KANSALLISARKISTO
RIKSARKIVET

2.3.2018

KANSALLISARKISTON VAATIMUKSET HÄVITTÄMISEEN TÄHTÄÄVÄÄN DIGITOINTIIN (LUONNOS)

Sisältö	Kansallisarkiston vaatimukset digitoinnille, joka mahdollistaa analogisen manifestaation hävittämisen sen digitaaliseen muotoon muuntamisen jälkeen.
Rajaukset	Tässä 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 manifestaation hävittämisprosessiin.
Tarkoitus	Varmistaa kansalliseen kulttuuriperintöön kuuluvien asiakirjojen sisältämän tietosisällön säilyminen ja niiden käytettävyys analogisen manifestaation hävittämisen jälkeen.
Kohderyhmä	Tässä asiakirjassa esitettävät vaatimukset on tarkoitettu Kansallisarkistolle sekä muille julkishallinnon toimijoille, jotka tähtäävät analogisen manifestaation hävittämiseen sen digitoinnin jälkeen.

Säännökset, joihin toimivalta määräyksen antamiseen perustuu

Arkistolaki (831/1994) 14a §

Voimassaoloaika	Toistaiseksi

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

Taulukossa 2 ilmaistaan, mitä tässä asiakirjassa seuraavilla käsitteillä tarkoitetaan:

Taulukko 2: Käsitteistö

KÄSITE	SELITE
Digitaalinen ilmentymä	Analogisesta asiakirjasta digitointiprosessilla tuotettu sähköinen versio.
Digitaalinen manifestaatio	Digitoitavaksi päätetyn analogisen teoksen/asiakirjakokonaisuuden digitaalinen ilmiasu.
Digitointiprosessi	Joukko toimintoja joiden avulla analoginen manifestaatio muunnetaan

[1] https://www.ietf.org/rfc/rfc2119.txt Viitattu 22.2.2018

	digitaaliseksi.
Analoginen manifestaatio	Digitoitavaksi päätetyn analogisen asiakirjakokonaisuuden analoginen ilmiasu. Tässä asiakirjassa analoginen manifestaatio tarkoittaa asiakirjakokonaisuutta, joka koostuu pääsääntöisesti A4/foliokokoisista – paperiasiakirjoista, mutta se voi sisältää myös sitä suurempia tai pienempiä asiakirjoja.
Tallekappale[2]	Tallekappale tarkoittaa digitointiprosessissa tuotettua bittikarttakuvaa, joka on teknisiltä ominaispiirteiltään laadukkain digitointiprosessissa tuotettu digitaalinen ilmentymä.
Käyttökappale[3]	Käyttökappale tarkoittaa digitointiprosessissa tuotettua bittikarttakuvaa, joka tarjotaan käytettäväksi esimerkiksi verkkokäyttöliittymän kautta. Yleisesti käyttökappale on tietosisällöltään tallekappaleen kanssa identtinen, mutta informaatio esitetään pakatussa tiedostoformaatissa.
Tuotantovuorokausi	Vuorokausi, jonka aikana laitteella tuotetaan digitaalisia ilmentymiä
Päälukusuunta	Mahdollistaa asiakirjan tietosisällön tulkitsemisen kuvatiedostoa kääntämättä. Mikäli asiakirjassa esiintyy tietosisältöä useampaan lukusuuntaan, tarkoittaa päälukusuunta sitä suuntaa, jossa suurin osa asiakirjan tietosisällöstä on luettavissa.

2 Johdanto

Hävittämiseen tähtäävällä digitoinnilla tarkoitetaan analogisen manifestaation hävittämistä digitointiprosessin päätteeksi. Kyse ei ole analogisten asiakirjojen tietosisällön hävittämisestä, vaan tietosisällön muuttamisesta toiseen säilytysmuotoon. Kun pysyvään säilytykseen määrätty asiakirja muutetaan digitointiprosessissa sähköiseen muotoon, säilyy sen pysyvään säilytykseen määrätty tietosisältö edelleen. Analogisen manifestaation hävittäminen edellyttää, että asiakirjan digitointiprosessi on toteutettu menetelmillä, jotka eivät heikennä asiakirjan todistusvoimaisuutta, eheyttä ja autenttisuutta.

