Tampereen yliopisto

Henrik Keskiniva

ONGELMIA JA NIIDEN RATKAISUJA ARKISTOJEN DIGITOIMISPROSESSISSA

Informaatioteknologian ja viestinnän koulukunta
Kandidaatintutkielma
Marraskuu 2021

TIIVISTELMÄ

Henrik Keskiniva: Ongelmia ja niiden ratkaisuja arkistojen digitoimisprosessissa
Kandidaatintutkielma
Tampereen yliopisto
Viestinnän monitieteinen kandidaattiohjelma
Marraskuu 2021

Digitaaliset arkistot tarjoavat laajan joukon etuja niin aineiston säilömisen kuin myös sen käytettävyyden suhteen ja viimeisein 20 vuoden aikana onkin toteutettu lukemattomia laajaskaalaisia digitointiprojekteja. Digitaaliset arkistointityökalut tarjoavat myös ennennäkemättömän mahdollisuuden välittää tietoa maailmastamme jälkipolville sekä tutkia syvemmin historiallista aineistoa. Aineiston digitoimisprosessi sisältää kuitenkin vieläkin laajan joukon ongelmia, joista monet ovat olleet läsnä digitaalisten arkistojen historian alusta asti. Tämän tutkimuksen päämääränä olikin tunnistaa nämä ongelmat sekä niiden mahdolliset ratkaisut läpi digitaalisten arkistojen historian.

Tutkimus suoritettiin soveltamalla narratiivista kirjallisuuskatsausta. Tutkimusta varten kerätty tutkimusaineisto koostui digitaalisia arkistoja ja niiden ongelmia monista eri näkökulmista lähestyvistä tieteellisistä julkaisuista, joiden pohjalta tutkimuksessa pyrittiin yleiskatsauksellisesti kokoamaan yhteen erinäisiä aineistossa esille tulleita ongelmia sekä mahdollisia ratkaisuehdotuksia niihin. Johtuen kerätyn aineiston sirpaleisesta luonteesta, tutkimus päätyi käsittelemään useita eri arkisto- ja aineistokonteksteja

Tutkimusta varten digitointiprosessi jaettiin karkeasti kolmeen vaiheeseen: digitointia edeltävä vaihe, digitointiakti ja digitoinnin jälkeinen vaihe. Digitointia edeltävää vaihetta koskien aineistosta kävi ilmi digitointiprosessiin allokoidun rahoituksen merkitys sekä tämän pohjalta syntyvä kysymys aineiston valinnasta: jos kaikkea säilöttyä aineisto ei kyetä digitoimaan, millä perusteilla kokoelman digitoitava osa pitäisi valita ja miten se pitäisi digitoida? Tämä johti käyttäjän ja aineiston itsensä ominaisuuksia painottavaan digitoimisprosessiin, jossa digitoitava aineisto valitaan yhtä lailla sen historiallista arvoa kuin myös mahdollista käyttäjäkuntaa painottaen, ja digitoimisprosessi suoritetaan näiden tekijöiden ehdoilla. Digitointiaktin yhteydessä nousi esiin kysymys aineiston integriteetistä ja esimerkkinä käytetyn OCR-teknologian vaikutuksista tähän. OCR-teknologia ainakin nykyisessä tilassaan tuottaa aina hieman virheitä, ja digitaalisissa arkistoprojekteissa vaaditaankin pragmaattisuutta näiden virheiden suhteen: aineistosta ja sen käyttötavoista riippuen tietty määrä virheitä on siedettävä taloudellisten rajoitteiden johdosta. Digitoinnin jälkeisestä vaiheesta kävi ilmi digitointiprosessin todellinen aikajana, sillä digitaalisen aineiston säilömiseen on sitouduttava teoriassa ikuisesti. Todellisuudessa digitaalisen aineiston säilöminen on paljon resursseja kuluttava huoltamisen prosessi, jossa digitaalista aineistoa säilövää laitteistoa ja ohjelmistoa täytyy alituisesti huoltaa ja uusia. Tämän lisäksi aineiston integriteetti voi kärsiä myös tahallisista tai tahattomista inhimillistä toimista johtuen digitaalisten työkalujen helppokäyttöisyydestä ja toisaalta voimakkuudesta. Tässä osiossa myös tuli kaikista selvimmin ilmi organisaationallisen kontekstin merkitys digitaalisille arkistointiprojekteille: kaikki arkistointi tapahtuu organisaatioiden sisällä, ja täten organisaationalliset toimintahäiriöt heijastuvat suoraan arkiston stabiiliuteen ja täten myös sen integriteettiin. Digitaalisen arkistoinnin kannalta on siis tärkeää keksittyä digitointiin liittyvien teknologisten seikkojen lisäksi myös organisaationallisiin ja taloudellisiin tekijöihin.

Avainsanat: Digitaalinen arkisto, digitointi, pääsy, etäkäyttö. valinta, tekstintunnistus, OCR, säilöminen, kovalevy, integriteetti, digitaalinen aineisto

SISÄLLYS

1 JOHDANTO

Digitaalinen arkistointi on suhteellisen tuore ilmiö ja sen historia on nivoutunut tiukasti ja erottamattomasti tietotekniikan historiaan. Yli kaiken, digitaaliset arkistot ovat riippuvaisia digitaalisen informaation tallennusvälineistä kuten kovalevyistä, ja näiden kehittyessä myös digitaaliset arkistot ovat yleistyneet. Puhtaasti idealistisella ja huomattavasti yksinkertaistetulla tasolla arkistoinstituution päämääränä on säilöä informaatiota niin eheänä ja niin pitkään kuin mahdollista sekä nykyisten läpinäkyvyyden ja tiedon demokratisoinnin ideaalien mukaisesti mahdollistaa pääsy tähän informaatioon niin vaivattomasti kuin mahdollista. Digitaalinen arkisto tarjoaa aikaisempaa, puhtaasti fyysistä arkistoa tehokkaammat työkalut tämän päämäärän toteuttamiseksi: yhdelle kovalevylle voi mahtua lukemattoman arkistokaapin verran dokumentteja, ja jokaiseen näistä dokumenteista voi päästä käsiksi toiselta puolelta maapalloa (Arthur et al. 2004, 8). Nämä digitaalisen arkiston tarjoamat edut eivät olekaan jääneet huomaamatta ja viimeisen 30 vuoden aikana on suoritettu huomattava määrä digitointiprojekteja. Esimerkiksi Iso-Britannian kansalliskirjaston British Libraryn digitaalisessa 1800-luvulla julkaistuja sanomalehtiä sisältävässä sanomalehtiarkistossa British Library Newspaper Database on huikeat 3 miljoona erillistä sivua 70:stä eri nimikkeestä (Kettunen, Pääkkönen & Koistinen 2016, 3).

Näiden edistysaskelten mukana on tullut kuitenkin oma joukko ongelmia, joista osa on uniikkeja digitaalisille arkistoille ja osa taas jatkumoa fyysisten arkistojen perimmäisimmille ongelmille. Tässä tutkielmassa pyrin tunnistamaan näitä ongelmia läpi digitaalisten arkistojen historian käyttäen edeltävää digitaalisia arkistoja koskevaa tieteellistä kirjallisuutta tutkimusaineistonani. Toivon, että katsomalla taaksepäin ja tunnistamalla digitaaliseen arkistointiin liittyneitä ongelmia sekä mahdollisia ratkaisuja näihin ongelmiin pitkältä aikaväliltä, voisin paremmin ymmärtää digitaalista arkistointia nykyhetkessä rajoittavia tekijöitä. Koska tämä tutkielma on luonteeltaan yleiskatsauksellinen, toivon tutkimustulosten valaisevan mahdollisia, tarkemmin rajattuja jatkotutkimuskohteita liittyen tutkimuksessa tunnistettuihin ongelmiin.

2 TEOREETTINEN TAUSTA

Tätä tutkimusta varten aineiston digitointiprosessi jaetaan hyvin karkeasti kolmeen eri vaiheeseen: digitointia edeltävä vaihe, digitointiakti ja digitoinnin jälkeinen vaihe. Jokaiseen näistä vaiheista sisältyy valtava määrä rinnakkaisia tai vaihtoehtoisia prosesseja, ja tämän jaottelun tarkoituksena ei olekaan mallintaa totaalisesti digitointiprosessin eri vaiheita, mutta sen sijaan helpottaa tämän prosessin ongelmakohtien kategorisointia. Esimerkkinä digitointia edeltävän vaiheen prosesseista voi toimia vaikkapa digitoitavan aineiston valinta kokonaisaineiston joukosta (Astel & Muir 2002, 68) tai fyysisen aineiston valmisteleminen digitointiprosessia varten (Arthur et al. 2004, 10). Vaiheista keskimmäinen, digitointiakti, sisältää kaikki mahdolliset tavat muuntaa aikaisemmin fyysisessä formaatissa ollut aineisto digitaaliseksi. Tämä voi tarkoittaa yksinkertaisimmillaan objektin valokuvaamista digitaalisella kameralla (Astel & Muir 2002, 67) tai sitten monimutkaisemmassa tapauksessa vanhojen tekstidokumenttien skannaamista digitaalisen muotoon tekstintunnistusteknologiaa käyttäen (Kettunen et al. 2016, 2). Vaiheista viimeiseen, digitoinnin jälkeiseen vaiheeseen, kuuluu laaja joukko aineiston rikastamiseen, säilömiseen ja mahdolliseen huoltamiseen liittyviä prosesseja kuten metadatan liittäminen aineistoon ja aineiston periodillinen siirtäminen uudelle tallennusvälineelle. (Arthur et al. 2004, 12).

