KANSALLISARKISTON VAATIMUKSET HÄVITTÄMISEEN TÄHTÄÄVÄÄN DIGITOINTIIN

Sisältö	Kansallisarkiston vaatimukset digitoinnille, joka mahdollistaa analogisen ilmentymän hävittämisen
Tarkoitus	Varmistaa kansalliseen kulttuuriperintöön kuuluvien asiakirjojen sisältämän tietosisällön säilyminen ja niiden käytettävyys analogisen ilmentymän hävittämisen jälkeen. Tämän asiakirjan tarkoituksena on toimia määrityksenä siitä minkälainen digitointiprosessi mahdollistaa hävittämisen.
Kohderyhmä	Asiakirja on tarkoitettu Kansallisarkistolle sekä muille julkishallinnon toimijoille, jotka tähtäävät arkistolain (831/1994) nojalla pysyvään säilytykseen määrättyjen asiakirjojensa analogisten ilmentymien hävittämiseen digitoinnin jälkeen. Analogisten asiakirjojen hävittäminen edellyttää Kansallisarkiston seulontapäätöstä.
Rajaukset	Tätä asiakirjaa sovelletaan, kun digitoidaan takautuvasti asiankäsittelyn tai muun toiminnan tuloksena muodostuneita aineistokokonaisuuksia tai arkistoituja aineistoja. Vaatimukset eivät koske digitointia, joka toteutetaan asiakirjan laatimisen tai asiankäsittelyn yhteydessä taikka organisaatiolle saapuvien analogisten asiakirjojen muuttamiseksi digitaaliseen muotoon. Asiakirjassa kuvataan digitointiprosessia ja sen lopputulosta. Asiakirjassa ei kuvata varsinaista säilytyspakettia, joka tallennetaan pitkäaikaissäilytysjärjestelmään. Pitkäaikaissäilytysjärjestelmään tallennettava paketti on mahdollista muodostaa tässä asiakirjassa esitetystä digitointiprosessin lopputuloksesta. Tässä asiakirjassa ei oteta kantaa analogisen ilmentymän hävittämisprosessiin, vaan kuvataan ne kriteerit, joiden täytyttyä analogisen asiakirjan hävittäminen olisi mahdollista.
Sovelletut säännökset	Arkistolaki (831/1994) 14a §
Voimassaoloaika	Toistaiseksi, asiakirjan päiväyksestä eteenpäin

Sisällys

1 Termit ja käsitteet

Dokumentissa käytetty termistö perustuu Internet Engineering Task Force:n toimesta tehtyyn määrittelyyn [RFC 2119].[1] Taulukossa 1 ilmaistaan, mitä käännöksiä termistöstä tässä asiakirjassa käytetään.

Taulukko 1: Tässä asiakirjassa käytetyt käännökset

ENGLANTI	SUOMI
MUST	PITÄÄ
MUST NOT	EI SAA
REQUIRED	PAKOLLINEN
SHOULD	PITÄISI
SHOULD NOT	EI PITÄISI
MAY	SAA
OPTIONAL	VAPAAEHTOINEN

Alla olevassa taulukossa (Taulukko 2: Käsitteistö) ilmaistaan, mitä tässä asiakirjassa taulukossa esitetyillä käsitteillä tarkoitetaan:

Taulukko 2: Käsitteistö

KÄSITE	SELITE
Analoginen asiakirja	Paperisessa tai muussa käsin kosketeltavassa muodossa laadittu, säilytetty ja/ käytettävä asiakirja.
Analoginen ilmentymä	Digitoitavaksi päätetyn analogisen asiakirjakokonaisuuden analoginen ilmiasu. Tässä asiakirjassa analoginen ilmentymä tarkoittaa asiakirjakokonaisuutta, joka koostuu pääsääntöisesti A4/foliokokoisista –

[1] https://www.ietf.org/rfc/rfc2119.txt Viitattu 6.3.2019

KÄSITE	SELITE
	paperiasiakirjoista (analogisista asiakirjoista), mutta se voi sisältää myös sitä suurempia tai pienempiä asiakirjoja.
Digitaalinen ilmentymä	Digitoitavaksi päätetyn analogisen teoksen/asiakirjakokonaisuuden digitaalinen ilmiasu.
Digitaalinen objekti	Digitaalinen tiedosto, joka joko yksin tai muiden digitaalisten tiedostojen kanssa muodostaa digitoidun asiakirjan. Tässä asiakirjassa digitaalinen objekti on joko kuvatiedosto tai XML-tiedosto.
Digitointiprosessi	Joukko toimintoja joiden avulla analoginen ilmentymä muunnetaan digitaaliseksi.
Digitoitu asiakirja	Analogisesta asiakirjasta digitointiprosessilla tuotettu sähköinen versio, joka voi koostua n-määrästä digitaalisia objekteja.
Kuvatiedosto	Kuvatiedosto tarkoittaa digitointiprosessissa tuotettua bittikarttakuvaa (digitaalinen objekti). Tässä asiakirjassa esitetty kuvatiedosto voi olla joko TIFF tai JPEG formaatissa ja se toimii tallekappaleena (master).[2]
Päälukusuunta	Mahdollistaa asiakirjan tietosisällön tulkitsemisen kuvatiedostoa kääntämättä. Mikäli asiakirjassa esiintyy tietosisältöä useampaan lukusuuntaan, tarkoittaa päälukusuunta sitä suuntaa, jossa suurin osa asiakirjan tietosisällöstä on luettavissa.
Siirtopaketti	Digitointiprosessissa muodostettu kokonaisuus, joka on siirrettävissä Kansallisarkiston tietojärjestelmiin.
Tuotantovuorokausi	Vuorokausi, jonka aikana laitteella tuotetaan digitaalisia objekteja