Tässä asiakirjassa esitettyjä kriteerejä PITÄÄ noudattaa, kun viranomainen digitoi pysyvään säilytykseen määrättyjä asiakirjoja, joiden analogisessa muodossa oleva kappale hävitetään digitoinnin jälkeen. Digitoidun aineiston vastaanottaminen Kansallisarkiston tietojärjestelmiin edellyttää, että aineisto täyttää tässä asiakirjassa esitetyt vaatimukset. Aineistoja, jotka eivät täytä tässä asiakirjassa esitettyjä vaatimuksia, ei vastaanoteta Kansallisarkiston tietojärjestelmiin.

Tämän asiakirjan laadinnassa on huomioitu arkistosektorilla yleisesti käytössä olevat standardit sekä muiden Kansallisarkistojen laatuvaatimukset digitoinnille. Lisäksi luvuissa: 5. Hyväksyttävät formaatit ja 6. Digitointiprosessissa muodostettava paketti, on huomioitu KDK:n PAS-palvelun määritykset pitkäaikaissäilytettäville tiedostoille ja niiden metatiedoille.[4]

[2] Federal Agencies Digital Guidelines Initiative -> Archival Master.
http://www.digitizationguidelines.gov/term.php?term=archivalmasterfile Viitattu 28.12.2017
[3] Federal Agencies Digital Guidelines Initiative -> Derivative file.
http://www.digitizationguidelines.gov/term.php?term=derivativefile Viitattu 28.12.2017
[4] Kansallinen digitaalinen kirjasto -> Pitkäaikaissäilytys -> Määrittelyt ja Dokumentit
http://www.kdk.fi/fi/pitkaaikaissailytys/maeaerittely-ja-dokumentit Viitattu 19.2.2018

Tämä asiakirja kohdentuu erilaisten asiakirjojen digitointiin kuvatiedostoiksi sekä niistä erilaisia tekniikoita hyväksikäyttäen tunnistettujen sisältöjen prosessointiin ja tallennukseen. Asiakirja ei käsittele esimerkiksi äänen tai elävän kuvan digitointia. Tämä asiakirja on Arkistolain (1994/831) 14a § :n nojalla velvoittava.[5]

3 Yleiset digitointiprosessin vaatimukset

Digitoinnin kohteena olevalla analogisella aineistolla PITÄÄ olla Kansallisarkiston säilytyspäätös, jossa määrätään asiakirjatiedon säilytysmuodosta. Mikäli kyseistä päätöstä ei ole, asiakirjoja EI SAA hävittää digitoinnin jälkeen, vaikka prosessi ja lopputulos olisivat tässä asiakirjassa kuvattuja.

Analogisen aineiston digitaaliseksi muuntaminen on prosessi (digitointiprosessi), jota PITÄÄ dokumentoida tässä asiakirjassa ilmaistuin tavoin ja tarkkuuksin. Prosessin tavoitteena on tuottaa autenttisia ja eheitä digitaalisia manifestaatioita.

Digitointiprosessissa PITÄÄ varmistua, että digitoitavaksi tarkoitettu kokonaisuus tulee digitoitua kokonaisuutena ja sisällöllisesti täydellisenä. Tämä tarkoittaa käytännössä sitä, että kaikki kokoelmat/sarjat/yksiköt/asiakirjat PITÄÄ digitoida siten, että mitään informaatiota ei jää teknisen tai toiminnallisen virheen takia digitoimatta.

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

Digitointiprosessissa PITÄISI poistaa asiakirjojen tyhjät kääntöpuolet. Tyhjällä kääntöpuolella tarkoitetaan asiakirjan sivua, joka ei sisällä minkäänlaisia merkintöjä. Merkintöjä sisältäviä sivuja EI SAA poistaa. Digitointiprosessissa tuotetut kuvatiedostot PITÄÄ olla käännetty päälukusuuntaan. Digitointiprosessissa tuotettuja tiedostoja SAA kääntää niiden skannauksen jälkeen vain 90 asteen portaissa.