3 TUTKIMUSASETELMA

3.1 Tutkimuskysymykset

Tätä tutkimusta varten kerätty tutkimusaineisto koostuu tieteellisistä artikkeleista ja esseistä, joista jokainen käsittelee digitaalisia arkistoja joko suoraan tai ainakin aihetta sivuten. Aineisto on luonteeltaan sirpaleinen ja tarjoaa useita eri näkökulmia digitaalisiin arkistoihin sekä digitoituun aineistoon. Tämän aineiston pohjalta tutkimuksessa pyritään vastaamaan näihin kysymyksiin:

1. Mitä ongelmia arkistojen digitointiin on liittynyt ja miten niitä on ratkottu?
2. Missä vaiheessa digitointiprosessia nämä ongelmat ilmenevät ja millaisia ne ovat luonteeltansa?

Tässä tutkimuksessa esille tulleita digitaalisia arkistoja koskevia ongelmia pyritään tarkastelemaan soveltaen edellä esiteltyä jaottelua aineiston digitointiprosessin eri vaiheiden välillä. Tämän jaottelun lopullinen tarkoitus on tulosten kategorisoinnin lisäksi tunnistaa digitaalisten arkistojen

niin tämänhetkiset kuin myös historiallisetkin kipupisteet mahdollisimman tarkasti, mikä toivottavasti johtaa selkeämpään kuvaan digitaalisten arkistojen tulevaisuudesta.

3.2 Tutkimusmetodi

Tämän tutkimus suoritettiin kuvailevana kirjallisuuskatsauksena käyttäen Salmisen vuonna 2011 julkaistussa artikkelissa ”*Mikä kirjallisuuskatsaus?: Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin*” esittelemää määritelmää narratiiviselle kirjallisuuskatsaukselle. Artikkelissa Salminen luonnehtii kuvailevaa kirjallisuuskatsausta ”yleiskatsaukseksi ilman tiukkoja ja tarkkoja sääntöjä” (Salminen 2011, 6). Aineiston sallittava laajuus sekä sen valintakriteerit ovat myös väljempiä kuin systemaattisemmissa kirjallisuuskatsauksen muodoissa. Narratiivisen kirjallisuuskatsauksen Salminen taas määrittelee kuvailevan kirjallisuuskatsauksen alaorientaationa, jossa ”epäyhtenäistä tietoa järjestetään jatkuvaksi tapahtumaksi” (Salminen 2011, 7). Johtuen tämän tutkimuksen tutkimusaineiston monimuotoisuudesta niin aihepiirinsä kuin myös julkaisuajankohtansa puolesta sekä tutkimuksen ajallisesti taaksepäin katsovasta luonteesta, narratiivinen kirjallisuuskatsaus kuten edellä määriteltynä tarjoaa parhaan mahdollisen metodisen viitekehyksen tutkimuksen suorittamisen kannalta. Narratiivinen kirjallisuuskatsaus on myös tutkimuksen laajuuden kannalta sopivin metodinen vaihtoehto, sillä narratiivinen kirjallisuuskatsaus ei välttämättä vaadi kaikkien aiheesta aikaisemmin kirjoitettujen tieteellisten tutkimusten analysointia toisin kuin vaikkapa tilastollisia metodeja soveltava systemaattinen kirjallisuuskatsaus.

3.3 Tutkimusaineisto

Tämän tutkimuksen tutkimusaineisto on laajalti kerätty erinäisiä avainsanahakuja toteuttamalla eri hakumoottoreissa, ja johtuen tästä hajanaisesta aineiston keräysstrategiasta tutkimusaineisto ei muodosta järin koherenttia kokonaisuutta. Avainsanahakujen lisäksi mahdollista lisäaineistoa kerrytettiin esimerkiksi helmenkasvatusstrategiaa käyttäen, minkä seurauksena aineiston sisällä on havaittavissa pienempiä kokonaisuuksia, joista selkein koostuu tekstintunnistusteknologiaa koskevista lähteistä. Johtuen tutkimuksen laajuudesta ja luonteesta, kerätyn aineiston hajanaisuus ei kuitenkaan ole ongelma: sirpaleinen ja aiheiltaan moninainen aineisto sivuaa useita eri arkistointikonteksteja ja niissä syntyviä uniikkeja ilmiöitä, minkä seurauksena aineisto antaa kokoonsa suhteutettuna odotettua laajemman kuvan digitaalisista arkistoista.

Vaikka tämä tutkimusaineisto käsittelee useita eri arkisto- ja aineistotyyppejä, selkeinten aineistossa esillä ovat historiallista, tekstipohjaista aineistoa säilövät julkiset arkistot. Tämä on monelle varmastikin tutuin arkistomuoto, ja ensimmäinen asia, joka tulee mieleen termin ”arkisto” kuultaessa. Aineistossa esille tuodut, tästä mallista selkeästi poikkeavat esimerkit kuitenkin

paljastavat arkistomaailman olevan huomattavasti monimuotoisempi. Esimerkiksi Proscovia Svärdin vuonna 2017 julkaistu artikkeli ”*The woes of Swedish private archival institutions*” käsittelee yksityisiä arkistoinstituutioita, joiden säilömä materiaali voi olla jo valmiiksi digitaalista johtuen sen syntymäkonteksteista. Arkistolla voi usein olla puhtaan historiallisen tarkoitusperän lisäksi myös funktionaalinen päämäärä, kuten vaikkapa organisaation toiminnan mahdollistaminen tiedonhallinnallisten prosessien suorittamisen muodossa. Tämä ei tarkoita, etteikö arkistoidulla aineistoilla olisi historiallista arvoa, mutta sillä on pääsääntöisesti organisaatioille itsellensä relevantti käyttötarkoitus, ja se on arkistoitu tätä varten. Tässä ja muutamassa muussa aineiston artikkelissa mainittuja arkistoja erottaa muusta aineistoista myös niiden yksityisyys, joka selkeästi vaikuttaa niiden toimintamenetelmiin ja täten synnyttää oman joukkonsa tutkittavia ongelmia. Myöskin artikkelien arkistoissa säilötyn materiaalin puhtaasti digitaalinen elämänkaari erottaa artikkelin loppuaineistosta: etenkin organisaationallisissa arkistoissa huomattava osa säilötystä materiaalista on syntynyt digitaalisena, eli digitointiprosessia sinänsä ei koskaan tapahdu. Edellä esitelty jaottelu digitointiprosessin eri vaiheiden välillä ei siis välttämättä ole relevantti, ainakaan kokonaisuudessaan, kaikille digitaaliselle arkistoinnin eri konteksteille.

Arkistoidun aineiston luonteen suhteen huomattavin poikkeus tutkimusaineistossa on Rielle Navitskin vuoden 2014 essee ”*Reconsidering the Archive: Digitization and Latin American Film Historiography”*, joka käsittelee latinalaisamerikkalaisen filmipohjaisen aineiston digitointia. Esseen huomattava painopiste on tätä aineistoa säilövissä open access-arkistoissa ja toisaalta niiden rajoitteissa. Essee tarjoaa oivalluksia etenkin digitointia edeltävän vaiheen ongelmiin, joskin se sivuaa myös muita digitointiprosessin vaiheita sekä niiden ongelmia.

Aineistossa on myös läsnä mainintoja digitaalisista kirjastoprojekteista, ja vaikka kirjastoilla ja arkistoilla instituutioina onkin useita kriittisiä eroja, tämän tutkimuksen kannalta digitaalisia kirjastoja koskeva tutkimus on arvokasta. Astlen ja Muirin vuonna 2002 julkaistussa artikkelissa ”*Digitization and preservation in public libraries and archives*” tutkitaan niin julkisten arkistojen kuin myös kirjastojenkin digitointia. Artikkeli pohjautuu osittain Gouldin ja Ebdonin vuonna 1999 julkaistuun tutkimukseen ”*Survey on digitisation and preservation*”, joka on tämän tutkimuksen tutkimusaineistoista varhaisin. Vaikka Astlen ja Muirin artikkelin fokus on jaettu näiden kahden instituution välillä ja artikkeli on huomattavan vanha, siinä esille tuodut ongelmat tulevat esille muissa tuoreemmissa ja eksklusiivisesti arkistoihin keskittyvissä lähteissä, mikä viittaa näiden ongelmien relevanttiuteen tämänkin hetkisille digitointiprojekteille sekä jonkin tasoiseen yhteyteen digitaalisten kirjastojen ja digitaalisten arkistojen ongelmien välillä. On myös tärkeää huomata, että itse digitointiprosessin suhteen digitaalinen kirjasto ja digitaalinen aineisto ovat lähestulkoon

identtisiä: aineisto pitää kummankin instituution tapauksessa valita ja valmistella digitointiprosessia varten, jonka jälkeen digitaalinen aineisto pitää formatoida ja säilöä tallennusvälineelle (Astle & Muir 2002).

Selkein kokonaisuus tämän aineiston sisällä koostuu tekstintunnistusteknologiaa (*OCR = optical character recognition*) käsittelevistä lähteistä. Kolmea neljästä tekstintunnistusteknologiaa käsittelevistä lähteistä yhdistää yksi yhteinen tutkija, Kimmo Kettunen, ja nämä kolme artikkelia käsittelevätkin suomenkielisen aineiston digitointia ja tekstintunnistusteknologian ongelmia luonnollisen kielen kanssa. Kokonaisuuden neljäs lähde, Kazimierz Chorośin ja Joanna Jaroszin vuoden 2018 artikkeli "*Most Frequent Errors in Digitization of Polish Ancient Manuscripts*", tutkii myös tekstintunnistusteknologiaa ja luonnollisen kielen sille asettamia haasteita, mutta tässä tapauksessa puolan kielen näkökulmasta. Nämä artikkelit ovat kaikista relevanteimpia tälle tutkimukselle havaintoesimerkkeinä digitointiaktin aikana mahdollisesti tapahtuvista virheistä ja näiden virheiden mahdollisista seurauksista, mutta ne myös valaisevat laajempaa ongelmaa nykyisten digitointityökalujen luotettavuudessa.