2 Johdanto

Hävittämiseen tähtäävällä digitoinnilla tarkoitetaan analogisen ilmentymän hävittämistä digitointiprosessin päätteeksi. Kyse ei ole analogisten asiakirjojen tietosisällön hävittämisestä, vaan tietosisällön muuttamisesta toiseen säilytysmuotoon. Analogisen ilmentymän hävittäminen edellyttää, että digitaaliseen muotoon muuttaminen eli digitointi on toteutettu menetelmillä, jotka eivät heikennä asiakirjan todistusvoimaisuutta, eheyttä ja autenttisuutta.

Digitoinnin kohteena olevalla analogisilla ilmentymillä PITÄÄ olla Kansallisarkiston seulontapäätös, jossa määrätään asiakirjatiedon säilytysmuodosta. Päätös vahvistaa onko analogisella ilmentymällä kulttuurihistoriallista arvoa, jonka vuoksi analogista ilmentymää EI SAA hävittää sen digitoinnin jälkeen. Yleisesti ottaen seulontapäätös PITÄISI olla olemassa jo ennen kuin asiakirjoja lähdetään digitoimaan, koska tällöin on helpompaa suunnitella, minkä määritysten mukaan asiakirjat digitoidaan. Mikäli seulontapäätöstä ei ole tehty, analogisia ilmentymiä EI SAA hävittää digitoinnin jälkeen, vaikka digitointi olisi suoritettu tässä asiakirjassa esitettyjen vaatimusten mukaisesti.

Tässä asiakirjassa esitettyjä kriteerejä PITÄÄ noudattaa, kun viranomainen digitoi arkistolain nojalla pysyvään säilytykseen määrättyjä analogisia asiakirjoja, joiden analogisessa muodossa oleva kappale on määrä hävittää Kansallisarkiston seulontapäätöksen perusteella digitoinnin jälkeen. Digitoitujen ilmentymien vastaanottaminen Kansallisarkiston tietojärjestelmiin edellyttää, että digitaaliset ilmentymät täyttävät tässä asiakirjassa esitetyt vaatimukset.

[2] FADGI, tallekappale: http://www.digitizationguidelines.gov/term.php?term=productionmasterfile Viitattu 6.3.2018

Tämän asiakirjan laadinnassa on huomioitu arkistosektorilla yleisesti käytössä olevat standardit sekä muiden Kansallisarkistojen laatuvaatimukset digitoinnille. Lisäksi luvuissa hyväksyttävät formaatit ja digitointiprosessissa muodostettava paketti, on huomioitu kansalliset pitkäaikaissäilytyspalveluiden (PAS-palvelut) asettamat vaatimukset säilytettäville aineistoille.[3]

Tämä asiakirja kohdentuu erilaisten analogisten asiakirjojen digitointiin kuvatiedostoiksi sekä niistä erilaisia tekniikoita hyväksikäyttäen tunnistettujen sisältöjen prosessointiin ja tallennukseen. Asiakirja ei käsittele esimerkiksi äänen tai elävän kuvan digitointia.

3 Yleiset digitointiprosessin vaatimukset

Analogisten ilmentymien digitaaliseksi muuntaminen on prosessi (digitointiprosessi), jota PITÄÄ dokumentoida tässä asiakirjassa ilmaistuin tavoin ja tarkkuuksin. Prosessin dokumentoinnilla tarkoitetaan, että skannauksesta sekä kuvien mahdollisesta käsittelystä tallennetaan näitä toimenpiteitä dokumentoivat metatiedot. Näiden metatietojen lisäksi digitaaliseen muotoon muuttamisen prosessista SAA tallentaa metatiedoiksi myös muita toimenpiteitä.

Digitointiprosessissa PITÄÄ varmistua, että digitoitavaksi tarkoitettu kokonaisuus tulee digitoitua kokonaisuutena ja sisällöllisesti täydellisenä. Tämä tarkoittaa käytännössä sitä, että kaikki digitoitavaksi päätetyn kokonaisuuden analogiset asiakirjat PITÄÄ digitoida siten, että mitään informaatiota ei jää teknisen tai toiminnallisen virheen takia muuntamatta digitaaliseen muotoon.

