CSC

ICT Solutions for Brilliant Minds

Aineistot kohti pitkäaikaissäilytystä

Digitaalisten aineistojen säilyttäminen Kulttuuriperintö-PAS-palvelussa

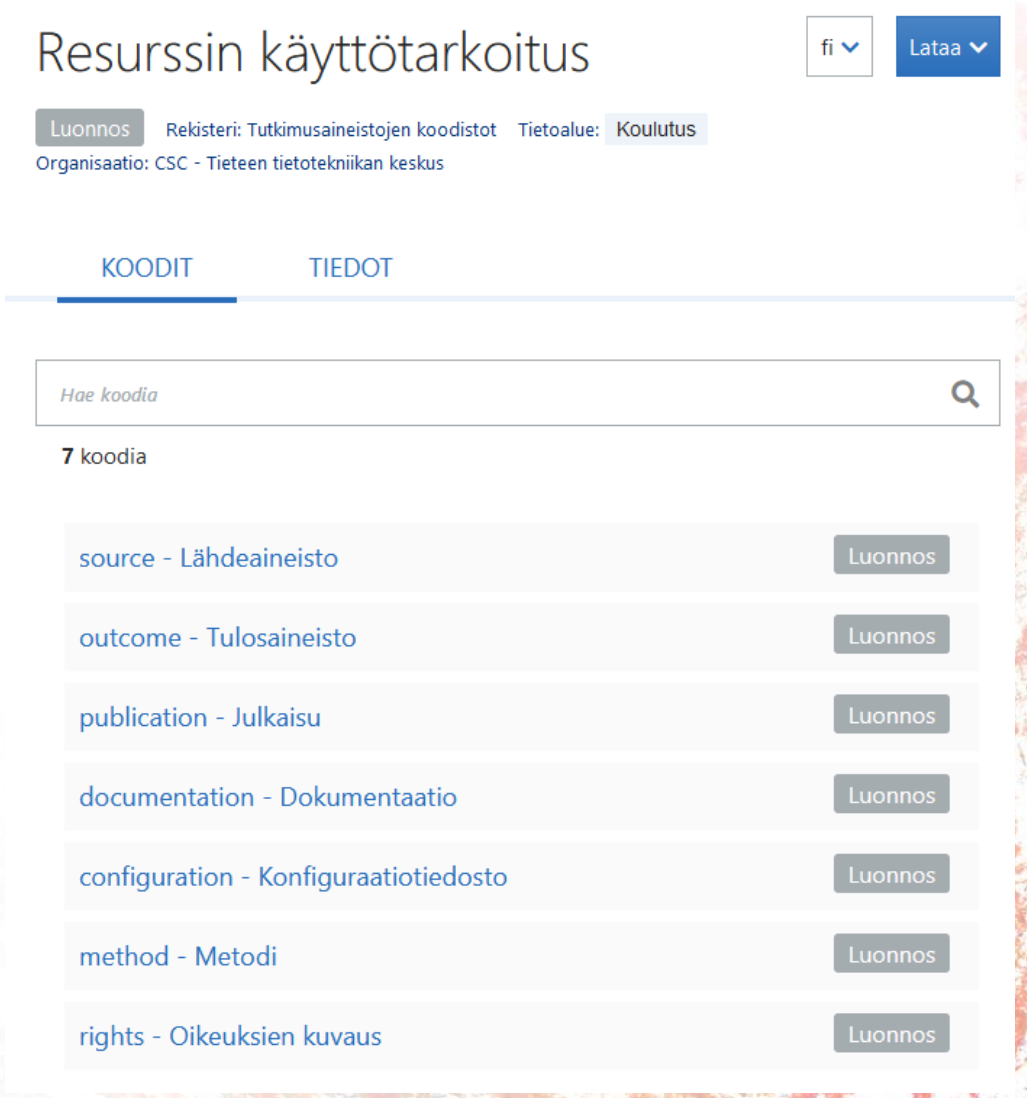
4.10.2022

Johan Kylander

Digitaalisen aineiston hallinta

Digitaalisten aineistojen jäsentäminen

- Digitaalisen aineiston hallinta lähtee siitä, että tunnistetaan mitä digitaalista aineistoa omistetaan ja missä muodossa se on
- Aineistoa pitää dokumentoida, varustaa metatiedoilla, jotka kertovat mistä on kyse
- Digitaalista aineistoa pitää osata jäsentää, esimerkiksi aineistokokonaisuus voi koostua useasta erityyppisestä tiedostosta
- Esimerkiksi oheisdokumentaatioita, joka kuvaa tai täydentää digitaalista aineistokokonaisuutta, on syytä säilyttää