Vaikka aineisto on kohtalaisen suppea, siitä käy selkeästi ilmi digitaalisten arkistojen moninaisuus. Aineisto myös todistaa, että digitaalisia arkistoja ja arkistojen digitointia on tutkittu läpi niiden historian useasta eri näkökulmasta. Erilaisia aineistoja, arkistoja, työkaluja sekä näitä koskevia ongelmia on tutkittu globaalisti ja tutkimusaineiston lähteet sijoittuvatkin huomattavan pitkälle aikavälille (1999–2018). Tutkimuksissa on kuitenkin toistuvia teemoja, ja lähteistä aikaisimmassa esiin tuodut ongelmat ovat vieläkin nähtävissä lähteistä kronologisesti tuoreimmassa. Tiedonhakuvaiheen aikana tuli kuitenkin selväksi, että kattavaa, yleistason tutkimusta ei ole juurikaan tehty. Jopa lähteistä luonteeltaan yleiskatsauksellisemmatkin keskittyvät loppujen lopuksi vain yhteen tai kahteen digitointiprosessin vaiheista, jättäen paljon tilaa jatkotutkimuksille.

4 TULOKSET

4.1 Digitointia edeltävä vaihe

Digitointia edeltävä vaihe tapahtuu nimensä mukaisesti ainoastaan fyysisenä aineistona elämänkaarensa aloittaneen aineiston yhteydessä. Tätä aineistoa on usein kerääntynyt pitkän aikavälin yli ja se voi olla säilöttynä useassa eri formaatissa, kuten vaikkapa paperille tai vastaavalle materiaalille kirjotettuna tai painettuna tekstidokumenttina sekä mikrofilminä. Kaikki nämä eri

formaatit omaavat omat haasteensa digitoinnin suhteen sekä tarjoavat mahdollisia digitointiaktissa menetettäviä etuja (Astle & Muir 2002, 67). Eniten aineistossa esiin tullut tätä vaihetta koskeva ongelma on kuitenkin valintaongelma: ei-digitoitua materiaalia voi olla enemmän kuin kapasiteettia sen digitoimisille, jolloin tarvitaan jonkinlainen standardi tai perustelu digitoitavan aineiston valinnalle.

Yksi tämän tutkimusaineiston kronologisesti varhaisimmista lähteistä, Peter J. Astlen ja Adrienne Muirin vuoden 2002 artikkeli ”*Digitization and preservation in public libraries and archives*”, havainnollistaa hyvin digitaalisten arkisto- ja kirjastoprojektien kasvukipuja sekä toisaalta tiettyjen ongelmien ajattomuutta. Kaikista aikaisimmissa digitointiprojekteissa 1990-luvun alussa fokus oli selkeästi säilönnässä ja digitaalisten jäljennösten tarjoamissa eduissa esimerkiksi kopioiden tuottamisen suhteen (Prescott & Hughes 2018, 3). Tietoverkkoteknologian ja etenkin Internetin kehittyessä pelkän säilönnän rinnalle ilmestyi myös kasvava kiinnostus etäkäyttöä (engl. remote access) kohtaan. Internet ja kasvavat latausnopeudet mahdollistivat sen, että digitoituun aineistoon pystyi päästä käsiksi mistä tahansa päin maailmaa niin monta ihmistä kerrallaan kuin palvelin vain salli. Tietotekniikan kehittyminen ja yleistyvät digitointiprojektit toivat mukanaan myös uusia, digitaaliselle aineistolle täysin uniikkeja toiminnallisuuksia, kuten mahdollisuuden suorittaa useita eri dokumentteja kattavia tekstihakuja tai digitaalisesti korostaa valokuvan eri ominaisuuksia. Vuoteen 2002 mennessä nämä tekijät olivatkin johtaneet huomattavan määrän rahoitusta keränneisiin julkisiin digitointiprojekteihin Iso-Britanniassa (Astle & Muir 2002, 67). Johtuen jo olemassa olevan aineiston määrästä ja digitointiprosessin kustannuksista, kaikkea arkistojen ja kirjastojen taltioimaa aineistoa ei voitu digitoida, jolloin heräsi digitointiprosesseja alusta asti riivannut kysymys digitointiprosessin läpikulkevien aineistokappaleiden valinnasta. Aiheesta aikaisemmin suoritetussa tutkimuksessa aineiston valinnalle tarjottiin useita kriteerejä, kuten kyseiseen aineistoon pääsyn (engl. access) lisääminen, aineiston historiallinen tai kulttuurillinen arvo, aineiston akateeminen tärkeys, aineiston tehostettu säilöminen, fyysisen tilan säästäminen ja digitoidun aineiston mahdollinen taloudellinen hyödyntäminen (Gould & Ebdon 1999, 12). Näistä kriteereistä aineiston historiallinen ja kulttuurillinen arvo sekä siihen pääsyn lisääminen olivat kummatkin todettu painavimmiksi syiksi aineiston digitoinnille (Gould & Ebdon 1999, 12). Yksi digitointiprosessia mutkistava tekijä on myös tekijänoikeuksia koskeva lainsäädäntö, ja tämän lainsäädännön vaikutus näkyi käänteisesti digitoitavan aineiston valintaprosessissa: Astlen ja Muirin vuoden 2002 tutkimuksessa havaittiin, että edellä mainittujen Gouldin ja Ebdonin vuoden 1999 tutkimuksessa esille tulleiden tekijöiden lisäksi myös tekijänoikeussuojan puute aineistosta vaikutti huomattavan positiivisesti päätökseen sen digitoinnista. Tämä tekijänoikeuslainsäädännön

käänteinen vaikutus digitointiin on nähtävillä myös digitaalisesti arkistoidun elokuvamateriaalin saatavuudessa: Latinalaisen Amerikan elokuvatuotanto on historiallisesti ollut sidottu maiden kulttuurillisiin ja kansallisiin identiteetteihin ja täten rahoitettu valtion toimesta. Tämän julkisen omistajuuden johdosta huomattava määrä digitoidusta Latinalaisen Amerikan elokuvatuotannosta onkin saatavilla ilmaiseksi erinäisten valtion rahoittamien digitaalisten arkistojen kautta (Navitski 2014, 122—123). Viimeinen Astlen ja Muirin tutkimuksessa esille tullut huomattava valintaprosessiin vaikuttava kriteeri on aineistoon kohdistuva mahdollinen julkinen kysyntä, joskin sen tarkka mittaaminen ennen digitointiprosessin itsensä suorittamista on kyseenalaista. Tämän julkisen kysynnän muuntaminen tulonlähteeksi oli osoittautunut tutkimuksen mukaan myös ongelmalliseksi, joskin taloudellinen hyöty ei ollut yhdenkään tutkimuksen kohteena olleen organisaation perimmäinen syy digitoinnille (Astle & Muir 2002, 73). Kaikkien edellä mainittujen valintaprosessiin liittyvien harkintojen taustalla on loppujen lopuksi kaikkia tässä maailmassa teetettyjä digitointiprojekteja yhdistävä tekijä: rahoituksen rajallisuus. Itse digitointiaktiin käytettävien työtuntien ja sitä varten tarvittavan teknologian lisäksi aineistoa digitoidessa on pakonomaisesti sitouduttava aineiston säilömiseen läpi sen digitaalisen elinkaaren, mikä tarkoittaa mm. säilöntää varten hankitun laitteiston ja ohjelmiston vuosittaisten ylläpitokustannusten maksamista. Tämän lisäksi etenkin kirjastojen osalta digitointiprosessissa vaaditaan huomattava määrä ammattitasoista tietämystä digitoitavasta aineistosta sen osuvaa luettelointia varten, mikä tarkoittaa lisää projektille omistettuja inhimillisiä resursseja (Astle & Muir 2002, 70). Nämä taloudelliset tekijät muodostavat siis pohjan valintaongelmalle: rajallisten resurssien, niin teknologisten kuin myös taloudellisten, varassa digitoitavan aineiston valinnalle on löydettävä vahvimmat mahdolliset perustelut. Vaikka historiallinen ja kulttuurillinen merkitys ovat vieläkin tärkeitä valintakriteerejä, digitointiprojektien fokuksen siirtyminen pelkästä säilömisestä myös pääsyyn johti uusiin valintakriteereihin, kuten vaikkapa aineistoon kohdistuvaan julkiseen kysyntään. Valintaongelmaa kronologisesti huomattavasti myöhemmin sivuava Prescottin ja Hughesin vuoden 2018 artikkeli ”*Why do we digitize? The case for slow digitization*” pyrkii tuomaan esiin pääsyä ja julkista kysyntää painottavan valintaprosessin mahdolliset negatiiviset vaikutukset sekä tarjoaa uuden, digitaalisia työkaluja painottavan näkökulman aineiston digitoinnille.

