

Kansallisarkisto

Luovutuskelpoiset digitaaliset aineistomuodot ja näiden
käsittely

27.12. 2022



Sisällysluettelo

1. Ohjeen käyttötarkoitus.....	3
2. Tiedostomuotojen säilytyskelpoisuus	3
3. Tiedostojen hallinta ennen arkistoluovutusta	4
4. Tiedostomuodot ja näiden eri versioiden yksilöinti.....	5
5. Alkuperäisen tiedostomuodon hyödyntäminen	6
6. Muuntaminen erilliseen PDF-arkistomuotoon	7
7. Kuvatiedostot.....	8
8. AV tiedostomuodot	10

1. Ohjeen käyttötarkoitus

Ohjeessa käsitellään digitaalisen säilyttämiseen vaatimuksia luovutettaviin digitaalisessa muodossa oleviin aineistokokonaisuuksiin sisältyvien tiedostomuotojen näkökulmasta.

Ohjeessa esitetään sekä yleisiä linjauksia tiedostomuotojen hallintaan sekä yksityiskohtainen määrittely Kansallisarkiston hyväksymistä tiedostomuotojen formaateista.

Ohjetta voidaan hyödyntää aineistoa luovuttava tahon toimesta sekä siinä vaiheessa, kun aineistoja digitoidaan tai tuotetaan digitaaliseen muotoon, että viimeistään siinä vaiheessa, kun aineiston luovuttamista Kansallisarkistoon valmistellaan.

2. Tiedostomuotojen säilytyskelpoisuus

Sähköisen pitkäaikaissäilyttämisen tavoitteena on varmistaa, että digitaalisessa muodossa olevat aineistot ovat pitkän ajan kuluttua yhä hyödynnettävissä todistusvoimaisina ja käyttökelpoisina. Ajan mittaan nykyisin käytössä olevia tiedostomuotoja hyödyntävät sovellukset sekä tarpeet hyödyntää aineistoja muuttuvat, jolloin arkistoon luovutettua tiedostomuotoa ei ilman Kansallisarkiston vastuulla olevia säilytystoimia kyetä hyödyntämään. Näiden säilytystoimien onnistumisen edellytys on hyväksytyjen ja laadukkaiden tiedostomuotojen noudattaminen.

Digitaalisessa muodossa olevien aineistojen luotettavan arkistoinnin keskeinen tehtävä hallita sekä tiedostomuotojen että eri tallennusmenetelmien kehittymisestä muodostuvaa muutospainetta teknologisin keinoin. Digitaalisessa muodossa oleva aineisto muodostuu *bittijonoista*, joilla on aina määrättyä standardia noudattava *tiedostomuoto* eli *formaatti*. Standardit ovat sekä määrityksiä itse tiedoston teknisestä rakenteesta että määrittely sille, miten tuotetaan sovelluksia tiedoston muodostamiseen ja hyödyntämiseen. Standardeja noudattavat sovellukset mahdollistavat sen, että bittijonoista muodostuu hyödynnettäviä tekstidokumentteja, sähköposteja, valokuvia, äänitteitä, videoita tai rekistereitä ja muita tietoaineistoja.

Kansallisarkisto ei suorita tiedostojen käsittelyyn liittyvien sovellusten säilyttämistä, minkä vuoksi ohjeistuksen painopiste on tiedostojen muotoon sekä laatuun liittyvät kysymykset. Tämä vuoksi tiedostojen pitää olla hyödynnettävissä itsenäisesti ilman näitä alkuperäisiä sovelluksia. Tiedoston hyödyntäminen tai tulkinta ei toisin sanoen saa olla riippuvaista

sellaisista ulkoisista tietovarannoista, muista sovelluksista tai palveluista, joita ei säilytetä arkistoon tiedoston yhteydessä.

Itse standardit, formaatit ja niitä noudattavat tiedostot eivät huolellisesti toimittaessa katoa tai muutu käyttökelvottomiksi, jos eri tallennusmedioihin ja ratkaisuihin liittyvät riskit on asianmukaisesti huomioitu. Ennen aineiston luovuttamista tapahtuvasta luotettavasta tallennuksenhallinnasta on Kansallisarkiston erillinen ohje, jossa on esitetty keskeiset vaatimukset sen varmistamiseksi, että tiedostot säilyvät ehyinä myös tallennusratkaisujen mahdollisista riskitekijöistä huolimatta.

