KANSALLISARKISTO

Massadigitoinnin Proof of Concept (PoC)

Tavoitteet ja tulokset tiivistetysti

Julkinen versio 1.0

Kansallisarkisto
4.3.2019

1 JOHDANTO

Kansallisarkiston ensimmäisessä, vuonna 2017 toteuttamassa massadigitoinnin suunnitteluprojektissa tehtiin luonnos massadigitoinnin prosessikokonaisuudesta, joka mahdollistaisi viranomaisten analogisten aineistojen muuntamisen digitaaliseen muotoon tehokkaasti mutta laadukkaasti asiakirjatietoja vaarantamatta. Projektin yhteydessä tuotettiin myös luonnos kriteereistä, jotka mahdollistaisivat analogisen aineiston hävittämisen digitoinnin jälkeen. Lisäksi tehtiin esitys massadigitoinnin tuotantovaiheessa hyödynnettävistä laitteista. Suunnitelmien mukaan tarkoituksena oli hyödyntää pääosin suurtehoskanneria viranomaisaineistojen keskitetyssä digitoinnissa.

Massadigitoinnin suunnittelua jatkettiin vuonna 2018 maaliskuussa käynnistyneessä massadigitoinnin jatkosuunnitteluprojektissa. Yhdeksi projektin tehtäväksi asetettiin massadigitoinnin Proof of Concept (PoC), jotta Kansallisarkisto saisi paremman käsityksen, mitä nykyhetken skannaukseen liittyvä teknologia mahdollistaa – erityisesti koskien suurtehoskannereita ja niihin kytkettyjä ohjelmistoja ja digitointiprosessissa syntyviä siirtopaketteja.

PoC:n avulla oli tarkoitus myös testata käytännössä, miten osa massadigitoinnin ensimmäisessä suunnitteluprojektissa suunnitelluista prosessivaiheista toimivat.

2 POC-PROJEKTIN TOTEUTUS

PoC-projekti toteutettiin maalis- ja joulukuun 2018 välisenä aikana. Projekti käynnistyi suunnitelman laadinnalla, hävittämiseen tähtäävien kriteerien eli teknisen vaatimusmäärittelyn tarkentamisella, rekrytoinneilla ja hankintojen kilpailutuksella (avoin kilpailutusmenettely) ja eteni infrastruktuurin pystytykseen sekä viranomaisaineistojen lainauksen toteutukseen. Käytännön testaus toteutettiin kolmen eri testikierroksen yhteydessä. Testikierrosten välillä tehtiin testikierroksen analysointi, jonka perusteella laadittiin suunnitelmat ja ohjeistus seuraavaa testikierrosta varten.

Aineistoja valittiin PoC:iin massadigitoitavan aineiston tiekartan 1. ryhmästä, jotta saataisiin kokemusta ensimmäisenä massadigitointiin suunniteltujen aineistojen käsittelystä. Viranomaisista projektiin mukaan lähtivät Terveyden- ja hyvinvoinnin laitos (THL), Lääkealan turvallisuus- ja kehittämiskeskus Fimea, Sosiaali- ja terveysalan lupa- ja valvontavirasto Valvira, Tilastokeskus ja Verohallinto. Aineistoa lainattiin viranomaisilta yhteensä noin 9,5 hyllymetriä. Edellä mainittujen lisäksi PoC:ssa hyödynnettiin Kansallisarkiston hallussa olevia viranomaisaineistoja. Tavoitteena oli saada mahdollisimman kattava ja monipuolinen kokonaisuus erityyppisiä aineistoja: paperiasiakirjoja, kortistoja ja sidoksia.

PoC:n alkuperäisenä tarkoituksena oli testata ennen kaikkea suurtehoskanneria sekä sen vaikutuksia valmisteluun ja skannaukseen. Ennen käytännön toteutuksen aloittamista päätettiin kuitenkin, että tuloksia halutaan myös dokumenttiskannerista sekä eri digitoinnin toteutusmalleista käsitellä aineistoa. Aineistoa käsiteltiin testikierroksilla kolmella eri toteutusmallilla:

- toteutusmalli 1: aineiston valmistelu etukäteen ja skannaus suurtehoskannerilla
- toteutusmalli 2: aineiston valmistelu etukäteen ja skannaus dokumenttiskannerilla
- toteutusmalli 3: aineiston valmistelu ja skannaus dokumenttiskannerilla yhtäaikaisesti

Lisäksi PoC:n yhteydessä tehtiin pieniä testejä sidosten skannauksesta mastoskannerilla.