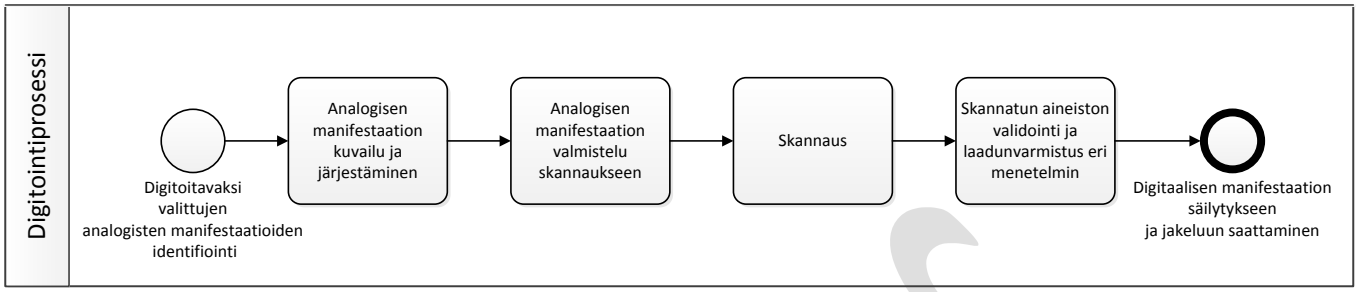
Ennen skannaustapahtumaa PITÄISI digitoinnissa käytetyn infrastruktuurin suorituskyky optimoida. Optimoinnin jälkeen PITÄÄ infrastruktuurin tuottamien digitaalisten tallekappaleiden laatu todentaa käyttämällä tähän tarkoitukseen tarkoitettuja mittatauluja. Laatu PITÄÄ todentaa kerran tuotantovuorokaudessa ja laatuarvojen on läpäistävä referenssiarvot. Hyväksyttäviä referenssiarvoja ovat Metamorfoze Extra Light ja FAGDI 2 Star (Unbound).[6] Referenssiarvot tarkentuvat vaatimusten jatkovalmistelussa vuoden 2018 aikana.

[5] Arkistolaki https://www.finlex.fi/fi/laki/ajantasa/1994/19940831 viitattu 19.2.2018

[6] Arvot on määritelty seuraavissa dokumenteissa: FADGI (Documents Unbound: General collections, 2 Star): http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf ((viitattu 10.1.2018) ja Metamorfoze (Extra light): https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf (viitattu 10.1.2018)

4 Yleiset digitointiprosessin suositukset ja hyvät käytänteet

Digitointi käsitetään yleisesti prosessina, joka sisältää kuvassa 1 esitetyt vaiheet:



Kuva 1: Sähköiseen muotoon muuntamisen prosessi (yleinen)

Tässä luvussa ei esitetä vaatimuksia, vaan kuvataan yleisiä hyviä käytänteitä, jotka liittyvät skannaukseen ja skannauksen laadunvarmistukseen.

Kuvassa 1 esitetty prosessi sisältää useita vaiheita. Yleisesti voidaan todeta, että digitointi suoritetaan siten, että asiakirjoja on kuvailtu metatietojärjestelmään ennen niiden sähköiseen muotoon muuntamisen aloittamista. Tämä osaltaan mahdollistaa myös sen, että analogisen aineiston käsittelyketjua voidaan dokumentoida. Skannauksen jälkeen aineiston metatietoja voidaan rikastaa joko ihmisen toimesta tai automaattisin menetelmin. Nämä toimet sisältävät skannauksessa tuotetun bittikarttakuvan analysointia.

Skannausinfrastruktuuri rakentuu yleensä erilaisista laitteista ja ohjelmista, jotka mahdollistavat erilaisten analogisten aineistojen digitoimisen. Tämän lisäksi työnkulkuun on usein sidottuna laaja joukko järjestelmiä ja teknologiaa, joilla kaikilla on erilainen rooli prosessin mahdollistamisessa.