Aineiston synty

- Aineiston syntyhistoria kertoo paljon siitä, miksi digitaalisella aineistolla on tiettyjä piirteitä
  - Esim. skannattu kuva on erilainen verrattuna kameralla otettuun kuvaan
  - Eri PDF-ohjelmistot tuottavat piirteitään erilaisia PDF-dokumentteja
- Syntyhistoriaa ei välttämättä voida rekonstruoida myöhemmin
- Digitaalisen aineiston laatua ja kestävyyttä pitää miettiä jo alusta saakka
  - Tiedostomuodot hallintaan jo luontivaiheessa
- Korkealaatuinen digitaalinen aineisto alkaa jo suunnittelusta

Pitkäaikaissäilyttäminen osana digitointivaihetta

Aineiston säilyttämisen suunnittelu osaksi digitaalisen aineiston syntyä:

1. Laatu
   - Korkeampi laatu kestää pidempään
2. Tiedostomuodot
   - ”Turhia” normalisointeja kannattaa välttää
3. Metatiedot
   - Aineiston synnyn ja digitointiprosessin dokumentointi
4. Käyttötarkoitus
   - Soveltamisohjeilla voidaan päästä haluttuun lopputulokseen

Mikä on tiedostomuoto?

- Tiedostomuoto ilmaisee tietokoneiden tallennusmedialle, esimerkiksi kiintolevylle talletetun tiedoston tallennusmuodon eli rakenteen (wikipedia)
  - Jono ykkösiä ja nollia tietyssä järjestyksessä, joille tiedostomuoto antaa merkityksen miten ne tulee tulkita
- Käyttäjät ”tunnistavat” tiedoston yleensä tiedostopäätteen avulla
  - .pdf, .docx, .mov, …
  - Käytännössä kuitenkin täysin riittämätöntä
- Tietokoneohjelmistot osaavat lukea ja käsitellä tukemiaan eheitä tiedostoja

