CSC

ICT Solutions for Brilliant Minds

Perustietoa validoinnista

PAS-palveluiden hyödyntäminen käytännössä -koulutus

25.4.2023

Johan Kylander

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

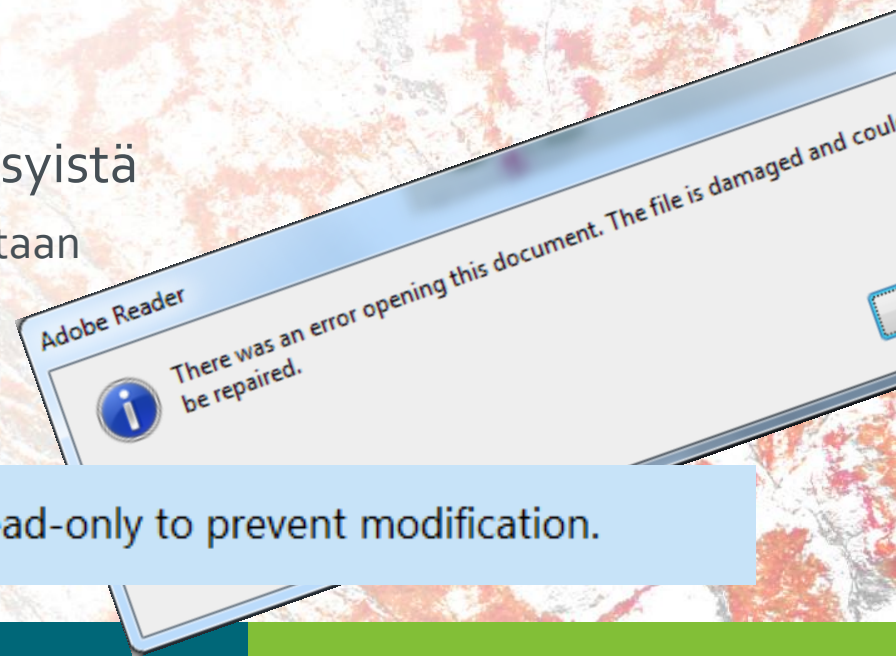
Tiedostomuodot

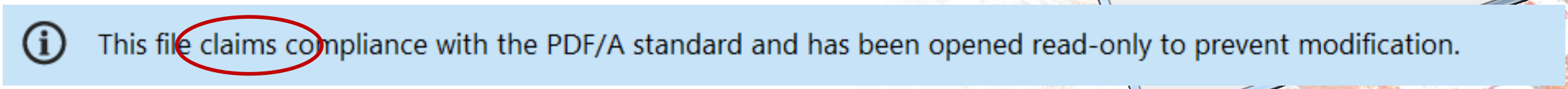
Meitä on enemmän kuin tarpeeksi…

- 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

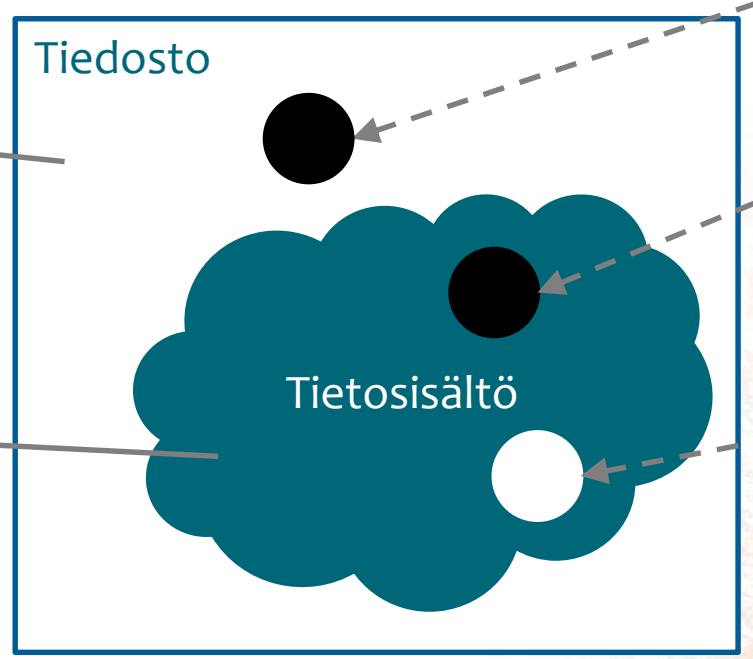




Virhe tiedostossa

Tietokoneen tarvitsema hyödyntäjälle näkymätön sisältö
- Yleensä liittyy tiedostomuotoon

Kohdeyleisön hyödynnettävissä oleva pitkäaikaissäilytettävä tietosisältö



Virhe tiedostossa, mutta ei tietosisällössä

Virhe tiedostossa ja tietosisällössä

Virhe tietosisällössä, mutta ei tiedostossa
- Semanttisen säilyttämisen asia
- Ei ole validaattorin mielestä virhe