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

Kuten luvussa 3 mainitaan, pitäisi infrastruktuurin suorituskyky optimoida ennen skannaustapahtumaa siten, että sen tuottama digitaalinen ilmentymä edustaa parasta mahdollista ilmentymää, joka kyseisellä kokoonpanolla voidaan tuottaa. Optimoinnin jälkeen pitää infrastruktuurin suorituskykyä tarkkailla suunnitellusti, jotta prosessissa tuotettavien digitaalisten ilmentymien laatu säilyy tasaisena. Tarkkailua varten tarvitaan mittataulu, mittataulun referenssiarvot ja analysointiohjelmisto. Kuvanlaadun lisäksi laiteinfrastruktuurissa pitäisi kiinnittää huomiota siihen, että voidaan varmistua asiakirjakokonaisuuksien muuntuvan sähköiseen 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 asiakirjan menemisestä laitteen läpi (läpisyöttöskannerit, avorataskannerit ja muut skannausratkaisut, missä asiakirjoja skannataan ”massana”).

Skannauksen jälkeen voidaan validointia suorittaa 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 siitä nykytekniikalla tunnistaa teksti OCR (optical character recognition) -menetelmin. Tätä vaihetta voidaan käyttää myös skannauksen onnistumisen mittarina, mikäli käytettävään sovellukseen voidaan asettaa tunnistuksen onnistumiselle rajoja.

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

5 Hyväksyttävät formaatit

Formaattiosio on jaettu kolmeen alaluokkaan:

1. Kuvaformaatit
2. Tunnistetun tekstin tallennusformaatti
3. Tallekappaletta ja tallekappaleen prosessointia kuvaavat metatiedot ja rakenne

5.1 Kuvaformaatit

Digitointiprosessissa PITÄÄ tuottaa erilaisiin käyttötarkoituksiin soveltuvia tiedostoformaatteja. Kaikki tiedostot PITÄÄ tallentaa 24 bittisinä RGB-kuvina. Sekä käyttökappaleen että tallekappaleen PITÄÄ olla tuotettu siten, että kumpikaan ei missään käsittelyvaiheessa ole ollut laadultaan heikompi kuin mitä luvussa 5.1.1. ja 5.1.2. on määritelty. Taulukoissa 3 ja 4 ilmaistaan pakolliset tiedot, jotka kuvatiedostoissa PITÄÄ olla koneymmärrettävässä muodossa. Mikäli ”Elementti” –saraketta ei tarkenneta, on tieto ilmaistava, mutta tiedolle ei ole tässä yhteydessä määritelty vaadittavaa kenttää. Käyttökappaleen PITÄÄ visuaalisella tarkastelulla sisältää sama informaatio kuin tallekappaleen. Tämä mahdollistaa sen, että tallekappaleeseen ei tarvitse kohdistaa toimenpiteitä lukuun ottamatta migraatioita tai muita mahdollisia erityistapauksia.

5.1.1 Käyttökappale

Alla olevassa taulukossa 3 on esitetty metatietoja, jotka PITÄÄ kirjoittaa digitointiprosessissa syntyvään käyttökappaleeseen (JPEG-tiedosto). Taulukossa esitettyjen tietojen lisäksi käyttökappale SAA sisältää muita metatietokenttiä.