PoC:ssa valmisteltiin ja skannattiin eri toteutusmalleilla aineistoa yhteensä 44,3 hyllymetriä.

3 PoC:n tavoitteet ja tulokset

Seuraavissa luvuissa numeroituihin laatikoihin on koottu PoC:n alkuperäiset tavoitteet. Lisäksi luvuissa on esitetty mahdolliset tarkennukset tai muutokset tavoitteisiin sekä lopulliset tulokset tiiveistetysti.

3.1 Digitoinnin valmistelu

	Alkuperäinen tavoite
1	Mitä vaatimuksia suurtehoskannerin käyttäminen aiheuttaa valmistelulle? Luodaan luonnos digitoinnin valmistelun minimitoimenpiteistä.

Digitoinnin valmistelun tarkoituksena on omalta osaltaan varmistaa digitoinnin lopputuloksena syntyneiden kuvien laatua ja luettavuutta, edesauttaa aineiston sujuvaa skannausta, ehkäistä skannerin pysähtyminen sekä vähentää aineiston vaurioitumisen riskiä skannauksen aikana. PoC:n yhteydessä pyrittiin tehostamaan valmistelua ja löytämään tarvittavat minimitoimenpiteet. Valmistelutoimenpiteiden tarkentamisessa ei tarvinnut ottaa huomioon arkistokelpoisuutta, sillä PoC:n yhteydessä tehdyt selvitykset liittyivät aineistoon, joka lähtökohtaisesti hävitetään digitoinnin jälkeen analogisessa muodossa.

PoC:ssa saatiin tuloksia suurtehoskannerille valmistelun lisäksi dokumenttiskannerille valmistelusta. Taulukkoon 1 on koottu tiivistetysti ehdotus minimitoimenpiteistä, jotka digitoitavalle aineistolle pitää tehdä tarpeen mukaan – aineistosta riippuen – kun aineisto valmistellaan dokumentti- tai suurtehoskannerilla skannattavaksi. Jos digitointi toteutetaan toteutusmallilla 2 tai 3, voidaan hyödyntää dokumenttiskanneriin yhdistettyä tasoskanneria. Tällöin osan valmistelutoimenpiteistä voi jättää tietyissä tilanteissa pois.

Taulukko 1. Digitoinnin valmistelun minimitoimenpiteet.

Digitoinnin valmistelun minimitoimenpiteet
Aineisto poistetaan säilytysyksiköstä.
Aineiston alkuun, ensimmäiseksi arkiksi sijoitetaan viivakoodillinen ohjausarkki, joka toimii aineiston tunnisteena digitointiprosessissa.
Aineisto tasataan, skannerivaatimukset huomioiden, käyttäen tasauskulmaa, minkä yhteydessä jokainen arkki käydään läpi informaatiohävikin minimoimiseksi. Kortistoaineisto käydään läpi selaamalla kortit läpi säilytysvälineessään tai ottamalla ne valmisteltavaksi pieni nippu kerrallaan palauttaen ne säilytysvälineeseensä tai erilliseen digitointilaatikkoon. Tasaamisen yhteydessä tehdään samalla muita tarvittavia toimenpiteitä arkeille.
Digitoinnin estävät taitokset suoristetaan käsin.
Niitit ja muut vastaavat liittimet (esim. muovitaskut, kumilenkit) poistetaan.
Liian pienet arkit kiinnitetään erilliselle arkille skannauksen mahdollistamiseksi. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.
Liian suuret arkit erotetaan skannattavasta yksiköstä. Erotetun arkin yhteyteen asetetaan erikoisaineiston ohjausarkki sekä toinen ohjausarkki siihen kohtaan, josta se erotetaan. Jos liian suuri arkki on maksimissaan A2-kokoinen ja jos informaatio ei jatku molemmilla puolilla, leikataan arkki kahteen osaan skannaamisen mahdollistamiseksi.
Liian paksut arkit erotetaan skannattavasta yksiköstä. Erotetun arkin yhteyteen asetetaan erikoisaineiston ohjausarkki sekä toinen ohjausarkki siihen kohtaan, josta se erotetaan. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.