Jokaisesta digitoitavaan kokonaisuuteen liittyvästä yksittäisestä kuvatiedostosta PITÄÄ olla visuaalisella tarkastelulla saatavissa sama informaatio kuin sen analogisesta vastineesta. Kuvatiedosto EI SAA sisältää mitään elementtejä, joita ei ilmene analogisessa vastineessa. Tästä poikkeuksen muodostavat mahdolliset samaan kuvatiedostoon skannattavat/kuvattavat digitaalisen objektin värejä, harmaasävyjä, mittasuhteita ja resoluutiota todentavat skannaustekniset mittataulut. Kyseiset mittataulut PITÄÄ asetella siten, että ne eivät peitä digitoitavaa kohdetta.

Digitointiprosessissa EI SAA poistaa merkintöjä sisältäviä sivuja. Digitointiprosessissa tuotetut kuvatiedostot PITÄÄ olla käännetty päälukusuuntaan. Digitointiprosessissa tuotettuja kuvatiedostoja SAA kääntää niiden skannauksen jälkeen vain 90 asteen portaissa.

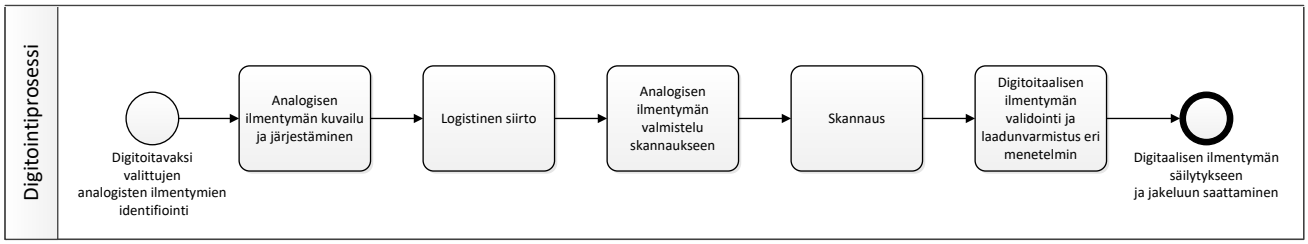
Ennen skannaustapahtumaa PITÄISI digitoinnissa käytetyn infrastruktuurin suorituskyky optimoida. Optimoinnin jälkeen PITÄISI infrastruktuurin tuottamien kuvatiedostojen laatu todentaa käyttämällä tähän tarkoitukseen tarkoitettuja mittatauluja. Laatu PITÄISI todentaa kerran tuotantovuorokaudessa.

[3]Kansalliset pitkäaikaissäilytyspalvelut -> Määritykset -> Säilytys- ja siirtokelpoiset tiedostomuodot http://digitalpreservation.fi/files/PAS-tiedostomuodot-1.6.1.pdf Viitattu 6.3.2018

4 Yleiset digitointiprosessin suositukset ja hyvät käytänteet

Tässä luvussa kuvataan yleisiä hyviä käytänteitä, jotka liittyvät skannaukseen ja skannauksen laadunvarmistukseen.

Digitointi käsitetään yleisesti prosessina, joka sisältää alla olevassa kuvassa (Kuva 1, Sähköiseen muotoon muuntamisen prosessi - yleinen) esitetyt vaiheet.



Kuva 1, Sähköiseen muotoon muuntamisen prosessi - yleinen

On hyvä korostaa, että digitointiprosessissa laadunvarmistusta ei voi kohdentaa vain tiettyyn vaiheeseen prosessissa, vaan digitoinnin lopputuloksen luotettavuus muodostuu siitä, että laadunvarmistusta tehdään prosessin eri vaiheissa. Tässä asiakirjassa keskitytään erityisesti skannauksen laadunvarmistukseen.

Digitoitavat asiakirjat on hyvä kuvailla metatietojärjestelmään ennen niiden sähköiseen muotoon muuntamisen aloittamista, jolloin analogisen ilmentymän käsittelyketjua voidaan dokumentoida kattavammin ja esimerkiksi tietoja aineiston valmistelusta voidaan dokumentoida. Skannauksen jälkeen aineiston metatietoja voidaan rikastaa joko ihmisen toimesta tai automaattisin menetelmin.

Skannauksen laadunvarmistus voidaan karkeasti jakaa ennen skannaustapahtumaa tapahtuvaan toimintaan ja sen jälkeiseen laadunvarmistamiseen eli validointiin.

Kuten luvussa 3 esitetään, pitäisi skannausinfrastruktuurin suorituskyky optimoida ennen skannaustapahtumaa siten, että sen tuottama digitaalinen ilmentymä edustaa parasta mahdollista ilmentymää, joka kyseisellä teknisellä kokoonpanolla voidaan tuottaa. Optimoinnin jälkeen infrastruktuurin suorituskykyä pitäisi tarkkailla suunnitellusti, jotta prosessissa tuotettavien digitaalisten ilmentymien laatu säilyy tasaisena. Tarkkailua varten tarvitaan yleisesti mittataulu, mittataulun referenssiarvot ja analysointiohjelmisto. Kuvanlaadun lisäksi laiteinfrastruktuurissa pitäisi kiinnittää huomiota siihen, että voidaan varmistua analogisten asiakirjojen muuntuvan digitaaliseen muotoon tietosisällöltään täydellisinä. Tämä tarkoittaa muun muassa sitä, että laitteistoa hankittaessa pitäisi kiinnittää erityistä huomiota laitteen kykyyn erotella asiakirjat toisistaan, jotta voidaan välttyä kahden päällekkäisen analogisen asiakirjan menemisestä laitteen läpi (läpisyöttöskannerit, avorataskannerit ja muut skannausratkaisut, missä asiakirjoja skannataan ”massana”).