Taulukko 3: Käyttökappaleen pakolliset metatiedot

Elementti	Tarkenne	Vaadittu arvo	Metatieto-skeema	Metatietokenttä
Formaatti	Ei yksiselitteistä kenttää. Tässä esimerkkinä Dublin core.	jpeg	Dublin core	dc:format
Kuvan nimi				
Kuvatiedoston koko				
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)
Tiedoston pakkaaminen	JPEG – laatu 60%			
Kuvan leveys	Kertoo kuvan leveyden pikselien määrällä per rivi		Exif.Image	ImageWidth (256)
Kuvan korkeus	Kertoo kuvan korkeuden pikselirivien määrällä kuvassa		Exif.Image	ImageLength(257)
Digitointilaite (skannaus tai kuvaus)	Kertoo minkä valmistajan laitteella analoginen ilmentymä 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			Exif.Image	CameraSerialNumber (50735)
Kuvatiedoston luomisessa käytetty ohjelma	Sovellus ja versio, millä käyttökappale on luotu		Exif.Image	Software (305)
Digitaalisen kuvatiedoston luontipäivämäärä ja aika (skannauspäivämäärä)	Ilmaistaan muodossa: YYYY:MM:DD HH:MM:SS		Exif.Image	DateTimeOriginal (36867)
Lukusuunta	Tiedoston lukusuunta (horisontaalinen tai vertikaalinen)		Exif.Image	Orientation (274)
Resoluution mittausyksikkö	tuumaa		Exif.Image	Image.ResolutionUnit (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 kuvatiedostoa käsitellään skannauksen jälkeen tallennetaan käsittelyohjelman nimi		Exif.Image	Image.ProcessingSoftware (11)

5.1.2 Tallekappale

Alla olevassa taulukossa 4 on esitetty metatietoja, jotka PITÄÄ kirjoittaa digitointiprosessissa syntyvään tallekappaleeseen (TIFF-tiedosto). Taulukossa esitettyjen tietojen lisäksi tallekappale SAA sisältää muita metatietokenttiä.

Taulukko 4: Tallekappaleen pakolliset metatiedot

Elementti	Tarkenne	Vaadittu arvo, mikäli ilmaistavissa yksiselitteisesti	Metatieto-skeema	Metatietokenttä
Formaatti	Ei yksiselitteistä kenttää. Tässä esimerkkinä Dublin	TIFF 6.0	Dublin Core	dc:format

	core.			
Kuvan nimi				
Kuvatiedoston koko				
Väritila	Kuvatiedoston väritila	2 = RGB	TIFF tag, baseline	PhotometricInterpretation (262)
ICC-profiili		sRGB, eciRGB v2, ProPhoto RGB, AdobeRGB (1998)	ICC	ICC Profile (34675)
Bittisyvyys	Bittien määrä pikselin kanava-arvossa	8 8 8	TIFF tag, baseline	BitsPerSample (258)
	Kanava-arvojen määrä pikselissä	3	TIFF tag, baseline	SamplesPerPixel (277)
Tiedoston pakkaaminen		5 = LZW	TIFF tag, baseline	Compression (259)
Kuvan leveys	Kertoo kuvan leveyden pikselien määrällä per rivi		TIFF tag, baseline	ImageWidth (256)
Kuvan korkeus	Kertoo kuvan korkeuden pikselirivien määrällä kuvassa		TIFF tag, baseline	ImageLenght (257)
Digitaalisen kuvatiedoston tekijä	Organisaatio (pakollinen), henkilö (suositeltava)		TIFF tag, baseline	Artist (315)
Digitointilaite (skannaus tai kuvaus)	Kertoo minkä valmistajan laitteella analoginen ilmentymä on muutettu sähköiseen muotoon (valmistajan nimi)		TIFF tag, baseline	Make (271)
Digitointilaitteen malli (skannaus tai kuvaus)	Tarkentaa digitointilaitetta kertomalla valmistajan mallin nimen		TIFF tag, baseline	Model (272)
Digitoinnissa käytetyn laitteen sarjanumero			Private TIFF tags	CameraSerialNumber (50735)
Digitaalisen kuvatiedoston luomisessa käytetty ohjelma	Sovellus ja versio, millä tallekappale on luotu (pakollinen). Mahdollinen tiedoston käsittelyohjelma erotettuna ";" (suositeltava).		TIFF tag, baseline	Software (305)
Digitaalisen kuvatiedoston luontipäivämäärä ja aika (skannauspäivämäärä)	Ilmaistaan muodossa: YYYY:MM:DD HH:MM:SS		TIFF tag, baseline	DateTime (306)
Lukusuunta	Tiedoston lukusuunta (vaaka tai pysty)		TIFF tag, baseline	Orientation (274)
Resoluution mittayksikkö	Mittayksikkö, jota käytetään tulkitessa X ja Y resoluutiota	2 = inch	TIFF tag, baseline	ResolutionUnit (296)
XResoluutio	Pikselien määrä resoluution mittayksikköä kohti leveyssuunnassa.	300/1	TIFF tag, baseline	XResolution (282)
YResoluutio	Pikselien määrä resoluution	300/1	TIFF tag, baseline	YResolution (283)