Artikkelissaan Prescott ja Hughes argumentoivat, että julkiseen kysyntään ja suurimpaan mahdolliseen digitoidun aineiston käyttäjämäärän nojaavalla digitoimisprosessilla voi olla odottamattomia haittavaikutuksia. Näiden kriteerien suhteen parhaiten onnistunut digitointiprojekti on sellainen, joka digitoi kaikista tehokkaimmin kaikista suosituimman aineiston. Kilpailussaan

rahoituksesta arkistot ja museot kiirehtivät digitoimaan kokoelmiensa tunnetuimmat osat, mikä on johtanut tämän tiedon saatavuuden paranemisen lisäksi kuitenkin myös olemassa olevien taiteen ja kulttuurien kaanonien vahvistumiseen (Prescott & Hughes 2018, 1–2). Etäkäytön sallimisen ja pääsyn lisäämisen suhteen arkistojen digitointiprosessit ovat olleet huomattavan menestyksekkäitä, mutta digitointiprosessi voi tarjota muitakin etuja. Digitaalisella arkistolla tai kirjastolla on potentiaalia tuoda esille unohdettua tai kulttuurillisissa narratiiveissa usein sivutettua informaatiota helposti ja esteettömästi, mutta pelkkiin käyttäjä- ja kysyntäpohjaisiin metriikkoihin nojaava digitoimisprosessi ei pysty toteuttamaan tätä potentiaalia. Vuoteen 2018 mennessä digitoitavan aineiston valintaan ei ollut vieläkään löydetty yleistä standardia, ja tämän seurauksena digitoidun aineiston muodostama kokonaisuus oli hyvin hajanainen: suurten ja tunnettujen teosten lisäksi oli digitoitu myös hyvin spesifisiä kokoelmien osia, kuten British Libraryn kaikki 900 kreikkalaista manuskriptiä (Prescott & Hughes 2018, 4). Päätös näiden manuskriptien digitoinnista voi aluksi vaikuttaa täysin sattumanvaraiselta, mutta tämän digitointiprojektin taustalla oli näistä manuskripteistä kiinnostuneelta säätiöltä tullut suora rahoitus (Prescott & Hughes 2018, 4). Kuten vuonna 2002, rahoitus oli yhäkin valintaprosessin kannalta tärkein tekijä. Näissä edellä mainituissa digitointiprojekteissa ei myöskään otettu tarpeeksi huomioon digitaalisen aineiston tarjoamia edistysaskelia akateemiselle tutkimukselle. Tällainen massadigitointi, jossa kokonaisia kokoelmia digitoidaan kerrallaan mahdollisimman nopean ja tehokkaan digitaalisen näkyvyyden saavuttamiseksi, vaatii rinnalleen harkitun, yhteen aineiston osaan kerrallaan kohdistuvan iteratiivisen digitointiprosessin, jota Prescott ja Hughes nimittävät ”hitaaksi digitoinniksi” (engl. slow digitization). Yksittäinen skannaus saattaa riittää vaikkapa kirjan digitointiin, jos haluttu lopputulos on vain kirjan tekstinmuodossa olevan tietosisällön siirtäminen digitaalisen muotoon, mutta jos tämä kyseinen kirja olisi luonteeltaan historiallisesti merkittävä, tämä ei välttämättä riittäisi historioitsijoille. Syvällinen historiallisesti arvokkaaseen aineistoon kohdistuva tutkimus vaatii aineiston fyysistä olemusta monesta eri näkökulmasta lähestyvää tutkimusmetodia, niin kuvainnollisesti kuin myös kirjaimellisestikin. Tämä voi tarkoittaa esimerkiksi useita eri skannauksia eri valaistustekniikkoja soveltaen (Prescott & Hughes 2018, 7). Hyvä esimerkki digitoinnin mahdollistamasta, aikaisempaa syvemmästä aineistoon kohdistuvasta tutkimuksesta on *”Codex Sinaiticus”*, kreikankielinen Raamatun manuskripti 300-luvulta jKr. Digitaalinen versio manuskriptistä pystyi tuomaan yhteen pakkaukseen eri museoissa ympäri maailmaa säilöttyjä osia itsestänsä, joista monet olivat olleet tutkijoille aikaisemmin saavuttamattomia. Tämän lisäksi manuskriptin sivuista on tarjolla mahdollisia virheitä ja raaputuksia paljastavan valaistuksen alla otettuja skannauksia, ja manuskriptin tekstiin sekä sen käännökseen kumpaankin voi myös suorittaa avainsanahakuja (Prescott & Hughes 2018, 5–6). Esimerkkinä taas laadultaan heikommasta

digitointiprojektista toimikoon toinen historiallisesti merkittävä kristitty manuskripti, tällä kertaa ”*Benedictional of St Æthelwold*” 900-luvulta jKr. Kyseinen manuskripti sisältää useita, alastomalla silmällä todella vaikeasti havaittavissa olevia merkintöjä, joista on tiedetty jo vuodesta 1994 alkaen. Tästä valmiiksi omatusta tietämyksestä huolimatta ainakin vielä vuonna 2018 British Libraryn virallisilla sivuilla esillä ollut versio ei tarjonnut mahdollisuutta nähdä näitä merkintöjä johtuen osittain massadigitointiprojekteissa käytettyjen LED-valojen kirkkaudesta, mutta toisaalta myös tarvittavien, huomattavasti kalliimpien digitointitekniikkojen sovelluksen puutteesta (Prescott & Hughes 2018, 7). Hidas digitointi ei ole siis nimestään huolimatta pelkästään tapa muuntaa fyysistä aineistoa digitaaliseksi, mutta myös itsessään tieteellisen tutkimuksen metodi. Hidas digitointi vaatii huolellista, iteratiivista sekä aineiston historiallisen ja kulttuurillisen kontekstin huomioonottavaa lähestymistapaa samaan tapaan kuin aineiston kohdistuva tutkimus, ja joskus nämä kaksi voivat limittyä. Vuonna 1989 alkaneen *The Canterbury Tales* -projektin perimmäisenä tarkoituksena oli transkriboida kaikki kyseisestä kuuluisasta teoksesta olemassa olevat manuskriptit ja vertailla niitä tietokoneohjelmia käyttäen. Tämä projekti vaati manuskriptien pieniinkin yksityiskohtiin keskittymistä, minkä seurauksena jo ennen itse transkriboimista havaittiin, että monet manuskripteistä olivat itseasiassa vanhempia kuin alun perin oletettiin (Prescott & Hughes 2018, 10–11). Hitaan digitoinnin voi siis ymmärtää prosessina, jossa sitoudutaan pelkän aineiston digitoinnin lisäksi sen läpikotaiseen tutkimiseen. Pelkän pääsyn ja yleisen kysynnän suhteen tällainen digitointi ei varmastikaan ole niin kustannustehokasta kuin massadigitointi, ja hidas digitointi vaatikin perspektiivin muutosta pelkistä kysyntä-tarjonta-pohjaisista malleista aineiston käytön suhteen avarakatseisempaan valintaprosessiin. Tämä tarkoittaa uusien valintakriteerien soveltamista: pelkän potentiaalisen käyttäjämäärän (kysyntä) lisäksi tulee myös pohtia näiden käyttäjien luonnetta sekä heidän käyttäjäkohtaisia tarpeitansa. Esimerkiksi edellä mainitun ”*Codex Sinaiticuksen”* tapauksessa on hyväksyttävää olettaa, että huomattava osa tästä manuskriptistä kiinnostuneista käyttäjistä on tutkijoita. Huomattava osa digitoitavasta aineistosta, etenkin arkistojen puolelta, onkin pääsääntöisesti historioitsijoiden ja muiden tutkijoiden käyttämää. Historioitsijoilla voi olla usein hyvinkin kekseliäitä tapoja käyttää aineistoa, esimerkiksi vaikkapa vihkimätodistusten soveltaminen 1800-luvulla tapahtuneen kaupungistumisprosessin etenemisen kartoittamiseksi (Maxwell 2010, 25). Tällaisia aineiston käyttötapoja on vaikea ennustaa, eli käyttäjälähtöinen valintaprosessi historialliseen aineistoon sovellettuna vaatii siis välttämättäkin historioitsijoiden tuomista osaksi valinta- ja digitoimisprosessia (Prescott & Hughes 2018, 9). Tämä vaatii myös aineiston luonteen huomioonottamista. Edellä mainittujen manuskriptien kohdalla vaadittiin hidasta digitointiprosessia, mutta tämä ei pidä paikkaansa kaikkien historian tutkimuksen suhteen tärkeiden aineistokappaleiden kohdalla. Esimerkiksi sanomalehtiarkistojen tapauksessa

digitoitavan aineiston mahdollisimman suureen määrään tähtäävä massadigitointi vastaa historioitsijoiden tällaiseen aineistoon kohdistuvia tarpeita riittävän hyvin (Maxwell 2010, 25).