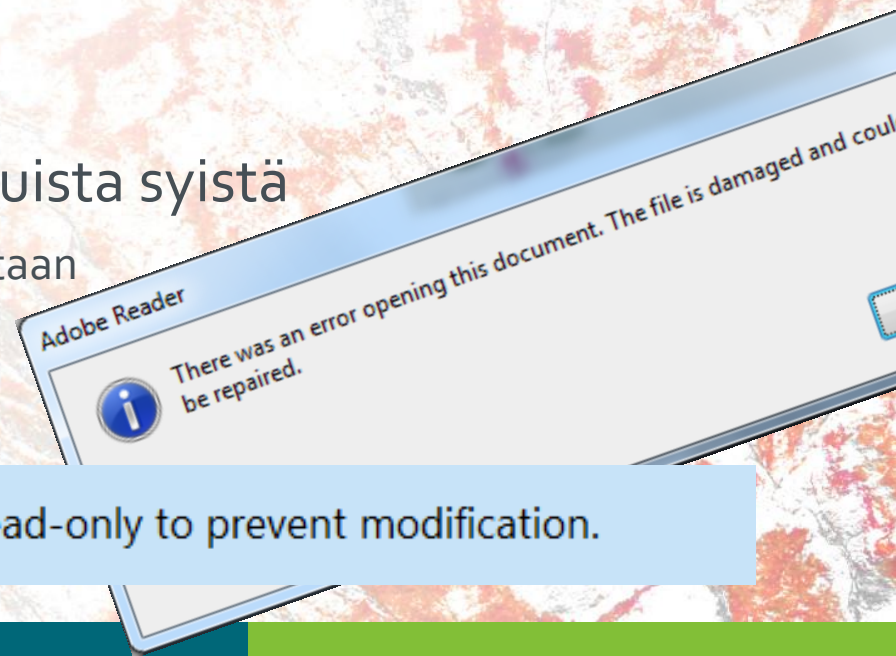
Tiedostomuodot

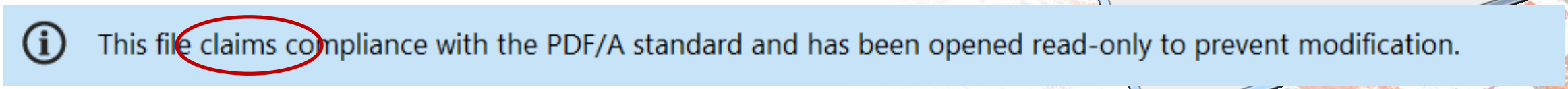
Meitä on enemmän kuin tarpeeksi…

- Mime-tyypit (~2000 rekisteröityä tiedostomuotoa (IANA))
  - Tunniste on merkkijono: application/pdf, text/plain, audio/x-wav …
  - Ei versioita, joten siltä osin kattaa huomattavasti suuremman joukon tiedostomuotoja
  - Omia mime-tyyppejä on mahdollista käyttää: application/x-mun-oma-formaatti
- PRONOM tiedostomuotokirjasto (~1400 rekisteröityä tiedostomuotoa)
  - Pysyvät tunnisteet tiedostomuodoille: fmt/431, fmt/569, …
  - Tiedostomuodon eri versioille omat tunnisteet
  - Pysyvät tunnisteet UK:n kansallisarkiston (TNA) myöntämiä; omien käyttäminen ei mahdollista
- Lisäksi lukuisia tiedostomuotoja joille ei ole mime-tyyppiä eikä PRONOM tunnistetta
  - Erityisesti laitevalmistajien omat eksoottiset suljetut tiedostomuodot
  - Osittain myös uudemmilta tiedostomuodoilta puuttuu toinen tai molemmat

Tunnistaminen & validointi

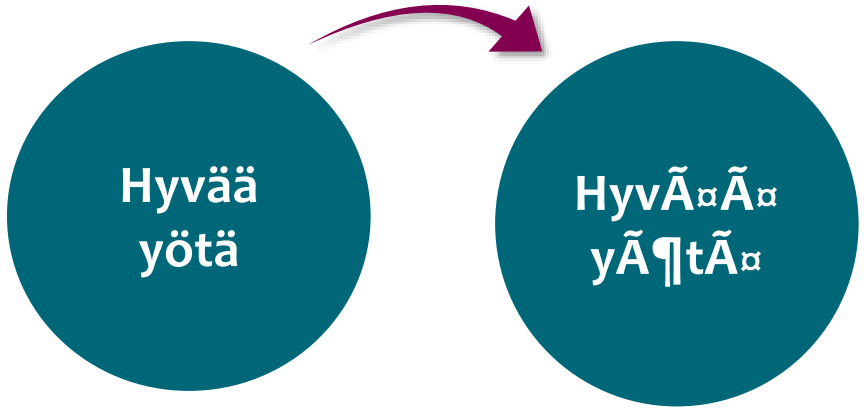
- Tiedostomuodon tunnistaminen ei normaalisti ole riittävä toimenpide
  - On myös varmistuttava, että tiedosto on ko. tiedostomuodon määrittelyn mukainen
  - Jotkut ohjelmistot avaavat tiedoston normaalisti vaikka se olisi pilkulleen oikein
- PAS-palvelu validoi kaikki palveluun siirrettävät tiedostot…
  - Ja pienenkin virheen löytyessä PAS-palvelu ei ota säilytysvastuuta tiedostosta
  - Mahdollistetaan automaattiset massamigraatiot tulevaisuudessa
  - Validointi ei aina ole täydellistä (puutteet validointityökaluissa)
- Tietyissä tapauksissa validointi voidaan ohittaa perustelluista syistä
  - Korjaaminen voi olla mahdotonta, mutta tiedoston säilyttäminen katsotaan välttämättömäksi/tarpeelliseksi
  - Tällöin tiedosto otetaan vain bittitason säilyttämiseen