Liian ohuet arkit erotetaan. Erotetun arkin yhteyteen asetetaan erikoisaineiston ohjausarkki sekä toinen ohjausarkki siihen kohtaan, josta se erotetaan. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.
Huonokuntoiset arkit erotetaan skannattavasta yksiköstä. Erotetun arkin yhteyteen asetetaan erikoisaineiston ohjausarkki sekä toinen ohjausarkki siihen kohtaan, josta se erotetaan. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.
Muun muotoiset kuin suorakulmion malliset arkit kiinnitetään erilliselle arkille skannauksen mahdollistamiseksi. Jos arkki on A4-kokoinen tai sitä suurempi, erotetaan arkki. Erotetun arkin yhteyteen asetetaan erikoisaineiston ohjausarkki sekä toinen ohjausarkki siihen kohtaan, josta se erotetaan. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.
Uudelleen kiinnitettävistä viestilapuista otetaan kopio alkuperäisellä paikallaan, minkä jälkeen hävitetään alkuperäinen viestilappu ja liitetään A4-kopio osaksi valmisteltavaa yksikköä oikealle paikalle. Toteutusmallissa 2–3 voidaan hyödyntää tasoskanneria.
Digitoinnin estävät merkittävät repeytymät ja irronneet palaset paikataan mattapintaisella teipillä. Seuraavat paikkaukset on tehtävä: kahteen tai useampaan osaan repeytyneen arkin paikkaus, arkin keskellä olevien repeytymien paikkaus sekä isojen repeytymien paikkaus, jos niitä on vain muutamia. Merkittävästi repeytyneitä arkkeja ei paikata, vaan ne skannataan sellaisenaan erikoisskannerilla, kuten tasoskannerilla. Pieniä, reunassa esiintyviä repeytymiä ei paikata.
Informaation päällä olevat vanhat paikkaukset poistetaan. Poistamista kokeillaan yhden kerran. Jos paikka ei lähde, jätetään se paikoilleen.
Taitetut arkit (vaippa-, kansilehdet yms.) ja vihot leikataan irtoarkeiksi käsileikkurilla, jos informaatio ei jatku taitoksen molemmilla puolilla. Leikkaaminen tehdään mahdollisuuksien mukaan yhdellä leikkauksella myös vihkojen tapauksessa. Epätasaisen tai taittuneen leikkauslinjan muodostumista tulee välttää.
Sidosmuodossa oleva aineisto puretaan käsin ja/tai sähköleikkurilla irtoarkeiksi. Tietyissä tapauksissa sidokset skannataan mastoskannerilla, jolloin sidosta ei pureta ollenkaan tai sitä puretaan vain osittain.
Kaiken kokoiset taitetut arkit leikataan taitoksesta auki, jos lukujärjestys tätä vaatii.
Aineisto laitetaan tarvittavien toimenpiteiden jälkeen odottamaan skannausta. Toteutusmallissa 3 valmisteltu aineisto voidaan laittaa heti skannattavaksi, minkä aikana jatketaan muun aineiston valmistelua.

Lisäksi PoC:ssa tehtiin seuraavat tärkeät huomiot liittyen aineistolle sopivan toteutusmallin valintaan:

- kiinnitetyt arkit (esimerkiksi liian pienet arkit) hidastavat merkittävästi skannausta, mistä syystä toteutusmallin 3 hyödyntäminen tämän tyyppisten aineistojen kohdalla on erittäin perusteltua.
- jos sidosaineisto täytyy purkaa kokonaan käsin (esimerkiksi kapeiden marginaalien takia) ja sidoksien sisältämät arkit ovat monen kokoisia, sidoksia ei kannata purkaa ja skannata dokumentti- tai suurtehoskannerilla. Tässä tapauksessa sidosten skannaus mastoskannerilla niitä purkamatta on nopeampi vaihtoehto. Mastoskannerilla skannatessa on kuitenkin kiinnitettävä huomiota, että syntyneistä kuvista ei rajaudu pois informaatiota esimerkiksi sidoksen kapeiden marginaalien takia. Ratkaisuna tähän on sidosrakenteen löystyttäminen / osittainen purkaminen ennen skannausta mastoskannerilla. Jos aineiston purkamisessa voidaan käyttää sähköleikkuria puoleen tai sitä suurempaan osaan ja jos sidoksen sisältämät arkit ovat samankokoisia, sidosten purkaminen ja skannaus dokumentti- tai suurtehoskannerilla on mastoskannerilla skannausta kannattavampaa.

