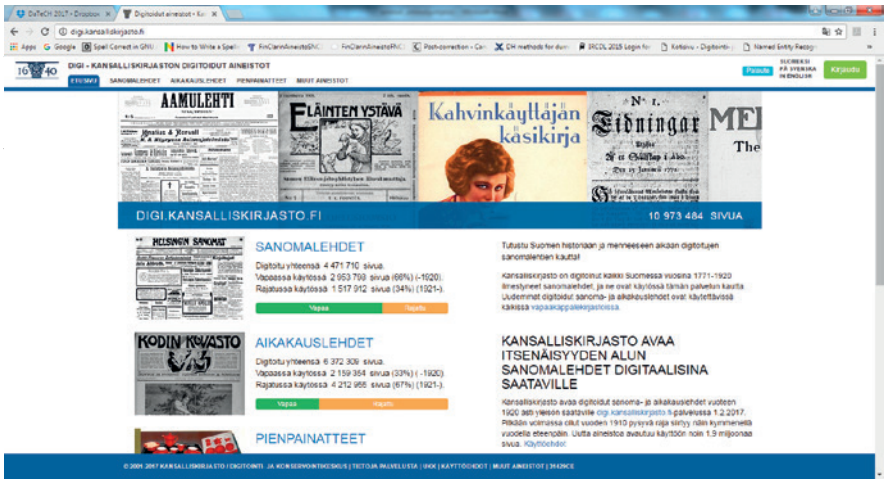
Sinnemäki, K & Tolonen, M. 2015. Digitaaliset ihmistieteet tutkimuskartalle. Tieteessä tapahtuu, [S.l.], v. 33. 2015. ISSN 1239-6540. Saatavissa: http://journal.fi/tt/article/view/51172 [viitattu 27.4.2017]

Tolonen, M. & Lahti, L. 2018. Digitaalinen humanismi historiantutkimuksessa. Teoksessa Matti Hannikainen et al. (toim.) Puut ja metsä – teoriat historiantutkimuksessa. Helsinki: Gaudeamus.

KANSALLISKIRJASTO DIGITALIA-HANKKEESSA

Kimmo Kettunen, *FT, tutkimuskoordinaattori, Kansalliskirjasto*
Mika Koistinen, *DI, tietojärjestelmäasiantuntija, Kansalliskirjasto*
Teemu Ruokolainen, *TkT, tutkijatohtori, Kansalliskirjasto*
Tuula Pääkkönen, *FM, tietojärjestelmäasiantuntija, Kansalliskirjasto*

Kansalliskirjaston Mikkelin digitointi- ja konservointikeskus on digitoinut Suomessa julkaistuja sanoma- ja aikakauslehtiä sekä pienpainatteita kattavasti vuodesta 1998 alkaen. Koko aineisto on käytettävissä Kansalliskirjaston verkkopalvelussa digi.kansalliskirjasto.fi. Aineiston määrä on kasvanut vuosien aikana tasaisesti, samoin aineiston käyttäjien määrä. Alkuun lehtiaineistoa oli verkkopalvelussa vapaasti saatavilla vuosilta 1771–1910. Vuoden 2017 helmikuun alussa aineisto avattiin saataville vuoden 1920 loppuun saakka. Koko aineisto vuosilta 1771–1910 on saatavilla nyt myös avoimen datan jakelupakettina verkkopalvelun sivuilta (Pääkkönen ym., 2016).



Kuva 1. *Digi.kansalliskirjasto.fi:n etusivu*

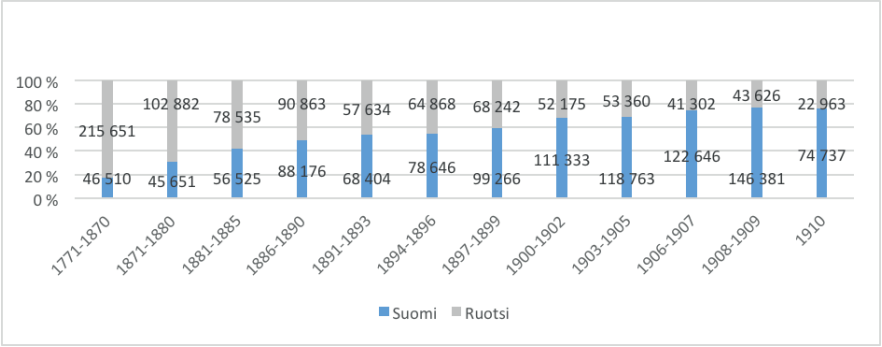
Kansalliskirjaston digitoitu lehtikokoelma on osa maailmanlaajuista laajenevaa historiallisten lehtien digitoitua tarjontaa. Europeana-projekti arvioi vuonna 2012 Euroopassa olevan digitoituja sanomalehtiä noin 129 miljoonaa sivua ja noin 24 000 nimikettä (Dunning, 2012). Varovainen arvio digitoitujen lehtinimikkeiden määrästä maailmanlaajuisesti on yli 45 000. Suurin osa aineistosta on digitoitu Euroopassa ja Yhdysvalloissa. (The ”State of the Art”, 2015).

Digitalia-hankkeessa Kansalliskirjasto kehitti ja tutki digi.kansalliskirjasto.fi -palvelua ja sen aineistoja eri tavoin. Sanomalehtiaineiston suomenkieliselle osalle tehtiin sa-

natason laatuarviota, aineiston optisen luvun tasoa kehitettiin ja aineiston jälkikorjauksen mahdollisuuksia selvitettiin yhdessä kieliteknologisen FIN-CLARIN-konsortion tutkijoiden kanssa. Sanomalehtiaineistosta tehtiin nimien tunnistamisen koeaineisto, jolla tutkittiin erityisesti henkilön- ja paikannimien tunnistamisen mahdollisuutta kokoelmassa. Digitoitujen sanomalehtien sivuilta aloitettiin myös kokeellinen artikkelien eristäminen. Palvelun käyttöliittymään tehtiin erilaisia käyttäjiä palvelevia muutoksia. Näistä kaikista tehdään lyhyt katsaus tässä artikkelissa. Lopuksi kerrotaan lyhyesti hankkeen aikana tehdystä tutkimusyhteistyöstä.

LEHTIAINEISTON LAATUARVIO JA LAADUN PARANTAMINEN

Digi-verkkopalvelun aineisto on pääasiassa suomen- ja ruotsinkielistä, mutta seassa on myös vähäisiä määriä saksaa, venäjää, ranskaa ja muita kieliä. Kuvassa 2 esitetään sanomalehtiaineiston jakautuminen suomen- ja ruotsinkieliseen materiaaliin sivuina vuosina 1771–1910. Aikavälit kuvassa ovat aineistosta tehdyn avoimen datan jakelupaketin mukaisia (Pääkkönen ym., 2016). Vuoteen 1890 saakka suurempi osa aineistosta on ruotsinkielistä, siitä eteenpäin julkaistaan enemmän suomeksi. Suomenkielisten sivujen kokonaismäärä aikajaksolla on 1 063 648, ruotsinkielisten 892 101.

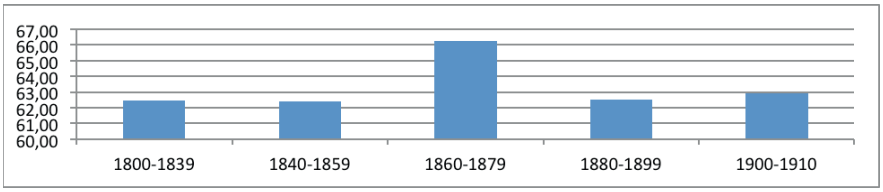


Kuva 2. *Suomen ja ruotsin osuus sanomalehtiaineistossa sivuina vuosina 1771–1910*

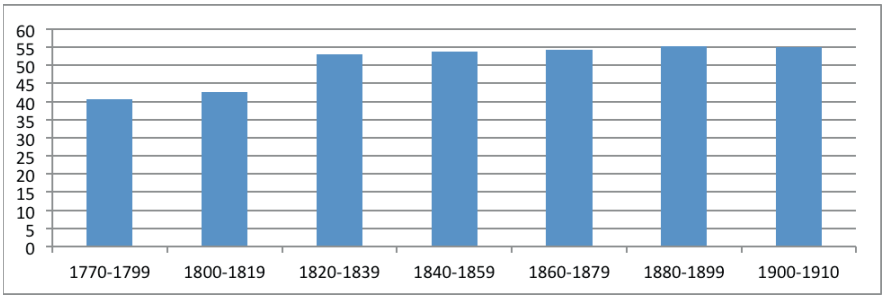
Digitoidun lehtiaineiston laadun parantaminen oli yhtenä työpakettina Digitalia-hankkeen Kansalliskirjaston osiossa. Työhön on sisältynyt seuraavat osat: tekstien nykylaadun arviointi, uuden optisen luvun mahdollistaminen ja aineiston jälkikorjauksen yrittäminen.

Suomenkielinen lehtiaineisto on laaja. Sen arvioitu sanamäärä vuosilta 1771–1910 on noin 2.4 miljardia sanaa. Suomenkielisen lehtiaineiston laatua arvioitiin empiirisesti käyttäen avuksi kieliteknologisia sanatason analyysiohjelmia. Tehtyjen analyysien perusteella päädyttiin siihen, että noin 70–75 % lehtiaineiston sanoista on tunnistettavia. Tunnistamaton 25–30 % on joko optisen luvun tuottamia virheitä tai tunnistusohjelmille vierasta sanastoa. Tarkemmin tehtyä analyysia ja tuloksia kuvataan lähteissä Kettunen & Pääkkönen (2016a,b), sekä Kettunen ym. (2016a,b).

Aineiston ruotsinkielinen osa on myös analysoitu yleisellä tasolla. Ruotsia aineistossa on noin 5.56 miljardia sanaa. Niistä suurin osa, noin 5.19 miljardia sanaa, on sanomalehdissä. Kuvissa 3 ja 4 esitetään sanojen tunnistus ruotsinkielisessä aineistossa.



Kuva 3. *Sanojen tunnistusprosentit ruotsinkielisissä sanomalehdissä*



Kuva 4. *Sanojen tunnistusprosentit ruotsinkielisissä aikakauslehdissä*

Digitoitujen vanhojen lehtien melko matala sanatason laatu voi johtua monesta tekijästä: huonosta alkuperäisten lehtien painolaadusta, kehnosta paperista, mikrofilmin laadusta, skannauksen resoluutiosta, hankalista kirjasinlajeista jne. (vrt. Holley 2008; Klijn 2008; Piotrowski, 2012), ja ilmiö on yleinen kaikissa laajoissa vanhojen lehtien digitointiprojekteissa. Esimerkiksi British Libraryn *19th Century Newspaper Project* on arvioinut noin yhden prosentin otoksella kahden miljoonan projektissa digitoidun painosivun laadun ja saanut tulokseksi, että noin 78 % aineiston sanoista on oikein, loput virheellisiä (Tanner ym., 2009). Tulosta ei voi pitää laadullisesti hyvänä, mutta se on realismia. Niklasin (2010) aineisto kattaa *The Times of Londonin* digitoidut vuosikerrat 200 vuoden ajalta vuosilta 1785–1985. Aineistossa on noin 7 miljardia sananmuotoa ja 8 miljoonaa artikkelia. Sanojen tunnistettavuus *The Times of Londonin* aineistossa vaihtelee 55–80 % välillä. Kautta koko 1900-luvun sanojen tunnistettavuus pysyttelee pääasiassa 70–80 prosentissa.

Digitoidut lehtiaineistot tuotetaan kuvaamalla lehtien sivut ja tunnistamalla sivujen sisältö optisella merkkien tunnistuksella. Optinen merkintunnistusohjelma vastaa viime kädessä siitä, miten hyvin kuvatun sivun teksti tunnistuu. Koska kirjaston käytössä ollut optisen luvun ohjelmisto alkaa olla vanha eikä sitä ole mahdollista päivittää lisenssien kalleuden vuoksi, olemme tehneet Digitalia-hankkeen aikana työtä uuden avoimeen lähdekoodiin perustuvan Tesseract-ohjelmiston[2] käyttöön saattamisessa.

Aineiston uutta optista lukemista varten luotiin ensin noin 500 000 sanan otos, josta tehtiin ihmistyönä tarkistettu mahdollisimman virheetön versio. Tätä aineistoa voidaan käyttää tekstien laadun parantumisen arvioimisen testiaineistona.

Erityisesti sanomalehdet on painettu Suomessa vielä 1900-luvun alkuvuosikym-

[2] https://github.com/tesseract-ocr

meniin saakka pääosin fraktuura-kirjasimella. Kirjasin on tunnetusti hyvin hankala tulkittava optisille lukuohjelmille (Holley, 2008; Piotrowski, 2012). Tesseractia varten luotiin hankkeessa malli suomen kielen fraktuura-kirjasimelle ja samalla parannettiin ohjelman kykyä käsitellä sanomalehtien digitoituja sivukuvatiedostoja.

Kehitystyön jälkeen Tesseractilla tehdyssä koeaineiston merkintunnistuksessa sanojen virheprosentti saatiin laskemaan 19.02:een alkuperäisestä 26.23 prosentista, mikä on 7.21 prosenttiyksikön parannus alkuperäiseen (Koistinen ym., 2017). Tesseract-ohjelmisto integroidaan Kansalliskirjaston tuotantojärjestelmään ja jatkossa koko lehtiaineisto tunnistetaan uusiksi Tesseractilla.

Toinen mahdollisuus parantaa tekstien laatua on tehdä aineistolle optisen luvun jälkeen jälkikorjaus ohjelmallisesti. Tällainen työ perustuu kieliteknologian tilastollisiin malleihin. Aineiston jälkikorjausta on valmisteltu yhteistyössä Helsingin yliopiston kieliteknologian tutkijoiden ja FIN-CLARINin kanssa. Jälkikorjauksesta on arvioitu kolme eri ohjelmaversiota, ja mahdollisuuksien mukaan jokin näistä pyritään liittämään Kansalliskirjaston tuotantojärjestelmään. Uuden optisen luvun ja jälkikorjauksen yhdistelmällä tekstien laatua voidaan todennäköisesti parantaa niin, että arviolta noin 80 % suomen sanoista aineistoissa on tunnistettavia.

NIMIEN ERISTÄMINEN LEHTIAINEISTOSTA

Ihmisten, paikkojen ja yritysten nimiä pidetään yleisesti tärkeinä erilaisissa historiallisissa aineistoissa. Aineistojen käyttäjät sijoittavat monesti myös itsensä aineiston maailmankartalle sekä sosiaalisesti että maantieteellisesti nimien kautta. Tiedonhaun tutkimuksesta tiedetään, että historialliseen tekstiaineistoon tehdään paljon hakuja erilaisia nimiä käyttäen. Tämän vuoksi nimiä voidaan käyttää myös auttamaan aineiston selailukäyttöä eristämällä tekstiaineistosta nimiä ja tekemällä niiden avulla hakemistoja aineistoon (Neudecker ym., 2014).

Kansalliskirjastossa on luotu Digitalia-hankkeen aikana nimien eristämisen tutkimiseen testiaineisto. Sen avulla on tutkittu erilaisten käytettävissä olevien nimien tunnistamisen ohjelmistojen kykyä löytää nimiä aineistosta. Testitulosten mukaan nimien löytyminen paikoin huonolaatuisen tekstin seasta on mahdollista, mutta tulokset eivät ole erityisen hyviä. Kun tekstien laatua saadaan parannetuksi, parantuu myös nimien löytyvyys. Nykyiset tarjolla olevat nimiä etsivät ohjelmat eivät kuitenkaan ole optimaalisia vanhan tekstiaineiston kanssa. Siksi Kansalliskirjastossa on ryhdytty toteuttamaan uutta erityisesti vanhoille aineistoille soveltuvaa nimien tunnistusohjelmaa, joka perustuu tilastolliselle koneoppimiselle. Uusi ohjelma vaatii laajan opetusaineiston, jonka tuottaminen on aloitettu.

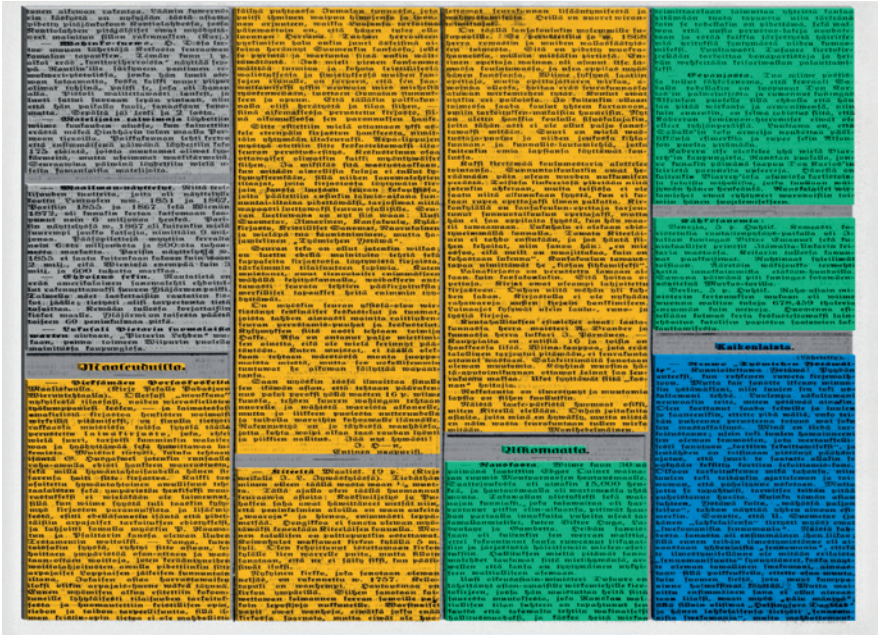
Nimien löytymistä ja nimien käyttömahdollisuuksia kokoelmassa on kuvattu tarkemmin lähteissä Kettunen ym. (2016, 2017) sekä Kettunen & Ruokolainen (2017). Hyvän käsityksen nimien käytöstä digitoidun sanomalehtiarkiston käyttöliittymässä antaa esimerkiksi italialaisen La Stampan historiallinen kokoelma. Siinä nimiä käytetään artikkelien selauksen tarkennuksessa sen jälkeen, kun kokoelmaan on tehty ensin haku jollain hakusanalla.

ARTIKKELIEN ERISTÄMINEN AINEISTOSTA

Kansalliskirjaston lehtiaineiston nykyinen esitysmuoto perustuu kokonaisiin digitoituihin painettuihin sivuihin. Aineistoon tehtävän tiedonhaun ja muun käytön kannalta tämä ei ole paras mahdollinen ratkaisu, ja siksi Digitalian yhtenä työpakettina on ollut artikkelien eristäminen digitoiduilta sivuilta. Artikkelit ovat käyttäjän kannalta lehtien perustava informatiivinen sisältö. Mahdollisuus kohdistaa tekstihaut suoraan artikkeleihin parantaisi digitoitujen lehtien käytettävyyttä. Myös kaikenlainen muu aineiston poiminta ja käsittely hyötyisi artikkelien eristämisestä.

Olemme selvittäneet erilaisia mahdollisuuksia toteuttaa artikkeleiden eristäminen lehdille. Tarjolla olevat kaupalliset ratkaisut ovat olleet joko liian kalliita tai laadultaan riittämättömiä ja liikaa jälkikorjausta vaativia. Selvitystyön aikana on löydetty yksi lupaava tutkimusohjelmisto, Rouenin yliopiston LITIS-laboratoriossa tehty PIVAJ (Hebert ym., 2014). Se on ollut Kansalliskirjastossa koekäytössä helmikuusta 2017 alkaen.

Koekäytön aikana olemme selvittäneet ohjelman kykyjä käyttäen kokeiluaineistona yhtä verkkopalvelun suosituimmista sanomalehdistä, Uutta Suometarta. Lehteä ilmestyi tällä nimellä vuosina 1869–1918 yhteensä 86 068 sivua. Lehdessä oli aluksi kolme palstaa, mutta palstojen määrä lisääntyi enimmillään yhdeksään, josta se palasi takaisin kuuteen palstaan. Palstojen määrän vaihtelun lisäksi artikkelien eristämiselle tuottavat ongelmia muun muassa mainosten lisääntyminen sekä erilaiset muut sivujen ulkoasun muutokset. Kuvassa 5 on PIVAJ'n tuottama visuaalinen mallisivu Uuden Suomettaren sivun sisällön jakautumisesta eri artikkeleihin. Kutakin ohjelman tunnistamaa artikkelia vastaa oma väri.



Kuva 5. *PIVAJ'n erottelemat artikkelit Uuden Suomettaren sivulla*

Kuvasta näkyy, että PIVAJ kykenee erottamaan sivulla olevat artikkelien osastot toisistaan, mutta osaston sisällä olevia lyhyitä artikkeleita ohjelma ei kykene erottamaan.

DIGI.KANSALLISKIRJASTO.FI –KÄYTTÖLIITTYMÄÄN TEHDYT UUDISTUKSET

Digi.kansalliskirjasto.fi -palvelun hakuliittymää on selkeytetty ja monipuolistettu Digitalia-hankkeen aikana. Hakurajauksista, kuten lehtien ilmestymispaikkakunnista, on tehty listavalinta, kuten myös nimikkeistä. Uusittu haku mahdollistaa myös tarkkojen tekstihakujen teon ja fraasihaun käytettävyyttä on parannettu.

Palvelun sanomalehtisivuille on lisätty toiminto, jolla yksittäisen sivun tekstisisällön saa näkyviin käyttöliittymään. Käyttöliittymästä on mahdollista ladata sivun teksti joko rakenteisessa XML-muodossa tai normaalina tekstinä. Sivulta saa myös ladattua suoraan aineiston viitetiedot, joko tekstitiedosto-, BibTex- tai RefTex-muodossa, jotka soveltuvat erilaisiin käyttöihin.