Skannauksen jälkeinen validointi voidaan toteuttaa otannoilla. Otannan määrä on riippuvainen skannausprosessin luotettavuudesta. Yleisiä viitearvoja ja suosituksia on julkaistu runsaasti. Validoinnin tavoitteena on varmistua siitä, että luvussa 3 esitetyt vaatimukset täyttyvät. Mikäli aineisto on konekirjoitettua, voidaan digitointiprosessissa luoduista kuvatiedostoista tunnistaa

tietosisältöä erilaisin menetelmin (esimerkiksi OCR = Optical Character Recognition). Tätä vaihetta voidaan käyttää myös skannauksen onnistumisen mittarina, mikäli käytettävään sovellukseen voidaan asettaa tunnistuksen onnistumiselle rajoja.

Jos kuvatiedostoja käsitellään skannaustapahtuman jälkeen, pitäisi yksityiskohtainen kuvankäsittelyhistoria tallentaa ainakin kuvatiedostojen metatietoihin. Mahdollisuuksien mukaan myös digitaalisen objektin syntyä kuvaileviin XML -tietoihin.

5 Hyväksyttävät formaatit

Formaattiosio on jaettu kolmeen alaluokkaan:

1. Kuvatiedosto
2. Kuvatiedostosta tunnistettu tietosisältö
3. Kuvatiedostoa ja kuvatiedoston prosessointia kuvaavat metatiedot ja rakenne

5.1 Kuvatiedosto

Digitointiprosessissa tuotettava kuvatiedosto PITÄÄ tallentaa 24 bittisenä RGB-kuvana. Kuvatiedosto EI SAA missään käsittelyvaiheessa olla tässä luvussa esitettyjä vaatimuksia heikkolaatuisempi. Kuvatiedosto tallennetaan joko häviöttömästi pakatussa TIFF-muodossa tai häviöllisesti pakatussa JPEG-muodossa (ei molemmissa). [4] Mikäli prosessissa ensimmäisenä tuotettu kuvatiedosto on pakattu, sitä EI SAA käsitellä sen ensimmäisen tallennuskerran jälkeen ja tallentaa tämän jälkeen uudelleen.

Alla olevissa taulukoissa 3 ja 4 esitetään pakolliset tiedot, jotka kuvatiedostossa PITÄÄ olla koneymmärrettävässä muodossa. Taulukossa 3 esitetään pakolliset metatiedot, mikäli kuvatiedoston tallennusmuoto on TIFF. Taulukossa 4 taas esitetään pakolliset metatiedot, mikäli kuvatiedoston tallennusmuoto on JPEG. Mikäli taulukoiden ”Elementti” – saraketta ei tarkenneta, on tieto ilmaistava, mutta tiedolle ei ole tässä yhteydessä määritelty vaadittavaa kenttää. Taulukossa esitettyjen tietojen lisäksi kuvatiedosto SAA sisältää muita metatietokenttiä.

Taulukko 3: Kuvatiedoston (TIFF) pakolliset metatiedot

Elementti	Tarkenne	Vaadittu arvo, mikäli ilmaistavissa yksiselitteisesti	Metatieto-skeema	Metatietokenttä
Formaatti	TIFF	image/tiff		MIME Type
Versio		6.0		
Kuvan nimi				
Kuvatiedoston koko				
Väritila	Kuvatiedoston väritila	RGB	Exif.Image	PhotometricInterpretation (262)
ICC-profiili		sRGB, eciRGB v2, ProPhoto RGB, AdobeRGB (1998)	TIFF tag, private	ICC Profile (34675)

[4] JPEG: https://jpeg.org/jpeg/index.html Viitattu 6.3.2019
TIFF: https://www.itu.int/itudoc/itu-t/com16/tiff-fx/docs/tiff6.pdf Viitattu 6.3.2019