3. Tiedostojen hallinta ennen arkistoluovutusta

Tiedostomuotoja koskevia ohjeita ja tietoja pitkäaikaissäilytykseen kelpaavista tiedostomuodoista päivitetään säännöllisesti. Tätä kehitystä on hyvä seurata ja suunnitella omaa digitaalisen tiedon säilyttämistä myös tästä näkökulmasta.

Arkistoitavien tietoaineistojen elinkaaren eri vaiheissa pitää varmistaa, että aineisto säilyy teknisesti ja tietosisällöltään eheänä ja hyödynnettävänä. Elinkaaren eri vaiheissa tietojen vastaanottamisesta tai luomisesta, eri käsittely- ja jakeluvaiheiden jälkeen lähestyttäessä sitä vaihetta, jossa aineistot luovutetaan arkistoon, tulee huolehtia, etteivät muunnokset eri tiedostomuotoihin muodosta riskejä myöhemmälle käytölle. Erityisesti siinä vaiheessa, kun aineistoa luovutetaan arkistoon pitää varmistaa, ettei tiedostojen avaamista tai tallentamista uuteen muotoon ole teknisesti rajoitettu tai estetty - tiedostoja ei siis ole esimerkiksi salattu tai teknisesti lukittu muutoksilta.

Tiedostomuunnoksiin (tiedoston tallennusformaatin muuttaminen) liittyy riskejä, joissa voidaan mahdollisesti menettää joitain alkuperäisen tiedoston tai sisällön ominaisuuksia. Esimerkiksi muunnettaessa tekstinkäsittelysovelluksessa tuotettua dokumenttia, joka sisältää graafisia elementtejä, taulukoita tai muita vastaavia, PDF/A muotoon tulee varmistaa näiden elementtien luotettava tallentuminen uudessa muodossa. Muunnoksissa tulee siksi kiinnittää erityistä huomiota tietosisällön luotettavaan välittymiseen muunnosten ylitse ja ylimää räisiä rutiininomaisia muunnoksia tulisi välttää.

Kun harkittu tiedostomuunnos tehdään, se kannattaa sijoittaa laadunvalvonnan osalta tarkoituksenmukaiseen elinkaaren vaiheeseen. Laadua olisi hyvä pystyä tarkastelemaan pitkällä aikavälillä ja turvautua tarvittaessa laadun silmämääräiseen arviointiin. Erityisen vaikeita laadun varmistuksen osalta ovat suuria tiedostomääriä koskevat "massamuunnokset".

Tiedostojen käsittelyn tulisi aina olla valvottua ja muutosten teko tulisi olla mahdollista vain siihen oikeutettujen toimesta. Tiedostojen muuttumattomuuden seurannassa voidaan hyödyntää aineistosta erillään säilytettäviä eheystarkisteita. Eheystarkisteita hyödyntämällä voidaan osoittaa aineiston muuttumattomuus, mutta rikkoutunut tarkiste ei selvennä mitä

muutoksia dokumentissa on tehty. Erityisesti tulee huolehtia pöytäkirjojen ja muiden päätöksiä dokumentoivien tiedostojen hallinnasta siten, että voidaan varmistua tietosisällön todistusvoiman säilymisestä. Näissä dokumenteissa voidaan hyödyntää digitaalisia allekirjoituksia, sillä edellytyksellä etteivät nämä allekirjoitukset teknisesti estä myöhempiä arkistointitoimia. Tiedostojen muuttumattomuus voidaan varmistaa muun muassa säilyttämällä tiedostot järjestelmässä vain lukutilassa (*read-only*).

4. Tiedostomuodot ja näiden eri versioiden yksilöinti

Tiedostoformaatit uudistuvat ja kehittyvät nopeasti sovellusten kehittyessä. Sekä sovellukset että formaattien uudet versiot tuovat uusia mahdollisuuksia kehittyessään. Tiedostojen uudet formaattiversiot voivat tuoda uusia mahdollisuuksia tietojen esittämiseen ja formaatin sisäiseen toteutustapaan. Näiden piirteiden havaitseminen käyttäjän toimesta voi olla vaikeaa, mutta säilyttämistoimien osalta nämä erot voivat olla merkittäviä.