Digin käyttäjätutkimuksissa on havaittu, että palvelua käyttävät monet tutkijat (Hölttä 2016; Matres, 2016). Tutkijoiden tarpeita varten käyttöliittymään on lisätty hakujen hakusanojen ja tuloslinkkien tallennus Excel-taulukkona. Hakusanojen esiintymiä lehdissä voidaan tarkastella myös aikajanalla. Nämä ominaisuudet palvelevat kaikkia sivuston käyttäjiä.

Kuvitusten poimintaan lehdistä on toteutettu omat välineet. Ne tallentavat digitoinnin jälkikäsittelystä saatavat kuvitustiedot tietokantaan ja sitä käyttävään hakuindeksiin. Käyttäjille ominaisuus näkyy uutena *kuvitushaku*-toimintona.

TUTKIMUSYHTEISTYÖ

Kansalliskirjasto on tehnyt Digitalia-hankkeessa tutkimusyhteistyötä useiden tahojen kanssa. Merkittävintä yhteistyö on ollut Suomen Akatemian rahoittaman Turun ja Helsingin yliopiston historiantutkijoiden COMHIS-projektin kanssa, jossa Kansalliskirjasto on myös mukana. Hankkeessa tutkitaan julkista keskustelua historiallisissa sanomalehdissä ja suomalaista tiedontuotantoa osana eurooppalaista kehitystä. COMHISia varten kirjaston lehtiaineistosta tuotettiin paketti, joka luovutettiin COMHISin käyttöön maaliskuussa 2016 (Pääkkönen ym., 2016). Pakettia kehitetään myöhemmin tutkijoilta tulleiden toiveiden ja tarpeiden mukaan. Paketti avattiin maaliskuussa 2017 Kansalliskirjaston omilla sivuilla kaikkien saataville.

Kieliteknologian soveltamisessa on tehty yhteistyötä FIN-CLARIN-konsortion kanssa. Konsortiolta on saatu käyttöön muun muassa nimien eristämisen ohjelmistoja ja sanankorjausmalleja. FIN-CLARIN on käyttänyt Kansalliskirjaston aineistoja sanojen jälkikorjausalgoritmien kehittämisessä (Silfverberg ym., 2016).

LÄHTEET

Dunning, A. 2012. European Newspaper Survey Report. PDF-dokumentti. Saatavissa: http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf [viitattu 12.4.2017].

Hebert, D., Palfray, T., Nicolas, T., Tranouez, P. & Paquet, T. 2014. Automatic article extraction in old Newspapers Digitized Collections. Teoksessa Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 3–8. Saatavissa: http://dl.acm.org/citation.cfm?id=2595195 [viitattu 12.4.2017].

Holley, R. 2009. How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, 15(3/4). Saatavissa: http://www.dlib.org/dlib/march09/holley/03holley.html [viitattu 12.4.2017].

Hölttä, T. 2016. Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat. Pro gradu, Informaatiotutkimuksen ja interaktiivisen median tutkinto-ohjelma, Tampereen yliopisto. PDF-dokumentti. Saatavissa: https://tampub.uta.fi/handle/10024/98714 [viitattu 12.4.2017].

Kettunen, K., Mäkelä, E., Kuokkala, J., Ruokolainen, T. & Niemi, J. 2016. Modern Tools for Old Content – in Search of Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. Teoksessa Krestel, R., Mottin, D. & Müller, E. (toim.) LWDA 2016, Lernen, Wissen, Daten, Analysen 2016. Saatavissa: http://ceur-ws.org/Vol-1670/ [viitattu 12.4.2017].

Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J. & Löfberg, L. 2017. Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. PDF-dokumentti. Saatavissa: https://arxiv.org/abs/1611.02839 [viitattu 12.4.2017].

Kettunen, K. & Pääkkönen, T. 2016a. Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. Teoksessa Calzolari, N. ym. (toim.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) Saatavissa: http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf [viitattu 12.4.2017].

Kettunen, K. & Pääkkönen, T. 2016b. How to do lexical quality estimation of a large OCRed historical Finnish newspaper collection with scarce resources. PDF-dokumentti. Saatavissa: https://arxiv.org/abs/1611.05239 [viitattu 12.4.2017].

Kettunen, K., Pääkkönen, T. & Koistinen, M. 2016a. Between Diachrony and Synchrony: Evaluation of Lexical Quality of a Digitized Historical Finnish Newspaper and Journal Collection with Morphological Analyzers. Teoksessa Skadina, I. & Rozis, R. (toim.) Human Language Technologies – The Baltic Perspective. Saatavissa: http://ebooks.iospress.nl/ISBN/978-1-61499-701-6 [viitattu 12.4.2017].

Kettunen, K., Pääkkönen, T. & Koistinen, M. 2016b. Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen. Informaatiotutkimus 3, 3–14. Saatavissa: http://journal.fi/inf/article/view/59433 [viitattu 12.4.2017].

Kettunen, K. & Ruokolainen, T. 2017. Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. DATeCH 2017 (ilmestyy).

Klijn, E. 2008. The Current State-of-art in Newspaper Digitization. A Market Perspective. D-Lib Magazin 14(1/2). Saatavissa: http://www.dlib.org/dlib/january08/klijn/01klijn.html [viitattu 12.4.2017].

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. Nodalida 2017 (ilmestyy).

Matres, I. 2016. Digi.kansalliskirjasto.fi information source and tool for academic research. Informaatiotutkimus, 3, 50–51. Saatavissa: http://journal.fi/inf/article/view/59437/20618 [viitattu 12.4.2017].

Neudecker, C., Wilms, L., Faber, W. J. & van Veen, T. 2014. Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition. IFLA 2014 Saatavissa: http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf [viitattu 12.4.2017].

Niklas, K. 2010. Unsupervised Post-Correction of OCR Errors. Diploma Thesis, Leibniz Universität, Hannover. Saatavissa: www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf [viitattu 12.4.2017].

Piotrowski, M. 2012. Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K. & Mäkelä, E. 2016. Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August. Saatavissa: http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html [viitattu 12.4.2017].

Silfverberg, M., Kauppinen, P., & Linden, K. 2016. Data-Driven Spelling Correction Using Weighted Finite-State Methods. Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata, 51–59. Saatavissa: https://aclweb.org/anthology/W/W16/W16-2406.pdf [viitattu 12.4.2017].

The "State of the Art": A Comparative Analysis of Newspaper Digitization to Date. 2015. PDF-dokumentti. Saatavissa: http://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf [viitattu 12.4.2017].

Tanner, S., Muñoz, T. & Ros, P. H. 2009. Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine, (15/8). Saatavissa: http://www.dlib.org/dlib/july09/munoz/07munoz.html [viitattu 12.4.2017].

DIGI.KANSALLISKIRJASTO.FI JA DIGITAALISET PALVELUT TUTKIJOILLE

Tuula Pääkkönen, *FM, tietojärjestelmäasiantuntija, Kansalliskirjasto / Digitointi- ja konservointikeskus*

Kansalliskirjasto kehittää digitoitujen ja digitaalisten aineistojen taustajärjestelmää digi.kansalliskirjasto.fi jatkuvasti. Digitalia-hankkeessa päätavoite oli toimia yhdessä Digitalian tutkimusryhmän kanssa ja löytää uusia tapoja hyödyntää tutkimustuloksia taustajärjestelmän kehityksessä. Tavoite oli myös tunnistaa ja toteuttaa toimintoja ja palveluita, joista erityisesti tutkijat muiden käyttäjien ohella voisivat hyötyä.

Digitalia-hankkeessa verkostoja tutkimusryhmiin alettiin muodostaa. Käyttäjäkyselyissä (Hölttä, 2016; Matres, 2016) erilainen ja laaja-alainen tutkijakäyttö tuli mm. sukututkijoiden, historiantutkijoiden ja opettajien vastausten myötä ilmi merkittävänä käyttäjäsegmenttinä. Näitä toimintoja saatiin myös palvelussa luotua lisää.

Digitalian yhteydessä tehdyn ohjelmistokehityksen käytössä ovat jo usemman vuoden ajan olleet ketterät ohjelmistokehitysmenetelmät. Erinomaisen ulkoisen toimittajan, digitointi- ja konservointikeskuksen IT-ryhmän ja projektityöryhmän kanssa kehitystä pyöritettiin noin kahden viikon pituisissa kehitysjaksoissa eli sprinteissä. Kunkin sprintin yhteydessä pohdittiin, mitä toimintoja tehdään seuraavaksi, sekä tehtiin ja testattiin ne.

Uudet toiminnot olivat kuitenkin laajoja ja ulottuivat aina digitointiprosessin jälkivaiheisiin, joten ne aika ajoin vaativat myös tausta-aineistojen esiprosessointia, ennen kuin toimintoja voitiin tuoda näkyväksi käyttöliittymään asti. Digitalia-hankkeessa uusia toimintoja julkaistiin melko paljon, esimerkiksi sivutekstin näyttäminen sivukuvan yhteydessä, kehittynyt hakulomake, kuvitushaku, erilaiset excel-exportit ja merkittävimpänä aineistopaketit. Luonnollisesti myös joitakin taustakomponentteja ja -korjauksia tehtiin. Nämä pitävät toiminnot tehokkaana ja antavat mahdollisuuksia tuoda uusia toimintoja myös käyttäjätasolle.

PIENI SANA SISÄLLÖISTÄ

Kansalliskirjaston digitoitujen aineistojen järjestelmä digi.kansalliskirjasto.fi sisältää tällä hetkellä suomalaisen julkaisutuotannon aina vuodesta 1771 vuoden 1920 loppuun saakka. Avoinna olevat aineistot olivat Digitalia-hankkeen alussa tarjolla vuoteen 1910, mutta Suomen 100-vuotisjuhlien yhteydessä aineistot avattiin helmikuussa 2017 aina vuoteen 1920 asti. Näin käyttäjille saatiin tutkittavaksi 1,9 miljoonaa uutta sivua, jotka oli aiempina vuosina digitoitu.

Digitaaliset aineistot digi.kansalliskirjasto.fi -palvelussa kattavat sanomalehdet, yleiset aikakauslehdet ja teollisuuden pienpainatteita. Tämän lisäksi Kansalliskirjastolla

on mm. eri kirjastojen, museoiden ja arkistojen aineistojen hakuliittymä finna.fi. Kansalliskirjasto toimii näin keskipisteenä kaikelle kansalliselle kulttuuriperintöaineistolle. Kansalliskirjaston muita digitaalisia aineistoja löytyy myös doria.fi -palvelusta, jossa mm. digitoitujen kirjojen, kuten erityiskokoelmien Klassikkokirjaston ja Kirjahistorian materiaalit ovat löydettävissä.

Kuva 1. *Sisältösivu Digi.kansalliskirjasto.fi –palvelusta (Pellervo 18.2.1882)*

DIGITOINNISTA JA JÄLKIKÄSITTELYSTÄ

Kansalliskirjaston digitointia ohjaa digitointipolitiikka, jota vuosittain tarkennetaan digitointisuunnitelmilla. Niiden avulla toteutetaan pidemmän tähtäimen digitointitavoitteita, joita ohjaavat Kansalliskirjaston tavoitteet, tutkijoiden tai muiden käyttäjien toiveet. Tavoitteiden kanssa tasapainotellaan pyrkien sekä suurten tekstikorpusten muodostamiseen (koska niillä nähdään käyttöä tutkimuksessa mm. digitaalisen humanismin parissa), että uniikkien erikoisaineistojen digitointiin. Herkät aineistot, jotka ovat vaarassa vaurioitua, on myös tärkeää digitoida, jotta niiden saatavuus paranee. Verkon välityksellä saadaan avattua niitä laajemmin.

Luonnollisesti digitointi kohdistuu myös aineistoon, jota käytetään paljon. Tällä voidaan perustella sanomalehtien ja aikakauslehtien digitointia. Voimavarat ovat tähän yksi syy – aiemmin mikrofilmattu aineisto on otollista valikoimaa digitoinnille. Toki aika ajoin aineistoa digitoidaan myös suoraan alkuperäisestä paperiasustaan.

Digitointia tehdään myös liiketoimintana, joko Kansalliskirjaston aineistoista tai asiakkaan omista aineistoista. Yhteistyötä on tehty mm. lehtitalojen kanssa, joilla on ollut tarve esimerkiksi omien aineistojensa digitaaliselle versiolle. Lopputuloksen avulla lehtitalot ovat voineetkin luoda omia digitaalisia palvelujaan, kuten esimerkiksi Suomen Kuvalehden digipalvelu (Pulsa, 2016).

Vähitellen kasvavaa ovat myös sähköisesti syntyneet aineistot, e-lehdet, joita voidaan myös ottaa vastaan esimerkiksi pdf-muodossa. Tällöin säästetään osassa työvaiheista, mutta aineisto voidaan kuitenkin ajaa jälkikäsittelyprosessien läpi. Tämä mahdollistaa sen, että kaikki aineisto saadaan muokattua yhteneväiseen muotoon ja pyritään

minimoimaan sitä, ettei eri tavalla luotu aineisto kovin paljon eroa peruslaadultaan. Näin tulevilla tutkijoilla on helpompi aineistomassa tutkittavanaan. He voivat nähdä sanomalehtien ja muiden aineistojen kautta nykypäivän arkielämää ja lähteä tutkimaan 2000-luvun alun ihmeellisyyksiä.

Kuva 2. *Digitointipalvelujen esittely*

Digitoinnin lisäksi tärkeää on myös se, kuinka digitoidut aineistot esitetään. Digi.kansalliskirjasto.fi -palvelulla on, hieman laskentatavasta riippuen, vuositasolla ollut noin 100.000 käyttäjää. Määrä on kasvanut hiljalleen. Suurin osa on aiemmin tehdyissä käyttäjätutkimuksissa tunnistettu sukututkijoiksi, historiantutkijoiksi tai muista aiheesta innostuneesta harrastajiksi, joko liittyen työhön tai kiinnostukseen tietystä alueesta.

Yksi mahdollistaja on ollut digiin kehitetyt joukkoistamistoiminnot, joiden suosio on yllättänyt. Käyttäjien tekemiä leikkeitä on nykyään jo yli 60.000, suurin osa sanomalehdissä. Suurimmat leikekokoelmat ovat tuhansien leikkeiden suuruisia. Yksi yksittäinen leike voi olla kokonainen artikkeli, kuvituskuva tai kiinnostava ilmoitus, joka kuvaa aikaansa omalla tavallaan. Leikkeitä – joko digissä luotuina tai muutoin – tulee vastaan mm. sosiaalisessa mediassa sekä perinteisten uutistoimitusten että erilaisten yhteisöjen toimesta. Digitalia-hankkeessa aiemmin toteutettuja joukkoistamistoimintoja hyödynnettiin jonkin verran naapuriprojekti Aviisin puolella Ristiinan sotakoulun artikkelien löytämisessä. Näiden artikkelien avulla sotakoulun tilannetta analysoiva työryhmä sai taustatietoa menneistä sotakoulun vaiheista ja ehkäpä kerättyä sisältöä tulevia suunnitelmia varten.

ESITYSJÄRJESTELMÄN JA PALVELUIDEN KEHITYS

Aineistojen käytettävyyden kannalta oleellista on löydettävyys ja hakutoiminnot. Digitalia-hankkeessa parannettiin aineistojen hakuominaisuuksia digi.kansalliskirjasto.fi -palvelussa mahdollistamalla haun tarkempi kohdistaminen haluttuun sisältöön. Esimerkiksi haun kohdistamista tiettyyn sanomalehteen helpotettiin luomalla uusi valikko, josta oikean lehden voi valita nimellä tai vaikka ISSN-tunnisteella, mikä auttaa samannimisten lehtien erottamisessa. Lisäksi alueellisia lehtiä voi hakea suoraan valitsemalla ilmestymispaikan lehdestä. Lehtien aiempaa nimekenäkymää muokattiin myös. Se näyttää nykyään myös digitointimäärät eli kuinka paljon tietystä nimekkeestä on digitoitu, ja miltä vuosilta.

Yksi kiintoisa kokeilu jatkoa varten oli myös aineistojen kuvitustietojen tuominen esiin. Digitoinnin jälkikäsittelyssä taustalla tehdään aineistojen rakenteellinen analyysi, jossa tunnistetaan sivuilla olevat elementit. Nyt tätä tietoa hyödynnettiin kuvitustietojen erottelemisessa. ALTO XML-tiedostosta etsittiin GraphicalElement-elementeiksi löydetyt objektit, ja tässä prosessissa hyödynnettiin British Libraryn aiemmin kirja-aineistoihin toteuttamaa vastaavanlaista ratkaisua, joka oli käytettävissä avoimena lähdekoodina.

Saaduista kuvituksista tiedot siirrettiin tietokantaan. Digin käyttöliittymään liitettiin tämän jälkeen mahdollisuus hakea kuvituksellisia sivuja, jolloin käyttäjä voi löytää kiinnostavia kuvia. Tekstihaku kohdistuu kuitenkin sivuun. Tämä useimmiten auttaa myös aiheeseen liittyvien kuvien löytämisessä, vaikka kuvan sisällön analysointia ei vielä tehtykään.

Koko aineiston läpi menevän kerta-ajon lisäksi kaikelle nykyiselle aineistolle toteutettiin myös seurannan kuvitustietojen haku. Se luo yksinkertaisen raportin, josta voi kertasilmäyksellä nähdä, paljonko kuvituksista puuttuu suhteessa tulleeseen uuteen aineistoon. Kuvitushakua tehdään ajastettuna ajona digitoinnin edistyessä ja sitä mukaa, kun kuvitustietoja tulee. Kun tämä tieto on saatavilla myös digin indeksin kautta, haku voi löytää kuvitukselliset sivut.

Haasteena kuvitushaulle on se, että jälkikäsittelyohjelman toimintoihin ei kuulu kuvien luokittelu, koska tavoite on käsitellä aineistot mahdollisimman tehokkaasti. Jälkikäsittelyohjelmiston kannalta mainoksesta löytyvä koriste on samanarvoinen kuin koko sivun piirros. Lehdistä riippuen esimerkiksi perheilmoitukset saattavat löytyä kuvitushaun myötä osin paremmin, koska niissä on yleensä käytetty koristekuvia. Toisaalta kuvitushaussa hyvin usein ensimmäiset sivut korostuvat, mikä johtuu myös sivujen mainoksista. Kuvitusten luokittelusta joko sisällön tai koon mukaan olisi jatkossa iloa tietynaiheisten kuvien löytämisessä, mutta tämä vaatii lisää arviointia.

HAKUTOIMINTOJEN MUU KEHITYS

Hakutoiminnoissa aiempi yhden kentän haku muutettiin perinteisempään kehittyneeseen hakuun. Kaikille eri hakuelementeille on oma kenttä, jota käyttää. Suurelle osalle palautteita lähettäneistä käyttäjistä tämä oli tervetullut muutos, sillä hausta on tämän muutoksen jälkeen tullut huomattavan paljon vähemmän kysymyksiä.

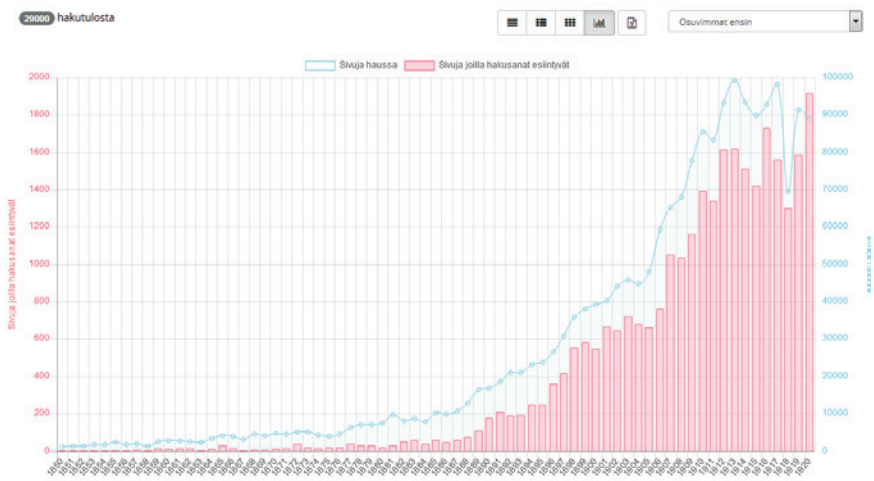
Lisäksi olemme toteuttaneet jonkin verran tilastointia eri hakutoimintojen käytölle.

Uudessa kehittyneemmässä haussa käyttäjät käyttävät eri hakuoptioita jonkin verran paremmin kuin ennen. Kunhan dataa kertyy hiukan enemmän, voimme tehdä tästä tarkemman analyysin. Yhden kentän haussa on toki puolensa, joten jatkossa tuleekin harkita, kuinka järjestelmässä voitaisiin yhdistää yhden kentän moderni tyyli ja kehittyneen haun parhaimmat puolet.

Digitalia-hankkeessa tehtiin myös alustassa uudistustöitä ja taustalla oleva hakumoottori ElasticSearch uusittiin tuoreempaan versioon. Tämä parantaa joitakin taustalla olevia hakutoiminnallisuuksia. Pystyimme myös lisäämään hakutoimintoihin uusia toiminnallisuuksia kuten läheisyyshaun. Läheisyyshaulla sanomalehtien sivutekstistä voi etsiä sanoja tietyltä etäisyydeltä toisistaan. Esimerkiksi hakulausekkeella `"maanviljelijä Virtanen"~16` etsittäisiin näitä kahta sanaa kuudentoista sanan etäisyydeltä – erittäin tehokas haku tietyssä tilanteissa, jos etunimi ei ole tiedossa tai tiedetään hakulausekkeeseen pari yksityiskohtaa, joilla hakea.

OCR-virheiden vaikutusten vähentämiseksi ennen ElasticSearchia hyödynnetään myös tiettyjä peruskorjauksia, joilla haettavista sanoista muodostetaan eri taivutusmuodot ja eri versiot esimerkiksi `v -> w` muunnoksin. Näin hakutuloksia saadaan enemmän eivätkä OCR-virheet estä sanojen löytymistä. ElasticSearch on jatkuvasti kehittyvä ohjelmisto, joten ko. komponenttia kannattaa jatkossakin päivittää sopivin aikavälein.