Elementti	Tarkenne	Vaadittu arvo, mikäli ilmaistavissa yksiselitteisesti	Metatieto-skeema	Metatietokenttä
Bittisyvyys	Bittien määrä pikselin kanava-arvossa	8 8 8	Exif.Image	BitsPerSample (258)
	Kanava-arvojen määrä pikselissä	3	Exif.Image	SamplesPerPixel (277)
Tiedoston pakkaaminen		5 = LZW	Exif.Image	Compression (259)
Kuvan leveys	Kertoo kuvan leveyden pikselien määrällä per rivi		Exif.Image	ImageWidth (256)
Kuvan korkeus	Kertoo kuvan korkeuden pikselirivien määrällä kuvassa		Exif.Image	ImageLenght (257)
Digitaalisen kuvatiedoston tekijä	Organisaatio, joka on luonut kuvatiedoston analogisesta ilmentymästä		Exif.Image	Artist (315)
Digitointilaite (skannaus tai kuvaus)	Kertoo minkä valmistajan laitteella analoginen objekti on muutettu sähköiseen muotoon (valmistajan nimi)		Exif.Image	Make (271)
Digitointilaitteen malli (skannaus tai kuvaus)	Tarkentaa digitointilaitetta kertomalla valmistajan mallin nimen		Exif.Image	Model (272)
Digitoinnissa käytetyn laitteen sarjanumero	Yksilöi käytetyn laitteen		Exif.Image	CameraSerialNumber (50735)
Kuvatiedoston luomisessa käytetty ohjelma	Sovellus ja versio, millä digitaalinen tiedosto on luotu		Exif.Image	Software (305)
Kuvatiedoston luontipäivämäärä ja aika (skannauspäivämäärä)	Ilmaistaan muodossa: YYYY:MM:DDTHH:MM:SS		Exif.Image	DateTimeOriginal (36867)
Lukusuunta	Tiedoston lukusuunta (vaaka tai pysty). Lukusuunta ei ota kantaa kuvan tietosisällön lukusuuntaan, vaan tässä ilmaistaan tiedoston lukusuunta.		Exif.Image	Orientation (274)
Resoluution mittayksikkö	Mittayksikkö, jota käytetään tulkitessa X ja Y resoluutiota	2 = inch	Exif.Image	ResolutionUnit (296)
XResoluutio	Pikselien määrä resoluution mittayksikköä kohti leveyssuunnassa.	300	Exif.Image	XResolution (282)
YResoluutio	Pikselien määrä resoluution mittayksikköä kohti pystysuunnassa.	300	Exif.Image	YResolution (283)
Tavujärjestys		big endian tai little endian		ByteOrder

Elementti	Tarkenne	Vaadittu arvo, mikäli ilmaistavissa yksiselitteisesti	Metatieto-skeema	Metatietokenttä
Kuvatiedoston käsittelyohjelma	Mikäli digitointiprosessissa luodaan ensin pakkaamaton tiedosto, jota käsitellään skannauksen jälkeen, tallennetaan käsittelyohjelman nimi ja versio		Exif.Image	Image.ProcessingSoftware (11)

Taulukko 4: Kuvatiedoston (JPEG) pakolliset metatiedot

Elementti	Tarkenne	Vaadittu arvo	Metatieto-skeema	Metatietokenttä
Formaatti	JPEG	image/jpeg		MIME Type
Versio	JPEG part 1 versio	1.00 tai1.01 tai 1.02		JFIF Version
Kuvan nimi				
Kuvatiedoston koko				
Väritila	Kuvatiedoston väritila	RGB	Exif.Image	PhotometricInterpretation (262)
ICC -profiili	Kuvatiedoston metatietoihin tallennettu väriprofiili.	sRGB	ICC	profileDescription
Bittisyvyys	Bittien määrä pikselin kanava-arvossa	8 8 8	Exif.Image	BitsPerSample (258)
	Kanava-arvojen määrä pikselissä	3	Exif.Image	SamplesPerPixel (277)
JPEG-laatu	JPEG-pakkauksen laatu asteikolla 0 -100%	90%		
Kuvatiedoston tekijä	Organisaatio, joka on luonut kuvatiedoston analogisesta ilmentymästä		Exif.Image	Artist (315)
Kuvan korkeus	Kertoo kuvan korkeuden pikselirivien määrällä kuvassa		Exif.Image	ImageLength(257)
Kuvan leveys	Kertoo kuvan leveyden pikselirivien määrällä kuvassa		Exif.Image	ImageWidth(256)
Digitointilaite	Kertoo minkä valmistajan laitteella analoginen objekti on muutettu digitaaliseksi (valmistajan nimi)		Exif.Image	Make (271)
Digitointilaitteen malli	Tarkentaa digitointilaitetta kertomalla valmistajan mallin nimen		Exif.Image	Model (272)
Digitoinnissa käytetyn laitteen sarjanumero	Tarkentaa mallia ja yksilöi laitteen, jonka avulla analoginen objekti on muunnettu digitaaliseksi		Exif.Image	CameraSerialNumber (50735)
Kuvatiedoston luomisessa käytetty ohjelma	Sovellus ja versio, millä digitaalinen tiedosto on luotu		Exif.Image	Software (305)
Kuvatiedoston luontipäivämäärä ja aika	Ilmaistaan muodossa: YYYY:MM:DD HH:MM:SS		Exif.Image	DateTime (306)
Lukusuunta	Tiedoston lukusuunta (horisontaalinen tai vertikaalinen)		Exif.Image	Orientation (274)