	mittayksikköä kohti pystysuunnassa.			
Tavujärjestys		big endian tai little endian		ByteOrder

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[7] (versio 3.0 tai 3.1). Jokaisesta sähköiseen muotoon muunnetusta asiakirjasta PITÄÄ tallentaa oma ALTO-tiedostonsa.

5.3 Tallekappaletta ja tallekappaleen prosessointia kuvaavat metatiedot ja rakenne

Tässä luvussa määritellyt metatiedot kuvaavat tallekappaleen syntyhistoriaa, joka osaltaan todentaa myös prosessissa syntyneen digitaalisen manifestaation autenttisuutta. Tallekappaleiden pakolliset tekniset metatiedot PITÄÄ esittää MIX-metatietoskeeman version 2.0 mukaisesti.[8]

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.

Taulukko 5: Tallekappaletta 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 ilmentymän yksilöivä ID on uniikki.	P
objectIdentifierValue	Digitaalisen ilmentymän identifioiva merkkisarja.	P
fileSize	Tiedoston koko tavuissa, esimerkiksi 72839.	P
formatName	Tiedoston formaatti, esimerkiksi image/tiff	P
formatVersion	Tiedoston versio, esimerkiksi 6.0	P, jos mahdollista
byteOrder	Dataelementti, joka määrittää tavujen tallennusjärjestyksen. Arvo on joko big endian tai little endian.	P

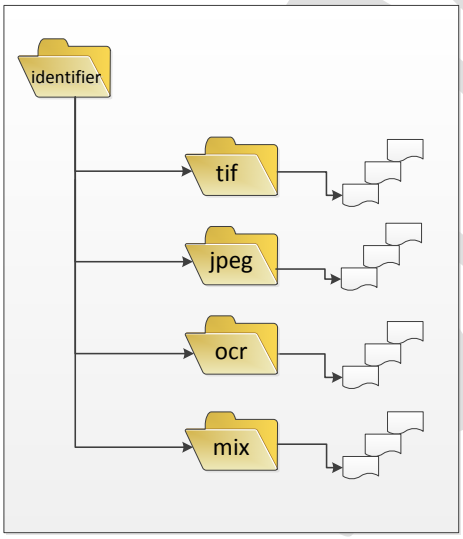
[7] The Library of Congress » Standards » ALTO. Kongressin kirjaston verkkosivu. Viitattu 19.12.2017. https://www.loc.gov/standards/alto/
[8] The Library of Congress » Standards » MIX. Kongressin kirjaston verkkosivu http://www.loc.gov/standards/mix/