Tämän lisäksi hakutoimintoihin lisättiin mahdollisuus visualisoida hakutuloksia aikajanalle suhteessa kaikkeen digitoituun aineistoon. Vaikka aineistosta voi löytyä tekstintunnistusvirheitä, joita Kettunen ym. edellisessä artikkelissa käsittelevät, aineiston suuri määrä auttaa visualisoinnissa paikallistamaan, milloin jokin ilmiö tai uusi keksintö ilmestyy joko sanomalehti- tai aikakauslehtien keskusteluun.



Kuva 3. *Mikkeli-sanan esiintyminen aikakauslehdissä 1850–1920*

Tutkijakäyttöön lisäsimme myös mahdollisuuden poimia hakutulokset Exceliin, jossa yllä mainitun muotoisia kaavioita voi luoda omien kiinnostusaiheiden mukaan tai

yhdistämällä eri hakutuloksia. Hakutulosten aikajana auttaa ainestomassojen hahmottamisessa ja auttaa tutkimuskysymyksen kohdistamisessa kiinnostaviin ajanhetkiin.

Koneellista aineistojen käsittelyä varten digiin lisättiin toiminto, jolla yksittäisen digitoidun sivun osalta saa näkyviin sekä tekstimuotoisen että digitoinnin jälkikäsittelystä saadun rakenteellisen analyysin sisältävän ALTO XML-tiedoston. Tekstimuoto on hyödyllinen, jos tekstistä haluaa ottaa lainauksen blogiin tai muuhun artikkeliin. Se soveltuu myös nopeisiin kokeiluihin, jos joko tekstiä tai sivun rakennetta haluaa käsitellä tietokoneella. Koska yksittäisten sivujen lataaminen on hidasta, kun tarvitaan suuria korpuksia, on myös kehitetty avoimen datan aineistopaketit, joissa koko aineisto on vuoteen 1910 asti saatavilla helposti ladattavana zip-pakettina.

AVOIN DATA JA DIGI.KANSALLISKIRJASTO.FI

Digitalia-hankkeen yhteydessä myös avoimen datan sivustoa alettiin valmistelemaan, jotta voisimme tarjota käyttäjille aineistopaketteja laajemmista kokonaisuuksista. Avoimeksi dataksi määritellään digitaalisesti tallennettu, koneellisesti luettavissa oleva informaatio, joka on julkaistu uudelleenkäytön sallivalla lisenssillä maksutta ja koneluettavassa muodossa (Sillanpää ym., 2016).

Digissä olevasta sanomalehti- ja aikakauslehtiaineistoista määriteltiin oma XML-tiedostomuotonsa, joka sisältää aineiston kolmella eri tavalla. Kansalliskirjaston oma XML-muoto sisältää aineiston metatiedot, sivukohtaisen rakenteisen analyysin ALTO XML-muodossa ja itse sivutekstin tekstimuodossa (Pääkkönen ym., 2016). Aineistopaketit on luotu eri vuosiväleiltä, pyrkien löytämään tasapaino paketin koon ja sen sisältämän aineistomäärän välillä. Niinpä vuoteen 1910 asti aineistopaketteja on nyt 12 sanomalehtipuolella ja yksi aikakauslehdissä. Mahdollisesti aineistoja tarjotaan jatkossa pidemmällekin, kunhan saamme palautetta nykyisten jakelupakettien toimivuudesta.

Lisäksi teimme kolme pienempää aineistopakettia, joista yksi sisältää Suomenkieliset Tieto-Sanomat. Se on ensimmäisenä suomenkielisenä lehtenä kiinnostava tapaus kompaktin sivumääränsä perusteella, mutta toisaalta näyttää myös OCR-tunnistuksen haasteet.

Aineistopaketit sisältävät siis yhden XML-tiedoston aina yhtä sivutiedostoa kohti. Sivu on käyttämämme aineiston pienin yksikkö, joka toimii haussa taustalla ja on pääloppututulos, joka käyttäjälle hakutuloksena näytetään. Digin sivulta pystyy myös katsomaan yksittäisen sivun tekstit tai ALTO XML-muodot, mutta aineistopaketissa voi ottaa suuremman aineistomassan helpommin käsittelyyn. Toiveissa olisi, että ainakin työkalu- ja menetelmäkehityksessä aineistosta olisi iloa, ehkä mahdollisesti OCR:n korjauksessa tai analysoimisessa, mitä Digitaliassa on jo tehtykin.

Toisaalta olisi kiinnostavaa nähdä, olisiko tekstiaineistoista mahdollista löytää uudenlaisia sovelluksia, kuten aikoinaan Digitointi- ja konservointikeskuksen ja Microtask-yrityksen kehittämässä Myyräpelissä, jossa käyttäjä näki pelin, mutta taustalla tehtiin OCR-korjausta. Aineistopakettien avulla on mahdollista tehdä kattavampia työkaluja, kun aineisto on nähtävissä ja kaikkien kiinnostuneiden pohdittavissa. Mahdollisesti jälkikäsittelyyn voidaan lisätä uusia vaiheita, joilla voidaan automaattisesti rikastaa aineistoa lisää. Uudet aineiston käyttötavat mahdollistavat myös uusien käyt-

täjätoimintojen luomisen, jolloin aineistoja pystyy käymään läpi monipuolisemmin.

Varsinainen aineistopakettien lataaminen on mahdollista digi.kansalliskirjasto.fi -palvelun opendata -sivun avulla: http://digi.kansalliskirjasto.fi/opendata. Aineistot saa nopeasti ladattua lyhyen kyselylomakkeen täyttämisen jälkeen. Aineiston XML-muoto voi vaatia erityisprosessointia. Odotamme kiinnostuneina, millaisia hyödyntämistapoja aineistolle löytyy eri yhteyksissä, sillä aineistopaketit ovat jo COMHIS-projektin käytössä. Mahdollisesti jatkossa aineistopaketteja voidaan kehittää eteenpäin lisäten uusia metadatatietoja, tai tarjota uudenlaisia aineistopaketteja. Kyselylomakkeen tarkoitus on auttaa meitä saamaan kuvaa siitä, millaisia käyttötarkoituksia aineistoille löytyykään. Käyttäjät voivat lomakkeen avulla ilmoittaa muistakin käyttötarkoituksistaan. Näitä kuvauksia voi myös lähettää digin palautelomakkeella, kuten ennenkin.

YHTEENVETO

Digitalia-hankkeessa digi.kansalliskirjasto.fi -palvelua kehitettiin eteenpäin hankkeen alkuperäisten suunnitelmien mukaan. Hyvää oli yhteistyö tutkijoiden kanssa, joilta saimme suoraa palautetta toivotuista toiminnoista. Näitä pystyttiin palveluun myös toteuttamaan. Tutkijoiden toiveet olivat samantyyppisiä kuin muillakin käyttäjillä – vaikka toiminto oli ensisijaisesti suunnattu tutkijoille, yhtä lailla palautetta on tullut myös muilta käyttäjiltä.

Voidaan sanoa, että palvelu parani huomattavasti ja loimme myös avauksia, joista voi pitkällä tähtäimellä tulla merkittäviä uusia lisiä palveluun. Toimintoja tuli lisää ja osa toiminnoista odottaa vielä löytämistään ja toteuttamistaan. Kun olemme muutamassa tapahtumassa palvelun uusia toimintoja esitelleet, esimerkiksi sivutekstin näyttäminen on löytänyt jo omat ystävänsä. Osa ideoista vaatii vielä jatkotutkimusta – niitä voidaan toteuttaa myöhemmin, kun idea on jalostunut selkeiksi vaatimuksiksi. Moninaisten käyttäjäryhmien palvelu erilaisilla toiminnoilla näyttää olevan hyvä ratkaisu. Näin kukin saa toimintoja, joilla heidän tärkeimmät toiveensa tulevat toteutettua.

Digi.kansalliskirjasto.fi -palvelu on kehittynyt hankkeen aikana sekä taustalla että käyttäjille asti ulottuvissa toiminnoissa. Hakuun on lisätty monipuolisempia hakutoimintoja, kuvituksia voi hakea ja hakutuloksia voi ottaa talteen Excel-muodossa.

Toteutetut toiminnot on nyt integroitu osaksi digi.kansalliskirjasto.fi -palvelun toimintopalettia ja niitä voidaan jatkossa kehittää edelleen käyttäjäpalaute huomioiden. Hakutoiminnot voivat tarvita lisäkehitystä mm. eri päätelaitteiden huomioinnissa. Joukkoistustoimintoja, kuten esimerkiksi suosittua leikepalvelua, voisi myös kehittää eteenpäin. Kaiken ytimessä kuitenkin on digitointi. Sisällöt ovat se, mihin käyttäjien kiinnostus kohdistuu. Tämän huomasimme 1910–1920 vuosivälin avaamisen yhteydessä. Suuret aineistomassat vaativat uusia keinoja päästä aineistojen ytimeen. Tätä on Digitaliassa tutkittu ja jatkossa tulokset ovat tuotavissa myös käyttäjien hyödynnettäväksi.

LÄHTEET

Hölttä, T. 2016. Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat. WWW- dokumentti. Saatavissa: http://urn.fi/URN:NBN:fi:uta-201603171337 [viitattu 27.3.2017].

Matres, I. 2016. Aviisi - Evaluation Report. Noudettu osoitteesta http://www.doria.fi/handle/10024/129898

Pulsa, T. 2016. Ainutlaatuinen palvelu nyt auki: 100-vuotiaan Suomen Kuvalehden kaikki numerot digitilaajien käytössä. WWW-dokumentti. Saatavissa: https://suomenkuvalehti.fi/jutut/kotimaa/ainutlaatuinen-palvelu-nyt-auki-100-vuotiaan-suomen-kuvalehden-kaikki-numerot-digitilaajien-kaytossa/ [viitattu 27.4.2017].

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

Sillanpää, L., Toikkanen, T., & Kanniala, E. 2016. P2PU | Datan avaaminen: Johdanto | 1 Mitä on avoin data? WWW-dokumentti. Saatavissa: https://courses.p2pu.org/en/courses/2486/content/5069/ [viitattu 4.9.2016].

PILKOTAAN PDFIÄ, MUTTA MIKSI?

Anssi Jääskeläinen, *TkT, TKI-asiantuntija, Kaakkois-Suomen ammattikorkeakoulu*

Kukapa ei olisi joskus ladannut PDF-tiedostoa netistä tai tulostanut tiedostoa suoraan PDF-muotoon. Miksi näin tehdään? Mitä etuja PDF tuo ohjelmien natiiveihin formaatteihin nähden?

Nämä ovat kysymyksiä, joihin harvempi osaa vastata ainakaan kunnolla. On tosiasia, että muokattavuuden kannalta PDF-tiedostot ovat vihoviimeinen vaihtoehto. Miksi kukaan siis haluaisi käyttää niitä?

Kysymystä edeltävä lause pitää sisällään vastauksen kysymykseen. Jos halutaan, että tiedosto pysyy muuttumattomana ja alustariippumattomana, on PDF ja etenkin sen arkistokelpoinen muoto PDF/A ainoa oikea vaihtoehto (PDF/A Competence Center, 2013). Kansalaisen kannalta siirrettävyys ja alustariippumattomuus lienee pääsyitä käyttää PDF-formaattia. Vastaani ei ole tullut käyttöjärjestelmää, jolle ei olisi saatavilla PDF-lukijaa. Sellainen löytyy jopa vanhoille Pocket PC -laitteille, joita nuorempi lukijasukupolvi tuskin edes tunnistaa.

Ehkäpä juuri edellä mainituista syistä jokainen itseään kunnioittava kansallisarkisto on määritellyt PDF/A-formaatin ja sen eri versiot yhdeksi suositelluista tavoista tallentaa tekstimuotoiset dokumentit pitkäaikaissäilytykseen. Näihin lukeutuu myös Suomen Kansallisarkisto, joka pohjaa KDK-PAS -määrityksiin pitkäaikaissäilytykseen soveltuvista formaateista (KDK, 2017). Muita arkistolaitosten arkistoihin hyväksymiä tekstimuotoista tietoa sisältäviä formaatteja ovat csv, xml, html, odt ja txt.

Näistä tiedostomuodoista vain html ja odt soveltuvat vastaavaan visuaaliseen tallentamiseen kuin PDF. Odt-formaatissa on se ongelma, että normaalit käyttäjät eivät tiedä mikä se on (näille lukijoille tiedoksi: odt on avoimien toimisto-ohjelmien käyttämä muoto tallentaa vastaavia dokumentteja, jotka MS Office tallentaa joko doc- tai docx-muotoon).

Html-esitysmuoto toimisi hyvin staattisten sivujen kanssa, mutta dynaamiset muualta sisältöä lataavat sivut ovat haasteellisia. Ongelman voisi kiertää lataamalla kaiken dynaamisen sisällön paikalliskopioksi samalla hetkellä kun html-arkistokappale luodaan, mutta kuinka syvälle linkkejä pitäisi seurata? Suurin osa Internetin sivuista on tällä hetkellä dynaamisia. Paras ratkaisu allekirjoittaneen mielestä onkin tallentaa tärkeät dokumentit tai selaimen renderöimä verkkosivu PDF- tai PDF/A-muotoon. Näin voidaan taata ainakin se, että sivusto tai dokumentti on tallennettu siinä visuaalisessa muodossa, missä käyttäjä sen tallennushetkellä näki.

Tämä artikkeli käsittelee sitä, kuinka näiden tallennettujen PDF-tiedostojen saatavuutta ja käytettävyyttä on pyritty parantamaan pilkkomalla tiedostoja pienemmiksi kokonaisuuksiksi. Tästä herää luonnollisesti kysymys, miksi niitä ei alun perin tallennettu riittävän pieniksi? Jälkiviisaus on helppoa ja syitä on monia. Niitä ovat huomioimatta jättäminen kilpailutuksessa, kustannukset, digitoiminen nidotusta kirjasta jne.

Toinen mieleen tuleva kysymys olisi se, onko kaikki pakko automatisoida? Nykyisen MikroBitin (entisen MikroPC -lehden) testaajia lainatakseni: jos jokin asia täytyy tehdä useammin kuin kerran, se pitää automatisoida. Tässä Helsingin kaupunginarkiston case-tapauksessa näitä kertoja olisi ollut 308 ja läpi selattavia sivuja reilusti yli sata tuhatta. Näiden tiedostojen käsittelyssä ainoa järkevä ratkaisu oli automatisointi.

Mainittavan arvoinen seikka on myös se, että automatisointi tehtiin yleishyödylliseksi. Mikä tahansa PDF-tiedosto, jossa on kirjanmerkit olemassa, pystytään katkaisemaan osiin tällä kehittämällämme menetelmällä.

PDF:N MONET KASVOT

Tavalliselle kansalaiselle PDF-tiedostot ovat PDF-tiedostoja riippumatta siitä, onko kyseessä A-, X-, UA- tai E-variantti. Arkistonäkökulmasta ja yleisestikin säilyttämisen näkökulmasta taas on hyvinkin tärkeää, missä variantissa PDF on tallennettu. Formaatti on ollut olemassa vuodesta 1993 asti. Kehitys alkoi jo vuonna 1991. Kypsyysasteen voidaan todeta olevan kohtalaisen ”mature” myös tietoteknisestä näkökulmasta.

- PDF/A: Pitkäaikaissäilytykseen kehitetty tallennusmuoto. Julkaistiin alun perin 1.10.2005 ISO standardina 19005-1:2005. Formaatti takaa, että tallennettu dokumentti voidaan näyttää vuosienkin päästä täsmälleen samalla tavalla kuin se tallennettiin. Kaikki käytetyt fontit, kuvat, piirrokset yms. sisällytetään tallennusvaiheessa dokumentin sisään. Lisäksi monia asioita, kuten äänet, videot, javascript, salaus, läpinäkyvyys, ulkoinen sisältö yms. on formaatissa standardin mukaan kielletty. Vuonna 2011 julkaistu PDF/A-2 toi edellä mainittuihin hieman helpotuksia ja antoi lisäksi mahdollisuuden liitetiedostojen käyttämiseen. Tuki tosin rajoittui PDF/A- tai PDF/A-2 -muotoisiin tiedostoihin. Vuonna 2012 standardia laajennettiin jälleen PDF/A-3 -muodolla, ISO 19005-3, joka sallii liitetiedostoiksi minkälaiset tiedostot tahansa.
- PDF/X: Normaalin PDF:n laajennus, jonka tarkoituksena on helpottaa graafisten objektien tallentamista. Tälläkin formaatilla on monia alalajeja, joiden vaatimukset poikkeavat toisistaan. Formaatti vaatii, että värit on tallennettu tietyssä muodossa ja liitettynä on myös ICC-väriprofiili.
- PDF/UA: UA tulee sanoista universal accessibility. Tämän mukaisesti se pyrkii takaamaan, että PDF formaatti ei olisi esteenä tiedoston käyttämiselle esim. ruudunlukijoiden tai muiden avustavien teknologioiden avulla.
- PDF/E: On tarkoitettu pääasiassa 3D-mallien ja CAD-kuvien tallennusformaatiksi. Se asettaa monia rajoitteita normaaliin PDF tiedostoon nähden, jotta siirrettävyys ja näytettävyys eri ohjelmien välillä voitaisiin taata.

Näiden lisäksi on olemassa muitakin variantteja. Jokaisesta määrittelystä on myös monia eri versioita. Siksi onkin tärkeää, että tallennusvaiheessa tiedettäisiin mikä tulevan PDF-tiedoston käyttötarkoitus on. Formaattien konversiot ovat toki mahdollisia ja niitä tehtiin tässäkin yhteydessä, esim. ScandAll PRO 1.8.1 -ohjelmalla 2014 luodut PDF-muotoiset tiedostot konvertoitiin pienemmiksi PDF-tiedostoiksi

GhostScript -ohjelmalla. Silti esimerkiksi alkuperäisen luontiympäristön ICC-väriprofiilia tai käytettyä fonttia on kovin haastavaa lisätä kymmenen vuotta dokumentin luomisen jälkeen, jos ympäristöstä ei ole muuta tietoa kuin luontihetki ja mahdollisesti ohjelma, jolla tiedosto on luotu.

CASE: HELSINGIN KAUPUNGINARKISTO

Esimerkki selventänee asiaa paremmin kuin teoreettiset selitykset. Yhteistyökumppanimme tässä tutkimuksessa oli Helsingin kaupunginarkisto, josta lähtökohdat kehitystyölle tulivat. Helkan arkistoissa on mm. 1085-sivuinen PDF-tiedosto, jossa ovat kaupunginhallituksen ehdotukset vuodelta 2003. Tämän tiedoston fyysinen koko on 466,7 Mt. Toisena esimerkkinä käytetään samaisessa arkistossa olevaa PDF-tiedostoa, joka pitää sisällään kaupunginhallituksen mietinnöt vuodelta 1969. Kuva 1 esittää tämän asiakirjan kannen. Asiakirjassa on sivuja 1228 ja sen fyysinen koko 235,3 Mt.


HELSINGIN
KAUPUNGINVALTUUSTON
ASIAKIRJAT
KAUPUNGINHALLITUKSEN MIETINNÖT
1969

Kuva 1. *Helka: Helsingin kaupunginhallituksen mietinnöt vuodelta 1969*

Alla oleva taulukko 1 esittää esimerkkeinä olevien tiedostojen latausajat kotikäytössä yleisesti olevalla mobiiliyhteydellä, jonka keskimääräiseksi siirtonopeudeksi arvioin kokemusteni perusteella noin 10 Mbit/s. Nopeampia ja hitaampiakin nettiyhteyksiä on toki käytössä ja silloin latausajat joko pienenevät tai kasvavat. Vertailun vuoksi taulukkoon on otettu mukaan Helkan casen keskimääräinen alkuperäinen tiedostokoko sekä keskimääräinen pilkottu tiedostokoko.

Taulukko 1. *Tiedostojen latausaikoja alkuperäisinä ja pilkottuina*

Koko	Latausaika 10Mbit/s	Pilkottujen osien ka. koko	Pilkotun osan latausaika 10Mbit/s	Latausajan pudotus%
466,7 Mt	6min 31s	38,9 Mt	32s	~92%
235,3 Mt	3min 17s	9,6 Mt	8s	~96%
45 Mt	37s	5 Mt	4s	~86%

Vertailun vuoksi otan tässä erikseen esiin vielä Viestintäviraston kaikille ”takaaman” kahden megan laajakaistan, jonka mukaan keskimääräisen latausnopeuden on oltava vähintään 1,5 Mbit/s 24h mittausjakson aikana ja vähintään 1Mbit/s minkä tahansa 4h mittausjakson aikana (Viestintävirasto). Jos tällä 1 Mbit/s sekuntiarvolla lasketaan 466,7 Mt tiedoston latausaika, päästään lukemaan 1h 3min 48s. Pilkotulla tiedostolla lataus kestää ”vain” 5min 19s, joka sekin on kaukana hyväksyttävästä, mutta kuitenkin paljon parempi kuin yli tunnin odottaminen.

Kun taulukossa ja tekstissä esitettyjä lukemia tarkastelee mistä näkökulmasta tahansa, edut tiedostojen koon pienentämiseen ovat selvät. Tiedostojen, tai ladattavan sivuston koko on yleisestikin ottaen tärkeä asia SEO:n (Search Engine Optimization) kannalta. Onnistunut SEO edesauttaa tuloksien näkymistä hakukoneiden hakutuloksissa ja siten tuo tiedon paremmin yleisön tietoisuuteen.