	Alkuperäinen tavoite
2	Miten kauan aikaa kuluu digitoinnin valmisteluun riippuen aineistotyypistä? Luodaan luonnos aineistoluokista.

Tulosten perusteella laadittiin esitys aineistoluokista, joiden avulla voidaan tehdä etukäteisarvioita läpimenoajoista digitoitaville aineistoille ja tehostaa koko digitointiprosessia. Paperiaineistoja (paperiasiakirjat ja kortistot) koskevat aineistoluokat ovat:

- **Aineistoluokka 1**
  Aineistoluokka koostuu fyysisiltä ominaisuuksiltaan tasalaatuisesta aineistosta. Aineiston muoto, koko ja muut ominaisuudet eivät vaihtele aineiston sisällä. Aineisto ei sisällä huonokuntoista, kuten repaleista tai muuten haurasta aineistoa. Aineistoon ei kohdistu valmistelun toimenpiteitä ollenkaan tai juuri lainkaan. Valmistelu koostuu aineiston tarkastamisesta ja tasaamisesta. Asiakirjoja ei tarvitse käydä valmistelussa yksittäin läpi. Aineiston ei pitäisi aiheuttaa skannauksessa pysähdyksiä.
- **Aineistoluokka 2**
  Aineistoluokka koostuu fyysisiltä ominaisuuksiltaan tasalaatuisesta aineistosta. Aineistossa on vähän paperiliittimiä, avattavia taitoksia tai muita valmisteluntoimenpiteitä vaativia elementtejä. Aineisto ei sisällä huonokuntoista, kuten repaleista tai muuten haurasta aineistoa. Aineisto ei juuri aiheuta pysähdyksiä skannauksessa.
- **Aineistoluokka 3**
  Aineistoluokka koostuu fyysisiltä ominaisuuksiltaan lähes tasalaatuisesta aineistosta. Arkkien koko voi vaihdella aineiston sisällä. Aineisto sisältää paperiliittimiä. Aineistossa voi olla avattavia ja leikattavia arkkeja, kuten vaippa- ja kansilehdet. Aineistossa voi esiintyä arkkeja, jotka on kiinnitettävä tyhjälle taustapaperille tai käsiteltävä jotenkin muuten, jotta ne ovat skannattavissa. Aineisto ei sisällä huonokuntoista, kuten repaleista tai muuten haurasta aineistoa. Aineisto aiheuttaa pysähdyksiä skannauksessa jonkin verran.
- **Aineistoluokka 4**
  Aineisto koostuu fyysisiltä ominaisuuksiltaan poikkeavista aineistoista. Arkkien koko vaihtelee aineiston sisällä. Aineisto on kiinnitetty paperiliittimillä. Aineistossa on avattavia ja leikattavia arkkeja, kuten vaippa- ja kansilehdet. Aineistossa on arkkeja, jotka on kiinnitettävä tyhjälle taustapaperille tai käsiteltävä jotenkin muuten, jotta ne ovat skannattavissa. Aineistossa on julkaisuja, lehtiä tai muita vastaavia, jotka on joko leikattava irtoarkeiksi tai skannattava tasolla. Aineistossa voi olla harvakseltaan repaleisia reunoja tai paikattavia repeämiä. Aineiston valmisteluun kuluu aikaisempia aineistoluokkia enemmän aikaa toimenpiteiden moninaisuuden vuoksi. Aineisto aiheuttaa pysähdyksiä skannauksessa.
- **Aineistoluokka 5**
  Aineistoluokka koostuu selkeästi monimuotoisesta aineistosta. Arkkien koko ja muoto vaihtelevat aineiston sisällä. Aineistossa on arkkeja, jotka on kiinnitettävä tyhjälle taustapaperille tai käsiteltävä jotenkin muuten, jotta ne ovat skannattavissa. Aineisto voi sisältää osittain huonokuntoista aineistoa. Aineisto sisältää tasoskannattavia tai aineiston joukosta erotettavia arkkeja. Aineisto on valmistelulla mahdollista saada skannattavaksi dokumenttiskannereilla ja vastaavilla. Valmistelun toimenpiteet vievät huomattavasti enemmän aikaa. Aineisto aiheuttaa merkittävästi pysähdyksiä skannauksessa.
- **Aineistoluokka 6**
  Aineistoluokka koostuu huonokuntoisesta aineistosta ja muuten fyysisiltä ominaisuuksiltaan sellaisista aineistosta, että niiden valmistelu on arvion mukaan hyvin hidasta. Valmistelun toimenpiteillä ei välttämättä saavuteta sellaista tasoa, että aineisto olisi skannattavissa dokumenttiskannereilla ja vastaavilla. Aineisto vaatii suurimmaksi osaksi tasoskanneria tai muuta erikoisskanneria.