Elementti	Tarkenne	Vaadittu arvo	Metatieto-skeema	Metatietokenttä
Resoluution mittausyksikkö	Mittayksikkö, jota käytetään tulkitessa X ja Y resoluutiota	2 = inch	Exif.Image	Image.Resolution Unit (296)
XResoluutio	Pikselien määrä resoluution mittayksikköä kohden kuvan leveyssuunnassa.	300	Exif.Image	Image.XResolution (282)
YResoluutio	Pikselien määrä resoluution mittayksikköä kohden kuvan korkeussuunnassa.	300	Exif.Image	Image.YResolution (283)
Kuvatiedoston käsittelyohjelma	Mikäli digitointiprosessissa luodaan ensin pakkaamaton tiedosto, jota käsitellään skannauksen jälkeen, tallennetaan käsittelyohjelman nimi ja versio		Exif.Image	Image.Processing Software (11)

5.2 Tunnistetun tekstin tallennusformaatti

Mikäli kuvatiedostoista tunnistetaan tekstiä esimerkiksi OCR (konekirjoitetun tekstin tunnistus) tai HTR (käsinkirjoitetun tekstin tunnistus) menetelmin, PITÄÄ se tallentaa Analyzed Layout and Text Object (ALTO)-formaattiin.[5] Hyväksyttävät ALTO versiot ovat 3.0 tai uudemmat. Jokaisesta kuvatiedostosta PITÄÄ tallentaa oma ALTO-tiedostonsa, mikäli kuvatiedoston sisältämä tietosisältö on pääsääntöisesti konekirjoitettua.[6]

5.3 Kuvatiedostoa ja kuvatiedoston prosessointia kuvaavat metatiedot ja rakenne

Tässä luvussa määritellyt metatiedot kuvaavat kuvatiedoston syntyhistoriaa, joka osaltaan todentaa myös prosessissa syntyneen digitaalisen ilmentymän autenttisuutta. Kuvatiedostojen pakolliset tekniset metatiedot PITÄÄ esittää MIX-metatietoskeeman version 2.0 mukaisesti.[7]

Alla olevassa taulukossa 5 ilmaistaan vasemmalta oikealle MIX -kentän nimi, kentän tarkoitus vapaasti käännettynä ja velvoite. Velvoite- kentässä ilmaistaan kyseisen kentän ja sen skeeman mukaisen tiedon pakollisuus seuraavalla tavalla:

- P = pakollinen – tämä tieto PITÄÄ kuvata
- V = Vapaaehtoinen – tämä tieto PITÄISI kuvata, mutta se ei ole pakollista

MIX-metatietoskeemassa on kahdenlaisia kenttiä: säiliöitä ja dataelementtejä. Dataelementit sisältävät tietyn arvon, kun taas säiliöt sisältävät yhden tai useamman dataelementin ja ne voivat sisältää toisia säiliöitä dataelementteineen. Taulukossa 5 ilmaistaan vain tietyn arvon sisältäviä kenttiä eli dataelementtejä.

[5] The Library of Congress » Standards » ALTO. Kongressin kirjaston verkkosivu https://www.loc.gov/standards/alto/ Viitattu 6.3.2019
[6] Pääsääntöisesti konekirjoitettu asiakirja sisältää sekä konekirjoitettua että käsinkirjoitettua tekstiä, mutta pääsääntöisesti konekirjoitettua tekstiä. Tämän lisäksi asiakirjassa voi esiintyä kuvia tai muuta sisältöä.
[7] The Library of Congress » Standards » MIX. Kongressin kirjaston verkkosivu http://www.loc.gov/standards/mix/ Viitattu 6.3.2019

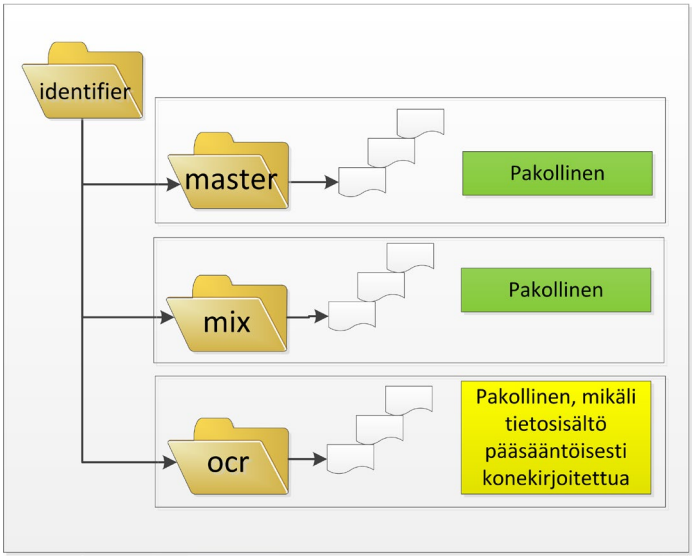
Taulukko 5: Kuvatiedostoa ja sen prosessointia kuvaavat metatiedot (taulukossa on ilmaistu vain tietoa sisältävät kentät, jotka PITÄÄ esittää MIX-metatietoskeeman version 2.0 mukaisessa rakenteessa)