Tekstitiedostojen merkistöistä

- Monet tiedostomuodot sisältävät metatietoja miten tiedosto tulee tulkita
- Tekstitiedostoissa (plain text) ei mitään metatietoja ole, mutta käytettävissä olevia erilaisia merkistöjä on paljon
  - ANSI, UTF-8, ISO-8859-#, Windows-1257, …
- Käytettävä merkistö on tallennettava johonkin muualle, jotta tiedosto voidaan tulkita oikein



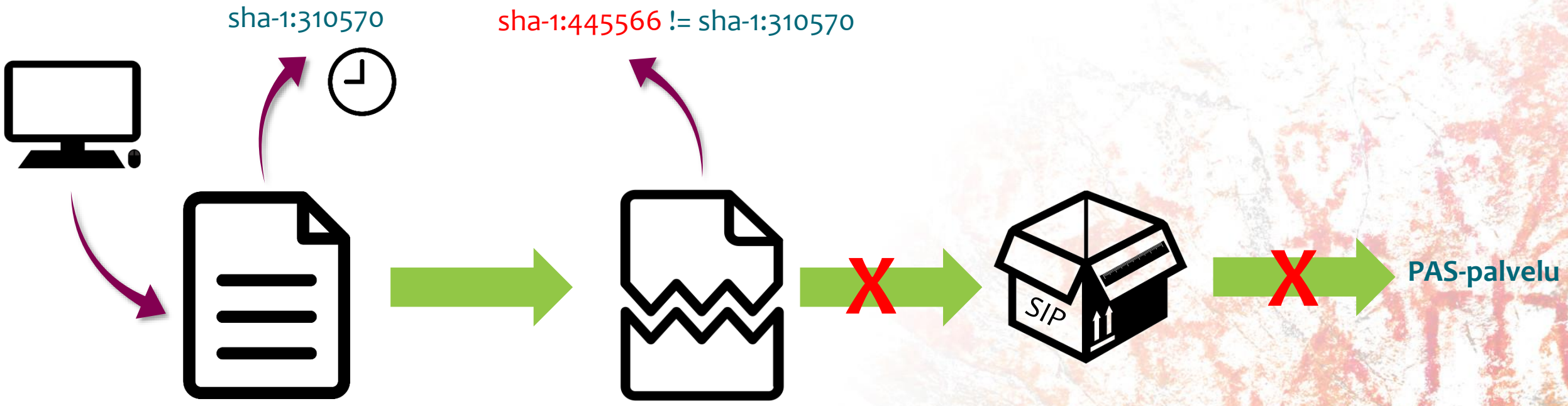
Tarkistussummat

Tiedostomuodosta riippumaton sormenjälki

- Tiedoston tarkistussummalla voidaan varmistua, että tiedosto ei ole ajansaatossa (tahattomasti) muuttunut
  - Mutta ei siis takaa etteikö se voisi muuttua, mutta mahdolliset muutokset voidaan huomata koneellisesti
  - Jos muutoksia on tapahtunut, niin muuttunut tiedosto voidaan korvata eheällä kopiolla varmuuskopiosta (…jos sellainen on…)
- Tarkistussumma tulee laskea tiedostolle mahdollisimman aikaisessa vaiheessa
  - Aikaleima myös talteen
- PAS-palvelun tukemat tarkistussummat
  - md5, sha1, sha224, sha256, sha384 ja sha512
  - Kaikille käyttöjärjestelmille löytyy näiden laskemiseen valmiit ohjelmistot, joten ei tarvitse ymmärtää algoritmien toimintaa