compressionScheme	Käytetty pakkaus. Esimerkiksi uncompressed tai LZW	P
compressionRatio	Dataelementti, joka kertoo käytetyn pakkauksen tason	P, jos mahdollista
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, esimerkiksi RGB.	P
iccProfileName	Dataelementti, joka määrittää yleisesti käytetyn ICC-profiilin nimen, esimerkiksi eciRGB.	P
iccProfileVersion	Dataelementti, joka kertoo käytetyn ICC-profiilin version, esimerkiksi v.2 [eli eciRGB v.2]	P
iccProfileURL	Jos ICC-profiili ei ole hyvin dokumentoitu, profiilin URL/URN tallennetaan tähän kenttään.	P jos mahdollista
dateTimeCreated	Dataelementti, joka kertoo digitaalisen ilmentymän luontiajan. Ilmaistaan muodossa: YYYY-MM-DD HH:MM:SS	P
imageProducer	Dataelementti, joka identifioi digitaalisen ilmentymän 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 2=inch	P
xSamplingFrequency	Pikselien määrä resoluution mittayksikkö kohden leveyssuunnassa. Vaadittu arvo 300/1	P
ySamplingFrequency	Pikselien määrä resoluution mittayksikkö kohden pystysuunnassa. Vaadittu arvo 300/1	P
bitsPerSampleValue	Dataelementti, joka määrittelee jokaissa kanavassa olevien bittien määrän, esimerkiksi 8 tai 8 8 8	P
bitsPerSampleUnit	Dataelementti, joka määrittää bittien tulkintatavan. Arvo on joko integer tai floating point.	P
samplesPerPixel	Dataelementti, joka määrittää kanava-arvojen määrän pikselissä.	P
targetType	Dataelementti, joka kertoo onko skannauksen laatua todentava mittataulu osa kuvaa vai skannattu	V

	erilliseen kuvaan.	
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 paketti

Luvussa 5 ja sen alaluvuissa mainitut digitointiprosessissa tuotetut erilaiset tiedostot PITÄÄ tallentaa kuvassa 2 esitettyyn hakemistorakenteeseen, jotta ne voidaan ottaa vastaan Kansallisarkistoon. Digitaalinen manifestaatio PITÄÄ tuottaa hakemistorakenteeseen riippumatta siitä, milloin se siirretään Kansallisarkistoon. 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 tiedostot PITÄÄ nimetä kuvassa 2 esitetyn hakemistorakenteen sisällä. Prosessissa tuotettujen ilmentymien PITÄÄ kohdata keskenään. Toisin sanoen käyttökappaleen 0001.jpg PITÄÄ sisältää sama tietosisältö, joka on tallekappaleessa 0001.tif. AltoXML -tiedoston 0001.xml PITÄÄ sisältää bittikarttakuvasta 0001 tunnistettuja sisältöjä. MIX-metatietoskeeman mukaisen 0001.xml-tiedoston PITÄÄ kuvailla tallekappaletta 0001.tif.

Taulukko 6 Siirtopaketin hakemistojen sisältö

Hakemisto	Selite
identifier	Tarkoittaa digitaalisen manifestaation yksilöivää tunnusta, jonka avulla PITÄÄ pystyä tunnistamaan, mistä asiakirjakokonaisuudesta on kyse (esimerkiksi arkistoyksikkö).[9] Hakemisto sisältää ”ilmentymähakemistot”.
tif	Hakemistoon PITÄÄ tallentaa taulukossa 3 esitetyt tallekappaleet yksittäisinä tiedostoina. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.tif
jpeg	Hakemistoon PITÄÄ tallentaa taulukossa 4 esitetyt käyttökappaleet yksittäisinä tiedostoina. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.jpg
ocr	Hakemistoon PITÄÄ tallentaa luvussa 5.2. esitetty AltoXML tiedosto siten, että jokaisesta digitoidusta asiakirjasta on oma XML-tiedostonsa. Tiedostot PITÄÄ nimetä nelinumeroisina alkaen 0001.xml
mix	Hakemistoon PITÄÄ tallentaa taulukossa 5 esitetyt pakolliset tiedot koskien kaikkia tif – hakemiston sisällä olevia tallekappaleita. 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.

7 Liitteet

1. Esimerkkipaketti.zip[10]

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

[10] Kuvaesimerkit eivät ole kuvanlaadullisia referenssejä. Tiedostoissa on tässä asiakirjassa pakollisiksi määritellyt metatiedot. AltoXML on esimerkki siitä, että jokaisesta tiedostosta PITÄÄ tehdä oma Alto-tiedostonsa. MIX.xml on esimerkkitiedosto tässä paketissa olevasta TIF-tiedostosta lukuun ottamatta elementtejä, missä toisin todetaan.. Hakemistoja ei ole paketoitu TAR-pakettiin.