Sähköisen pitkäaikaissäilyttämisen toimien edellytyksenä on arkistoitavassa aineistossa olevien tiedostojen formaattien tarkka tunnistaminen sekä sen varmistaminen että tiedostot täyttävät formaatin laatukriteerit.

Normaalissa käytössä pelkän tiedostopäätteen perusteella ei voi päätellä tarkasti formaattia. Esimerkiksi yleisesti käytetystä PDF-formaatista (*PDF – Portable Document Format*) on kymmeniä alaversioita, joilla kaikilla on sama PDF-tiedostopääte (*tiedostonimi.pdf*). Kansallisarkisto on hyväksynyt osa näistä alaversioista ja voi varmistaa niitä noudattavien tiedostojen säilyvyyden.

Eri formaattien laatuun ja piirteisiin liittyvien eroavaisuuksien hallitsemiseksi formaattien eri versiolla on oma yksilöllinen tekninen tunniste (eng. *Persistent Unique Identifier, puid*). Tämän tunnisteen avulla voidaan luotettavasti viitata eri version ominaisuuksiin ohjeissa ja teknisissä formaattirekistereissä. Formaattirekistereiden avulla pidetään yllä teknistä tietokantaa standardien piirteistä sekä voidaan keskitetysti ohjata sekä tukea käsittelyyn liittyen säilyttämistoimia. Tässä ohjeessa hyödynnetään Iso-Britannian Kansallisarkiston digitaalisten säilytyspalvelujen tukemiseksi kehittämää PRONOM¹ formaattitietokantaa. PRONOM:ista löytyvät tekniset tiedot tiedostomuotojen rakenteesta sekä niiden tukemiseen tarvittavista ohjelmisto- ja laitteistoympäristöistä.

¹ <https://www.nationalarchives.gov.uk/PRONOM/>

5. Alkuperäisen tiedostomuodon hyödyntäminen

Arkistoinnin ja pitkäaikaissäilytyksen näkökulmasta hyväksytyt tiedostomuodot voidaan jakaa karkeasti kahteen kategoriaan, eli hyväksyttävät alkuperäiset tiedostomuodot sekä tiedostomuodot, joihin näistä alkuperäisistä tiedostomuodoista voidaan muuntaa, mikäli alkuperäinen muoto ei täytä säilytysvaatimuksia.

Alkuperäisen tiedostomuodon hyödyntäminen lisää usein merkittävästi tietojen myöhempää käytettävyyttä. Tallennusmuodoissa on omat merkittävät vahvuudet, joilla voidaan alkuperäinen ulkoasu ja tulkinta säilyttämään sen kustannuksella, ettei dokumentteja voida myöhemmin samalla tavoin muokata.

Hyväksytyjä tiedostoformaatteja on tässä jaettu tekstinkäsittelyn, taulukkolaskennan, esitysmateriaalien sekä kaavojen ja piirrosten mukaisesti. Näiden kaikkien eri formaattien osalta tulee huomioida se, että kukin näistä muodoista voi pitää sisällään erilaisia graafisia elementtejä, ääntä tai videota sekä muista digitaalisia objekteja, joilla voi olla aktiiviajan käytön kannalta merkitystä joko visuaalisuuden, aineiston tulkinnan tai tietoaaineistoon liittyvien toimintojen tueksi. Näiltä osin esimerkiksi esitysmateriaaliin liittyvien kuvatiedostojen tulee täyttää säilytyksen kriteerit samalla tavoin kuin silloin kuin nämä kuvat ilmenevät yksittäisinä tiedostoina.