Digitoitavan aineiston valintaprosessi on ollut siis ongelmallinen läpi digitointiprojektien historian ja se sisältääkin useita eri harkintoja. Säilönnän suhteen digitointi tarjoaa paljon kieltämättömiä etuja, kuten vaikkapa alkuperäiskappaleen käytön vähentämisen (Astle & Muir 2002, 69), joten säilyvyyttä painottavista valintaprosesseista ollaan yleensä yhtä mieltä. Jos päämääränä on taas pääsyn maksimointi niin suurelle yleisölle kuin mahdollista, taloudellisesti järkevintä on soveltaa nopeita ja tehokkaita digitointimenetelmiä ja painottaa aineiston tunnetuimpia osia. Tämä lähestymistapa sisältää kuitenkin ristiriitoja: vaikka 1000-luvulla kirjoitetun eepoksen ”*Beowulfin”* alkuperäinen käsikirjoitus on todella kuuluisa, kuinka suuren määrän käyttäjiä tämä muinaisenglanniksi kirjoitetun manuskriptin digitaalinen jäljennös todellisuudessa houkuttelisi? Valintaprosessi selkenee, kun keskitytään syvemmin aineistoon itseensä sekä sen potentiaalisiin käyttäjiin. Tieteellisen tutkimuksen näkökulmasta digitoinnin voi nähdä osana tieteellistä prosessia (Prescott & Hughes 2018, 9), jolloin tutkimusta itseänsä koskevat harkinnat on otettava osaksi digitointiprosessia. Jos aineistolla on historiallista arvoa, kuten vaikkapa edellä mainitulla *Beowulfin* alkuperäiskäsikirjoituksella hyvin selkeästi on, pelkkä pääsyyn ja pinnalliseen etäkäyttöön perustuva digitointiprosessi tekee karhunpalveluksen niin aineistolle itsellensä kuin myös sen potentiaaliselle käyttäjäkunnalle. Toisaalta 1960-luvulla syntyneiden organisaationallisten paperidokumenttien digitointi tuskin vaatii useaa eri skannausta eri valaisumetodeja käyttäen. Kaikkia digitointiprojekteja yhdistää niiden riippuvaisuus rahoituksesta, ja tämä tosiseikka tulee aina rajaamaan näiden projektien mahdollista skaalaa. Järkevintä on siis valita digitoitava aineisto potentiaalisen käyttäjäkunnan tarpeet sekä aineiston luonteen huomioiden. Toisaalta puhtaasti taloudelliseen tehokkuuteen joko pääsyn tai akateemisen hyödyn suhteen nojaavat valintamallit voivat helposti sivuttaa vähemmän tunnetun tai unohdetun aineiston digitoinnin. Tällainen aineisto on historiallisen tietämyksen suhteen äärettömän arvokasta, sillä mitä enemmän tällaista negligoitua aineistoa tuodaan helposti saataville, sitä tarkemman ja rikkaamman kuvan menneisyydestä on mahdollista saada.

4.2 Digitointiakti

Digitoitavan aineiston valinnan ja valmistelun jälkeen alkaa itse digitointiakti. Digitointiaktissa fyysisestä objektista luodaan digitaalinen jäljennös, ja tämä voi tarkoittaa yksinkertaisimmillaan digitaalisen valokuvan ottamista objektista. Useimmiten digitointiakti on kuitenkin kohtalaisen monimutkainen ja vaatiikin useiden eri teknologien yhteistoimintaa. Kuten edellä kävi ilmi,

esimerkiksi digitointiaktissa käytetty valaistus saattaa johtaa tiettyjen alkuperäisen aineiston ominaisuuksien vahingoittumiseen tai katoamiseen sen digitaalisesta jäljennöksestä. Tällaisessa tapauksessa digitaalisen aineiston integriteetin voi nähdä vahingoittuneen. Tämä aineiston digitaalisen jäljennöksen integriteetin kärsiminen on yksi huomattavammista itse digitointiaktiin liittyvistä ongelmista, sillä se on kohtalaisen yleistä etenkin tekstintunnistusteknologian tapauksessa ja se vaikeuttaa aineiston luettavuuden lisäksi myös sen jälkikäyttöä: esimerkiksi virheellisesti digitoidussa dokumentissa voi olla virheitä sen tekstimuotoisessa tietosisällössä, jolloin siihen tehdyt avainsanahaut eivät välttämättä tuota luotettavia lopputuloksia (Järvelin et al. 2016, 2). Tässä tutkimuksessa keskityn erityisesti OCR:ään eli tekstintunnistusteknologiaan ja siihen liittyviin ongelmiin, sillä ne valaisevat todella tehokkaasti digitointiaktille ominaisia ongelmia juuri aineiston integriteetin näkökulmasta. OCR-teknologiaa (Optical Character Recognition) käytetään nimensä mukaisesti kirjoitettua kieltä sisältävän aineiston digitoimiseksi. OCR:n avulla voidaan tunnistaa merkkejä sekä merkkijonoja dokumentista, ja täten luoda aineiston tietosisällöstä digitaalinen jäljennös (Choros & Jarosz 2018, 171). Ideaalissa tapauksessa OCR:n avulla suoritettu digitointiakti johtaa sellaiseen digitaaliseen dokumenttiin, jonka tekstisisältö on täysin digitaalisten työkalujen saavutettavissa ilman ihmisen puuttumista prosessiin. Kuten tämän tutkimuksen aineistosta käy kuitenkin ilmi, tämä ei ole aina mahdollista, ja etenkin massadigitoinnin yhteydessä monet OCR:stä johtuvat virheet voivat päätyä lopulliseen digitaaliseen aineistoon. Melkein kaikki tämän tutkimuksen OCR:ää käsittelevä aineisto keskittyi historiallisiin sanomalehtiarkistoihin ainakin osittain, ja vanha, tekstipohjainen aineisto tarjoaakin oivallisen katsauksen OCR:n vahvuuksiin ja heikkouksiin: tapa, jolla historioitsijat pääsäännöllisesti käyttävät historiallisia sanomalehtiaineistoja ei vaadi yksityiskohtaista digitointia (Maxwell 2010, 25), jolloin OCR:n avulla toteutettu digitointi on kannattavaa. Toisaalta historiallinen, tekstipohjainen aineisto tarjoaa paljon haasteita OCR-teknologialle ja täten demonstroi automaatiopohjaiseen digitointiin liittyviä ongelmia. Historialliset sanomalehtiarkistot ovat myös suosittuja digitointikohteita digitoinnin suhteellisen vaivattomuuden ja aineiston historiallisen hyödyn johdosta, ja esimerkiksi Kansalliskirjaston lehtiarkistossa vuosien 1771–1910 väliltä on yhteensä noin 3 miljoonaa digitaalista sivua (Kettunen, Pääkkönen & Koistinen 2016, 3).

Ensimmäinen haaste, jonka tällainen aineisto esittää tekstintunnistusteknologialle on hyvin fundamentaalinen melkein kaikelle ihmisten kieltä käsittelevälle teknologialle: luonnollinen kieli ja sen näennäinen epäloogisuus. OCR:llä voi olla paljonkin hankaluuksia luonnollisen kielen kanssa, etenkin jos kieli sisältää paljon epäjohdonmukaisia taivutusmuotoja. Johtuen esimerkiksi suomen kielen morfologisesta moninaisuudesta, OCR:llä sekä automaattiseen indeksointiin käytetyllä

teknologialla voi olla huomattaviakin ongelmia suomenkielisen aineiston kanssa (Järvelin et al. 2016, 2930). Jotta saman sanan toisistaan mahdollisesti hyvinkin erilaiset taivutusmuodot voitaisiin yhdistää toisiinsa, tarvitaan erinäisiä keinoja, joilla tunnistaa kaikkien taivutusmuotojen taustalla oleva alkuperäinen sana. Tekniikat kuten stemmaus ja lemmatisointi voivat auttaa kaikkien tekstissä esiintyvien termin taivutusmuotojen yhdistämisen yhteiseen avainsanaan indeksointia varten, ja näitä sovelletaankin digitointiprosessin kaikissa vaiheissa, mutta nämä eivät ratkaise kaikkia historiallisiin aineistoihin liittyviä ongelmia (Järvelin et al. 2016, 2932). Kazimierz Chorosin ja Joanna Jaroszin vuoden 2018 artikkeli ”*Most Frequent Errors in Digitization of Polish Ancient Manuscripts”* käsittelee syvemmin historiallisesta aineistosta itsestänsä kumpuavia ongelmia muinaisten puolalaisten manuskriptien kautta. Pelkästään nykyisen luonnollisen kielen aiheuttamien ongelmien lisäksi historiallisia lähteitä digitoidessa pitää ottaa huomioon myös historiallisen kielen mukanaan tuomat ongelmat, kuten muutokset sanastossa, muutokset taivutuksessa ja mahdolliset epäjohdonmukaisuudet taivutuksessa (Järvelin et al. 2016, 2933). Esimerkiksi termi ”diakonissalaitos” voidaan olla kirjoitettu historiallisissa lähteissä monella eri tavalla, kuten ”diakonissa-laitos” tai ”diakonialaitos” (Järvelin et al. 2016, 2941). Räikeämpänä esimerkkinä historiallisen kielen mahdollisista eroista nykyisen kielen kanssa voi toimia puolan kieli: monet historiallisissa manuskripteissä esiintyvät termit eivät välttämättä ole modernissa puolan kielessä laisinkaan, ja jossain tapauksissa kirjain ’s’ voidaan olla kirjoitettu muotoon, joka muistuttaa lähemmin kirjainta ’f’ tai ’l’ (Choros & Jarosz 2018, 174–175). Etenkin tämä jälkimmäinen vaihtelu yksittäisen kirjaimen kirjoitusasussa esittää selkeitä ongelmia puhtaaseen hahmontunnistukseen pohjautuvalle OCR-teknologialle. Näiden kielellisten seikkojen lisäksi aineiston itsensä fyysiset ominaisuudet voivat tuottaa ongelmia. Aineiston mahdollinen heikentynyt kunto voi tehdä tekstistä epäselvää, joka johtaa heikompaan OCR:n tarkkuuteen. OCR:n tarkkuudella tarkoitetaan sen tuottaman tuloksen virheettömyyttä eli digitaalisen jäljennöksen vastaavuutta alkuperäiskappaleseen, ja tätä tarkkuutta mitataan yksinkertaisesti vertailemalla digitaalista jäljennöstä ja alkuperäistä aineistoa toisiinsa (Choros & Jarosz 2018, 172). Toinen tarkkuuteen vaikuttava tekijä voi olla dokumentin mahdollinen historiallinen käyttötarkoitus: palimpsesti on manuskriptin sivu, jonka alkuperäinen teksti on pesty tai raaputettu pois ja korvattu uudella tekstillä (Choros & Jarosz 2018, 173). Palimpsesteissa voi usein näkyä vieläkin kohtalaisen selkeästi vanhemman tekstisisällön jäänteitä, ja nämä voivat ”vuotaa” niiden päälle kirjoitetun tekstin digitointiaktin lopputulokseen. Yksi tapa ehkäistä mahdollisen taustatekstin sekä dokumentin heikon kunnon aiheuttamia ongelmia on dokumentista otetun skannauksen käsittely ennen OCR-lukua. Pelkästään skannauksen muuntaminen värillisestä mustavalkoiseksi ja tämän jälkeen

kirkkauden nostaminen voivat jo vähentää potentiaalisten virheiden määrä huomattavasti (Choros & Jarosz 2018, 177).