MIX -kentän nimi	Kentän tarkoitus	Velvoite
objectIdentifierType	Dataelementti, joka määrittää järjestelmän tai verkkotunnuksen, jossa digitaalisen asiakirjan yksilöivä ID on uniikki.	P
objectIdentifierValue	Digitaalisen objektin identifioiva merkkisarja.	P
fileSize	Tiedoston koko tavuissa, esimerkiksi 72839.	P
formatName	Tiedoston formaatti. Vaadittu arvo image/jpeg tai image/tiff	P
formatVersion	Tiedoston versio, esimerkiksi 1.01	V
byteOrder	Dataelementti, joka määrittää tavujen tallennusjärjestyksen. Vaadittu arvo on joko big endian tai little endian.	P
compressionScheme	Käytetty pakkaus. Vaadittu arvo JPEG (kun tallekappale on jpeg-formaatissa) tai LZW (kun tallekappale on TIFF-formaatissa)	P
compressionRatio	Dataelementti, joka kertoo käytetyn pakkauksen tason. Ilmaistaan käyttäen numerator ”90” ja denominator ”100”.	P (vain JPEG)
messageDigestAlgorithm	Dataelementti, joka identifioi algoritmin, jolla messageDigest-kentän arvo on luotu. Kentän arvo on jokin seuraavista: MD5, SHA-1, SHA256, SHA384, SHA512	P
messageDigest	messageDigestAlgorithm kentän määrittämän algoritmin tuottama merkki sarja, esimerkiksi e8064dc0	P
imageWidth	Kuvan leveys pikseleissä, esimerkiksi 1330.	P
imageHeight	Kuvan korkeus pikseleissä, esimerkiksi 1600.	P
colorSpace	Dataelementti, joka määrittää kuvan väriavaruuden. Vaadittu arvo RGB.	P
iccProfileName	Dataelementti, joka määrittää yleisesti käytetyn ICC-profiilin nimen. Vaadittu arvo JPEG-tiedostossa sRGB. Vaadittu arvo TIFF-tiedostossa sRGB, eciRGB v2, ProPhoto RGB tai AdobeRGB (1998)	P
iccProfileVersion	Dataelementti, joka kertoo käytetyn ICC-profiilin version, esimerkiksi v4 [eli sRGB v4]	P
iccProfileURL	Jos ICC-profiili ei ole hyvin dokumentoitu, profiilin URL/URN tallennetaan tähän kenttään.	V
dateTimeCreated	Dataelementti, joka kertoo kuvatiedoston luontiajan. Ilmaistaan muodossa: YYYY-MM-DDTHH:MM:SS	P
imageProducer	Dataelementti, joka identifioi digitaalisen objektin luoneen organisaation.	P
scannerManufacturer	Dataelementti, joka kertoo skannauksessa käytetyn laitteen valmistajan nimen.	P
scannerModelName	Dataelementti, joka kertoo käytetyn digitointilaitteen mallin nimen.	P
scannerModelNumber	Dataelementti, joka tarkentaa digitointilaitteen mallin nimeä sen tyyppinumerolla.	P
scannerModelSerialNo	Digitointilaitteen sarjanumero, jonka avulla tietty laite on mahdollista yksilöidä.	P
scanningSoftwareName	Käytetyn skannausohjelmiston nimi.	P
scanningSoftwareVersionNo	Käytetyn skannausohjelmiston version numero.	P
orientation	Dataelementti, joka kertoo kuvan lukusuunnan.	P
samplingFrequencyUnit	Dataelementti, joka kertoo mittayksikön, jota käytetään tulkittaessa X ja Y resoluutiota. Vaadittu arvo ”in.”	P

MIX -kentän nimi	Kentän tarkoitus	Velvoite
xSamplingFrequency	Pikselien määrä resoluution mittayksikkö kohden leveyssuunnassa. Vaadittu arvo 300	P
ySamplingFrequency	Pikselien määrä resoluution mittayksikkö kohden pystysuunnassa. Vaadittu arvo 300	P
bitsPerSampleValue	Dataelementti, joka määrittelee jokaisessa kanavassa olevien bittien määrän. Vaadittu arvo 8	P
bitsPerSampleUnit	Dataelementti, joka määrittää bittien tulkintatavan. Arvo on joko integer tai floating point.	P
samplesPerPixel	Dataelementti, joka määrittää kanava-arvojen määrän pikselissä. Vaadittu arvo 3	P
targetType	Dataelementti, joka kertoo onko skannauksen laatua todentava mittataulu osa kuvaa vai skannattu erilliseen kuvaan.	V
targetManufacturer	Dataelementti, johon merkitään mittataulun valmistaja.	V
targetName	Dataelementti, joka identifioi käytetyn mittataulun nimen.	V
targetNo	Dataelementti, joka sisältää käytetyn mittataulun sarjanumeron.	V
externalTarget	Dataelementti, joka kertoo mistä TargetID-säiliön yksilöidyn mittataulun digitaalinen kuva löytyy.	V
performanceData	Dataelementti, joka kertoo mistä TargetID-säiliön yksilöimän mittataulun mittausdata löytyy.	V