Toinen huomionarvoinen seikka on käytettävyys. Siinä missä 10 Mt kokoinen PDF-tiedosto aukeaa ja käyttäytyy vielä kivuttomasti koneella kuin koneella, 500 Mt:n PDF-tiedoston rivakkaan lukemiseen ja käsittelyyn vaaditaan koneeltakin jo kohtalaisen paljon. Oma lukunsa, jota emme tässä edes käsittele, ovat ne tapaukset, joissa selain päättää avata ladatun PDF-tiedoston suoraan omaan näkymäänsä.

Kolmas huomioitava seikka on käyttäjäkokemus, jonka merkitystä vielä nykypäivänäkin usein aliarvioidaan. Vaikka edellä mainitut tekijät ovatkin puhtaasti teknisiä, niillä on suuri merkitys käyttäjäkokemukseen. Nykykäyttäjät ovat tottuneet siihen, että asiat tapahtuvat netissä heti. Sama pätee myös ohjelmien käyttämiseen (Lowdermilk, 2013). Jos näkymän siirtymistä sivulta 1 esim. sivulle 562 joutuu odottelemaan useita sekunteja, käyttäjä tuskastuu. Sama pätee myös latauksiin, jotka helposti keskeytetäänm koska turhaudutaan hitauteen.

TEKNINEN NÄKÖKULMA CASE-TOTEUTUKSEEN

Käsitellyistä kaupunginvaltuuston asiakirjoista luettiin Digitalian kehittämällä Python-ohjelmalla ensin metatiedot hyödyntäen avoimen lähdekoodin pdftk-ohjelmaa. Metatietorakenteessa mahdolliset kirjanmerkit näyttäytyvät ”BookmarkTitle” ja ”BookmarkLevel” -tägeinä, joita seuraa ”BookmarkPageNumber” -tägi.

Näiden metatietojen avulla pdf-asiakirjoille luotiin katkaisupisteet. Tässä tehtiin oletus, että aiempi dokumentti päättyy edellisellä sivulla. Tässä tapauksessa oletus toimi, mutta kohtasimme myös nidotun kirjan aukeamilta skannattuja dokumentteja, joissa vasen sivu saattaa olla edellisen dokumentin loppu ja oikea sivu seuraavan dokumentin alku. Ohjelmallinen luku näkee sivut kuitenkin yhtenä ja näin ollen toimintaa olisi muutettava.

Kun katkaisulista on saatu luotua kokonaan, se syötetään moniprosessointiin kykenevälle työnkululle Python-ohjelmassa, josta katkaisupisteet ja muut tarvittavat parametrit syötetään eteenpäin GhostScript-ohjelmalle. GhostScript pilkkoo pdf-tiedostot pienempiin kokonaisuuksiin ja nimeää ne ennalta määrätyllä tavalla odottamaan takaisin arkistoon vientiä.

Lisäominaisuutena rakennettiin toiminto, joka lukee jokaisen luodun pdf-tiedoston sanasta sanaan läpi ja etsii sieltä suomalaisia erisnimiä. Mahdollisesti löydetyt

erisnimet kirjoitetaan .csv-tiedostoon tiedostonimen ja sivunumeron kanssa. Tällä mahdollistetaan sensitiivisten tietojen helpompi tunnistaminen ja häivyttäminen lopullisista dokumenteista.

JATKOKEHITYS JA TULEVAISUUDEN NÄKYMÄT

Kehitystyö keskittyi yhteistyökumppanimme toiveesta vain niihin aineistoihin, joissa kirjanmerkit olivat jo valmiina. Aloitimme samalla kuitenkin varautumisen sellaisiin tilanteisiin, joissa näitä kirjanmerkkejä ei ole valmiina. Käyttämällä OCR-tietoa ja avainsanojen tunnistamista pääsimme noin 80% tarkkuuteen kaupunginhallituksen pöytäkirjojen pilkkomisessa. Yksinkertaistettuna pöytäkirjan alkusivu perustui sanojen ”puhetta, kaupunginjohtaja, kokouspaikka, läsnä, kokoontui, johti, jäsenet” yhdistelmään. Se, miksi tunnistusprosentti jäi noinkin alhaiseksi, johtui kahdesta seikasta: OCR:n ajoittainen heikko laatu ja vuosien saatossa muuttuneet tavat aloittaa kokouksen pöytäkirja. Tämän dilemman parissa jatkamme työskentelyä Digitaliassa.

Sensitiivisten tietojen tunnistaminen aineistosta on tässä työkokonaisuudessa viety alkumetreille. Seuraavassa vaiheessa pyrimme siihen, että mahdollisesti sensitiivisiä tietoja pystytään tunnistamaan lisää ja sen jälkeen häivyttämään nämä tiedot automaattisesti.

LÄHTEET

KDK. Säilytys- ja siirtokelpoiset tiedostomuodot v1.5.1. 2017. WWW-dokumentti. Saatavissa http://www.kdk.fi/images/tiedostot/KDK-PAS-tiedostomuodot-v1.5.1.pdf [viitattu 29.4.2017]

Lowdermilk, T. 2013. User-Centered Design. A Developer's Guide to Building User-Friedly Applications. O'Reilly Media.

PDF/A Competence Center. 2013. PDF/A in a Nutshell 2.0. WWW-dokumentti. Saatavissa https://www.pdfa.org/wp-content/until2016_uploads/2013/05/PDFA_in_a_Nutshell_211.pdf [viitattu 29.4.2017]

Viestintävirasto. Jokaisella on oikeus kahden megan laajakaistaan. WWW-dokumentti. Saatavissa https://www.viestintavirasto.fi/internetpuhelin/oikeuspuhelin-jalaajakaistaliittymaan/oikeusmeganlaajakaistaan.html [viitattu 29.4.2017]

KANSALAISARKISTO – SUKUYHTEISÖN AARTEET TALTEEN DIGITAALISEEN ARKISTOON

Eero Kausalainen, *eläkkeellä oleva historian harrastaja*
Liisa Uosukainen, *DI, IT-asiantuntija, Kaakkois-Suomen ammattikorkeakoulu*

Muistitikut, ulkoiset kovalevyt ja pilvipalvelut ovat nykyään yleisiä tapoja säilyttää henkilökohtaista digitaalista aineistoa. Erilaiset tallennusvälineet ja -järjestelmät kehittyvät jatkuvasti, eivätkä ne automaattisesti takaa aineiston säilyvyyttä ja käytettävyyttä pitkällä aikavälillä. Kansalaisarkiston kehittäminen on lähtenyt ajatuksesta, että myös tavallisilla kansalaisilla on tarve luotettavalle digitaaliselle säilytyspalvelulle, jossa tärkeät tiedostot säilyvät käytettävässä muodossa yli vuosikymmenten (Jääskeläinen ym., 2017).

AVOIMEN LÄHDEKOODIN ARKISTOSTA KANSALAISARKISTOKSI

Kaakkois-Suomen ammattikorkeakoulussa on viime vuosina kehitetty sovellusta digitaalisen aineiston säilytykseen ja hallintaan. Sovelluksen toteutus aloitettiin vuonna 2013 OSA (Avoimen lähdekoodin arkisto) -hankkeessa, jossa hankekumppaneina oli arkistoalan toimijoita. Hankkeen lähtökohtana oli vastata lyhyellä aikasyklillä toimiviin kaupallisiin ratkaisuihin avoimeen lähdekoodiin perustuvalla ratkaisulla, jonka kehittäminen yhteisesti ja yhteisöllisesti takaa pidempiaikaisen jatkuvuuden eikä vaadi kalliita lisenssi- ja ylläpitomaksuja.

Hankkeessa kehitettiin ensimmäinen versio digitaalisen aineiston pitkäaikaiseen säilyttämiseen soveltuvasta arkistojärjestelmästä, joka voi toimia erilaisten muistiorganisaatioiden palvelu- ja jakelualustana (Uosukainen, 2014). Sitä seuranneissa hankkeissa ratkaisua on pilotoitu mm. järjestöorganisaatioiden kanssa tutkien ratkaisun soveltuvuutta järjestöjen sähköiseen arkistointiin ja mahdollisesti myös arkiston operatiiviseen käyttöön.

Digitalia-hankkeessa on keskitytty kansalaisten henkilökohtaisen aineiston tallentamiseen, järjestämiseen ja hallintaan digitaalisessa arkistossa. Hankkeessa on kehitetty OSA-arkistosovelluksesta seuraava versio, josta käytetään nimeä Kansalaisarkisto. Olemme tutkineet, mitä alun perin muistiorganisaatioiden käyttöön kehitetty sähköinen arkistopalvelu voi tarjota yksityisille henkilöille.

Myös yksityisille henkilöille muodostuu monenmuotoisten asiakirjojen ja muun tiedon kokonaisuus, joka on kertynyt henkilön toiminnasta ja erilaisten tehtävien hoitamisesta (Loponen & Kosonen, 2016). Digitaalisessa muodossa oleva aineisto kasautuu helposti joukoksi epäjohdonmukaisesti nimettyjä tiedostoja, joita ei ole juurikaan järjestelty. Näkemyksemme mukaan henkilökohtaisen aineiston tallentaminen keskitetysti yhteen arkistoon, jossa aineisto on ryhmitelty ja kuvailtu riittävin metatiedoin, tarjoaa luotettavan tavan hallita omaa aineistoa.

DIGITAALISEN ARKISTON ETUJA

Kansalaisarkiston arkkitehtuuri on suunniteltu siten, että yksi arkistosovellus voi palvella monia, eri tavalla konfiguroituja arkistoja samanaikaisesti. Jokaiselle arkistolle voidaan määritellä halutut aineistotyypit ja niiden metatietokentät. Arkiston omistaja määrittelee itse arkistohierarkian; miten tehtävät ja niiden sarjat kuuluvat päätehtäviin.

Arkistosovellus noudattaa ns. Capture-tietomallia, jonka perusideana on kytkeä aineisto arkistossa erikseen määriteltävien toimijoiden (esim. henkilöiden), paikkojen ja tapahtumien avulla (Lampela, 2016). Näiden ns. kontekstuaalisten tieto-objektien avulla arkiston aineistot linkittyvät toisiinsa, mikä tekee aineiston löydettävyydestä helpompaa ja tarkempaa.

Kun aineisto on riittävällä tasolla kuvailtu ja metatiedot lisätty yhdenmukaisesti, arkistossa olevaa tietoa on mahdollista havainnollistaa monella tapaa, karttapohjat ja aikajanat mukaan lukien. Saman aineiston erilaiset esitystavat palvelevat etenkin niitä, jotka käyttävät ja etsivät tietoa arkistopalvelusta.

Arkistosovelluksen pilotointien yhteydessä on ilmennyt tarvetta myös jo arkistoidun aineiston rutiininomaiselle käsittelylle, kuten sarjojen massamuokkaukset, formaattikonversiot ja aineiston lataustoiminnot. Sovelluksen osana toimivan työnkulkumoottorin avulla on mahdollista toteuttaa ketjumaisten ja luonteeltaan toistettavien tehtävien suorittaminen arkistossa ns. mikropalveluina. Tarvittaessa uusia arkistokohtaisia mikropalveluita on helppo kehittää, sillä ne on suunniteltu toimimaan itsenäisinä pieninä kokonaisuuksina.

Digitaalisella aineistolla voi olla arvoa paitsi arkiston omistajalle itselleen myös esim. perheelle tai lähisukulaisille. Digitaalisen materiaalin voi halutessaan jakaa verkkoselaimen kautta käytettävästä arkistosta eri kohderyhmille. Tämä on tehokas tapa esim. kollektiivisen muistitiedon keräämiseen. Laajojen asiakirja- tai valokuvakokoelmien metatietoja voidaan koko käyttäjäryhmän avustuksella täydentää. Arkiston aineisto on kuitenkin aina arkiston omistajan kontrollissa ja omistaja päättää, missä määrin hän haluaa jakaa arkiston sisältöä. Jokaiselle käyttäjälle voidaan määritellä tiedosto- ja kansiokohtaiset käyttöoikeudet.

Syntysähköisen aineiston arkistoinnissa noudatetaan helposti periaatetta säilyttää kaikki ilman tarkempaa seulomista. Aikaa myöten henkilökohtaisen aineiston määrä kasvaa ja sitä on hankala enää hallita. Osa kertyneestä aineistosta on julkista, osa salaista, osa erityisen tärkeää (pysyvästi säilytettävää) ja osa melko merkityksetöntä. Kaikkea aineistoa ei ole kuitenkaan tarkoituksenmukaista arkistoida pysyvästi. Säilytysaikojen määrittäminen metatietona auttaa hallitsemaan laajaa aineistoa.

Kansalaisarkisto siirtää automaattisesti poistolistalle sellaiset asiakirjat, joiden säilytysaika on umpeutunut. Tämä ominaisuus auttaa myös yksityisarkiston omistajaa pitämään laajaa aineistoaan ajan tasalla. Esimerkiksi vakuutusasiakirjat vanhenevat voimassaoloajan jälkeen. Tarvittaessa poistolistalta voi palauttaa aineiston takaisin arkistoon. Kansalaisarkisto pyrkiikin tarjoamaan palvelun, jossa alati kasvavaa digitaalista omaisuutta voi hallita niin, että kaikkea ei tarvitse arkistoida pysyvästi, mutta arvokas aineisto säilyy luotettavassa tallennusympäristössä.

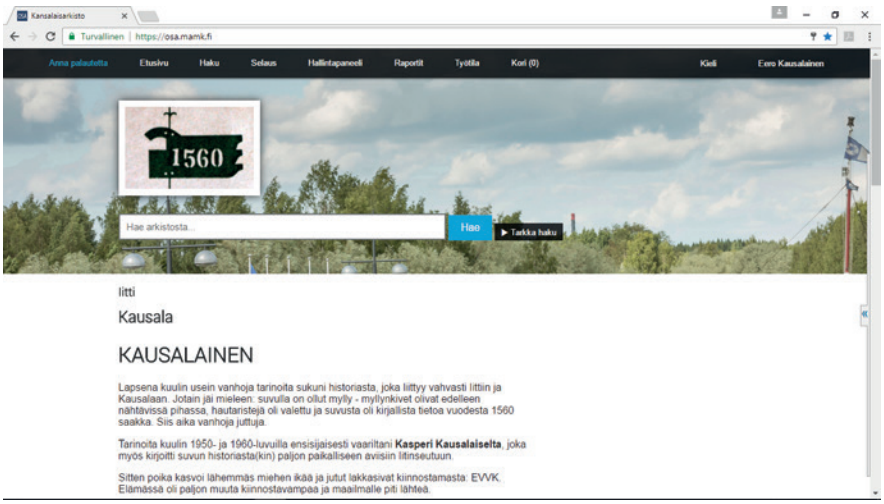
Kansalaisarkiston käyttöliittymän kehityksessä on kiinnitetty huomiota helppokäyttöisyyteen ja käyttäjäystävällisyyteen. Erityisesti arkiston käyttöön liittyvää oh-

jeistusta on parannettu ja arkistotermejä selitetty tooltip-tekstein. Myös aineiston selausominaisuuksia on kehitetty aineiston käyttäjien näkökulmasta.

On kuitenkin huomattava, että aineiston kerääminen, järjestäminen ja metatietojen lisääminen on sähköisestä alustasta huolimatta aikaa vievää, jos sen haluaa tehdä perusteellisesti. Tuorein materiaali voi olla syntynyt sähköiseen muotoon, mutta esimerkiksi syntysähköisiä, tiedostoina käsiteltäviä valokuvia on olemassa vasta parin viime vuosikymmenen ajalta. Jo laajan aineiston digitoiminen voi olla oma prosessinsa.

Kansalaisarkiston kehitystyötä on vauhdittanut pilottiryhmä, joka on päässyt käyttämään ja antamaan palautetta palvelusta oman henkilökohtaisen arkiston käyttökokemuksen kautta. Seuraavassa kuvataankin Kansalaisarkiston pilotointiin osallistuneen Eero Kausalaisen kokemuksia.

PILOTOIJA - TAUSTA JA ARKISTOINTITARPEET



Kuva 1. *Pilottiarkiston pääsivu Kansalaisarkistossa*

Pilotoijaksi valikoitunut Kausalainen on taustaltaan eläkkeellä oleva valtion virkamies, joka itse arvioi IT-taitonsa ”keskinkertaisiksi valtion virkamiehen taidoiksi”. Hänellä oli jonkin verran aiempaa kokemusta digitoidusta arkistosta hänen vetämästään laskuvarjourheilun historiaan liittyvästä projektista. Tämän lisäksi hän oli jo ennen Kansalaisarkisto-projektiin liittymistään koonnut sukunsa kirjallista ja suullista aineistoa, jota hän oli digitoinut ja järjestänyt systemaattiseksi kokonaisuudeksi. Aineisto oli alkuaan tarkoitus digitoinnin ja organisoinnin jälkeen luovuttaa Kausalaisen veljille muistitikuilla.

Kansalaisarkiston pilottiprojektiin liittyminen avasi kuitenkin uusia mahdollisuuksia, jolloin alkuperäinen tavoite muuttui näitä vastaavaksi. Kansalaisarkisto mahdollisti valokuvien ja asiakirjojen lisäksi myös liikkuvan kuvan tallennusmahdollisuuden sekä aineiston jakamisen samanaikaisesti useille käyttäjille. Arkistosovellus tuntui sopivan erinomaisesti pilottihenkilön tarpeisiin. Hänen arkistotarpeensa oli ensisijaisesti taltioida oman suvun pääosin paperilla tai valokuvina oleva aineisto sekä tuottaa

jonkin verran arkistoa täydentävää ja selittävää kirjallista materiaalia – kollektiivista muistitietoa, kuten esimerkiksi yhteenveto suvun historiasta ja myös muuta suvun historiaa täydentävää materiaalia.

Valokuvat ja paperilla olevan materiaalin Kausalainen oli suurelta osin digitoinut jo ennen pilotoinnin alkua. Tämä aineisto oli taltioitu hänen tietokoneelleen. Aineiston säilyvyyden turvaamiseen käytettiin sekä pilvipalvelinta että erillistä kovalevyä.

Edellä mainittu täydentävä ja selittävä materiaali on ja tulee olemaan lähes kokonaan syntysähköistä aineistoa. Se kuitenkin muodostaa vain melko vähäisen osan arkistokokonaisuudesta. Syntysähköistä aineistoa on myös jonkin verran liikkuvana kuvana: videoita ja DVD-aineistoa. Nämä on arkistoa varten jouduttu muuntamaan arkistoon paremmin soveltuvaan MP4-formaattiin. Samaan formaattiin on myös muunnettu jonkin verran Super8-formaatissa olevia kaitafilmejä.

Jäljempänä kuvattujen pilotoinnista saatujen kokemuksien ja niiden perusteella muodostettujen mielipiteiden arvioinnissa on huomioitava tarkoitukset ja lähtökohdat: suvun historian taltioiminen ja jakaminen rajatun kohderyhmän käyttöön. Tämän vuoksi pilottiprojektissa ei ole saatu kokemuksia, kuinka järjestelmä toimisi tavallisen ”peruskansalaisen” apuna hänen taltioidessaan arkiseen elämäänsä liittyvää aineistoa, kuten vakuutuskirjoja, takuutodistuksia, kauppakirjoja tai muita tärkeiksi koettuja asiakirjoja ja tallenteita.

Aineiston taltiointi on aloitettu vanhimmasta materiaalista. Kun vähitellen siirrytään uudempaan aineistoon, tulee syntysähköisen materiaalin määrä lisääntymään. Uudempi materiaali on niin laaja, että sitä on välttämättä karsittava sekä jättämällä arkistoimatta vähäarvoista materiaalia että asettamalla aineistolle erilaisia tallennusaikoja. Vanhimman materiaalin osalta tällaiseen karsintaan tai luokitteluun ei ole nähty tarvetta.

PILOTTIPROJEKTIN ANTAMIA KOKEMUKSIA KANSALAISARKISTOSTA

Arkistosovellus oli Kausalaisen liityttyä projektiin vielä uusi ja osin keskeneräinen, sitä varten ei ollut vielä käyttöohjeistusta eikä pääosa sovelluksen ns. tooltipeistä vielä toiminut. Siksi oli tarpeen järjestää ohjausta sovelluksen käyttöön. Tätä tarkoitusta varten järjestettiin kaksi muutaman tunnin tapaamista Kausalaisen ja projektia ohjaavan Liisa Uosukaisen kanssa. Sen lisäksi järjestettiin yksi puhelinneuvottelu. Koska käyttöohjeistus oli alussa puutteellista, kaikkia sovelluksen tarjoamia monipuolisia mahdollisuuksia ei ole taitojen puuttuessa vielä otettu käyttöön.

Pilotoija Eero Kausalainen on koonnut ja järjestänyt sukuyhteisönsä aineistoa Kansalaisarkistoon noin vuoden ajan. Hänen hallussaan olevasta aineistosta ja/tai arkistoinnin arvoiseksi katsottavasta aineistosta lienee arkistossa tätä kirjoitettaessa noin kaksi kolmasosaa. Arkistossa on yli 200 asiakirjaa (kirjoituksia, todistuksia ja kirjeenvaihtoa), noin 500 valokuvaa ja 25 liikkuvan kuvan tiedostoa.

Sovellus mahdollistaisi edellä kuvatun tyyppisten tiedostojen lisäksi myös karttoja, äänitallenteita ja sähköposteja. Karttoja arkistoon tulee myöhemmin jonkin verran, mutta pilottihenkilöllä ei ole arkistosovelluksen mahdollistamia digitoituja äänitallenteita, joskin osa liikkuvan kuvan tallenteista sisältää musiikkiesityksiä. Aineiston luonteen vuoksi ja koska arkistointi on tähän mennessä käsitellyt vain vanhempaa

aineistoa, arkistoon ei ole ainakaan toistaiseksi taltioitu sähköposteja.