Näistä edellä mainituista seikoista johtuen OCR-pohjaisessa digitointiaktissa voi tapahtua virheitä tekstintunnistuksessa. Jos näitä virheitä ei korjata manuaalisesti ja niiden annetaan jäädä aineistoon, digitointiprosessin voidaan nähdä teoreettisella tasolla epäonnistuneen, sillä digitaalinen jäljennös on tietosisällöltään vähäisempi kuin alkuperäinen. Todellisuudessa kuitenkin tällaiset virheet ovat yleisiä: esimerkiksi edellä mainitun Kansalliskirjaston vuosien 1771–1910 välinen digitaalinen sanomalehtiarkisto sisältää 70–75 % sanatarkkuuden, tarkoittaen, että noin neljäsosa OCR:n avulla digitoiduista sanoista ovat jollakin tavalla virheellisiä (Kettunen et al. 2016, 1). Digitointiaktin aikana tapahtuvaan integriteetin menetykseen kannattaakin siis ottaa käytännönläheisempi perspektiivi ja pohtia integriteettivahingon aiheuttamaa haittaa aineiston kahdelle pääkäyttäjäkunnalle, aineisto käyttävälle ihmiselle ja aineistoa käsitteleville tietojärjestelmille. Kimmo Kettusen, Tuula Pääkkösen ja Mika Koistisen vuoden 2016 artikkelissa ”*Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen*” tulee esille historioitsijoiden näkökulma OCR:n tuottamiin virheisiin. OCR:n tuottamat virheet kyllä hämmentävät tutkijoita, mutta nämä virheet eivät tee aineistoista kuitenkaan käyttökelvotonta. Todellinen OCR:n virheiden tuottama ongelma on niiden vaikutus tiedonhakuun. Virheellinen tekstisisältö voi johtaa halutun aineiston täydelliseen puuttumiseen tai poikkeuksellisen alhaiseen sijoitukseen tiedonhakujärjestelmän hakutuloksissa. Tämän lisäksi OCR-virheet aiheuttavat ongelmia kieliteknologian ja tekstin louhinnan sovelluksille, kuten nimellisiä entiteettejä tekstistä tunnistavalla NER-ohjelmistolle (Kettunen, Pääkkönen, Koistinen 2016, 5). NER-ohjelmisto etsii sille annetusta tekstistä automaattisesti entiteettejä kuten henkilöiden, organisaatioiden ja paikkojen nimiä indeksointia varten (Kettunen et al. 2016, 2). NER-ohjelmiston suoriutumisen ja OCR-virheiden taajuuden välillä on todettu ainakin jonkin tasoinen käänteinen korrelaatio: jos alun perin noin 73 % OCR-tarkkuuden omaava aineisto joko korjataan manuaalisesti tai oikoluetaan tarkempaa OCR-ohjelmistoa soveltaen, NER-ohjelmiston mahdollisuus löytää entiteettejä aineistosta kasvoi 10–20 % (Kettunen et al. 2016, 16–17). Johtuen kuitenkin olemassa olevan aineiston koosta ja virheiden määrästä, näiden edellä mainittujen askelien ottaminen kokonaisen kokoelman kohdalla ei ole mahdollista. Tätä tarkoitusta varten on kuitenkin kehitetty OCR-virheitä korjaavia algoritmeja, jotka teoriassa ainakin voivat korjata OCR:n tuottamia virheitä automaattisesti (Kettunen et al. 2016, 18).

Suurin digitointiaktiin liittyvä ongelma on perustavanlaatuinen kaikille digitointiprojekteille: digitointiprosessin lopputulos voi olla laadultaan heikkoa. Etenkin OCR:llä suoritetut, isoskaalaiset

digitointiprojektit päätyvät sisältämään huomattavan määrän virheitä (Kettunen, Pääkkönen, Koistinen 2016, 4). Nämä virheet ovat osittain prosessissa käytetyn teknologian aiheuttamia, mutta laajalti nämä ongelmat kumpuavat itse aineistosta ja sen tietosisällöstä. Ajan kulun vaikutus itse dokumentteihin kuin myös niiden sisältämään kieleen vaikeuttaa tekstinlukuohjelmiston toimintaa huomattavasti, ja lopputuloksena on usein virheellinen digitaalinen jäljennös. OCR-teknologian tarkkuus kehittyy jatkuvasti, mutta tämä ei ratkaise jo olemassa olevan digitoidun aineiston heikon integriteetin aiheuttamia ongelmia. Digitointiaktin aikana syntyneet virheet voidaan korjata manuaalisesti, mutta laajempien kokoelmien kohdalla tämä ei yksinkertaisesti ole mahdollista. Yksi vaihtoehto on suorittaa uusi optinen luku aineistolle tarkempaa OCR-ohjelmistoa soveltaen, mutta suurien kokoelmien kohdalla tämä on silti liian työlästä. Toinen vaihtoehto on suorittaa ohjelmallinen jälkikorjaus erinäisiä korjausalgoritmeja soveltaen, mutta näiden algoritmien tuottamat lopputulokset olivat ainakin vielä vuonna 2016 riittämättömiä. Digitointiaktia on siis lähestyttävä pragmaattisesti: prosessi tuottaa aina ainakin hieman virheitä, ja kuten aikaisemmin jo todettiin, aineistokohtaisia valintoja hyväksyttävän laadun sekä digitointiaktiin sovellettavien tekniikkojen kohdalla on tehtävä.

4.3 Digitoinnin jälkeinen vaihe

Aineiston digitointiprosessin viimeinen vaihe on myös vaiheista pisin, sillä etenkin historiallisesti merkittävän aineiston kohdalla sen säilyvyydestä tulisi huolehtia vuosisatoja, jos ei vuosituhansia. Digitaalisen aineiston säilömistä tarkasteltaessa on hyvä ottaa huomioon sen uniikki, digitointiaktin johdosta syntynyt olemus. Digitoinnin seurauksena aineiston voi nähdä kokevan eräänlaisen muodonmuutoksen: digitointiprosessin johdosta aikaisemmin täysin fyysinen arkistoitu aineisto näennäisesti monistuu, jättäen jälkeensä fyysisen alkuperäiskappaleen, jonka rinnalla on nyt syntynyt uusi, digitaalinen objekti (Astle & Muir, 2002, 67). Tämä digitaalinen aineisto omaa niin fyysisen, aineellisen ulottuvuuden sitä säilövän tallennusvälineen muodossa kuin myös virtuaalisen, aineettoman ulottuvuuden tallennusvälineiden ja tietojärjestelmien sisällä. Nämä ulottuvuudet eivät ole kuitenkaan toisistaan täysin erillisiä, ja niiden välillä onkin vahva yhteys: jos aineiston aineellinen ulottuvuus, esimerkiksi kovalevy, vahingoittuu kriittisesti, sen sisällään pitämä aineeton ulottuvuus voi myös vahingoittua, mikä voi johtaa aineiston pysyvään menettämiseen. Yhtä lailla, jos aineiston aineetonta ulottuvuutta muokataan tietojärjestelmän kautta ja nämä muutokset tallennetaan ilman varmuuskopiointia, aineiston aineellisen ulottuvuuden voi nähdä myös muuttuvan: kovalevyllä ei enää ole sinne alun perin talletettua versiota aineistosta, vaan sen sijaan uusi, muokattu versio (Baker et al. 2006, 2–3). Digitaalisen aineiston säilömiseen kohdistuvia

ongelmia voikin tarkastella näiden kahden ulottuvuuden sekä niiden välisen yhteyden näkökulmasta, mikä auttaa hahmottamaan digitaalisen säilömiseen liittyviä laajempia ongelmia.