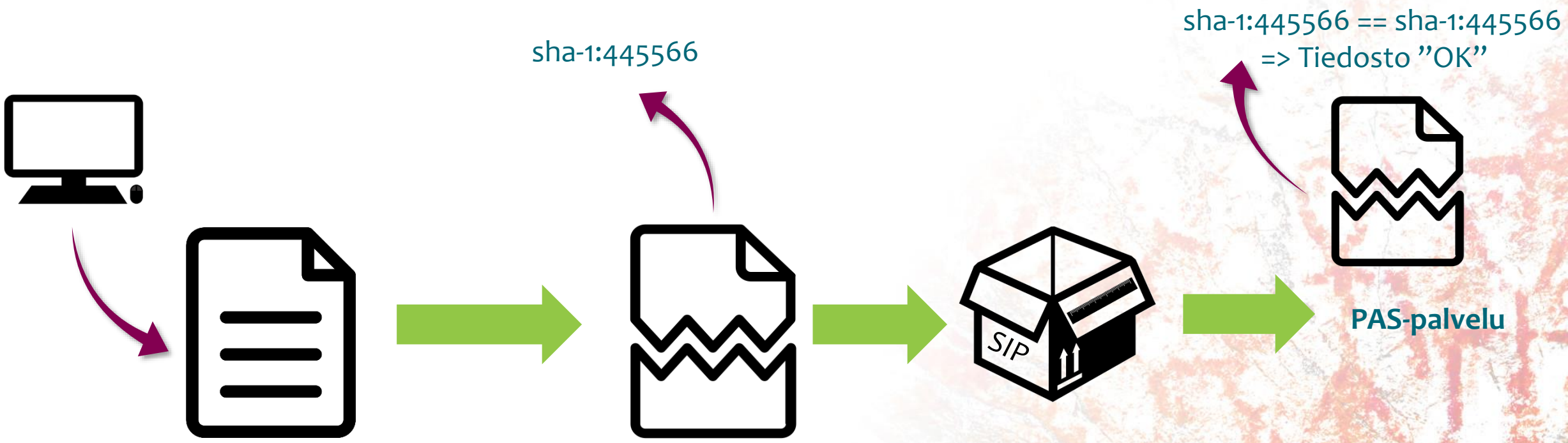
Tarkistussummat

Näin...



Tarkistussummat

Ei näin...



Tekninen metatieto

- Tiedostomuoto (mimetyyppi ja versio)
- Tiedoston eheystieto (tarkistussumma)
- Aineistotyyppikohtaiset tiedot (merkistö, kuvan korkeus, äänen näytteenottotaajuus jne.)

- (Lähes) kaikki tekninen metatieto on luettavissa tiedostoista
- Tarkistussumma on kuitenkin eheyden kannalta hyvä ottaa talteen mahdollisimman aikaisessa vaiheessa

Tapahtumahistoria

- Aineiston tapahtumahistorialla kerrotaan mitä aineistolle on tapahtunut ja milloin
- Tapahtumahistoria esitetään tyypillisesti tapahtumina
- Tapahtumahistoria selittää miksi digitaalinen aineisto on juuri tämänkaltainen (onko se digitoitu, alkujaan digitaalinen, onko aineistoa muokattu, millä ohjelmistoilla tiedosto on käsitelty)
- Tapahtumahistoriaa on käytännössä mahdotonta luoda retroaktiivisesti
- Varsinkin aineiston synty (laite, ohjelmisto) on yleensä vaikeaa luoda jälkikäteen

Hyödyntäviä organisaatioita ohjaavat PAS-määrittely

- Yksi PAS-palvelun näkyvimpiä osia hyödyntäville organisaatioille
- PAS-määrittelyt on tehty tiiviissä yhteistyössä hyödyntävien organisaatioiden kanssa