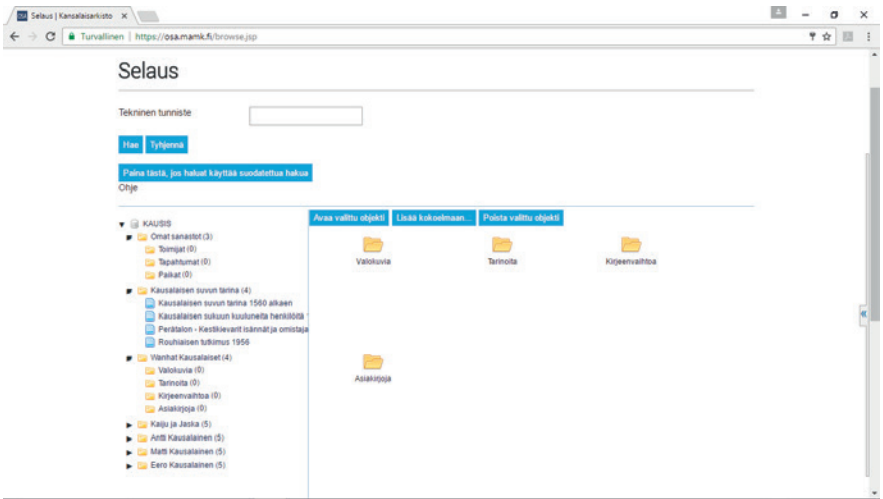
Nyt aineiston koko on suuruusluokkaa 9 Gt (noin 800 tiedostoa). Vaikka nyt on kysymys vain yhdestä arkistokokonaisuudesta, saattaa se antaa viitteitä siitä, minkä kokoisia tällaiset arkistot saattavat olla.

Aineiston luonteesta johtuen ei ohjelmiston mahdollistamia rajoitettuja säilytysaikoja ole käytetty. Tähän mennessä taltioitu aineisto on tarkoitettu säilymään pitkään, jolloin taltiointiajajaksi on määritetty ”pysyvästi”. Aineisto on määritetty salaiseksi eli se ei ole avoimesti kenen tahansa selattavissa.

Varsinaisia käyttäjiä Kausalaisen arkistolla on kahdeksan, joista yksi on hallinnoija (admin) ja muut vain luku- ja selausoikeudet omaavia (public, browser). Käyttäjästä pääosa on yli 60-vuotiaita, joiden IT-taidot ovat oleellisesti rajallisemmat kuin pilottikäyttäjän. Aineistoa arkistoitaessa on kuitenkin ajateltu, että tämän suppean käyttäjäjoukon kaikki käyttäjät voivat käyttää kaikkia aineistoja. Jos käyttäjäkuntaa nykyisestä laajennettaisiin, saattaisi olla tarpeellista rajata jokin aineisto tai sen osa muiden käyttäjien saatavilta.

Tähänastisten kokemusten perusteella voidaan arvioida, että tämän artikkelin alussa esitellyt tavoitteet ja niiden perusteella räätälöity sovellus näyttäisivät vastaavan toisiaan ainakin sukuyhteisön aineiston taltioinnin osalta. Sovellusta ja sen toimivuutta ei ole vielä tähänastisten käyttökokemusten perusteella mahdollista kaikilta osin arvioida, koska aineistoa taltioidessa on merkittävä osa metatiedoista jätetty eri syistä täyttämättä. Puuttuvat metatiedot rajoittavat monipuolisten hakuvaihtoehtojen käyttöä, joten siltä osin on tietoja täydennettävä ja saatava lisää käyttäjäkokemuksia.

Arkistosovelluksen käytön helppouden kannalta merkittävä osuus työstä jää käyttäjälle itselleen. Aineiston järjestäminen johdonmukaisiin kokonaisuuksiin on erityisen merkittävää aineiston käytettävyydelle silloin, kun sovelluksella on useita käyttäjiä. Jos aineiston luokitteluperusta ja -logiikka eivät ”avaudu” käyttäjille, se vaikeuttaa arkiston käytettävyyttä. Sovelluksen hakutoiminnoilla voidaan kuitenkin ainakin jonkin verran kompensoida arkistointilogiikan ja – järjestelyjen puutteita.



Kuva 2. *Esimerkki pilottiarkiston järjestetystä aineistosta*

Pilottiprojektissa, jossa keskityttiin suvun historiaan, tiedostojen järjestäminen ja hierarkia ratkaistiin kuvan 2 mukaisesti. Tässä mallissa aineisto on luokiteltu yleiseen ja kokoavaan osaan (Omat sanastot, Kausalaisen suvun tarina) ja sen jälkeen rakenteeltaan keskenään samansisältöisiin, samat kansiot käsittäviin osa-alueisiin. Luokittelu tässä perustuu sukupolviin ja sen kautta aikaan: aikakaudet ennen pilottihenkilön isoisän aikaa, isovanhempien aika, vanhempien aika, jne.

Kunkin kansion sisällä saattaa olla useampia alikansioita, esimerkiksi valokuva-albumeita. Sitä, kuinka toimiva tällainen luokittelu on, ei ole vielä arvioitu yhdessä käyttäjien kanssa, koska sovellus on ollut heidän käytettävissään vasta melko lyhyen ajan.

Kansalaisarkisto on mahdollistanut Kausalaisen sukuyhteisön tiedostojen – runsas valokuva- ja asiakirja-aineisto 1800-luvun loppupuolelta alkaen ja kirjalliset tekstit vuodesta 1560 alkaen – jakamisen useille käyttäjille tavalla, joka ei olisi muutoin ollut mahdollista. Samoin on syntynyt mahdollisuus jatkuvasti täydentää ja korjata aikaisempia tietoja, jotka käyttäjät saavat käyttöönsä reaaliajassa.

AJATUKSIA KANSALAISARKISTOSTA

Pilottihenkilönä toiminut Kausalainen kertoo joutuneensa pohtimaan arkistointia uudella tavalla tultuaan kosketuksiin aivan uudenlaisia mahdollisuuksia avaavan Kansalaisarkiston kanssa.

Nyt saatujen kokemusten perusteella voidaan arvioida, että palvelu voisi antaa täysin uuden mahdollisuuden kansalaisten omien aineistojen arkistoinnille, kunhan sen jatkuvuudesta ja luotettavuudesta on voitu varmistua. Kansalaisarkisto antaisi erinomaiset ja monipuoliset mahdollisuudet hyvinkin erityyppisten arkistojen luomiseen. Sovellus mahdollistaisi paitsi sukuyhteisölle tarkoitetun palvelun myös yksittäisen kansalaisen omiin yksityisiin tarpeisiinsa ja arkisten asioiden hallinnointiin luodun arkiston.

Digitaalisia ”arkistopalveluja” – pilvipalveluita – tarjotaan tällä hetkellä jo melko yleisesti. Niiden tietoturvasta, aineiston säilyvyydestä, erityisesti pilvipalveluyrityksen mahdollisen konkurssin tai muun kriisitilanteen aikana, voidaan perustellusti olla huolissaan. Luotettavia ja vastuullisia kaupallisia palveluita epäilemättä jo löytyy, mutta niillä tuskin on näin valmiiksi pohdittua, monipuolista kansalaisen arkistotarpeisiin räätälöityä sovellusta.

On kuitenkin oltava realisti ja muistettava, että kansalaisen digitaalinen arkisto edellyttää paljon käyttäjältään. Vaikka monia toimintoja voidaan automatisoida, paljon jää myös käyttäjän vastuulle – hän on itse oman arkistonsa paras asiantuntija. Kaikki ihmiset eivät ole riittävän pedanttisia ja järjestelmällisiä. Sitä minkä tahansa arkiston, perinteisen tai digitaalisen, luominen ja ylläpito edellyttää. Siksi on vaikea kuvitella, että huomattavan suuri osa kansalaisista innostuisi tällaisen palvelun käyttöön.

Inhimillisten rajoitteiden lisäksi esteinä voivat myös olla arkistosovelluksen ominaisuudet. Pilottiprojektiin osallistunut henkilö sai kolmessa eri koulutussessiossa noin kymmenen tunnin perehdytyksen, joskin kehitystyön ollessa vasta alussa sen tarvetta lisäsi ohjeistuksen puuttuminen kokonaan ja ohjelmiston näppäinpainalluksin avautuvien vihjeiden (tooltip) puuttuminen. Perehdytyksen jälkeenkään eivät kaikki toimenpiteet vielä sujuneet, joten sovelluksen kaikkia ominaisuuksia ei tämän vuoksi käytetty.

Edellä esitetyn perusteella on sovellusta kehitettäessä panostettava erityisesti sen helppokäyttöisyyteen. Ilman sitä tällaisella sovelluksella ei tule olemaan laajempaa merkitystä. Sovelluksen tulisi olla käyttövalmis suoraan, mahdollisen erillisen yleisohjeen ja ohjelmiston itseohjaavuuden avulla, ilman erityistä perehdytystä.

Jos arkistosta saadaan hyvä, yksinkertainen ja itseohjaava, se voisi kokeilussa mukana olevan Kausalaisen mielestä antaa todellisen mahdollisuuden yksityisten digitaalisten arkistojen luomiseen virallisten arkistojen (vrt. julkinen hallinto, Kansallisarkisto, Kansalliskirjasto, jne.) rinnalle. Sukujen ja kansalaisten arkistot voivat täydentää virallisia arkistoja. Ne auttavat suoraan kansalaisia omien tietojen käsittelyssä, hallinnoinnissa ja esimerkiksi sukututkimuksessa. Lisäksi ne voisivat aikanaan tarjota erinomaista, helposti käsiteltävää digitaalista aineistoa historiantutkijoille.

LÄHTEET

Jääskeläinen, A., Kosonen, M. & Uosukainen, L. 2017. My precious information – how to preserve it? In Proceedings of IS&T Archiving Conference 2017, Riga, Latvia, May 15–18, 2017.

Lampela, A. 2016. Johdatus suomalaisen arkistokuvailun historiaan. TRIM Research Reports 20. Tampereen Yliopisto, Informaatiotieteiden yksikkö. Saatavissa: https://tampub.uta.fi/bitstream/handle/10024/98693/978-952-03-0152-1.pdf?sequence=3/ [viitattu 30.3.2017].

Loponen, M. & Kosonen, M. 2016. Miten säilytän merkityksellistä muistitietoa ja tarinoita? Mikkelin ammattikorkeakoulun verkkolehti REaD 1/2016. Saatavissa: http://www.mamk.fi/read/2016/artikkeli/miten-sailytan-merkityksellista-muisti-tietoa-ja-tarinoita/ [viitattu 28.3.2017].

Uosukainen, L. 2014. Web application development in the open source archive project. In research reports A94, Uosukainen, L. (ed.) Open Source Archive, Towards open and sustainable digital archives. Mikkeli University of Applied Sciences, 63–71.

SÄHKÖPOSTIEN IHANUUS JA KURJUUS

Anssi Jääskeläinen, *TkT, TKI-asiantuntija, Kaakkois-Suomen ammattikorkeakoulu*
Tenho Kokkonen, *FM, tutkija, Päivälehden arkisto*

Kuva 6. *”My preciousss…emails”*

Ne ovat meille tärkeitä, meillä kaikilla on niitä ja usein vielä monessa eri paikassa. Niitä on tuhansia, kymmeniä tuhansia, joskus jopa satojatuhansia. Tässä massassa voi olla ensimmäiset työpaikkahakemukset, ensimmäiset varovaiset treffipyynnöt, matkustusasiakirjat, kauppalistat, työsopimukset sekä kaikki joutavanpäiväiset ”ok”-viestit.

Tätä ja paljon muuta ovat sähköpostit ja juuri kukaan ei tiedä, miten ne pitäisi säilyttää. NARA eli USAn kansallisarkisto on kehittänyt Capstone Approach –menettelyn, koska sähköpostien hallinta on ollut ongelma lähes jokaiselle valtion virastolle. Menettelyssä jokaisen tiettyä virkatasoa korkeammalla olevien henkilöiden kaikki työsähköpostit arkistoidaan niiden sisällöstä riippumatta (Nara, 2015).

Yleisesti ottaen voi todeta, että suuret kansallisarkistot ovat luoneet säännöt sähköpostien arkistointiin. Kotiseutuliiton ohjeissa yhdistysten arkistoille puolestaan todetaan, että *”tärkeistä sähköpostiviesteistä ja julkilausumista tehdään paperitulosteet arkistoa varten”* (Kotiseutuliitto, 2017). Viimeisimmän tiedon mukaan KDK-PAS ei edelleenkään mainitse sähköposteja millään tavalla (KDK, 2017), joten sähköpostien oikeaoppinen säilyttäminen Suomessa on edelleen lapsenkengissä. Hyviä poikkeuksiakin on, kuten tässä artikkelissa esiteltävä Päivälehden arkisto.

Missä me tavalliset kansalaiset sähköpostejamme säilytämme? Luonnollisesti siellä mihin ne ovat tulleetkin, eli sähköpostilaatikossa. Rajoitteena on vain laatikon koko ja työympäristössä tietohallinnolta ajoittain tulevat tiukkasävyiset viestit koskien sähköpostilaatikoiden koon pienentämistä.

Kuinka ihmiset tällaiseen vaatimukseen sitten reagoivat, riippuu täysin henkilöstä itsestään. Itse joko 1. järjestän laatikon koon mukaan ja poistan suurimmat viestit tai vaihtoehtoisesti 2. käytän Outlookin arkistoi-toimintoa poistamaan vanhimpia sähköposteja. Molemmat vaihtoehdot ovat suoraan sanoen todella huonoja.

1. Suuret viestit sisältävät usein liitetiedostoja ja viestiä poistaessa en koskaan jaksa tarkistaa, onko liite jossakin tallessa vai ei.

2. Outlookin arkistoi-toiminto ei tee mitään muuta kuin tallentaa valitut viestit & kansiot .pst-tiedostoon tietokoneen kovalevylle ja samalla poistaa ne sähköpostiohjelmasta. Nämä .pst tiedostot ovat luettavissa Outlookilla, mutta kuten tässä artikkelissa myöhemmin selviää, tämä ei ole aivan yksiselitteistä. Lisäksi, jos tietokoneen kovalevy hajoaa, nämä kallisarvoiset sähköpostiarkistot ovat poissa. Arkistoi-toiminto poistaa viestit myös sähköpostipalvelimelta eikä ainoastaan paikalliselta tililtä.

SÄHKÖPOSTIT, OSA DIGITAALISTA PERINTÖÄ

Miksi sähköpostit sitten ovat niin tärkeitä, ettei niistä osaisi luopua, tai tahdosta riippumaton luopuminen olisi ongelma? Aluksi mainittakoon se fakta, että arkistolaitoksia tai muita virallisia arkistoja eivät kiinnosta tavallisten ihmisten materiaalit, joihin myös sähköpostit kuuluvat. Perillisille, sukututkijoille tai esimerkiksi historian/politiikan tutkijoille nämä sähköpostilaatikot voisivat olla kultaakin kalliimpia. Lisäksi ihmisten kiinnostus henkilökohtaiseen arkistointiin on kasvanut vuosien saatossa, ja tarve jättää itsestään jonkinlainen digitaalinen perintö tuleville sukupolville on lisääntynyt (Hawkins, 2013).

Kenkälaatikoissa aiemmin säilytetyt perintökalleudet ovat vaihtumassa syntysähköiseen digitaaliseen perintöön, valokuviin, dokumentteihin, videoihin, sähköposteihin, Whatsapp-viesteihin yms. Tähän digitaaliseen perintöön liittyvä kansalaisten tuottaman digitaalisen tiedon määrä kasvaa jatkuvasti, kamerat tuottavat isompiresoluutioisia valokuvia ja sosiaalista mediaa käytetään ahkerasti. Jo edesmenneeltä vaariltani esimerkiksi jäi tietokoneen kovalevyllinen talvi- ja jatkosodasta kertovaa materiaalia, jota hän viimeiset kymmenen elinvuottansa tarmokkaasti keräsi. Tämän kovalevyn sisältöä ei ole onneksi levitetty pitkin sosiaalista mediaa, vaan se jaettiin meille jälkeläisille muistitikuilla.

Toisin kuin edellä Eero Kausalaisen ja Liisa Uosukaisen artikkelissa, tässä ei kuitenkaan keskitytä digitaaliseen materiaaliin kokonaisuutena, vaan sähköposteihin. Ne ovat yksi merkittävimmistä sisällöistä tuoreemmassa digitaalisessa perinnössä. Internet Live Stats -sivuston[3] mukaan maailmassa lähetetään noin 2.6 miljoonaa sähköpostia joka sekunti (vrt. noin 60 000 Google-hakua sekunnissa ja 7600 tweettiä sekunnissa). On siis ilmiselvää, että sähköposti on edelleen dominoiva viestintämuoto, eikä sille lähitulevaisuudessa löydy todellista uhkaajaa.

Omassa Gmail-tilissäni on vain 15125 viestiä ja käytettävissä olevasta tilasta on vapaana vielä noin 70%. Työsähköpostini, tietohallinnon siivoamispyynnöistä huolimatta, pitää sisällään edelleen 8219 viestiä. Tässä massassa on mukana niin tärkeitä kuin täysin hyödyttömiäkin viestejä ja jossakin ne pitää säilyttää. Uskoakseni en ole ainoa, jolla tällainen ongelma on ottaen huomioon sähköpostiviestien määrän.

Tätä tukee myös Digitalian yhteistyökumppanin Päivälehden arkiston näkemys siitä, että Helsingin Sanomien entisten päätoimittajien sähköpostit tulisi saada arkistokelpoiseen muotoon ja samalla tutkijoiden helpommin hyödynnettäviksi. Tällä hetkellä

[3] http://www.internetlivestats.com/

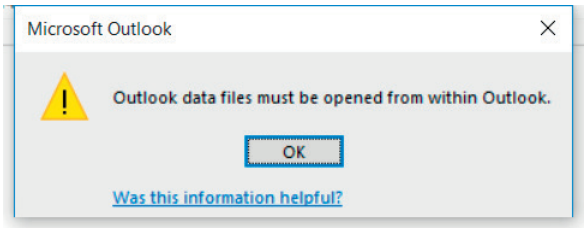
nämä sähköpostit makaavat asianhallintajärjestelmässä .pst-tiedostoina. Näistä, ja monista muista syistä, Digitaliassa päädyttiin kehittämään automaattista ratkaisua, jolla voidaan helposti toteuttaa siirtymä Outlookista oikeaan arkistoon.

ONGELMA NIMELTÄ OUTLOOK

Otsikko on provosoiva, koska Outlook on kuitenkin maailman eniten käytettyjä sähköpostiohjelmia yrityspuolella ja sähköpostiohjelmana se toimii erinomaisesti. Mutta kuten tiedämme, Outlook kuuluu Microsoftin tuoteperheeseen ja on siis kaupallinen tuote. Kaupalliset tuotteet ja niiden formaatit voivat muuttua hyvinkin nopeasti ja siksi kaupallisten tuotteiden tallennusformaatteja ei yleisesti ottaen pidetä pitkäaikaissäilytykseen soveltuvina. Poikkeuksen tekevät avoimet formaatit ja tällaiseksi pst-formaatti avoimien määritystensä takia kuuluu[4]. Tästä syystä esimerkiksi Englannin, USA:n, Australian ja Kanadan arkistolaitokset hyväksyvät sen myös pitkäaikaissäilytykseen.

Toisaalta KDK- PAS (2017) tiedostomuodot -dokumentissa ei hyväksytä mitään sähköpostiformaattia edes siirtomuotona, puhumattakaan pitkäaikaissäilytyksestä. Lisäksi se, mitä eri maiden kansallisarkistot hyväksyvät pitkäaikaissäilytykseen, ei liity oikeastaan millään tavalla siihen, missä muodossa kansalaisten tulisi sähköpostejaan säilyttää. Kansallisarkistot kun eivät ole kiinnostuneita tavallisten kansalaisten sähköposteista.

Kun peruskansalaisen näkökulmasta huomioidaan pst-formaatissa vuosien kuluessa tapahtuneet muutokset sekä mahdolliset samassa tiedostossa olevat kalenterimerkinnät, tehtävät sekä osoitekirjat, ongelma on valmis. Vanhalla Outlookilla tallennettu .pst-arkisto ei välttämättä aukeakaan Outlookin tuoreimmilla versioilla.



Kuva 2. *Outlookin virheilmoitus - .pst-tiedostot pitää avata Outlookilla*

Outlookin ”informatiiviset” virheilmoituksetkaan (kuva 2) eivät tavallista käyttäjää kauheasti naurata. IT-orientoituneemmalle käyttäjälle tuo tosin tarkoittaa Outlookin import-toimintoa. Pst-formaatti on ollut käytössä kohtalaisen pitkään, mutta 97-2002 tiedostot eivät ole yhteensopivia uudempien kanssa. Pääsyynä tähän on ANSI-merkistökoodauksen muuttaminen UNICODE-merkistökoodaukseksi uudemmissa versioissa. Lisäksi <2003 formaatti tukee maksimissaan 2 Gt tiedostoja, kun uudemmissa kokorajoitus on 50 Gt.

Näiden ongelmien ratkaisuun on olemassa työkaluja, mutta harvempi kansalainen tai harrastajatutkija hallitsee tietotekniset asiat niin hyvin, että lähtisi niitä etsimään tai käyttämään. Edellä mainittujen teknisten syiden, pitkäaikaissäilytykseen kelpaamattoman formaatin sekä avattavuuteen ja laitteistoriippumattomuuteen liittyvien

[4] https://msdn.microsoft.com/en-us/library/ff385210(v=office.12).aspx

syiden takia päätimme muuntaa sähköpostit yleisesti hyväksyttyyn, arkistolaitosten suosittelemaan PDF/A formaattiin, tarkemmin sanoen PDF/A-3b -muotoon.

Tässä yhteydessä on tosin mainittava, että KDK-PAS tiedostomuodot -dokumentissa ei hyväksytä tätäkään, vaan "tuki" loppuu jo aikaisempaan 2b-versioon. Tämä voi johtua myös siitä, että PDF/A-3 -formaatti ei rajoita mitenkään liitetiedostojen tyyppiä ja aiheuttaa näin ollen tietynasteisen ongelman.