Tekstinkäsittely ja muotoilemattomat tekstitiedostot

Formaatti	Standardi	Tunniste
odt, v.1.0	ISO 26300:2006	fmt/136
odt, v.1.1		fmt/290
odt, v.1.2	ISO 26300:2015	fmt/291
Microsoft Word	97-2003	fmt/40
Microsoft Word	2007-, ISO-29500-1	fmt/412

Muotoilematon tekstitiedosto formaattina ei sisällä ulkoasuun liittyviä muotoiluja, kuten fontteja, tehosteita tai graafisia elementtejä. Tekstissä voi olla asetteluja luettavuuden helpottamiseksi. Teksti noudattaa valittua merkistöä, joka määrittää käytössä olevien merkkien määrän sekä esitystavan. Aineistossa tulisi käyttää aina mahdollisimman laajaa merkistöä. Tekstimuotoisten aineistojen yhteydessä sallittuja ovat seuraavat merkistöt: ISO-8859-15, UTF-8, UTF-16 ja UTF-32. Käytettävän merkistön tunniste tulee toimittaa aineiston yhteydessä osana aineiston teknisiä metatietoja.

Muotoilematonta tekstiä voidaan hyödyntää rakenteisen tiedon esittämisessä. Rakenteen kuvaamisessa hyödynnettäviä tekstimuotoisia rakenteita ovat muun muassa XML, CSV, JSON, HTML tai vastaavat. Kun hyödynnetään edellä mainittuja rakenteita, tulee varmistaa, että tekstitiedosto itsessään on hyväksyttävässä muodossa. Lisäksi on varmistettava datarakenteen kuvaamisessa käytetyn esitystavan formaatinmukaisuus ja hyödynnettävyys. Yksityiskohtaiset vaatimukset rakenteisen tiedon esittämiseksi käyttäen tekstimuotoisia rakenteita on esitetty erillisessä ohjeessa.

Taulukkolaskenta

Esimerkiksi erityisesti taulukkolaskennan tietojen osalta... Mikäli tiedostot on muodostettu toisella taulukkolaskentasovelluksella ja ne tallennetaan säilytystä varten .ods-muotoon, tulee varmistaa sekä tietojen että käsittelyyn tarkoitettujen kaavojen toiminta.

Formaatti	Standardi	Tunniste
ods, v.1.0	ISO_26300:2006	fmt/137
ods, v.1.1		fmt/294
ods, v.1.2	ISO_26300:2015	fmt/295
Microsoft Excel	8	fmt/61
Microsoft Excel	8x	fmt/62

Esitysmateriaalit

Formaatti	Standardi	Tunniste
odp, v.1.0	ISO_26300:2006	fmt/138
odp, v.1.1		fmt/293
odp, v.1.2	ISO_26300:2015	fmt/293
Microsoft Powerpoint	97-2003	fmt/126
Microsoft Powerpoint	2007, ISO-29500-1	fmt/215

Kaavat ja piirto

Käyttötarkoitus	Formaatti	Standardi	Tunniste
Piirto	odg, v.1.0	ISO_26300:2006	fmt/139
"	odg, v.1.1		fmt/296
"	odg, v.1.2	ISO_26300:2015	fmt/297
Kaava	odf, v.1.0	ISO_26300:2006	
"	odf, v.1.2	ISO_26300:2015	

6. Muuntaminen erilliseen PDF-arkistomuotoon

PDF on standardoitu tekstimuotoisen dokumentin esittämiseen tarkoitettu, riippumaton ja avoin tiedostomuoto. Sen avulla alkuperäisen dokumentin ulkoasu sekä sisältö säilyvät muuttumattomina ja niitä voidaan tarkastella eri päätelaitteilla.

PDF:stä on eri käyttötarkoituksiin optimoituja aliversioita. Esimerkiksi PDF/A-versio poikkeaa tavallisesta perusmuodosta muun muassa siten, että dokumentissa käytettyjen fonttien esittämiseen tarvittavat tiedot on kapseloitu (eng. *embedded*) tiedoston sisään.

Formaatin vahvuus on visuaalisen ulkoasun ja rakenteen säilyttäminen. Ominaisuuksien rajoittamisen avulla varmistetaan, että dokumentin keskeinen rakenne ja sisältö säilyvät

käytettävänä pitkälle tulevaisuuteen. Rajoitettuja piirteitä ovat komentojonot (script), äänet, elävä kuva, salaus sekä hyperlinkit dokumentin ulkopuolelle.