Esimerkkejä eri virhetilanteista

- Tiedosto ja tietosisältö viallinen
  - Tiedoston siirto on jäänyt kesken
  - Videotiedoston kuvaframe ei ole tallentunut tiedostoon kokonaan (tavuja puuttuu)
- Tiedosto viallinen, mutta ei tietosisältö
  - EXIF-metatiedoista puuttuu pakollinen EXIF-määrityksen versionumero
  - Syntaksivirhe XML/HTML-tiedostossa
- Tietosisältö viallinen, mutta ei tiedosto
  - *Semanttisen tason asia: Validaattori ei ilmoita, ei ole tiedostovirhe*
  - Kirjoitusvirhe kirjan tekstissä
  - Videokuvassa näkyvät sisällölliset häiriöt (kohina, värienvaihtelu)
  - Valokuvassa osa kohteesta jäänyt kuvan ulkopuolelle

Esimerkkejä validointiin liittyvistä virheistä

(Joihin olemme PAS-palvelussa törmänneet)

- PDF, PDF, PDF, PDF, PDF, PDF, PDF, PDF, ...
  - PDF haasteita on suhteessa muihin haasteisiin huomattavasti
  - Tyypillisesti liittyvät versioihin, eli tiedosto väittää olevansa versiota X, mutta käyttää ominaisuuksia joita ko. versiossa ei ole
- Jotkut (jopa ammattitason) kamerat tekevät virheellisiä TIFF tiedostoja
- Videossa väärä kuvasuhde
  - Korjattavissa melko helposti, mutta ei kuitenkaan ilman dekoodausta, joka taas saattaa kasvattaa tiedoston kokoa huomattavasti (tai heikentää laatua)
- Tekstitiedostojen merkistöt ja niiden tunnistaminen