Lisäksi sidokset jaettiin viiteen luokkaan:

- Kokonaan sähköleikkurilla purettavat sidokset.
- Osittain sähköleikkurilla purettavat sidokset.
- Käsin helposti purettavat sidokset, joihin sähköleikkuria ei voida käyttää.
- Käsin vaikeasti purettavat sidokset, joihin sähköleikkuria ei voida käyttää.
- Sidokset, joita ei voida purkaa sähköleikkurilla eikä käsin.

Sidosten aineistoluokkien lisäksi on otettava huomioon sidosten käsittelyaikoja arvioitaessa myös edellä mainitut aineistoluokat 1–6, joiden avulla pitää arvioida sidoksen sisällön vaatimaa käsittelyaikaa.

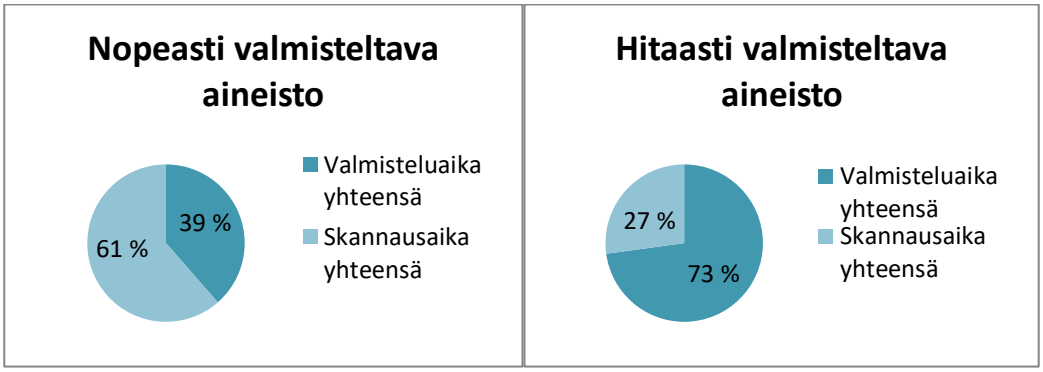
Testikierrosten jälkeen tehtiin matemaattisiin menetelmiin perustuva aineistoluokka-analyysi, jota varten laskettiin, paljonko käsittelyaikaan erityisesti vaikuttavia tekijöitä (esim. niitit) esiintyi senttimetriä kohden kussakin aineistossa. Koska osa ominaisuuksista vaikutti aineistojen käsittelyaikoihin muita enemmän, niille laskettiin jokaista toteutusmallia kohden pienimmän neliösumman menetelmällä sellaiset painotuskertoimet, että aineiston kokonaiskäsittelyaika pystyttiin arvioimaan mahdollisimman tarkasti. Kertoimien avulla laskettiin PoC-aineistoille käsittelyaika-arviot, joita voi vertailla toteutuneisiin käsittelyaikoihin. Aineistojen käsittelyaikarajoiksi annettiin seuraavat luvut:

- Aineistoluokka 1: 0–1 min/cm
- Aineistoluokka 2: 1–5 min/cm
- Aineistoluokka 3: 5–10 min/cm
- Aineistoluokka 4: 10–15 min/cm
- Aineistoluokka 5: 15–25 min/cm
- Aineistoluokka 6: yli 25 min/cm

PoC:ssa sidosten määrä ei ollut niin merkittävä, että niiden pohjalta olisi voinut tehdä vastaavanlaisia laskelmia kuin paperiaineistojen kohdalla. Voidaan kuitenkin todeta, että sidosten käsittelynopeuteen vaikuttavat sidostyyppi ja marginaalin leveys.

PoC:n perusteella todettiin myös, että aineistojen valmisteluajat vaikuttavat aineistojen kokonaiskäsittelyaikoihin enemmän kuin skannausajat. Tästä syystä erityisesti aineiston valmisteluun kuluvasta ajasta on tehtävä arvio etukäteen. Kuvassa 1 esitellään valmistelun ja skannauksen kokonaisaikojen suhdetta toisiinsa nopeasti ja hitaasti valmisteltavan aineiston tapauksissa.