VASTAUKSENA PDF/A-3

Digitalia päätyi "vastustuksesta" huolimatta kuitenkin PDF/A-3 -muotoon, pelkästään käytännönläheisestä syystä. PDF/A -standardia tarkemmin tuntevat tietävät, että alkuperäinen PDF/A-1 -formaatti ei salli liitetiedostoja lainkaan. PDF/A-2 -formaatti sallii liitetiedostoiksi ainoastaan tiedostoja, jotka ovat joko PDF/A-1- tai PDF/A-2 -standardin mukaisia. PDF/A-3 puolestaan sallii liitetiedostoiksi minkä tyyppisen tiedoston tahansa. Tämä on ainoa ero PDF/A-2:n ja PDF/A-3:n välillä.

Tiedostomuotorajoitteen poistuminen on sekä hyvä että huono asia. Mikään standardissa ei estä esimerkiksi laittamasta liitteeksi ohjelmanpätkää, joka avaa takaportin tietokoneelle. Koska sähköpostien liitetiedostot ovat usein muuta kuin PDF/A-1- tai PDF/A-2 -formaatin mukaisia, päätimme tukea suoraan mahdollisuutta liittää liitetiedostot sellaisina kuin ne olivat saapuessaan. Sähköpostiohjelmat sisältävät jo itsessään usein virustarkistuksen, mutta mikään ei estäisi lisäämästä myös tähän konversiovuohon virustarkastusta. Tällä hetkellä sitä ei kuitenkaan vuossa ole mukana.

Myöhemmässä kehitysvaiheessa käyttäjälle annettaneen mahdollisuus päättää mihin formaattiin hän sähköpostinsa haluaa muutettavan. Tosin sillä varauksella, että liitetiedostot saatetaan menettää, jos valitsee jonkin muun kuin PDF/A-3b -muodon. Tätä ajatellen vuohon on jo rakennettu mukaan toiminnallisuus, jossa liitetiedostoja pyritään muuttamaan PDF-tiedostoiksi ennen niiden liittämistä konvertoituun sähköpostiin. Tämä toiminnollisuus on kuitenkin sidoksissa liitetiedostojen alkuperäiseen formaattiin. Jos oikeaa konversiota ei voida taata, liitetiedosto liitetään alkuperäisessä muodossaan.

CASE: PÄIVÄLEHDEN ARKISTO

Päivälehden arkisto toimii Sanoma-konsernin Suomen-toimintojen historiallisena arkistona, ja se on osa Helsingin Sanomain Säätiötä. Arkisto kokoaa konserniin kuuluvien yhtiöiden ja toimintayksiköiden toiminnassaan tuottamat arkistoaineistot, mukaan lukien mediasisällöt ja säilyttää tämän lisäksi Sanomaan liittyvien keskeisten henkilöiden ja sukujen arkistoja. Sähköisten aineistojen määrän kasvu on ollut 2010-luvun aikana merkittävä, tiedostomäärän kasvun ollessa vuositasolla noin 9 %.

Manuaalista kirjeenvaihtoa täydentäväksi ja sen monilta osin kokonaan korvaavaksi on 1990-luvulta lähtien muodostunut sähköinen kohdeviestintä, mistä merkittävimpänä sähköposti. Päivälehden arkiston kokoelmiin on tähän mennessä luovutettu muutamia laajempia sähköpostiaineistoja, jotka luovuttajat ovat yleensä itse seuloneet ja siten valikoineet arkistoon lähetettävät, keskeiseksi katsomansa aineistot.

Luovuttajina ovat olleet lehtien päätoimittajat ja eräät muut yhtiössä keskeisessä

asemassa olleet henkilöt. Arkiston kehittämistoiminnan haasteena on ollut saada nämä aineistot pitkäaikaissäilytykseen liittyvien toimenpiteiden piiriin ja lisäksi valmistella mahdollisuutta aineistojen tutkimukselliseen hyödyntämiseen.

Sähköpostiaineistojen kannalta suurimman ongelman muodostaa viestinnän osapuolten moniulotteisuus ja erilaiset roolit. Perinteinen kirjeenvaihto suuntautuu joukkopostituksia lukuun ottamatta yleensä yhdeltä lähettäjältä yhdelle vastaanottajalle, eikä siihen sisälly useampia osapuolia. Sähköpostien osalta vastaanottajien määrä voi sitä vastoin olla hyvinkin suuri. Viestin osapuolet ovat erilaisissa rooleissa, joko varsinaisina vastaanottajina tai kopion/piilokopion saajina.

Lisäksi viestien muuntaminen ja edelleen välittäminen on teknisesti vaivatonta, jolloin aineistossa esiintyy samasta viestiketjusta useampia ilmentymiä. Näiden yhteyksien tunnistaminen ja tietojen käytettäväksi saattaminen ei onnistu perinteisin sähköisten asiakirjojen talteenottomenetelmin, joissa jokainen yksittäinen sähköpostiviesti tallennetaan säilytystietokantaan omana tiedostonaan ja lisätään samalla manuaalisesti sen metatiedot.

Sähköpostiviesteihin ja niiden sisältöön liittyy myös oikeudellisia ongelmia. Voimassa oleva tietoyhteiskuntakaari ei ota yksiselitteisesti kantaa siihen, miten viestinnän osapuolen oikeuksia yhtäältä omien ja toisaalta välitettyjen viestien edelleen luovuttamiseen säilytettäväksi historiantutkimuksen tarpeisiin olisi syytä tulkita.

Viestin lähettäjä on joissakin tapauksissa saattanut välittää vastaanottajalle tiedon siitä, että viestiä on käsiteltävä luottamuksellisena. Sähköpostiviestinnässä alkuperäinen lähettäjä ei voi mitenkään kontrolloida sitä, mitä vastaanottajat viestillä tekevät. Lähettäjä ei esimerkiksi voi mitenkään tietää, onko viestiä välitetty eteenpäin. Viestien mahdollinen arkaluontoisuus ei paljastu pelkästään yksittäisen viestin metatiedoista vaan vasta sen sisällöstä käsin.

Näin ollen aineiston tutkimuskäyttöön liittyvät oikeudet ja henkilörekistereitä koskevan lainsäädännön vaatimukset sekä eettiset kysymykset on joka tapauksessa selvitettävä huolellisesti. Sähköpostin tapauksessa kyse on varsin tuoreesta, alle 25 vuotta vanhasta aineistosta. Sama ongelmakenttä on kaikkien historiallista aineistoa sähköisessä muodossa vastaanottavien muistiorganisaatioiden pohdittavana.

Perinteisen kirjeenvaihdon tapauksessa viestien määrä on usein maltillinen paperiformaatin luonteen vuoksi. Sähköpostiviestien ominaispiirteisiin taas kuuluu se, että ne ovat helposti laadittavia ja helposti välitettäviä, jolloin viestien määrä saattaa yhden toimijan tietokannassa nousta kymmeniin tuhansiin tai jopa satoihin tuhansiin. Viestien joukkoon saattaa myös sisältyä massapostituksia, jotka käyttäjä on saanut postituslistojen välityksellä.

Tarvitaan siis automaattisia työkaluja sekä metatiedon keräämiseen että viestien mahdolliseen luokitteluun. Lähettäjien tunnistaminen ja jakaminen erilaisiin intressiryhmiin auttaa hahmottamaan toimijan verkostoja ja niiden viestinnän intensiteettiä. Sähköpostiviestiketjuista voidaan myös tutkia organisaation sisäistä päätöksentekotapaa ja keskustelukulttuuria.

Viestien liitetiedostot ovat kiinteä osa viestiä ja niidenkin pitkäaikaissäilytys on voitava varmistaa. Haasteeksi nousee se, että vaikka suurin osa liitetiedostoista on

tuotettu tavallisimmilla toimisto-ohjelmilla, on tiedostoformaattien kirjo yhdessä sähköpostiaineistossa silti hyvinkin suuri. Viestien liitetiedostojen metatiedot eivät ole välttämättä samoja kuin pääasiakirjan eli sähköpostiviestin metatiedot, jolloin on ehkä myös selvitettävä, missä määrin liitetiedostojen metatietoja on tarpeen ottaa talteen tiedon käytettävyyden varmistamiseksi.

Liitetiedostojen elinkaaren hallinnassa on historiallisen arkiston toiminnassa hyödynnettävä standardoituja ratkaisuja. Tämä tapahtuu käytännössä käyttämällä pitkäaikaissäilytyksessä yleisesti arkistokelpoisiksi tunnustettuja tiedostoformaatteja. Mikäli tätä ei voida automaattisesti toteuttaa kaikkien liitetiedostojen osalta, on näiden tiedostojen mahdollista konversiotarvetta jatkuvasti seurattava.

Viimeiseksi mutta ei suinkaan vähäisimmäksi ongelmakentäksi nousee sähköpostien sisältämien tietojen esitystapa. Kyse on siitä, miten ja missä muodossa tutkija voi saada käyttöönsä tietoja aineistosta ja millaisen käyttöliittymän avulla niitä hänelle esitetään. Analyysityökalujen ja hakutoimintojen tulee arkiston käyttäjien näkökulmasta olla riittävän tehokkaita ja niiden olisi myös riittävän hyvin esitettävä aineiston tutkijalle hakujen kattavuuteen liittyvät seikat. Mikäli aineistosta esimerkiksi rajataan tutkijoiden käytettäväksi vain ne viestit, joissa toimija on ollut aktiivisena osapuolena, mutta ei esimerkiksi kopion tai piilokopion saajana, hakutoiminnallisuuksien olisi ilmoitettava tästä viimeistään tulosten esittämisen yhteydessä.

Sähköpostien hallintaan ja pitkäaikaissäilytykseen liittyvä kehittämistyö vaatii arkistointiprosessin mallintamista ja arkistoitavien aineistojen käsittelyn vaiheistamista, jotta käytettävä työkalu voi parhaalla mahdollisella tavalla toimia. Tarvitaan työkalut sähköpostien talteenottoon, liitetiedostojen konvertointiin, analysointiin ja tiedonlouhintaan sekä formaattien elinkaaren hallintaan.

Sopivien työkalujen löytämiseksi ja kehittämiseksi Päivälehden arkisto on antanut Digitalian käyttöön yhden päätoimittajan aineiston, joka sisältää sekä saapuneita että lähetettyjä sähköpostiviestejä liitetiedostoineen kaikkiaan neljäntoista vuoden ajalta. Kyseinen henkilö on valikoinut arkistolle luovuttamansa viestit, jotka ovat hankkeen lähtötilanteessa olleet yhdessä Outlook-sähköpostitietokannassa ja sen yhdessä kansiossa.

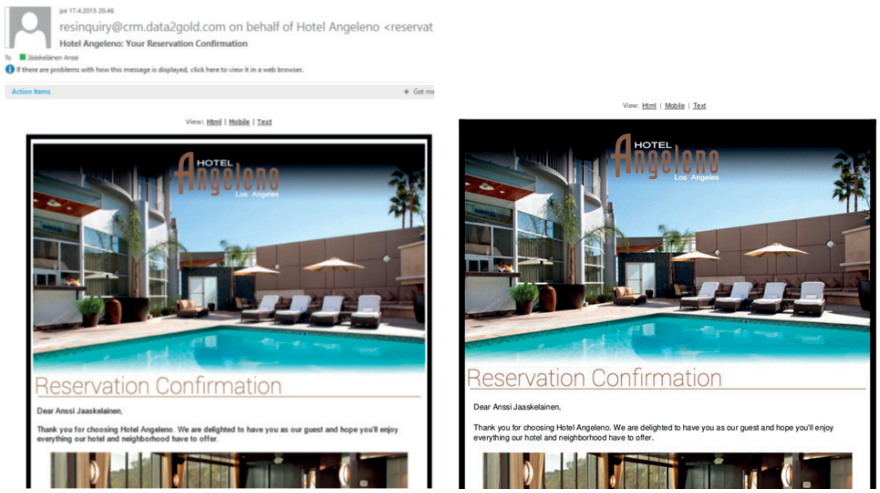
KONVERSIOVUO

Ennen ratkaisun tarkempaa avaamista on todettava, että se perustuu täysin jo olemassa oleviin avoimen lähdekoodin ohjelmistoihin. Mikään ei siis estä hyödyntämästä tätä esitettyä ratkaisua myös itsenäisesti: kaikki työkalut esitellään artikkelissa. Digitaliassa on kuitenkin rakennettu moniprosessointiin kykenevä työnkulku siten, että kaikki hyödynnyttävät ohjelmat toimivat automaattisesti ja saumattomasti yhteen, ja mahdollisimman tehokkaasti.

Loppukäyttäjän näkökulmasta sähköpostien arkistokonversion käyttäminen on hyvin yksinkertaista. Ottamalla käyttöönsä Kansalaisarkiston saa samalla myös tämän toiminnallisuuden. Käyttäjä pudottaa .pst- tai .ost-tiedostonsa Kansalaisarkiston käyttöliittymään ja näkee konversion valmistuttua kaikki konvertoidut sähköpostinsa Kansalaisarkiston käyttöliittymässä.

Samaisen käyttöliittymän kautta sähköpostit ovat myös etsittävissä mm. metatietoihin tai indeksoituun tekstisisältöön kohdistuvalla haulla. Päivälehden arkiston tapauksessa konversiovuo ajettiin täysin komentorivipohjaisena. Olemme rakentaneet myös tällä hetkellä testivaiheessa olevan itsenäisen verkkosivuston, johon käyttäjä voi pudottaa Outlookin tiedostonsa ja nähdä reaaliajassa, kuinka konversio etenee. Konversion valmistuttua käyttäjä saa ladata konvertoidut tiedostot tar.gz -pakettina. Testipääsyä tähän sivustoon voi tiedustella Digitaliasta.

Konversion aikana jokaisesta Outlookin .pst- tai .ost-tiedoston sisältämästä sähköpostista tehdään standardin mukainen PDF/A-3b -tiedosto, joka pitää sisällään kaikki alkuperäisen sähköpostin liitetiedostot sekä metatiedot. Huomioitavaa on myös se, että luodut tiedostot sisältävät OCR-tiedon, joten niiden sisältöön voidaan kohdistaa hakuja millä tahansa PDF-lukijalla. Kuva 3 esittää samaisen sähköpostin Outlookissa ja konvertoituna Adobe Acrobat Readerissa.



Kuva 3. *Sähköposti Outlookissa ja konversion jälkeen Acrobat Readerissa*

Digitalian näkökulmasta konversioprosessi on huomattavasti mutkikkaampi. Kun käyttäjän .pst- tai .ost-tiedosto on otettu vastaan, suoritetaan seuraava toimenpiteet automaattisesti.

1. Digitalian kehittämä Python-ohjelma huolehtii koko konversioprosessista.
2. Pffexport: Purkaa Outlookin tiedoston kansiorakenteeksi. Myös mahdolliset sähköpostin liitetiedostot, kalenterimerkinnät ja osoitekirjat puretaan, mutta näitä ei oteta konversioon mukaan.
3. Txt2html / unrtf
   a. Käytettävä ohjelma riippuu Pffexport-ohjelman tuloksesta. Jos Pffexport on saanut purettua html-muotoisen sähköpostin, tämä vaihe hypätään yli. Txt-tiedostot muutetaan .html-tiedostoiksi txt2html-ohjelmalla ja mahdolliset .rtf-tiedostot unrtf:n avulla latex-muotoisiksi .tex-tiedostoiksi.

b. pdflatex: Jos Pffexport on purkanut .rtf-tiedoston ja se edellisessä vaiheessa muutettiin onnistuneesti .tex-tiedostoksi, tällä ohjelmalla konvertoidaan .tex-tiedosto suoraan .pdf-muotoon.

4. Digitalian kehittämä Java-ohjelma. Sähköpostien metatietorakenne muutetaan helpommin analysoitavaan ja kirjoitettavaan muotoon.
5. Wkhtmltopdf. Nimensä mukaisesti tällä ohjelmalla muunnetaan joko pffexportin suoraan purkamat .html-tiedostot tai txt2html-ohjelman muodostamat .html- tiedostot .pdf-muotoon.
6. LibreOffice Writer. Pyrkii muuntamaan liitetiedostoista yleisimpien toimisto-ohjelmien formaatit (doc, docx, odt, ppt, pptx ja odp) PDF-muotoon.
7. ImageMagick. Pyrkii muuntamaan liitetiedostoina olevat ei-arkistokelpoiset kuvaformaatit arkistokelpoiseen .png-tiedostomuotoon.
8. Ghostscript. PDF-tiedosto yhdistetään metatietojen ja liitetiedostojen kanssa. Huomioitavaa on, että Ghostscript on vasta versiosta 9.19 asti tukenut PDF/A-3b -formaatin kirjoittamista ja useimpien Linux-jakeluiden ohjelmapankeista löytyy edelleen vanhempi versio, joten ohjelma on usein käännettävä itse lähdekoodeista.
9. VeraPDF. Validoidaan PDF/A-3b -tiedostojen standardinmukaisuus.

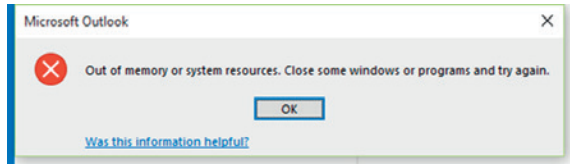
Ymmärrän lukijan mahdollisen ihmetyksen ja kummastelun siitä, miksi asiat pitää tehdä näin vaikeasti. Eikö tällaisia ratkaisuita ole jo olemassa? On, mutta maksullisina, ja monissa verkosta löytyvissä konversio-ohjelmissa on tässä vuossa tehtyihin toiminnallisuuksiin verrattuna paljon puutteita.

Lähimmäksi Digitalian toteuttamaa toiminnollisuutta pääsee maksullisella Lura-Techin PDF Compressor -ohjelmalla, mutta sekään ei tue PDF-formaatin lisäksi kuin tiettyjä kuvaformaatteja ja niiden käsittelyä. Born Digital -lisäosalla ohjelmalla saadaan konvertoitua myös Word-, Excel-, PowerPoint-, Visio-, txt-, html-, eml- ja msg-tiedostot.

Näistä kaksi viimeistä ovat formaatteja yksittäisten sähköpostien esittämiseen. Jos esimerkiksi raahaat Outlookista sähköpostin työpöydälle, se tallentuu .msg-tiedostona. Raahaamalla sähköpostit Outlookista kansioon ja käsittelemällä kansion PDF Compressor -ohjelmalla ja sen Born Digital -lisäosalla pääsee samaan tulokseen kaupallisilla työkaluilla. Tosin tätä toiminnallisuutta emme ole päässeet todentamaan, koska omistamme lisenssin ainoastaan PDF Compressoriin.

Sähköpostien tulostaminen Outlookista suoraan PDF tiedostoiksi on myös mahdollista. Kokeilin tätä toiminnallisuutta koko sähköpostilaatikolleni (noin 450 Mt) ja lopputuloksena oli 4 ytimen, 8 säikeen, 16Gt työmuistilla varustetun Core i7 koneen täydellinen hyytyminen ja lopulta ilmoitus, että resurssit loppuivat kesken (kuva 4).

Pienempää kokonaisuutta kokeillessani havaitsin, että kaikki kerralla PDF-muotoon tulostetut sähköpostit tallentuvat samaan tiedostoon. Yksittäisten sähköpostien tu-



Kuva 4. *Outlookin yritys tallentaa koko sähköpostilaatikko PDF-tiedostoiksi päättyi nolosti*

lostamiseen Outlookin tarjoama PDF-tallennus on hyödyllinen, mutta suuremman sähköpostimassan kanssa ongelmia tai turhautumista ei voine välttää.

Digitalia-hankkeessa lähtökohtana oli hyödyntää avoimen lähdekoodin tuotteita. Luonnollisesti ensimmäinen tutkimuskohde oli löytää yksittäinen avoimen lähdekoodin ohjelma, joka osaisi muuntaa .pst-sähköpostilaatikon suoraan PDF/A-3 -muotoisiksi tiedostoiksi.

Kuten arvata saattaa, tällaista ei tullut vastaan, eikä myöskään aiheeseen liittyviä tutkimuksia. Seuraavaksi tutkittiin mahdollisuutta muuttaa sähköpostit PostScript-muotoon, mutta tällaistakaan ohjelmaa ei löytynyt. Lukuisien kokeilujen ja välivaiheiden kautta päädyimme yllä esitettyyn työnkulkuun, joka tuntuisi toimivan kohtalaisen varmasti ja tehokkaasti.

Samainen tehtävä missä Outlook epäonnistui, muuntui PDF/A-3b -muotoon noin 8 minuutin ja 30 sekunnin aikana. Tässä ajassa 2783 sähköpostia muunnettiin PDF/A-3b -standardin mukaisiksi tiedostoiksi alkuperäisine metatietoineen ja liitetiedostoineen. Samanaikaisesti konversio loi käsiteltyjen sähköpostiviestien metatiedoista .csv -tiedoston, joka voidaan syöttää suoraan esim. verkostoanalyysiohjelmille, esimerkiksi Gephille.

Konversio tehtiin palvelimella, jossa oli käytössä 16 virtuaalista ydintä ja 16 Gt muistia. Konversionopeus riippuu toki monista eri asioista, kuten käytössä olevasta prosessointitehosta, sähköpostien pituudesta, merkistökoodauksesta, liitetiedostojen määrästä ja laadusta, jne. Suuntaa antavana arvona voinee pitää noin sekuntia per sähköpostikonversio per ydin.

JATKOKEHITYS JA TULEVAISUUDEN NÄKYMÄT

Kehitystyö on vielä käynnissä ja teemme Digitaliassa mieluusti yhteistyötä ja kokeita erilaisilla Outlookista ulostuoduilla sähköpostikokonaisuuksilla. Sähköpostissa mahdollisesti esiintyvät erikoiset merkistökoodaukset, esim. kyrilliset kirjaimet aiheuttavat vielä ongelmia, mutta länsimaiset merkistökoodaukset mukaan lukien ääkköset toimivat oikein.