Kokonaismäärään verrattuna näitä on kuitenkin melko vähän

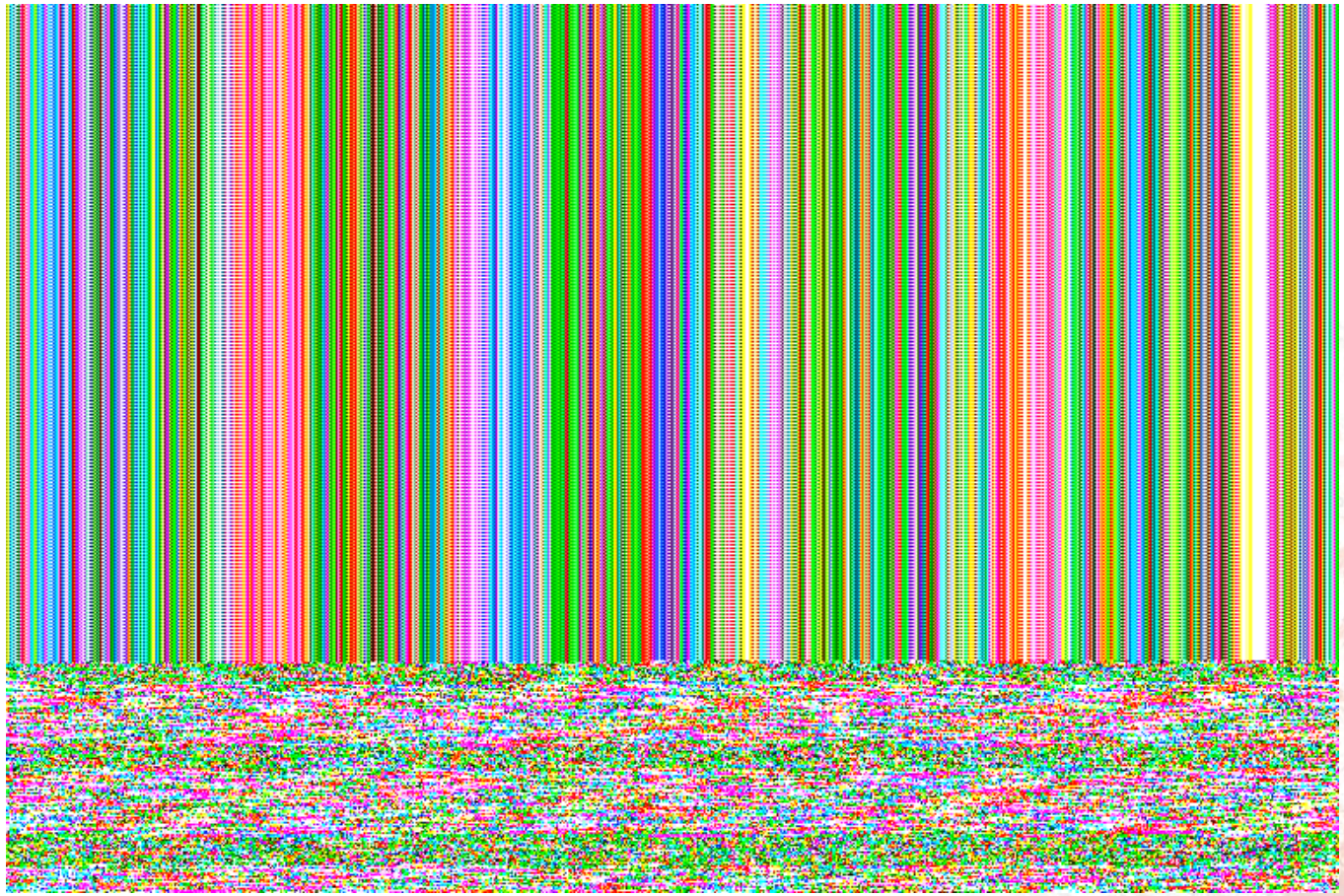
Esimerkki

Alkuperäinen kuva

Viallinen tiedosto avattuna:
Vain yksi bitti on muuttunut

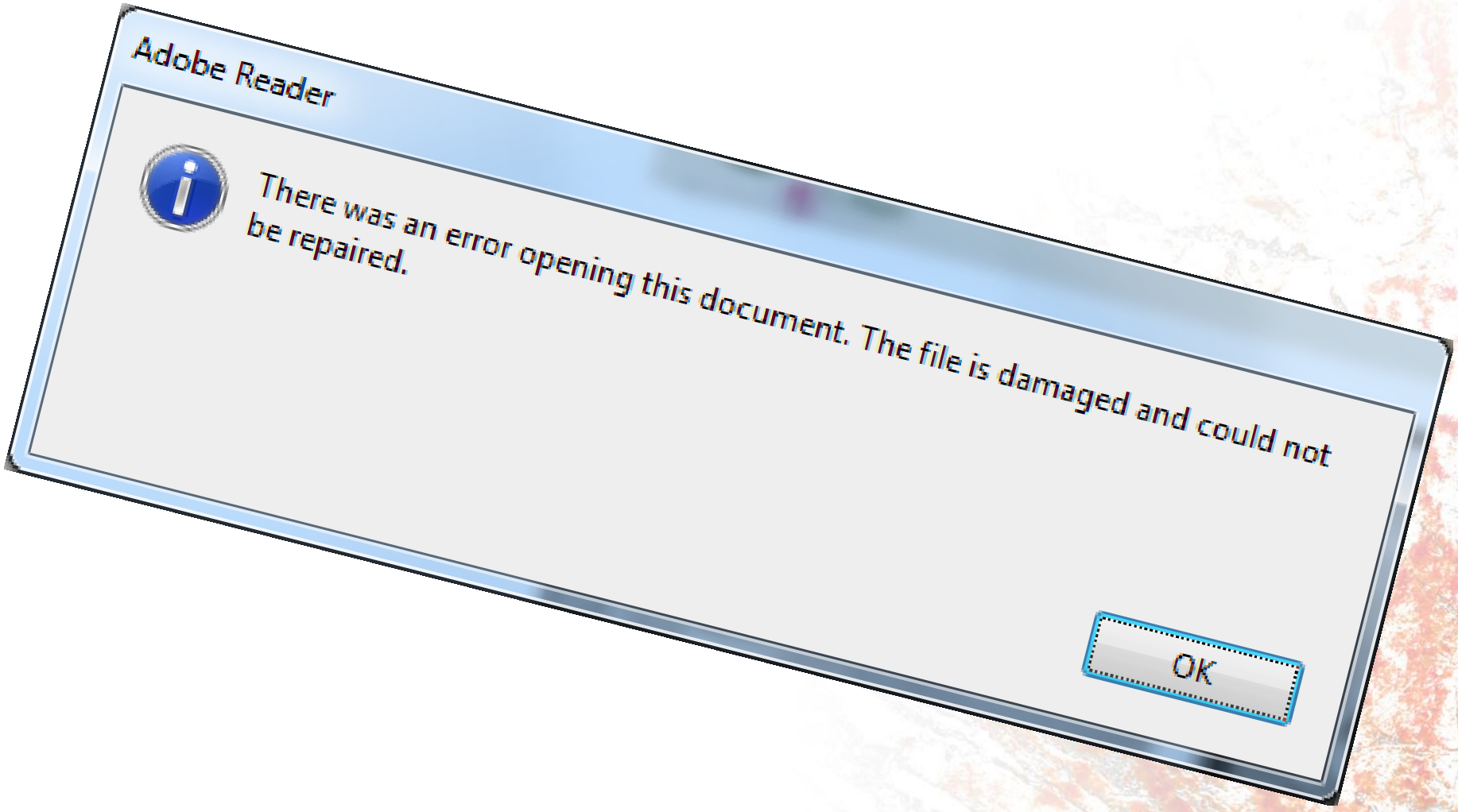
…molemmilla sama tiedostokoko…

Esimerkki viallisesta kuvasta



Source:
Ikou, "A corrupted PCD photo file, converted to png.", Wikimedia Commons, Creative Commons CC0 1.0 Universal Public Domain Dedication.