Kuva 1. Valmistelu- ja skannausaikojen välinen suhde nopeasti ja hitaasti valmisteltavalla aineistolla.



3.2 Skannaus

	Alkuperäinen tavoite
3	Miten tiekartan 1. ryhmän aineistot soveltuvat digitoitavaksi suurtehoskannerilla? Testataan viranomaisten hallussa olevia tiekartan 1. ryhmän aineistoja kattavasti. Tämän lisäksi testataan Kansallisarkistoon siirrettyjä aineistoja.
4	Mitä aineistoa suuretehoskannerilla pystytään skannaamaan? Tarkennetaan massadigitointiin suunniteltua laiteinfraa.

PoC:iin valitut lainatut sekä Kansallisarkiston hallussa olevat viranomaisaineistot pystyttiin digitoimaan kokonaisuudessaan. Ainoastaan yksi kortistoaineistoista (Kansallisarkiston aineistoja) havaittiin materiaaliltaan liian paksuksi sujuvaa skannausta ajatellen. Käytännössä osa aineistosta oli alkuperäisiin massadigitoinnin tavoitteisiin verrattuna hitaampaa käsitellä, minkä vuoksi aineistojen soveltuvuutta massadigitointiin on tarkennettava.

Massadigitoinnissa käytettäväksi suunniteltu suurtehoskanneri ei sovellu pääasialliseksi skanneriksi. Suurtehoskannerille soveltuu ominaisuuksiltaan tasalaatuinen aineisto, joka ei vaadi oikeastaan mitään valmistelutoimenpiteitä – eli esimerkiksi siisti A4-kokoinen aineisto. Jos aineisto sisältää etukäteen arvioituna erityisominaisuuksia, kuten posti- tai leimamerkkejä, ei suurtehoskanneri toimi parhaiten sen skannaukseen.

Kortistojen skannauksessa dokumenttiskanneri on saatujen tulosten perusteella suurtehoskanneria parempi vaihtoehto. Vaikuttaa myös siltä, että dokumenttiskanneri toimii paremmin, koostuipa kortistoaineisto sitten suurimmaksi osaksi samankokoisista tai erikokoisista korteista.

Toteutusmalli 3 (yhtäaikainen valmistelu ja skannaus dokumenttiskannerilla) on suurimmassa osassa tapauksista nopein käsittelymalleista. Erityisesti paljon valmistelua vaativan, monimuotoisen ja -kokoisen aineiston kohdalla ero muihin toteutusmalleihin on merkittävä.

	Alkuperäinen tavoite
5	Pystytäänkö PoC:n kokoonpanolla tuottamaan hävittämiseen tähtäävän digitoinnin kriteerien mukaisia siirtopaketteja (pois lukien TAR-paketointi)? Tarkennetaan prosessia ja päätetään sovelluksen toimivuudesta.
6	Miten laadukkaita kuvia suurtehoskannerilla saadaan tuotettua? Tarkennetaan hävittämiseen tähtääviä kriteereitä.

Teknisten vaatimusmäärittelyjen mukaisen skannauksen lopputuloksen saavuttamiseen käytettiin Proof of Concept -toteutuksessa paljon aikaa. Laitteet eivät tuottaneet lähtökohtaisesti PoC-kilpailutuksessa asetettujen vaatimusmäärittelyjen mukaisia siirtopaketteja ulos oikeaoppisesti. Kilpailutuksen vaatimusmäärittelyt peilasivat silloiseen luonnokseen hävittämiseen tähtäävän digitoinnin vaatimusmäärittelyistä. Testauksen aikana toimittajat tekivät kuitenkin laitteisto- ja ohjelmistokokoonpanoihin muutoksia, joiden perusteella vaadittuja metatietoja saatiin paremmin tuotettua.

Testatuilla kokoonpanoilla havaittiin, että skannauksessa tuotettavan raakakuvan kuvanlaadulla sekä formaatilla on vaikutuksia tuotantonopeuteen. Kaikissa tapauksissa skannerilla tuotettavan

raakakuvaan ei voitu vaikuttaa, vaan vaatimusten mukaisen kuvan muodostaminen oli mahdollista toteuttaa vasta skannauksen jälkikäsittelyssä.