6 Digitointiprosessissa muodostettava siirtopaketti

Luvussa 5 ja sen alaluvuissa mainitut digitointiprosessissa tuotetut erilaiset digitaaliset objektit PITÄÄ tallentaa alla olevassa kuvassa (Kuva 2, Digitointiprosessin vaadittu siirtopakettirakenne) esitettyyn hakemistorakenteeseen, jotta ne voidaan ottaa vastaan Kansallisarkistoon. Digitaalinen ilmentymä PITÄÄ tuottaa hakemistorakenteeseen riippumatta siitä, milloin se siirretään Kansallisarkistoon. Aineistoja luovutettaessa siirtopaketti EI SAA sisältää mitään muuta kuin kuvassa 2 esitettyjä hakemistoja. Mikäli aineistoja ei tulla missään vaiheessa siirtämään Kansallisarkistoon, on hakemistorakenteen noudattaminen VAPAAEHTOISTA. Tässä määritellyn hakemistorakenteen lisäksi organisaatio SAA tallentaa esimerkiksi käyttökappaleet omiin tietojärjestelmiinsä siinä tietorakenteessa, mitä kyseinen järjestelmä edellyttää. Tässä asiakirjassa määritelty rakenne ei siis sulje pois muiden mahdollisten tallennusrakenteiden käyttöä.



Kuva 2, Digitointiprosessin vaadittu siirtopakettirakenne

Taulukossa 6 kuvaillaan, miten digitaaliset objektit PITÄÄ nimetä kuvassa 2 esitetyn hakemistorakenteen sisällä. Prosessissa tuotettujen digitaalisten objektien PITÄÄ kohdata keskenään. Toisin sanoen AltoXML -tiedoston 0001.xml PITÄÄ sisältää kuvatiedostosta 0001.jpg tai 0001.tif tunnistettu tietosisältö. MIX-metatietoskeeman mukaisen 0001.xml-tiedoston PITÄÄ sisältää kuvatiedostoa 0001.jpg tai 0001.tif kuvailevia metatietoja.

Taulukko 6 Siirtopaketin hakemistojen sisältö

Hakemisto	Selite
identifier	Tarkoittaa digitaalisen ilmentymän yksilöivää tunnusta, jonka avulla PITÄÄ pystyä tunnistamaan, mistä asiakirjakokonaisuudesta on kyse (esimerkiksi arkistoyksikkö).[8] Hakemisto sisältää ”digitaalisten objektien hakemistot”.
master	Hakemistoon PITÄÄ tallentaa taulukoissa (Taulukko 3: Kuvatiedoston (TIFF) pakolliset metatiedot **tai** Taulukko 4: Kuvatiedoston (JPEG) pakolliset metatiedot) esitetyt kuvatiedostot yksittäisinä tiedostoina. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.tif tai 0001.jpg
ocr	Hakemistoon PITÄÄ tallentaa luvussa 5.2 esitetty AltoXML tiedosto siten, että jokaisesta kuvatiedostosta on oma XML-tiedostonsa. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.xml.
mix	Hakemistoon PITÄÄ tallentaa taulukossa 5 esitetyt pakolliset tiedot koskien kaikkia master – hakemiston sisällä olevia kuvatiedostoja. Tiedostoon PITÄISI tallentaa myös muut taulukossa esitetyt tiedot. Tiedostoon SAA tallentaa myös muita MIX-metatietoskeeman mukaisia tietoja, skeeman mukaisessa rakenteessa. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.xml

Mikäli aineisto toimitetaan Kansallisarkistolle, PITÄÄ jokainen siirtopaketti paketoida TAR-paketiksi. TAR-paketin sisältöä EI SAA tässä vaiheessa pakata. TAR-paketille PITÄÄ laskea tarkistesumma MD5-muodossa ja se PITÄÄ toimittaa siirron yhteydessä. Kansallisarkistoon aineistoa toimitettaessa tarkoitetaan identifier-hakemistolla AHAA -järjestelmän aineiston tunnistetta Ai01 (roolissa ahaa tekninen).

[8] Digitoitavaksi päätetyn analogisen ilmentymän pitäisi olla kuvailtuna (kuvaileva metatieto tuotettu) ennen sen digitointia. Identifierin avulla PITÄÄ pystyä yhdistämään digitointiprosessissa syntyneet digitaaliset ilmentymät edellä mainittuun kuvailevaan metatietoon.

7 Esimerkkipaketit

Esimerkkejä siirtopaketeista on kaksi:

1. Esimerkkipaketti_JPEG.zip
2. Esimerkkipaketti_TIF.zip

”master” hakemistojen kuvatiedostot eivät ole kuvanlaadullisia referenssejä. Kuvatiedostoissa on tässä asiakirjassa pakollisiksi määritellyt metatiedot. ”ocr”-hakemistojen tiedostot ovat esimerkkejä siitä, että jokaisesta tiedostosta PITÄÄ tehdä oma AltoXML-tiedostonsa. ”mix” – hakemistojen tiedostot ovat esimerkkejä liitteen paketissa olevista master-tiedostoista lukuun ottamatta elementtejä, joiden kohdalla todetaan toisin. Hakemistoja ei ole paketoitu TAR-pakettiin.

8 Allekirjoitukset

Pääjohtajan sijainen,
Tutkimusjohtaja — Päivi Happonen

Kehittämispäällikkö — Mikko Eräkaski