Tällä hetkellä ratkaisu kattaa ainoastaan Outlook-järjestelmän .pst- ja .ost-tietorakenteet, mutta suunnitteilla on sen laajentaminen kattamaan myös Gmail, Hotmail, AOLMail ja muut vastaavat sähköpostipalvelut. Oletettavasti tämä laajennus olisi helpompi, koska kaikki edellä mainitut tallentavat sähköpostit avoimempaan formaattiin kuin Outlook. Toiseen vaiheeseen kuuluu myös toiminnallisuus, jolla viestejä pystyttäisiin luokittelemaan automaattisesti niiden asiasisältöön perustuen.

LÄHTEET

Hawkings, D.T. 2013. Personal Archiving: Preserving Our Digital Heritage. Medford, NJ

KDK. Säilytys- ja siirtokelpoiset tiedostomuodot v1.5.1. 2017. WWW-dokumentti. Saatavissa http://www.kdk.fi/images/tiedostot/KDK-PAS-tiedostomuodot-v1.5.1.pdf [viitattu 30.4.2017]

Kotiseutuliitto. Yhdistyksen lyhyt arkistointiohje. 2017. WWW-dokumentti. Saatavissa: http://www.kotiseutuliitto.fi/tietopankki/jarjestoosaamisen-tietopankki/yhdistyksen-oma-arkisto [viitattu 30.4.2017]

Nara. White Paper on The Capstone Approach and Capstone GRS. 2015. WWW-dokumentti. Saatavissa: https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf [viitattu 30.4.2017]

VERKOSTOANALYYSI VIESTII VALLASTA JA SUHTEISTA

Miia Kosonen, *KTT, TKI-asiantuntija, Kaakkois-Suomen ammattikorkeakoulu*

Anssi Jääskeläinen ja Tenho Kokkonen kuvasivat edellä Digitalian tuottamaa ratkaisua sähköpostien prosessointiin ja arkistointiin. Erilaisten digitaalisen viestinnän aineistojen haltuunotto ei ole itsetarkoitus: tavoitteena on, että säilytettyä tietoa hyödynnetään, sovelletaan käytäntöön, analysoidaan, yhdistellään ja tuotetaan sen pohjalta uutta tietoa.

Tämän artikkelin tavoitteena onkin tuoda lisävalaistusta siihen, miten sähköpostiaineistoa voidaan Digitalian työn tuloksena hyödyntää tutkimuksessa. Esimerkkinä on verkostoanalyysi.

YLEISTÄ VERKOSTOANALYYSISTA

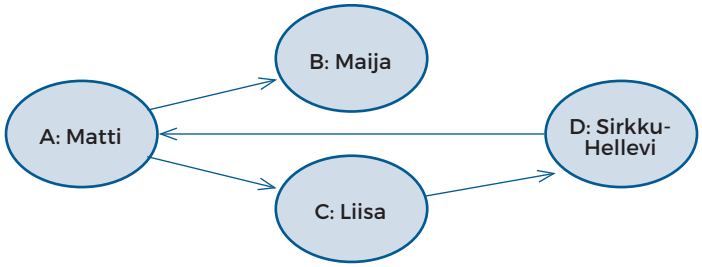
Verkostoanalyysi on joukko tutkimusmenetelmiä, joiden tarkoituksena on hahmottaa ja selittää sosiaalisia rakenteita ja sosiaalisten ilmiöiden riippuvuutta toisistaan (Johanson ym., 1995). Tutkimuskohteena ovat siis erilaisten sosiaalisten toimijoiden vuorovaikutussuhteet ja yhteydet. Suhdeverkoston perustana voi olla viestintä, sukulaisuus, omistus, samanaikainen osallistuminen tai toiminta.

Verkostoanalyysin juuret ovat matematiikassa, graafi- eli verkkoteoriassa. Sitä sovelletaan laajasti eri tieteenaloilla ja perusteisiin pääsee hyvin käsiksi ilman matemaatikon taustaa.

Verkostoanalyysin suosion taustalla ovat Johanson ym. (1995) mukaan muuttuvat yhteiskunnalliset rakenteet. Menetelmän perinteisesti vahvoja alueita ovat olleet sosiaalipsykologia ja viestintä, mutta viime vuosikymmeninä esimerkiksi organisaatiososiologit ovat kiinnostuneet siitä. Organisaatioita on alettu tarkastella avoimina järjestelminä, systeemeinä, jotka ovat tiiviissä vuorovaikutuksessa ympäristönsä kanssa ja muotoutuvat sen mukana. Verkostoanalyysi soveltuu mainiosti näiden suhteiden analysointiin. Sen avulla voidaan tarkastella myös omistus- ja valtarakenteita erilaisten instituutioiden taustalla.

Kuinka verkostot sitten käytännössä rakentuvat? Tätä havainnollistaa kuva 1, jossa on käytetty esimerkkinä sähköpostiviestejä. Arkikielessä Matti lähettää Maijalle sähköpostin. Verkostoanalyysin kielellä Matti olisi toimija A, Maija toimija B.

Näiden välille muodostuu viestin lähettämällä yhteys. Jos sama viesti lähtee cc:nä Liisalle, muodostuu samalla yhteys Matin ja Liisan välille. Maija ja Liisa taas ovat yhteydessä toisiinsa vain Matin kautta, eivät suoraan. Lopuksi Liisa välittää Matilta saamansa viestin edelleen Sirkku-Helleville, joka vastaa Matin alkuperäiseen sähköpostiin.



Kuva 1. *Yksinkertainen verkostokaavio*

Verkosto muodostuu siis toimijoista (nodes) ja toimijoiden välisistä **suhteista** (ties, edges, links). Suhteiden taustalla ovat yhteydet, joita on kahta lajia: *suunnattu* (directed) ja *suuntaamaton* yhteys (undirected).

Suunnatussa yhteydessä suhteella on alkuperä.

- Edellisessä esimerkissä Matti lähetti Maijalle sähköpostia, mutta Maija ei lähettänyt mitään Matille.
- Matti on Juulian isä, joten Juulia ei voi olla Matin äiti.

Suuntaamattomassa yhteydessä suhteella ei ole alkuperää. Jos Matti on Pekan naapuri, on Pekka väistämättä Matin naapuri. Esimerkiksi sopivat niin ikään Facebook- tai LinkedIn-kaveruus, tai Twitterissä kaksi henkilöä, jotka molemmat seuraavat toisiaan. Molempien on täytynyt omalta osaltaan tehdä päätös, että toinen osapuoli kuuluu heidän sosiaalisen median verkostoonsa.

Myös sähköpostiaineistoja voidaan käsitellä suuntaamattomana, vaikka niillä onkin lähettäjät ja vastaanottajat. Erottelu on tarpeellinen lähinnä silloin, jos tutkitaan tapoja viestiä.

Sähköpostiaineistojen tapauksessa on tarpeen määritellä myös *dikotominen* suhde: viestejä joko on lähetetty A:n ja B:n välillä, tai sitten ei. Verkostoanalyysi tuo esiin vain yhteydet niiden välillä, jotka ovat lähettäneet tai saaneet sähköpostia. Kokonaan toinen kysymys onkin, miksi kaksi henkilöä eivät ole lainkaan olleet yhteydessä toisiinsa, vaikka heidän voisi taustansa ja tavoitteensa puolesta olettaa niin tehneen. Vastauksen löytäminen edellyttää muiden tutkimusmenetelmien käyttöä.

Vielä on syytä nostaa esiin yksi verkostoanalyysin kannalta olennainen jaottelu eli **viralliset** ja **epäviralliset** verkostot. Ensin mainittu ilmenee esimerkiksi hierarkiseen valta-asemaan perustuvissa suhteissa ja jälkimmäinen nojaa esimerkiksi ystävyyteen tai tuttuuteen. Verkostoanalyysin tulosten tulkitsemisessa on hyödyllistä tietää kunkin toimijan virallinen asema. Kuvion 1 esimerkissä Maija ja Liisa voisivat työskennellä samassa suuressa organisaatiossa mutta sen eri yksiköissä, eivätkä he ole muutoin tuttuja toisilleen. Näin ollen he kytkeytyvät toisiinsa vain yhteisen esimiehen, Matin, kautta.

MITÄ VERKOSTOANALYYSI KERTOO

Verkostoanalyysin mittareista esimerkkinä voidaan mainita **tiheys** (density). Se vastaa kysymykseen siitä, kuinka monen toimijan välillä on tosiasiassa yhteys. Tiheys lasketaan jakamalla yhteyksien kokonaismäärä niiden teoreettisella maksimimäärällä.

Kuten tästä jo voi päätellä, kymmenien tuhansien viestien sähköpostiaineistossa tiheyslukema on hyvin pieni. Ei ole mielekäs lähtökohta, että kaikki aineistossa esiintyvät toimijat olisivat viestineet toistensa kanssa. Käytännössä tarkastellaan aktiivisimpia dyadeja eli pareja.

Satunnaiset toimijat eli silloin tällöin yksittäisen viestin lähettäneet tyypillisesti seuloutuvat matkan varrella pois analyysista – ellei sitten tutkijan huomio ole nimenomaan verkoston laidoilla, marginaalissa, ja siinä, millaisista lähteistä näitä viestejä on saatu. Jopa roskapostit ja huijaustarkoituksessa lähetetyt ns. nigerialaiskirjeet voivat olla joillekin arvokasta aineistoa, sikäli kun tällaiset viestit on ylipäätään säilytetty. Kaikki riippuu tutkimuskysymyksestä.

Sähköpostiarkistoissa kansion alkuperäinen haltija on keskeinen toimija. **Keskeisyys** (centrality) mittaa tietyn toimijan asemaa verkostossa: montako suoraa yhteyttä hänellä on muihin toimijoihin ja kuinka lyhyitä ovat polut muihin vaikutusvaltaisiin tahoihin. Keskeinen asema viestii valtaa ja tiedon virtauksen solmukohtana toimimista, mutta se kielii myös isosta vastuusta ja kuormituksesta. Aineistoa tarkemmin analysoimalla verkostosta voidaan tunnistaa myös muita keskeisiä toimijoita.

Tieteellisen tutkimuksen toteuttamista ajatellen vinoutunut aineisto on ongelma. Sähköpostiaineistot ovat monella tavoin vinouneita.

- Arkistoitu versio on käytännössä seulottu aineisto eli **näyte** henkilön sähköposteista, harvemmin kaikki hänen viestinsä.
- Vaikka viestejä olisi paljon, vain osa tutkimuksen tavoitteen kannalta relevantista **tiedosta** on liikkunut sähköpostin välityksellä, mikä voi johtaa kompromisseihin. Aineiston rajallisuus ei kuitenkaan saisi määritellä tutkimuskysymystä.
- Tiedon **konteksti** puuttuu. Usein keskustelun taustat eivät aukea lainkaan: *”Hei. Palatakseni eiliseen, olin väärässä.”* Mihin eiliseen? Mitä silloin tapahtui?
- Neljäs vinouma liittyy **postituslistojen** käyttöön. Kyseessä ei ole itsenäinen toimija, vaan pikemminkin verkosto verkoston sisällä. Case-esimerkissämme listat on kuitenkin rinnastettu toimijaan. Ei ole mahdollista selvittää, keitä henkilöitä on vuosia sitten ollut tietyllä jakelulistalla.

Luetteloa voisi jatkaa. Sähköpostien tapauksessa kyseessä on siis yksinkertaistus, näyte henkilön viestinnästä. Laatukriteereiltään vähemmän vaativaan tutkimuskäyttöön aineisto kuitenkin sopii.

VERKOSTOANALYYSIN SOVELLUSALUEITA

Verkostoanalyysi voi osaltaan auttaa tunnistamaan, mistä organisaatioiden arvokkain aineeton varallisuus on peräisin. Viralliset rakenteet voivat toimia tehokkaasti esimerkiksi silloin, kun pelivälineenä on helposti mitattavia ja hallittavia resursseja. Tiedolle ja osaamiselle rakentuvassa asiantuntijatyössä sen sijaan korostuvat epäviralliset verkostot. Ne ovat tutkimusten mukaan liima, joka sitoo organisaatioita yhteen. Sitoutumisen kohteena on useammin työpaikalle rakentunut suhdeverkosto kuin organisaatio itsessään. Epäviralliset verkostot myös tukevat oppimista, luovat perustan uusien tuotteiden ja palveluiden kehittämiselle ja edistävät henkilöstön tyytyväisyyttä työhönsä. (Cross ym., 2002)

Verkostoanalyysi voi pohjautua hyvinkin erilaisiin empiirisiin aineistoihin (Johanson ym., 1995). Lähteenä voivat olla haastattelut, kyselyt, toiminnan havainnointi, arkistoaineistot jne. Arkistot ovat siis vain yksi näkökulma verkostojen rakenteisiin.

Yksiulotteisessa aineistossa tutkimuskohteena ovat yhden toimijaryhmän keskinäiset suhteet. Tutkimuksessa voitaisiin tarkastella esimerkiksi tietyn yrityksen tuotekehitysyksikköä ja sitä, miten informaatio tässä verkostossa kulkee.

Kaksiulotteisessa aineistossa tarkastellaan kahden toimijaryhmän keskinäisiä suhteita. Edellisen esimerkin tuotekehitysyksikössä voitaisiin esimerkiksi verrata sitä, onko mies- ja naispuolisten tai eri-ikäisten työntekijöiden välille rakentuvissa informaatiovirroissa eroja.

Egosentrisessä aineistossa tutkitaan yksittäisen ihmisen ja muiden toimijoiden välisiä suhteita. Verkoston kaikki suhteet jäsennetään tämän yksittäisen ihmisen kautta.

Erilaisista aineistonkeruutavoista voidaan mainita **aineistolähtöinen**, jossa tutkimuksen kohde määrittää, millaisista osapuolista verkosto muodostuu. **Tutkijalähtöisessä** aineistonkeruussa määrittelijänä on tutkija oman kokemuksensa pohjalta. (Mattila & Uusikylä, 1999)

Kolmanneksi, ns. **lumipallo-otantaa** pidetään verkostoanalyysissa varsin luotettavana aineistonkeruutapana. Otanta toimii siten, että tutkimuksen kohdetta pyydetään nimeämään tahot, joihin hänellä on tutkimuksessa kuvatun mukainen suhde. Seuraavalla kierroksella pyydetään lisää nimiä aiemmin nimetyiltä henkilöiltä. Näin jatketaan, kunnes aineisto saturoituu eli uusia toimijoita ei enää tule esille. (Mattila & Uusikylä, 1999)

Sähköpostiarkistot edustavat aineistolähtöistä määrittelyä verkostolle. Aineistolähtöistä menetelmää soveltavat tutkijat ovat hyödyntäneet esimerkiksi Enronin mittavaa avointa aineistoa. Se koostuu 158 pääasiassa ylempään johtoon kuuluvan käyttäjän yli 600 000 sähköpostiviestistä. Shetty & Adibi (2004) tekivät sähköpostikorpuksesta MySQL -tietokannan. Aineistoa ovat sen julkistamisen jälkeen prosessoineet ja hyödyntäneet lukuisat tutkijat.

Tutkijat ovat myös rajanneet tästä laajasta aineistosta osia: esimerkiksi Kolli ja Narayanaswamy (2013) tarkastelivat runsaan 21 000 sähköpostin kokonaisuutta vuosilta 1999–2002 selvittääkseen, millä tavoin organisaatio viesti kriisitilanteessa. Sähköpostiviestejä analysoimalla luotiin kuvaus sosiaalisten verkostojen rakenteesta Enronilla. Tällä tavoin voitiin havaita, että isoja muutostilanteita (ts. kriisivaihetta) edelsi viestinnässä aina poikkeama totutusta.

Mitä nämä poikkeamat voivat olla? Esimerkiksi sisällöllisiä muutoksia viestien otsikoissa, termivalinnoissa ja sävyissä, mutta myös rakenteellisia muutoksia siinä, millaisella jakelulla viestit lähtevät eteenpäin, miten ja missä ajassa niihin reagoidaan, kuka reagoi jne. Kun tällaiset muutostilanteet opitaan arkistoitujen aineistojen avulla tunnistamaan, tietoa voidaan käyttää hyödyksi riskinhallinnassa ja ennakoinnissa.

Tällä hetkellä tietoa analysoidaan vielä pitkälti ihmisvoimin, mutta tulevaisuudessa yhä enemmän tekoälyn avustamana. Tästä koituu tutkimukselle merkittäviä hyötyjä. Ensinnäkin kapasiteetti tiedon käsittelyyn kasvaa räjähdysmäisesti (esimerkiksi Enronin yli 600 000 sähköpostiviestiä ovat ihmisaivoille mahdoton aineisto hallittavaksi, mutta eivät koneille). Toiseksi, tutkijat ja kehittäjät voivat keskittyä siihen, missä heitä eniten tarvitaan, eli aineistojen saattamiseen arkistoitavaan ja analysoitavissa olevaan muotoon, mielekkäiden tutkimusasetelmien rakentamiseen sekä tutkimuksen löydösten tulkitsemiseen ja selittämiseen.

Verkostoanalyysilla on myös heikkoutensa, kuten rajautuminen pelkästään verkoston rakenteeseen ja siihenkin vain tiettynä ajanhetkenä. Tutkimukselle onkin tyypillistä soveltaa verkostoanalyysin rinnalla muita menetelmiä. Myös sähköpostiaineistojen tapauksessa parhaaseen lopputulokseen päästään yhdistämällä tätä tietoa muihin aineistoihin ja asiakirjoihin. Esimerkiksi yritysjohdon, poliitikkojen ja median väliltä voi verkostoanalyysin avulla löytyä mielenkiintoisia kytköksiä. 20 vuoden kuluttua politiikan tutkijat kenties analysoivat lähemmin sitä, mitä Yleisradion ja pääministeri Sipilän välillä oikeastaan tapahtui marraskuussa 2016. Ketkä olivat yhteydessä toisiinsa? Kuka oli aloitteellinen osapuoli? Ketkä lopulta reagoivat viesteihin?

Löydökset toki vastaavat vain kysymyksiin ”kuka” ja ”milloin”. Syvempi analyysi edellyttää aina ympäröivän kontekstin ja tapaukseen liittyvän tietosisällön analysointia ja tulkintaa.

SÄHKÖPOSTIEN JURIDIIKKAA

Näiden huikaisevien mahdollisuuksien avaamisen jälkeen on syytä palata tutkimussfääreistä maan pinnalle ja tiedostaa digitaalisen viestinnän aineistojen käyttöön liittyvät rajoitteet. Näitä ovat käsitelleet Itä-Suomen yliopiston Tomi Voutilainen ja Denis Galkin raportissaan ’Oikeudet ja velvollisuudet Kansalaisarkiston tiedonhallinnassa’ (2016), jonka Digitalia on tilannut osana henkilökohtaisen tiedon arkistoinnin kehittämistä.

Sähköpostiaineistot eivät luonnollisestikaan ole julkisia. Ensinnäkin työntekijän sähköpostit kuuluvat viestintäsalaisuuden piiriin, eikä toinen henkilö saa lukea niitä ilman hänen lupaansa (Työelämän tietosuojalaki, aik. Laki yksityisyyden suojasta työelämässä, 759/2004). Sähköpostit rinnastuvat siis kirjeisiin ja kirjesalaisuuteen ja nauttivat luottamuksellisen viestin suojaa.

Kuten Päivälehden arkiston Tenho Kokkonen edellisessä artikkelissa huomautti, sähköpostiviestintä on kuitenkin perusluonteeltaan sellaista, että se venyttää paperisten kirjeiden aikakaudella laadittujen juridisten sääntöjen ja eettisten periaatteiden rajoja. Viestistä, jonka A on lähettänyt B:lle, tulee yhdellä klikkauksella myös sadan muun vastaanottajan viesti. Alkuperäisellä lähettäjällä ei ole mahdollisuutta vaikuttaa siihen,

kenelle hänen lähettämänsä tiedot välitetään edelleen ja mihin niitä käytetään.

Perustuslain takaama luottamuksellisen viestin suoja ei rajoitu ainoastaan viestin sisältöön vaan myös välitystietoihin. Nämä ovat käyttäjään yhdistettävissä olevia tietoja, joita käsitellään viestintäverkoissa viestien siirtämiseksi, jakelemiseksi tai tarjollapitämiseksi. Viestinnän välittäjä saa käsitellä näitä tietoja vain siinä laajuudessa kuin on toiminnan kannalta välttämätöntä, esimerkiksi laskutusta, teknistä kehittämistä tai tilastointia varten. (Voutilainen & Galkin, 2016)

Toiseksi, työsuhteen päätyttyä työntekijän sähköpostitili tulee sulkea mahdollisimman pian. Henkilön omalla suostumuksella on kuitenkin mahdollista luovuttaa sähköpostikansiot arkistoitavaksi. Tämä on perusteltua erityisesti silloin, kun kyseessä on merkittävä toimija, jonka viestit pitävät sisällään aimo annoksen organisaation historiaa ja tarinaa. Jos henkilö on työuransa aikana noudattanut tarkoin työnantajan ohjeita sähköpostin käytöstä, ei mahdollisesti arkaluontoista tietoa sisältävien henkilökohtaisten viestien siivoamiseen kulu lainkaan ylimääräistä aikaa – työsähköpostihan on tarkoitettu vain työasioita varten.