	Alkuperäinen tavoite
7	Miten PoC:n kokoonpano (skanneri ja ohjelmisto) kommunikoi (lähettää koneymmärrettäviä tietoja) muihin järjestelmiin? Tarkennetaan prosessia ja päätetään PoC:n kokoonpanon soveltuvuudesta.

PoC:ssa vertailtiin dokumentti- ja suurtehoskannereiden skannaus- ja prosessointiohjelmistoja. Siirtopaketteja ei kuitenkaan siirretty mihinkään erillisiin järjestelmiin. Operaattorien kannalta ohjelmistot toimivat skannereissa yhtä hyvin ja ne olivat helppoja käyttää. Lisäksi PoC:ssa tarkasteltiin automaattista kuvan kääntöä lukusuunnan mukaisesti, joka toimi vaihtelevasti. Ohjelmat eivät tunnistaneet suurimmassa osassa tapauksista käsinkirjoitettua tekstiä tai ohjelmistoilla oli hankaluuksia määrittää päälukusuuntaa, jos tekstiä oli useaan eri suuntaan.

	Alkuperäinen tavoite
8	Osaako PoC:n kokoonpano hyödyntää ohjausarkkeja? Tarkennetaan prosessia ja päätetään PoC:n sovelluksen toimivuudesta.

PoC:ssa todettiin, että skannauksessa voidaan hyödyntää ohjausarkkeja. Ohjelmistot asetettiin lukemaan viivakoodeja, jolloin valmistelussa asetetut viivakoodit ohjasivat tuotetut tiedostot oikeisiin arkistoyksikön mukaisiin kansioihin. Ohjelmistot tarkastivat viivakoodien tunnukset erillisistä Excel-taulukoista, jonka avulla varsinaiset viivakoodit myös luotiin.

	Alkuperäinen tavoite
9	Osaako PoC:n kokoonpano tuoda kuvia suurtehoskannerin skannauserän (batch) sisälle toisesta skannausprosessista eli osaako sovellus hyödyntää erikoisaineiston ohjausarkkeja? Tarkennetaan prosessia ja päätetään PoC:n sovelluksen toimivuudesta.

PoC:ssa erotettiin valmisteluvaiheessa dokumentti- ja suurtehoskannereille soveltumattomat erikoisaineistot ”erotettu”-ohjausarkilla. Sen sijaan PoC:ssa ei toteutettu erotettujen aineistojen skannausta erikoisskannerilla sekä näiden kuvien tuontia oikealle paikalleen skannauserän (batch) sisälle. Tähän liittyvää erillistä testausta kuitenkin tehtiin suurtehoskannerilla. Kuvien yhdistäminen saatiin onnistumaan ohjelmistossa viivakoodien avulla.

	Alkuperäinen tavoite
10	Mikä on suurtehoskannerin todellinen nopeus? Tarkennetaan aineistojen läpimenoaikoja.

Skannauksen tehokkuuteen vaikuttavat skannattavan aineiston ominaisuuksien lisäksi skannerin tekniset ominaisuudet. Suurtehoskanneri on dokumenttiskannereita herkempi pysähdyksille ja häiriöille. Vaikka dokumenttiskannerilla tulisi enemmän pysähdyksiä, on yksittäiseen pysähdykseen kulunut aika lyhyempi kuin suurtehoskannerilla.

PoC:n perusteella saatiin laskettua skannausnopeuksia sekä dokumentti- että suurtehoskannereille.

	Alkuperäinen tavoite
11	Miten suurtehoskannerin mekaaninen laadunvarmistus ja eheyden varmistaminen (esimerkiksi syöttötelat ja tuplasyöttösensorit) toteutuvat? Tarkennetaan laadunvarmistusta.

Osana laadunvarmistusta molemmissa laitteissa käytettiin ultraäänellä toimivia sensoreita arkkien tuplasyötön[1] (double feed) tunnistamiseen. Vaikka suurteho- ja dokumenttiskannereissa tuplasyötön tunnistus toimi herkästi, siitäkin huolimatta havaittiin, että skannereista meni ajoittain läpi samanaikaisesti useampi asiakirja, jolloin osa asiakirjoista jäi skannaamatta. Massadigitoinnin jatkosuunnittelussa on huomioitava, että skannauksen aikaiseen laadunvarmistukseen panostetaan riittävillä tuplasyöttösensoreilla. Sensorien on katettava koko skannausalue. Lisäksi kuvanlaadun validoinnin otannan osuutta alkuperäisiin massadigitoinnin suunnitelmiin (1. suunnitteluprojekti) verrattuna on nostettava. Näillä toimenpiteillä ehkäistään informaatiohävikin syntymistä.