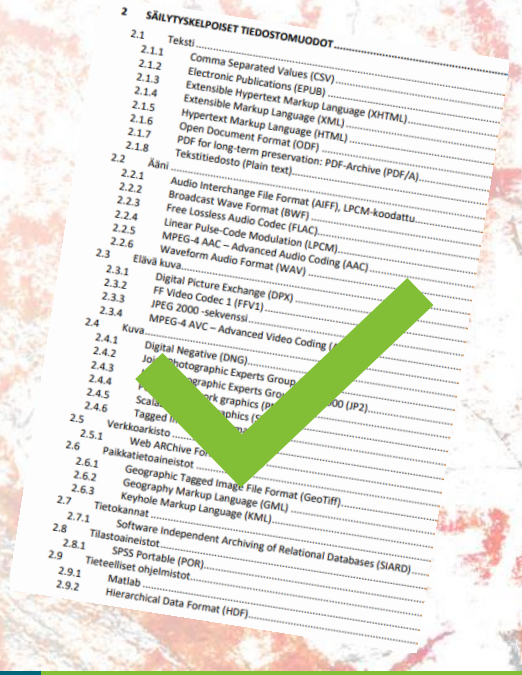
Kansallinen tiedostomuotojen määrittely

https://urn.fi/urn:nbn:fi-fe2020100578095

- Säilytys- ja siirtokelpoiset tiedostomuodot määrittely
  - ”Recommended” vs. ”acceptable for transfer”
  - Yleisesti hyväksytty jako kansainvälisesti
    - Toki muitakin lähestymistapoja on olemassa
- Tietyin ehdoin myös muita tiedostomuotoja voidaan ottaa säilytykseen
  - Organisaation pitää itse pystyä arvioimaan säilyttämisen riskit pitkällä tähtäimellä
  - Asiasta on kuitenkin aina sovittava erikseen
- Päivittyvä määrittely
  - Viime vuosina päivittynyt varsin maltillisesti
  - Määrittelyä päivitetään yhteistyössä hyödyntävien organisaatioiden kanssa
    - Mutta päivitykset aina tarve edellä

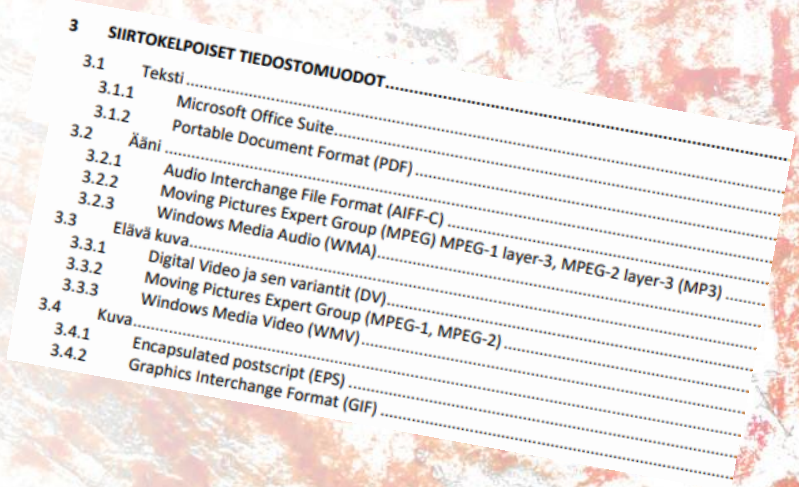
Säilytyskelpoiset tiedostomuodot

- Säilytyskelpoinen tiedostomuoto on sellainen, jonka tietosisällön säilyminen ja ymmärrettävyys voidaan taata pidemmällä aikavälillä
  - Tällä hetkellä yhteensä 32 tiedostomuotoa
- PAS-palvelu vastaanottaa säilytyskelpoisia tiedostomuotoja siirtopaketeissa
- PAS-palvelu lupaa voivansa migroida hallitusti ja suunnitellusti
- Tiedostomuodot on jaettu aineistotyypittäin, esimerkiksi
  - Kuva: JPEG, PNG, SVG, TIFF …
  - Teksti: plain text, CSV, ODF, PDF …
  - Tietokannat: SIARD, POR
  - Ääni: BWF, FLAC …
  - Elävä kuva: FFV1, AVC …



Siirtokelpoiset tiedostomuodot

- Siirtokelpoiset tiedostomuodot ovat sellaisia, joiden tietosisällön säilymistä tai ymmärrettävyys ei voida taata pidemmällä aikavälillä
  - Mutta joita on kuitenkin huomattavasti organisaatioilla
  - Siirtokelpoisia tiedostomuotoja ei pitäisi enää tuottaa
- Käytännössä siirtokelpoinen on pakko migroida (joskus) ensiksi säilytyskelpoiseen, eli se on askeleen säilytyskelpoisen jäljessä
- Käytössä vastaava jako aineistotyypittäin kuin säilytyskelpoisissa muodoissa
  - Kaikilla aineistotyyppiryhmillä ei kuitenkaan ole siirtokelpoisia muotoja