PDF/A:sta on olemassa eri versioita (1, 2 ja 3). Lisäksi eri versioiden yhteydessä on alatakenteita (a, b ja u)

- a = all, eli tiedosto täyttää kaikki määrittämisen vaatimukset. Lisättynä erityisesti saavutettavuuteen liittyvät dokumentin rakenneosat.
- b = basic, täyttää keskeiset vaatimukset ja on siis hieman löyhempi a-tasoon verrattuna. Olennaisin ero on tässä se, että b keskittyy säilyttämään visuaalisen rakenteen.
- u = unicode tarkoittaa sitä lisäystä, että kaikille merkeille tulee löytyä unicode vastine.

Hyväksytyt versiot ovat

Formaatti	Standardi	Tunniste	Huomio
PDF/A-1a	ISO 19005:2005	fmt/95	
PDF/A-1b	ISO 19005:2005	fmt/354	
PDF/A-2a	ISO 19005-2:2011	fmt/476	Suosittelua
PDF/A-2b	ISO 19005-2:2011	fmt/477	Suosittelua
PDF/A-2u	ISO 19005-2:2011	fmt/478	Suosittelua
PDF/A-3a	ISO 195005-3:2012	fmt/479	
PDF/A-3b	ISO 195005-3:2012	fmt/480	
PDF/A-3u	ISO 195005-3:2012	fmt/481	

Hyödynnettäessä PDF/A-3 formaattia tulee huomioida, että formaatti mahdollistaa sen, että tiedostoon voidaan niin sanotusti upottaa tieto-objekteja (esimerkiksi kuvia) jotka eivät ole itsessään säilytyskelpoisissa säilytysmuodossa. Formaatin muodostajan vastuulla on varmistaa, että tiedostot ovat sekä PDF/A rakenteen, että siihen paketoitujen tietosisällön osalta kokonaisuutena säilytyskelpoisia.

7. Kuvatiedostot

Kuvatiedostot jaetaan tässä ohjeessa bittikartakuviin sekä vektorikuviin. Bittikartapohjaisia kuvatiedostoja muodostuu sekä digitoinnissa että digitaalisessa valokuvauksessa. Näissä kuva muodostuu yksittäisistä kuvapisteistä. Vektorikuvat muodostuvat laskennallisista vektoreista, jotka esitetään sovelluksen avulla viivoina, kaarina ja väritettyinä pintoina.

Kuvatiedostojen kohdalla tulee formaatin lisäksi erityisesti varmistua tiedoston sisällön teknisestä laadusta ja tarkkuudesta. Teknisellä laadulla tarkoitetaan esimerkiksi digitoinnin yhteydessä bittikartakuvan resoluutiota sekä muita ominaisuuksia, jotka vaikuttavat kuvan sisällön hyödyntämiseen. Vektorimuotoisten kuvien osalta tekniset vaatimukset kohdentuvat aineiston erityispiirteiden myötä tietosisällön teknisten piirteiden dokumentointiin.

Bittikartta

Formaatti	versio ja tunniste	Lisätiedot
JPEG (Joint Photographic Experts Group)	v1.00, fmt/42, v1.01, fmt/43, v1.02, fmt/44	Erityisesti digikuvauksessa hyödynnetty tiedostomuoto, että pakkausmenetelmä sekä täysväri (RGB) että harmaasävykuville.
JPEG2000	x-fmt/392	Tiedostoon voidaan tallentaa joko häviöllisesti tai häviöttömästi sekä harmaasävy että RGB ja CMYK-värikuvia. Verrattuna JPEG muotoon, tarjoaa JPEG2000 paremman tallennuksen tilasuhteen että kuvanlaadun.
TIFF (Tagged Image File Format)	v.6.0, fmt/353	Tiedostoon tulisi tallentaa vain pakkaamattomia mustavalko-, harmaasävy tai väritiedostoja.
GeoTIFF (Geographic Tagged Image File Format)	fmt/155	TIFF tiedoston laajennus, joka määrittää paikkatietoelementtien tallennuksen TIFF-kuvan metatietorakenteeseen. Paikkatietojen säilytysmuodoista laajemmin erillisessä paikkatieto-ohjeessa.
PNG (Portable Network Graphics)	fmt/13	Tiedostoon voidaan tallentaa 16-bittisiä harmaasävy tai 48-bittisiä värikuvia.
DNG (Digital Negative)	fmt/438, fmt/730	Digitaalivalokuvauksessa käytetty RAW-raakakuvatiedostojen säilytysmuoto. DNG tiedoston lisäksi pitää toimittaa säilytykseen myös vastaava TIFF tai JPEG2000 tiedosto. RAW tiedostot ovat erityisesti järjestelmäkameroissa käytetty mahdollisuus tallentaa muokkaamattomana kameran kennon tuottama informaatio. RAW tiedostot ovat laitevalmistajariippuvaisia eivätkä itsessään säilytyskelpoisia.