Kuten edellä mainittiin, digitoidun aineiston aineellinen ulottuvuus on aina jonkinlaisen digitaalista informaatiota säilövän talletusvälineen muodossa. Loppujen lopuksi kaikki digitaalinen informaatio on säilöttynä jollekin yksittäiselle kovalevylle jossakin päin maailmaa, ja tähän kovalevyyn kohdistuu kaikki samat fyysisen maailman voimat kuin alkuperäiskappaleeseenkin. Bakerin ja kumppanien vuoden 2006 artikkeli ”*A fresh look at the reliability of long-term digital storage”* keskittyy digitaalisen, luonteeltaan funktionaalisen datan säilöntään ja tätä kautta valaisee useita fyysisen sekä virtuaalisen maailman asettamia rajoitteita digitaalisille arkistointiprojekteille. Huomattava osa historioitsijoita kiinnostavasta aineistoista on ollut alun perin luonteeltaan funktionaalista, eli sillä on ollut selkeä käyttötarkoitus yhteiskunnallisten tai organisaationallisten prosessien toteuttamisessa. Kuten edellä mainitut 1800-luvun vihkimätodistukset demonstroivat, arkisiin prosesseihin liittyvät dokumentit ovat tärkeitä historiantutkimuksen kannalta ja tämän johdosta yhteiskunnallisen ja organisaationallisten prosessien kautta syntyneitä dokumentteja tulisikin säilöä pelkän organisaationallisen funktion toimimisen takaamisen lisäksi myös historiallisista syistä. Johtuen kuitenkin digitaalisten prosessien, kuten vaikkapa sähköpostin, johdosta syntyvän datan määrästä, tällainen teoriassa ääretön säilöminen voi osoittautua hyvinkin kalliiksi. Pelkän energiakulutuksen lisäksi kovalevyihin ja niiden säilömän datan integriteettiin kohdistuu paljon välittömiä sekä teoreettisia uhkia, joihin pitää valmistautua onnistuneen säilömisen saavuttamiseksi. Kovalevyyn kohdistuva vahinko voi vahingoittaa sen säilömää dataa, jonka seurauksena aineiston integriteetti voi kärsiä pysyvästi. Kovalevyt ovat aina osa suurempaa järjestelmää, kuten tietokonetta tai monta tietokonetta kattavaa tietokonefarmia, ja kovalevyn sisältävän järjestelmän vaurioituneesta komponentista syntyvä ongelma voi levitä ja johtaa koko järjestelmää vahingoittavaan virhetilaan. Jossain tapauksissa itse kovalevy voi vaurioittaa sen sisältämää dataa, johtuen joko sen iästä tai muuten vain virheellisestä toiminnasta (Baker et al. 2006, 2–5). Näiden ongelmien torjuminen vaati jatkuvaa järjestelmän laitteiston hoitamista sekä mahdollista korvaamista, ja kovalevyjen tapauksessa myös niiden sisältävän datan siirtämistä uudelle tallennusvälineelle. Toinen datan migraatioon kannustava tekijä on myös teknologinen kehitys ja sen mukanaan tuoma teknologinen vanhentuminen: jossain vaiheessa vanhan laitteiston ja ohjelmiston säilöminen ei ole enää kannattavaa, ja data on siirrettävä uuteen digitaaliseen ympäristöön (Astle & Muir 2002, 69). Tässä prosessissa voidaan kuitenkin törmätä teknologisesta vanhenemisesta kumpuavaan ongelmaan: vanha ja uusi ohjelmista tai laitteisto eivät välttämättä pystykään enää kommunikoimaan keskenään (Baker et al. 2006, 2–3). Periodillinen datan

siirtäminen uudelle tallennusvälineelle sekä komponenttien päivittäminen ehkäisevät teknologisesta vanhenemisesta johtuvia ongelmia, mutta vaativat myös paljon aikaa sekä rahaa (Astle & Muir 2002, 69). Säilömisessä on myös varauduttava epätodennäköisempiin ja laajaskaalaisempiin uhkiin, kuten luonnonkatastrofeihin, sotatilanteisiin tai terrori-iskuihin (Baker et al. 2006, 2). Kaikki näistä voivat johtaa arkistoidun aineiston täydelliseen tuhoutumiseen, ellei kyseistä aineistoa ole varmuuskopioitu useaan maantieteellisesti eriävään sijaintiin.

Siinä missä aineiston aineellinen ulottuvuus on olemassa fyysisen, laitteistosta koostuvan järjestelmän sisällä, aineiston aineeton ulottuvuuden voi nähdä olevan olemassa virtuaalisen, ohjelmistosta koostuvan järjestelmän sisällä. Huolimatta digitaalisen aineiston ulottuvuuksien välisestä vahvasta sidoksesta sekä aineellisen ulottuvuuden perustavanlaatuisuudesta, erottelu sen ja aineettoman ulottuvuuden välillä on merkityksellinen eksklusiivisesti aineiston aineettomaan ulottuvuuteen kohdistuvien uhkien johdosta. Ulottuvuuksien välinen yhteys menee kumpaankin suuntaan, ja muutokset aineettomaan ulottuvuuteen muuttavat pysyvästi myös sen aineellista ulottuvuutta. Koska arkistoinnissa on loppujen lopuksi kyse aineiston eheydestä, tällaiset muutokset voivat vahingoittaa aineiston integriteettiä ja täten vaarantaa koko arkistointiprojektin. Digitaalisten prosessien johdosta syntyneiden dokumenttien muokkaaminen käyttäen digitaalisia työkaluja on suhteellisen helppoa, ja muutosten tekeminen historiallisten aineistokappaleiden digitaalisiin jäljennöksiin on huomattavasti helpompaa kuin niiden alkuperäiskappaleisiin. Yksi motiivi tällaiselle tahalliselle aineisto integriteetin vahingoittamiselle voi olla poliittinen sensuuri: esimerkiksi kommunistisen Tšekkoslovakian hallitus sensuroi Karl Marxin kirjoituksista tšekkiläisiä koskevia kommentteja (Maxwell 2010, 26). Digitaalista aineistoa voidaan myös poistaa samaan tapaan kuin fyysistä aineistoa voidaan tuhota, ja syinä tällaiselle toiminalle voi olla vaikkapa organisaation laittoman toiminnan peittely (Baker et al. 2006, 3). Johtuen kuitenkin tietojärjestelmien monimutkaisuudesta sekä aineiston silkasta määrästä, nämä edellä mainitut uhat aineiston integriteetille voivat useimmiten manifestoitua tahattomasti. Käyttäjä voi esimerkiksi vahingossa päällekirjoittaa tai poistaa osia aineistoista ilman, että virhettä huomataan, etenkin jos aineisto on kooltaan iso (Baker et al. 2006, 4). Toinen aineiston aineetonta ulottuvuutta uhkaava ongelma on sitä käsittelevä ja organsioiva ohjelmisto. Edellä mainittu järjestelmän fyysisistä komponenteista johtuva virhetila voi johtaa ohjelmiston virheelliseen toimintaan, joka taas puolestaan voi johtaa aineiston korruptoitumiseen. Myös käyttäjän tekemät virheet voivat johtaa ohjelmistosta johtuvaan aineiston integriteetin kärsimiseen tai ainakin aineiston väliaikaiseen menettämiseen: esimerkiksi ohjelmiston toiminnalta elintärkeiden ajurien poistaminen voi vahingoittaa aineistoa tai tehdä siitä väliaikaisesti saavuttamatonta, jolla voi olla

organisaationallisen toiminnan kannalta haitallisia seurauksia (Baker et al. 2006, 2). Kuten aineistoa säilövän järjestelmän laitteisto, myös sen ohjelmisto voi vanhentua, jonka seurauksena pahimmassa tapauksessa digitaalinen aineisto voi jäädä jumiin vanhentuneeseen tiedostoformaattiin: esimerkiksi digitaalisilla kamerayhtiöillä voi olla omia, heikosti dokumentoituja formaatteja raa'an datan tallentamista varten, ja näiden yritysten kaatuessa tai teknologian kehittyessä tästä datasta voi tulla täysin lukukelvotonta vaihtoehtoisille ohjelmistoille (Baker et al. 2006, 3).

Yksi tärkeä perspektiivi digitoinnin jälkeisen vaiheen ongelmiin on organisaationallinen: digitaalisen aineiston niin aineellisen kuin myös aineettoman ulottuvuuden ongelmat toteutuvat organisaationallisen kontekstin sisällä. Tämä organisaatio voi olla vaikkapa kirjaston IT-osasto, jonka vastuulla on digitaalisen arkiston ylläpitäminen (Astle & Muir 2002, 76), tai arkistopalveluita tarjoava yksityisen sektorin organisaatio (Svärd 2017, 276). Esimerkiksi organisaation syystä tai toisesta johtuva luhistuminen voi johtaa arkistoidun aineiston menettämiseen tai sen unohtumiseen, jolloin ajan myötä johtuen joko laitteiston tai ohjelmiston vanhenemisesta aineistosta tulee saavuttamatonta (Baker et al. 2006, 3). Tutkimusaineistosta käy selkeästi ilmi, että kaikille organisaationallisille konteksteille on yhteistä taloudellisten tekijöiden asettamat rajoitteet. Jotkin näistä rajoitteista, kuten pitkäntähtäimen digitaaliseen säilöntään liittyvät kulut, ovat universaaleja, kun taas toiset ovat organisaatiokohtaisia. Svärdin vuoden 2017 artikkeli "*The woes of Swedish private archival institutions"* käsittelee nimensä mukaisesti yksityisiä arkistoinstituutioita ja juurikin niiden organisaationallisen luonteen johdosta syntyneitä ongelmia. Näiden arkistoinstituutioiden etu suhteessa julkisen sektorin arkistoihin on niiden löyhät standardit niille säilöttäväksi luovutettavan aineiston suhteen (Svärd 2017, 277). Toisin kuin julkisen sektorin arkistot, nämä yksityiset organisaatiot rahoittavat toimintaansa puhtaasti niiden tarjoamien palveluiden tuottamista tuloista ja mahdollisten standardien asettaminen luovutettavalle aineistolla saattaisi ajaa asiakkaita pois. Tällaisen arkistoinnin ongelmat ovat kuitenkin ilmeisiä: ilman yhteistä standardia aineiston säilömisestä, huoltamisesta ja hallinnoimisesta tulee haastavaa ja osa aineistoista voi helposti juuttua vanhentuneeseen formaattiin (Svärd 2017, 275–276). Toistuva teema aineistossa on suunnitelmallisuuden tärkeys ja toisaalta sen yleinen puute arkistointiprojekteista. Organisaationallisella tasolla tämä tarkoittaa kaikkiin tässä tutkimuksessa edellä mainittuihin ongelmiin varautumista niin taloudellisesti kuin myös ajallisestikin, mutta myös exit-strategian formuloimista aineistolle. Etenkin yksityisen sektorin organisaatioiden tapauksessa tämä tarkoittaa aineiston jatkuvuuden takaamista siinä tapauksessa, että organisaatio enää kykene ylläpitämään arkistoansa ja täten takaamaan aineiston integriteetin eheyttä (Baker et al. 2006, 3).