Selvää on, ettei tätä arkistoitua aineistoa voi kuka tahansa kadunmies ryhtyä käymään läpi, niin mielenkiintoisia kuin neljännesvuosisadan mittaisen työuran tehneen juorutoimittajan sähköpostit epäilemättä olisivatkin. Tutkija tai kansalaisille suunnatun arkistopalvelun tarjoaja on sähköpostiviestinnän osapuolten näkökulmasta sivullisen asemassa. Laki kuitenkin määrittelee aineistoille erityisiä käyttötarkoituksia, kuten tieteellinen tutkimus tai sukututkimus. Myös yrityshistoriikit ovat historiallista tutkimusta, joka rinnastetaan henkilötietolaissa tieteelliseen tutkimukseen. Näitä tarkoituksia varten voi siis käsitellä henkilötietoja, mutta viestintäsalaisuudesta poikkeamiseen tämä oikeus ei vaikuta. (Voutilainen & Galkin, 2016)

SÄHKÖPOSTIAINEISTOT JA VERKOSTOANALYYSI – CASE PÄIVÄLEHDEN ARKISTO

Päivälehden arkistolla on mittava kokoelma muun muassa Helsingin Sanomien (aik. Päivälehden) arkistomateriaalia. Tähän lukeutuvat myös lehden entisten päätoimittajien sähköpostit, jotka nämä ovat luovuttaneet arkistoitaviksi. Sähköposteihin on varastoitunut 20 vuoden ajalta arvokasta tietoa organisaation vaiheista ja historiasta. Lähtötilanne oli kuitenkin se, että .pst -tiedostot ainoastaan lojuivat dokumentinhallintajärjestelmässä – sen paremmin arkistolla kuin aineistojen käyttäjilläkään ei ollut tarkkaa kuvaa siitä, miten tietoa voitaisiin hyödyntää ja tehdä sen käyttämisestä helpompaa.

Digitalia on vastannut tähän haasteeseen kahdella tasolla. Ensinnäkin tuottamamme ratkaisu mahdollistaa sähköpostiaineiston saattamisen automaattisesti arkistointikelpoiseen formaattiin ja aineiston jatkokäsittelyn haluttuun muotoon. Toiseksi, olemme hankkeen aikana demonstroineet Päivälehden arkistolle, kuinka aineistoa olisi mahdollista soveltaa verkostoanalyysissa ja millaisia tuloksia sen avulla voidaan saada.

Aineistojen käsittelyssä on hyödynnetty Gephi-sovellusta (gephi.org). Gephi on avoimeen lähdekoodiin perustuva verkostoanalyysin ja datan visualisoinnin työkalu.

Tang ym. (2014) esimerkin mukaisesti sähköpostiosoite muodosti tässä yhden

verkoston toimijan eli noodin. Yhteyden kahden toimijan välillä katsotaan syntyvän, mikäli viestien lukumäärä heidän välillään ylittää tietyn ennalta määritellyn kynnyksen. Koska sähköpostiviestejä oli esimerkkiaineistossamme kymmeniä tuhansia, käsittelyn helpottamiseksi yksittäiset viestinvaihdot karsittiin pois ennen analyysivaihetta.

Aineistoa prosessoidessa rinnakkaiset käyttäjän nimen ja sähköpostiosoitteen eri muodot myös yhdistettiin kansion haltijan osalta (Matti.Meikalainen@domain.com, mattimeik@domain.com, ”Matti Meikäläinen”), mutta ei muiden verkostoon kuuluvien. Näin on toimittu siksi, että eri sähköpostiosoitteet voivat olla tutkijoille hyvin olennainen tieto: kummassa roolissa henkilö esimerkiksi on viestinyt, kansanedustajana vai meppinä?

Aineistosta poimittaviin tietoihin lukeutuivat viestin lähettäjä, vastaanottaja, aikaleima ja otsikkokenttä. Tietojen perusteella muodostettiin lähettäjä-vastaanottaja -parit yksilöivine tunnisteineen (id-numero). Varsinaista viestisisältöä ei siis käsitelty. Tuotimme aineistosta kaksi erillistä csv-tiedostoa, nodes.csv ja edges.csv, jotka ovat Gephi-sovelluksen luettavissa.

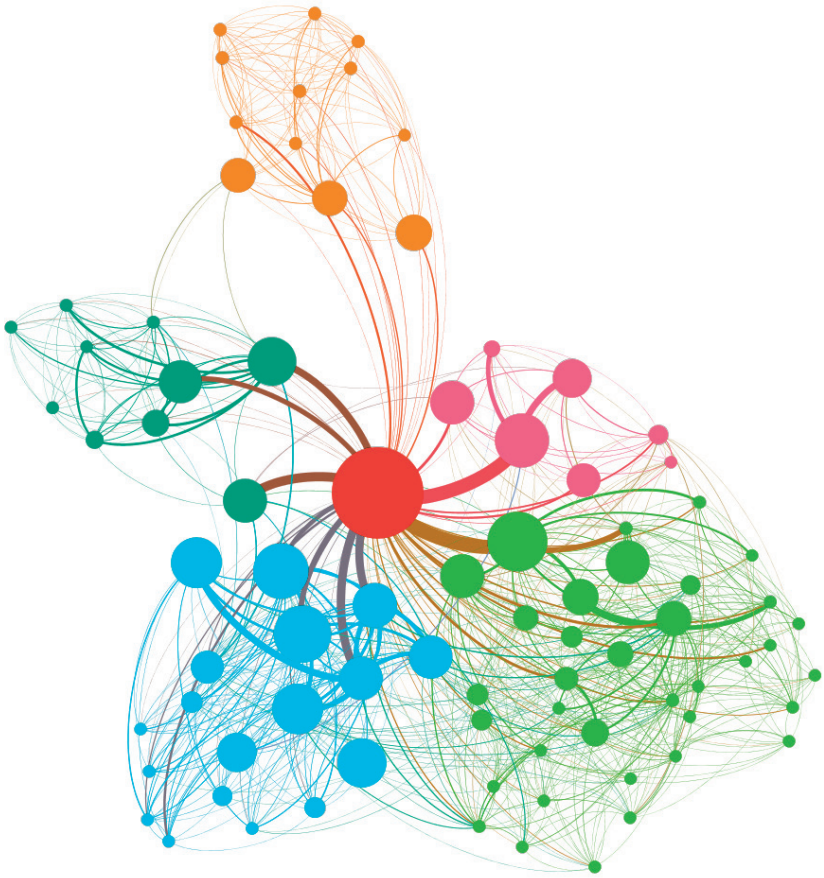
Kuvassa 2 on esimerkki suuntaamattomasta verkostokaaviosta Digitalian testaaman sähköpostiaineiston pohjalta. Kaavio havainnollistaa vain vahvimmat yhteydet eli ne toimijat, jotka ovat olleet sähköpostiviestien määrällä mitattuna eniten yhteydessä toisiinsa. Mukana on kaikkiaan 84 eri sähköpostiosoitetta.

Arkistoidun sähköpostikansion alkuperäinen haltija eli Helsingin Sanomien entinen päätoimittaja N.N. erottuu kuvan keskellä keskeisenä toimijana. Muut toimijat muodostavat toisiinsa kytkeytyviä ryhmiä, alayhteisöjä, joista esimerkkinä voidaan mainita suomalaispoliitikkojen klusteri kuvan ylälaidassa sekä kansainvälisten mediayhtiöiden klusteri kello kymmenessä. Loput merkittävät ryhmät kuvan alaosassa muodostuvat Sanoma-konsernin ja Helsingin Sanomien omiin organisaatioihin kuuluvista henkilöistä.

Verkostokaavioilla on monia käyttötarkoituksia. Tutkijat ja yrityshistoriikkien laatijat voisivat tarkastella esimerkiksi sitä, millaisia ovat olleet ylimmän johdon henkilösuhteet eri päätoimittajakausilla. Verkostoanalyysi antaa myös mahdollisuuden arvioida valtaa käyttävien instituutioiden keskinäisiä suhteita.

Sähköpostiarkistot avaavat erittäin kiinnostavia jatkokehitysmahdollisuuksia tiedon käytettävyyden parantamiseksi. Verkostorakenteen tarkastelua täydentäisi muun muassa viestien sisällönanalyysi tekstinlouhinnan avulla.

Ideaalitilanteessa sovellus ohjaa tutkijan verkostokaavion perusteella automaattisesti oikeaan kohtaan aineistossa, antaa mahdollisuuden poimia halutut aineiston osat tarkempaan analyysiin ja luonnollisesti myös tarjoaa ohjelmallisia ratkaisuja viesteissä käytettyjen termien, elementtien ja tunnesävyjen tunnistamiseen ja luokitteluun.



Kuva 2. *Verkostokaavio päätoimittajan sähköposteista*

DIGITALIA TUTKIJOIDEN TUKENA

Lopuksi on syytä huomauttaa, että Digitalian tehtävänä ei ole tehdä aineistoille varsinaisia analyyseja, niin mielenkiintoisia kuin ne sisällöltään olisivatkin. Digitalian painopiste on tiedon käytettävyyden parantamiseen liittyvässä kehitystyössä. Loppukäyttäjien – esimerkiksi yrityshistoriikkien kirjoittajien ja sukututkijoiden – on määriteltävä, mitä sähköpostiaineistosta kulloinkin halutaan selvittää. Tavoite määrittelee keinot ja tarkoitukseen parhaiten sopivat työvälineet.

Lisäksi tutkimuksen konteksti ratkaisee sen, millaista aineistoa on ylipäätään saatavilla ja millaisin edellytyksin sitä voidaan hyödyntää. Kaikkialla yleispätevää ratkaisua tai ohjeistusta ei ole mahdollista tuottaa.

Digitalia keskittyy käsittelemään aineistot mahdollisimman pitkälle sellaiseen muotoon, että ne ovat vaivattomasti loppukäyttäjien hyödynnettävissä ja yhteensopivia tavallisimpien analyysisovellusten kanssa. Kehitämme digitaalisen viestinnän aineistojen käytettävyyttä yhteistyössä kumppanien ja pilottiasiakkaiden kanssa.

Jos Sinulla on uusi näkökulma aineistojen hyödyntämiseen tai kehittämisidea omaa tutkimustyötäsi ajatellen, ota yhteyttä Digitaliaan. Luodaan yhdessä tarpeisiisi soveltuva ratkaisu!

LÄHTEET

Cross, R., Nohria, N. & Parker, A. 2002. Six Myths About Informal Social Networks – And How To Overcome Them. MIT Sloan Management Review, Vol. 43(3), Spring 2002.

Johanson, J., Mattila, M. & Uusikylä, P. Johdatus verkostoanalyysiin. 1995. Kuluttajatutkimuskeskus. Saatavissa: https://agoracenter.jyu.fi/projects/soca/jan-erik-johanson-mikko-mattila-petri-uusikyla-johdatus-verkostoanalyysiin [viitattu 29.3.2017]

Kolli, N. & Narayanaswamy, B. 2013. Analysis of e-mail communication using a social network framework for crisis detection in an organization. 8th Conference on Applications of Social Network Analysis. Procedia – Social and Behavioral Sciences 100 (2013), 57-67.

Mattila, M. & Uusikylä, P. (toim.) 1999. Verkostoyhteiskunta. Käytännön johdatus verkostoanalyysiin. Tampere: Gaudeamus.

Shetty, J. & Adibi, J. 2004. The Enron email dataset database schema and brief statistical report. Saatavissa: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.296.9477&rep=rep1&type=pdf [viitattu 4.4.2017]

Tang, G., Pei, J. & Luk, W. 2014. Email Mining: Tasks, Common Techniques, and Tools. Knowledge and Information Systems, Vol. 41(1), October 2014, 1-31. Saatavissa: https://www.cs.sfu.ca/~jpei/publications/EmailMining-KAIS.pdf [viitattu 12.4.2017]

Voutilainen, T. & Galkin, D. 2016. Oikeudet ja velvollisuudet Kansalaisarkiston tiedonhallinnassa. Mikkelin ammattikorkeakoulun julkaisuja. Saatavissa: http://www.theseus.fi/handle/10024/121274 [viitattu 6.4.2017]

DIGITALIAN TULEVAISUUS

Miia Kosonen, *KTT, TKI-asiantuntija, Kaakkois-Suomen ammattikorkeakoulu*
Noora Talsi, *YTT, tutkimusjohtaja, Kaakkois-Suomen ammattikorkeakoulu*
Mikko Tolonen, *FT, professori, Helsingin yliopisto*

Tässä artikkelikokoelmassa avattiin Digitaalisen tiedonhallinnan tutkimus- ja kehittämiskeskus Digitalia -hankkeessa vuosina 2015–2017 tehdyn kehitystyön tuloksia. Yhteenvetona voidaan todeta, että hanke saavutti suunnitellut tavoitteet digitaalisen tiedon saatavuuden ja käytettävyyden parantamisessa ja tehty työ on hyödyttänyt useita eri kohderyhmiä. Tutkimus- ja kehitystyö aineistojen käyttöön saattamiseksi jatkuu.

DIGITALIAN TYÖ ON TIETOJOHTAMISTA

Merkittävimmät digitaalisuutta edistävät mahdollistajat ovat tiedon hyödyntämisen edellytysten parantaminen ja tehokkaampi tiedonhallinta. Näihin Digitalian hanke on pureutunut. Yhteinen nimittäjä em. tavoitteille on tietojohtamisen kehittäminen. Digitalian seuraavissa vaiheissa tietojohtamista edistetään selvittämällä digitaalisuuden vaikutuksia ja kohdentamalla tutkimus- ja kehitystyötä niin yritysten, tutkijoiden kuin kansalaisten tarpeista lähtien.

Digitaalisuuskehitykselle asetetut vaatimukset saavutetaan vain hyödyntämällä modernia teknologiaa, virtuaalitodellisuutta, tekoälyä ja koneoppimista. Ne haastavat perinteiset käsitykset tiedonhallinnasta ja tiedon käsittelystä. Digitalialle tunnusomainen kokeileva, osallistava ja iteroiva tutkimus- ja kehitystoiminta parantaa niin yritysten, muiden yhteisöjen kuin kansalaisten mahdollisuuksia ottaa käyttöön näitä uudenlaisia ratkaisuja.

Digitalisaatioon ja digitaalisuuden tutkimukseen liittyy olennaisesti monitieteisyys. Kun mietitään itsenäisten ja eri lähtökohdista liikkeelle lähtevien tieteenalojen kohtaamista, on tärkeää ymmärtää erilaisten episteemisten kulttuurien merkitys. Eri tieteenaloilla tapahtuvaa kehitystä tulee kunnioittaa ja lähestyä myös sen omista lähtökohdista. Informaatiotulvan kiihtyessä suurena vaarana on tehdä ratkaisuja, joiden seurauksena tieteenalojen identiteetin kannalta olennainen yhteys aikaisempaan tutkimustraditioon ja tietoperustaan katkeaa. Digitaliassa onkin hyvä omistaa aikaa tämän näkökulman käsitteellistämiseen ja mahdollisesti myös operationalisoimiseen.

DIGITAALISESTA TIEDOSTA UUTTA ARVOA

Digitaalinen tieto ja sen käytön edistäminen ei ole itsetarkoitus. Kaiken tietojohtamisen ytimessä on yhteiskunnan ja sen verkostojen kyky luoda tiedosta uutta arvoa. Arvonluonti ei tässä tarkoita lyhytnäköisesti yksinomaan euroja, vaan eri toimijoiden yhdessä kehittämien tuotteiden ja palveluiden kykyä ratkaista käyttäjien ja asiakkaiden ongelmia paremmin kuin aikaisemmat tuotteet/palvelut sekä auttaa näkemään aiemmin piiloon jääneitä yhteyksiä ja merkityksiä.

Tiedon vaikuttavuuden näkökulmasta digitaalisen humanismin tutkimusprosessin ei tule päättyä akateemiseen julkaisuun tai tutkijoiden tekemään havaintoon, vaan juuri siitä alkaa mielenkiintoisin osuus. Tieto saa merkityksensä vasta tultuaan osaksi toimintaa ja muuttaessaan sitä.

Koska tiedolla on aina tietty konteksti, jossa nämä merkitykset muodostuvat, yksinkertaisia ”hopealuoteja” ja kaikille soveltuvia patenttiratkaisuja digitaaliseen tietoon on mahdotonta tuottaa. Digitalia voi ainoastaan pyrkiä parantamaan edellytyksiä tiedon hyödyntämiseen ja sen kautta mahdolliseen uuden arvon luomiseen. Digitalian työ on omalta osaltaan muutosjohtamista: autamme organisaatioita näkemään, millaisia mahdollisuuksia digitaaliset aineistot voivat avata ja edistämme samalla luopumista paperimaailman toimintamalleista.

Muutosagenttina Digitalia myös edistää erityisesti avoimeen ja saatavilla olevaan dataan perustuvien palvelujen kehittämistä. Kehitystyö parantaa tietojen vapaata saatavuutta ja laajempaa hyödynnettävyyttä. Näin uudet ja innovatiiviset palvelukonseptit tulevat mahdollisiksi. Pöyhimällä vanhoja normeja Digitalia voi osoittaa, millaisia mahdollisuuksia tietojen vapaa hyödyntäminen yhteiskunnalle tuottaa.

ALUEELLISTA, KANSALLISTA JA KANSAINVÄLISTÄ VAIKUTTAVUUTTA

Mikkelin seudun kannalta Digitalia mahdollistaa vahvemman profiloitumisen digitaalisessa tiedonhallinnassa. Se on niin ikään tukemassa Mikkelin erikoistumista entistä vahvemmin historiallisen digitaalisen muistin ja digitaalisten arkistoaineistojen hyödyntäjäksi ja osaajaksi. Tutkimuksen tuloksia voidaan hyödyntää Etelä-Savon alueen yrityksissä sekä kansalaisten ja yhteisöjen hyvinvoinnin lisäämisessä.

Myös yhteistyö Helsingin yliopistoon hiljattain perustetun digitaalisen humanismin keskuksen Heldigin kanssa tukee osaltaan näitä tavoitteita. Humanistisen osaamisen yhdistäminen laajeneviin tietoaineistoihin avaa mahdollisuuksia ymmärtää ja mallintaa digitalisoituvaa yhteiskuntaa ja sen rakenteita. Työhön tarvitaan digitaalisen tiedon, tietoaineistojen, niiden käsittelyn ja digitoinnin tuntemusta, joita kaikkia Digitalia tarjoaa.

Alueellisten näkökohtien ohella Digitalian tavoitteena on myös yhteiskunnallinen vaikuttavuus, kuten digitaaliseen tietoon liittyvien riskien ja mahdollisuuksien parempi ymmärtäminen. Alan keskeisenä toimijana Digitalia pyrkii myös vaikuttamaan kansallisiin tiedonhallinnan linjauksiin, joiden avulla digitaalisuutta ja digitalisaation hyödyntämistä halutaan Suomessa edistää. Digitaalisuus on hallitusohjelman läpileikkaava teema ja Suomeen haetaan digiloikkaa. Jos tiedon käytettävyyttä ei samalla paranneta, loikasta tulee lyhyt. Digitalian kehitystyön tulokset, suositukset ja hankkeen verkostomainen toimintatapa ovat yhteistyökumppaneidenkin käytössä.

Digitalian kansainvälisistä kumppanuuksista esimerkkeinä voidaan mainita neuvonantajan rooli digitaalista arkistointia edistävässä E-Ark -projektissa sekä yhteistyö Stanfordin yliopiston kirjaston kanssa sähköpostien arkistointisovelluksen kehittämisessä.

KATSE TULEVAISUUTEEN

Tulevaisuudessa Digitalian verkostoja laajennetaan ja myös kansainväliset kumppanuudet vahvistuvat. Esimerkkinä on digitaalisen tiedon kesäkoulun järjestäminen yhteistyössä kansainvälisten luennoitsijavieraiden kanssa. Digitalian vahvuus on tieteenala- ja oppilaitosrajat rikkova monialainen yhdessä tekeminen. Kuten jo tämän artikkelikokoelman kirjoittajien erilaiset taustat havainnollistavat, Digitalia tuo yhteen tutkimus- ja kehitystoimijoita tavalla, joka on yliopisto- ja ammattikorkeakoulukentässämme vielä harvinaisuus. Yhdelläkään tieteenalalla ei ole monopolia digitaalisuuteen: kyse on laajasta ja monimutkaisesta ilmiöiden kentästä, jossa monialaisuus ei ole pelkästään toivottavaa vaan välttämättömyys.

Digitalia jatkaa aktiivista yhteistyötä koulutuksen kanssa. Toiminta on kaksisuuntaista. Tutkimus- ja kehitystoiminnasta nousseita löydöksiä tuodaan osaksi Xamk:n ja Helsingin yliopiston opetusta, jolloin opiskelijoilla on mahdollisuus saada tuoreinta digitaalisuuteen ja digitaalisen tiedon hyödyntämiseen liittyvää tietoa. Digitalia voi myös nostaa esille haasteita ja kehittämistarpeita, joihin opiskelijat voivat omista lähtökohdistaan tuottaa yhteisesti uusia, erilaisia näkökulmia. Konkreettisia yhteistyömuotoja ovat digitaalisen tiedon kesäkoulun ohella erilaiset kurssityöt, opinnäytetyöt, harjoittelut ja asiantuntijaluennot. Yhteiskehittämistä opiskelijoiden kanssa voidaan toteuttaa myös hackathoneissa.

Digitaliassa luodut ratkaisut luodut ratkaisut avaavat sekä kiinnostavia jatkokehitysmahdollisuuksia että uutta liiketoimintapotentiaalia. Esimerkiksi sähköpostien arkistointiratkaisu on herättänyt kiinnostusta kansainvälisesti ja Kansalaisarkisto avaa mahdollisuuksia kaupallisen palvelun kehittämiseen. Lähitulevaisuudessa Digitalian kumppaneiksi tulee näkyvämmin myös alueen yrityksiä, jotka haluavat parantaa tietojohtamisen valmiuksiaan ja hyödyntää digitaalisia tietomassoja liiketoimintansa kehittämisessä. Digitalia edistää automaation laajempaa hyödyntämistä digitaalisten aineistojen käsittelyssä ja käyttöön saattamisessa. Työn kohteena ovat pdf-asiakirjat, digitoidut sanoma- ja aikakauslehdet, sähköpostit ja muut digitaalisessa viestinnässä syntyvät aineistot.

XAMK
KEHITTÄÄ