	Alkuperäinen tavoite
12	Miten suurtehoskanneri kestää käyttöä? Toteutetaan testijakso (rasitustesti), jonka aikana pyritään analysoimaan skannerin reagointia tuotantoa vastaavalle kuormalle (esim. skannerin mahdollinen kuumeneminen jne.). Tuotetaan analyysi skannerin kestokyvystä verrattuna suunnitteluhankkeen tuotantomääriin.

Suurtehoskannerilla toteutettiin pienimuotoista rasitustestausta ennen ensimmäisen testikierroksen alkamista sekä skannerin toimittajan tekemien päivitysten jälkeen. Rasitustestit olivat PoC:ssa pisimmillään tunnin yhtäjaksoisia skannaustestauksia. Myös skannerin kykyä luoda suuria eriä (batch) testattiin luomalla noin 10 000 kuvan eriä. Skanneri kesti hyvin nämä rasitustestit. PoC:n aikana ei kuitenkaan toteutettu pidempikestoisia rasitustestejä, jotka olisivat rinnastettavissa tuleviin tuotantomääriin.

	Alkuperäinen tavoite
13	Mikä on PoC:ssa tarjottavan OCR -sovelluksen laatu? Verrataan sovelluksia, tehdään päätös sovellusten välillä ja tarkennetaan prosessia.
14	Pystyykö PoC:ssa tarjottua OCR-sovellusta käyttämään suunnitelmien mukaisesti? Tarkennetaan prosessia.
15	Miten tarjotun ohjelmiston laadunhallinta toimii? Tarkennetaan prosessia.
16	Voidaanko tuoda muita OCR-sovelluksia osaksi prosessia? Tarkennetaan prosessia.

Alkuperäisiin tavoitteisiin 13–16 ei PoC:n laitteistokokonaisuuksien osalta tehty tarkempaa selvitystä. Varsinaisesti OCR-sovelluksiin liittyviä näkökulmia tarkasteltiin erillisessä, seuraavassa luvussa käsitellyssä, OCR-selvityksessä.

[1] Useampi arkki menee skanneriin samanaikaisesti, jolloin informaatiohävikkiä syntyy.

3.3 Erillinen OCR-selvitys

	Alkuperäinen tavoite
17	Miten erilaiset OCR -sovellukset eroavat keskenään? Verrataan konkreettisia tunnistustuloksia erilaisten sovellusten kesken.

Osana PoC-projektia selvitettiin optisen tekstintunnistuksen (OCR, Optical Character Recognition) tarkkuutta massadigitoinnin piiriin suunnitelluilla aineistoilla. OCR-selvitystä varten testiaineistoksi valittiin massadigitoinnin piiriin tulevien viranomaisaineistojen kaltaista – ei kuitenkaan PoC-testauksessa käytettyä – aineistoa. Testauksessa hyödynnettiin eri kaupallisia sekä avoimen lähdekoodin OCR-moottoreita.

Selvityksen perusteella OCR-tarkkuus jäi kokeiluissa alle tavoitellun tason. Selvityksen perustella jo pieni määrä käsinkirjoitettua tekstiä, kuten diaarinumerot, voivat vaikuttaa merkittävästi OCR-moottoreiden tarkkuuteen, minkä seurauksena tietoa menetetään arkistoaineistojen koneluettavasta versiosta. Tulosten perusteella parhaat tunnistustulokset saadaan häviöttömistä TIFF-tiedostoista tarkimmilla moottoreilla, mutta erot kuvaformaattien ja OCR-tarkkuuden välillä ovat hyvin pieniä.

Jatkotyössä tulee selvitettäväksi, onko OCR-tuloksia mahdollista parantaa vähäisellä optimoinnilla. Lisätutkimuksia pitäisi tehdä muun muassa XML-tiedostoilla sekä ei-kaupallisilla OCR-ohjelmilla. Laajemmin OCR-tulosten hyödynnettävyyttä oikeisiin käyttötilanteisiin pyritään selvittämään osana massadigitoinnin pilotointia.