Yleiset rajoitukset tiedostomuodoille

- Tiedostoissa ei saa käyttää salasanasuojauksia eikä mitään muita salaustekniikoita
- Tiedostoissa ei saa käyttää DRM (Digital Rights Management) -tekniikoita
- Tiedostoja ei saa allekirjoittaa digitaalisesti, jos se estää tiedoston käsittelyn
- Tiedostoja ei saa (tarpeettomasti) pakata
- Tiedostosta ei saa puuttua sen esittämiseen tarvittavia ulkoisia komponentteja

Tiedostomuoto vs. sen soveltaminen

- Pelkkä tiedostomuoto ei takaa, että säilytettävä sisältö säilyy tarkoituksen mukaisesti, vaan sitä osattava soveltaa käyttötarkoituksen mukaisesti
  - Skannatussa asiakirjassa värisävyjen laadulla ei välttämättä ole suurta merkitystä, jos kyseessä on esimerkiksi konekirjoitettu asiakirja (kuvan pakkausmenetelmä voi kuitenkin vaikuttaa automaattiseen tekstin tunnistamiseen!)
  - Valokuvassa värien laadulla on yleensä aivan toisenlainen merkitys
- Pakkaamaton vs. pakattu
  - Muutamat tiedostomuodot mahdollistavat pakkaamisen (esim. jpeg)
  - Kannattaa aina käyttää mahdollisimman hyvää laatua ja pakata tiedostoa mahdollisimman vähän, erityisesti jos kyseessä on häviöllinen pakkaaminen (kuten jpeg)
  - Hyvä laatuinen voidaan aina myöhemmin muuntaa heikompi laatuiseksi, mutta muunnos toiseen suuntaan ei yleensä ole mahdollista

Määrittelyt saatavilla

Myös englanniksi

- http://digitalpreservation.fi/specifications
  - Aineistojen ja niiden metatietojen paketointi pitkäaikaissäilytykseen
  - Säilytys- ja siirtokelpoiset tiedostomuodot
  - PAS-palveluiden rajapinnat
- Vuosittaiset laaturaportit ja muita julkaisuja
- Ohjelmistokoodia, skeemat & schematron säännöt saatavilla @GitHub
  - https://github.com/Digital-Preservation-Finland/

Mitä pitää ottaa huomioon jo ennen PAS-palveluun siirtymistä?

- ✓Laske tarkistussummat tiedostoille mahdollisimman aikaisessa vaiheessa
- ✓Huolehdi laadukkaista metatiedoista
- ✓Huomioi tiedostomuodot-määrittely ja mahdolliset soveltamisohjeet
- ✓Dokumentoi aineiston syntyhistoria (digitointi, …)

Ø Älä muodosta siirtopaketteja ”varastoon”

Kysyvä ei tieltä eksy

Muuta huomioitavaa

- Kaikki toimenpiteet tiedostoille pitää dokumentoida koko elinkaaren ajan
  - Mitä tehtiin, koska, miksi, kuka teki,…
  - Myös ennen säilyttämisen aloittamista (ja erityisesti silloin)
  - Tapahtumat sanasto: https://www.digitalpreservation.fi/specifications/vocabularies
- Tiedostojen nimeämiseen kannattaa kiinnittää huomiota
  - ”presentation1.pptx” vs ”pas-museot-webinaari-2022-10-04-jkyl.pptx”
  - Paljon muita konkreettisia ohjeita webinaarissa: https://youtu.be/Xkqkg1oiUOQ

Mitä taustajärjestelmään?

- Linkki aineiston kuvailun ja tiedostojen välillä
  - Mahdollisimman tarkka polku (ei pelkkä tiedostonnimi)
- Tiedoston tiedot:
  - Tarkistussumma (sekä käytetty algoritmi että laskettu summa)
  - Tiedostomuoto ja -versio
  - Tiedostokohtaiset tunnisteet
- Tapahtumahistoria
  - Tärkeimmät tapahtumat ja linkitykset tiedostoihin
  - Suorittaja (ohjelmisto) mukaan
- Kirjanpito siitä, mitä aineistoja on viety PAS-palveluun!


YOU ARE NOT ALONE...
pas-support@csc.fi
digitalpreservation.fi
@dpres_fi