Vektorimuotoiset kuvatiedostot

Vektorikuvia tuotetaan ohjelmallisesti sovellusten avulla. Vektorigrafiikkaa hyödynnetään muun muassa rakennuspiirustuksissa ja paikkatiedossa. Vektorikuvat ovat numeerista

dataa, jota tulkitaan ja esitetään sovelluksen avulla käyttäjälle. Formaattien osalta keskeinen haaste on niiden riippuvuus laatimiseen käytetystä sovelluksesta sekä sovellusten erilaiset käytänteet formaattien tulkitsemiselle. Kun arkistoitava tietoa-aineisto on vektorimuotoista ja halutaan erityisesti säilyttää sen esitystapaan liittyviä piirteitä, tulee vektoritiedoston lisäksi toimittaa riittävä valikoima bittikarttapohjaisia kuvia keskeisistä aineistoon liittyvistä näkymistä.

Formaatti	versio ja Standardi	Tunniste
Scalable Vector Graphics (SVG), kaksiulotteisten vektorikuvien esittämiseen.	SVG 1.1, fmt/92	
Geography Markup Language (GML)	GML 3.2.1, fmt/1047	XML-pohjainen formaatti paikkatiedon esittämiseksi
Keyhole Markup Language (KML)	fmt/244	XML-pohjainen formaatti kaksi ja kolmiulotteisten karttojen annotointiin ja visualisointiin

8. Audiovisuaaliset tiedostomuodot

Audiovisuaaliset (AV) tiedostomuodot sisältävät ääni sekä videotallenteita. Osa AV-tiedostomuodoista on niin sanottuja säiliömuotoja, mikä tarkoittaa, että sama tiedosto sisältää tällöin sekä kuvaa että ääntä, joissa näillä on omat muotonsa, jotka yhdistyvät nämä kokoavaksi formaatiksi. Videotiedostoa katseleva ei tätä erityispiirrettä huomaa, mutta teknisesti sillä on merkitystä. Säiliömuotoa käytetään äänen ja kuvan muodostaman kokonaisuuden yhdistämiseksi ja että videota katsottaessa sekä ääni että kuva ovat varmasti samassa tahdissa keskenään.

Äänitiedostot

Formaatti	versio ja tunniste	Huomautus
AIFF (Audio Interchange File Format)	x-fmt/135	Vain häviötön LPCM koodattu ääni, Esiintyä itsenäisenä
BWF (Broadcast Wave Format)	Versio 2 fmt/527	Voi sisältyä ASF, AVI ja MXF säiliömuotoihin. Vain häviötön LPCM koodattu ääni
FLAC (Free Lossless Audio Codec)	1.2.1, fmt/279	Voi sisältyä ASF tai MOV-säiliömuotoihin.
MPEG-4 AAC		Voi sisältyä ASF, AVI, MPEG-1, MPEG-2, MXF, MJ2 ja MOV säiliömuotoihin
WAV	fmt/141	Esiintyä itsenäisenä

Videotallenteet

Formaatti	versio ja tunniste	Huomautus
MPEG-4 AVC –Advanced Video Coding (AVC)	fmt/199	Voi sisältyä säiliömuotoihin: Material Exchange Format [MXF] MPEG-1 part 1 Program Stream MPEG-2 part 1 Program Stream MPEG-2 part 1Transport Stream [MPEG-4 part 14
Digital Picture Exchange(DPX)	DPX, fmt/541	
JPEG 2000sekvenssi	x-fmt/392	Voi sisältyä säiliömuotoihin: Material Exchange Format [MXF]Motion JPEG 2000 [MJ2]