Digitaalisen aineiston elämänkaaren näkökulmasta digitointia edeltävä vaihe sekä itse digitointiakti muodostavat potentiaalisesti vain sadasosan siitä. Digitoinnin jälkeinen vaihe, eli aineiston säilöntä ja ylläpitäminen, kestää ainakin teoreettisella tasolla äärettömästi. Siinä missä mikrofilmi voi potentiaalisesti säilyä jopa 100 vuotta eheänä, digitaalinen tallennusväline vaati paljon tiuhempaa korvaamista (Astle & Muir 2002, 67). Vuosittaiseen laitteiston ja ohjelmiston ylläpitoon sekä järjestelmän eri osien mahdolliseen uusimiseen ja tästä johtuvaan aineiston migraatioon menevien kulujen lisäksi organisaatioiden pitää myös varautua aineiston integriteettiä potentiaalisesti vaurioittaviin skenaarioihin. Yksi ongelma liittyen näiden skenaarioiden ehkäisemiseen on aineiston potentiaalisesti valtava koko: käyttäjän tai kovalevyn aiheuttama korruptio aineistossa voi jäädä huomaamatta hyvinkin pitkäksi aikaa, etenkin jos korruptoitunutta aineistokappaletta ei käytetä säännöllisesti (Baker et al. 2006, 4). Myös tahalliset digitaalisen aineiston integriteettiä vahingoittavat toimet kuten valtiollinen sensuuri voivat tapahtua ilman, että kukaan huomaa tai sitten niin, että organisaatio on voimaton kamppailemaan niistä vastaan. Toisaalta valtavan aineiston säännöllinen seulonta voi olla taloudellisesti ja ajankäytöllisesti mahdotonta, ja integriteettivirheiden estäminen vaatiikin alituisten varmuuskopioiden tekemistä siltä varalta, että tällainen virhe havaitaan. Aineiston digitoinnissa ja digitaalisen aineiston säilömisessä on siis todellisuudessa kyse sitoutumisesta aikaa ja rahaa vaativaan jatkuvaan huoltamisen prosessiin, ja tämä on alituisesti ristiriidassa tätä prosessia suorittavien organisaatioiden taloudellisten realiteettien kanssa. Tämä prosessi on kuitenkin välttämätön, sillä jopa näennäisesti tavanomaisen digitaalisen objektin tarjoama historiallinen arvo voi olla suunnaton, ja digitaaliset työkalut tarjoavat ennennäkemättömän mahdollisuuden välittää tietoa maailmastamme jälkipolville.

5 KESKUSTELU JA JOHTOPÄÄTÖKSET

Tässä tutkimuksessa pyrittiin tunnistamaan digitaaliseen arkistointiin sekä arkistoidun materiaalin digitointiin liittyneitä ongelmia sekä tuomaan esille niihin löydettyjä ratkaisuja. Johtuen digitaalisen arkistoinnin suhteellisen lyhyestä historiasta, suurin osa aineiston kronologisesti varhaisimmissa lähteissä esille tulleista ongelmista oli havaittavissa vielä myöhemmissäkin lähteissä, kuten digitoitavan aineiston valinta ja sen vaikutus digitaalisen aineiston koostumukseen sekä käytettävyyteen. Osa tutkimusaineistossa esiin tulleista digitointiprosessiin liittyvistä ongelmista olivat luonteeltaan puhtaasti teknologisia, kuten OCR-teknologian epätarkkuus ja digitaalisen

talletusteknologian lyhytikäisyys. Etenkin OCR-teknologian osalla tehdään kuitenkin alituisesti harppauksia eteenpäin, ja on hyvinkin mahdollista, että OCR-pohjainen digitointiakti on jossain vaiheessa niin tarkka, että manuaalista, ihmisen tekemää korjausoperaatiota tai korjausalgoritmien soveltamista ei enää välttämättä tarvita. Huomattava osa esille tulleista ongelmista kumpusi kuitenkin digitoitavan aineiston ja digitointiprosessin eri vaiheissa sovellettavan teknologian ulkopuolelta. Taloudelliset sekä organisaationalliset tekijät tulivat aineistossa kerta toisensa jälkeen esiin, ja rahoituksen rajallisuus sekä riittämättömät organisaationalliset käytännöt digitaalisen aineiston integriteetin takaamiseen läpi sen elinkaaren vaikuttavat olevan huomattavimpia digitointiprosessia haittaavia tekijöitä. Digitointia edeltävää vaihetta koskevan tutkimusaineiston kohdalla heräsi myös kysymys digitointiprosessin lähtökohdista: mikä osa aineistosta digitoidaan ja miten sekä ketä varten? Tämä kysymys kulkee käsi kädessä digitointiprosessin suunnitelmallisuuden kanssa, ja onnistunut digitointiprosessi vaatiikin siihen ryhtyvältä organisaatiolta huomattavaa prosessia edeltävää harkintaa ja suunnittelua sekä sitoutumista pitkän tähtäimen säilöntään niin rahallisesti kuin myös ajallisestikin. Ratkaisut näihin ongelmiin löytyivät siis organisaationallisten käytäntöjen kehittämisestä sekä mahdollisesti digitointiprosesseihin allokoidun rahoituksen kasvattamisesta. Potentiaalisen jatkotutkimuksen näkökulmasta olisi mielekästä tutkia mahdollista korrelaatiota organisaationallisten käytäntöjen kehittyneisyyden ja projekteihin allokoitujen taloudellisten resurssien sekä aineiston integriteetin välillä, joskin tällainen tutkimus vaatisi useiden arkisto-organisaatioiden vertailemista toisiinsa tuottaakseen aidosti hyödyllisiä tuloksia ja täten olisi skaalaltaan hyvin laaja.

Tämän tutkimuksen suhteen ongelmallista oli kandidaatintutkielman lopulta rajattu laajuus ja toisaalta tutkimusaiheen puutteellinen rajaus. Vaikka tutkimus onkin luonteeltaan yleiskatsauksellinen, aineiston sirpaleisuus yhdessä sen suppeuden kanssa johti loppujen lopuksi hieman pinnalliseen katsaukseen kaikkiin aineiston eri osa-alueisiin, ja etenkin digitointiaktin yhteydessä olisi ollut mielekästä käsitellä muitakin digitoinnin metodeja. Vaikka aineiston sirpaleisuus siis mahdollisti useamman eri arkistokontekstin käsittelemistä, se myös johti näiden eri kontekstien lievään trivialisointiin: kaikki erilaiset arkistokontekstit sisältävät laajan joukon ongelmia, osa jaettuja ja osa taas kontekstispesifisiä, ja tämän tutkimuksen laajan aiheellisen skaalan seurauksena tutkimuksen lopputulokset eivät täysin heijastele tätä. Tämä on osittain tutkimuksessa käytettyjen tiedonhakumetodien syytä, ja tiukempi metodisuus aineistoa kartuttaessa olisi voinut hyödyttää tutkimusta. Tutkimus olisi voinut hyötyä myös keskittymisestä joko yhteen arkisto- tai aineistokontekstiin, tai sitten yksittäiseen digitointiprosessin vaiheeseen, jolloin

tutkimuksen löydökset olisivat voineet tarjota yksityiskohtaisempia ratkaisuja pelkästään yleisten taloudellisten ja organisaationallisten sijasta.

6 LÄHDELUETTELO

Astle, P., & Muir, A. (2002). *Digitization and preservation in public libraries and archives.* Journal of Librarianship and Information Science, 34(2), 67–79.

Arthur, K., Byrne, S., Long, E., Montori, C. Q., & Nadler, J. (2004). *Recognizing digitization as a preservation reformatting method.*

Baker, M., Shah, M., Rosenthal, D. S., Roussopoulos, M., Maniatis, P., Giuli, T. J., & Bungale, P. (2006). *A fresh look at the reliability of long-term digital storage.* In Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006 (pp. 221-234).

Choroś, K., & Jarosz, J. (2018). *Most Frequent Errors in Digitization of Polish Ancient Manuscripts.* In Asian Conference on Intelligent Information and Database Systems (pp. 170-179). Springer, Cham.

Gould, S., & Ebdon, R. (1999). *Survey on digitisation and preservation*. IFLA Offices for UAP and International Lending.

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). *Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach.* Journal of the Association for Information Science and Technology, 67(12), 2928-2946.

Kettunen, K., Pääkkönen, T., & Koistinen, M. (2016). *Kansalliskirjaston digitoitu historiallinen lehtiaineisto 1771–1910: sanatason laatu, kokoelmien käyttö ja laadun parantaminen*. Informaatiotutkimus, 35(3), 3-14.

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*. In LWDA (pp. 124-135).

Maxwell, A. (2010). *Digital archives and history research: feedback from an end-user.* Library Review, 59(1), 24–39.

Navitski, R. (2014). *Reconsidering the Archive: Digitization and Latin American Film Historiography*. Cinema Journal, 54(1), 121–128.

Prescott, A., & Hughes, L. M. (2018). *Why do we digitize? The case for slow digitization*. Archive Journal.

Salminen, A. (2011). *Mikä kirjallisuuskatsaus?: Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin.*

Svärd, P. (2017). *The woes of Swedish private archival institutions*. Records Management Journal, 27(3), 275–285.