Tiedosto ei avaudu, koska viallinen



Tarkistussumma ja tietosisältö

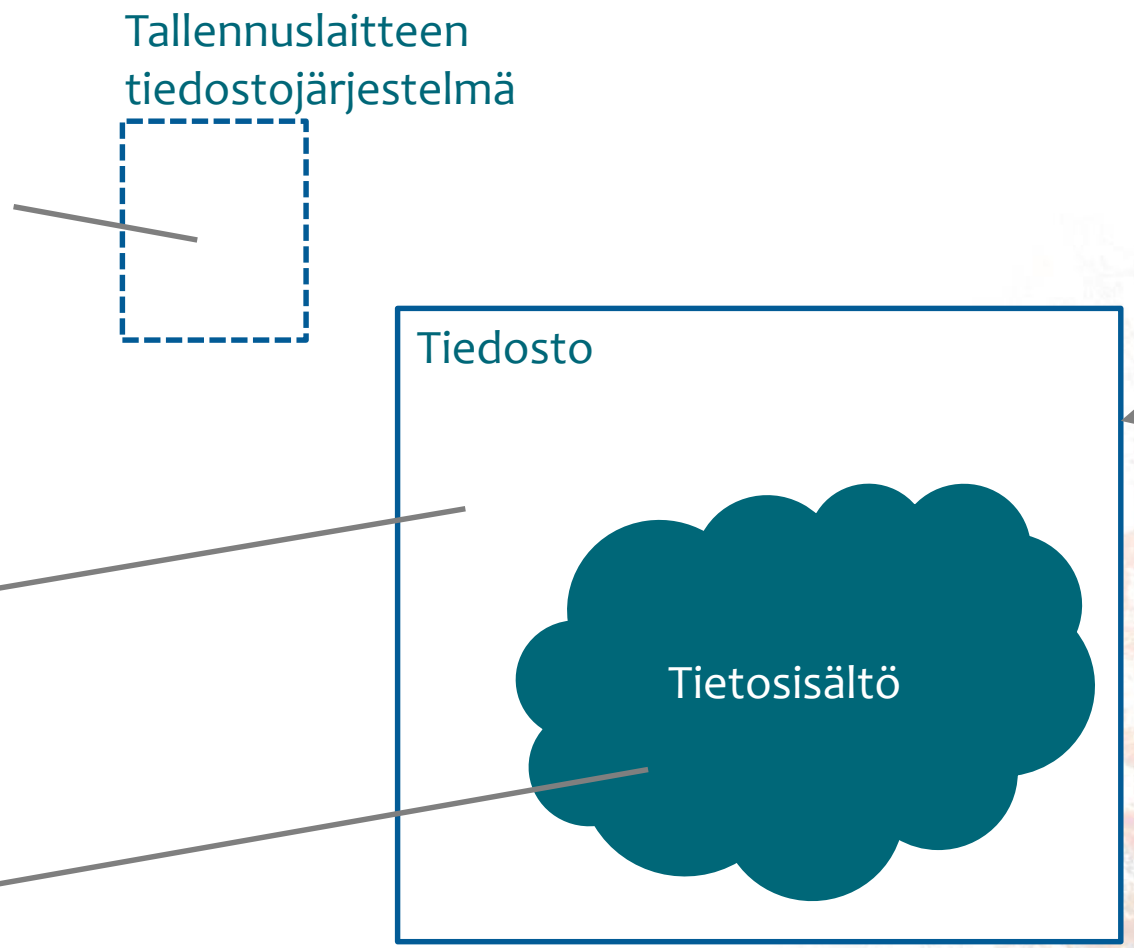
Tallennuslaitteen järjestelmän ylläpitämät tiedostoon sisältymättömät tiedostoa koskevat tiedot, kuten

- tiedostopolku ja –nimi
- tiedoston aikaleimat
- tiedoston luku- ja kirjoitusoikeudet

Tietokoneen tarvitsema hyödyntäjälle näkymätön sisältö

- Yleensä liittyy tiedostomuotoon

Kohdeyleisön hyödynnettävissä oleva pitkäaikaissäilytettävä tietosisältö



Tarkistussumma lasketaan koko tiedostosta. Summa muuttuu minkä tahansa tiedostossa olevan tiedon muuttuessa.

- Summa voi muuttua, vaikka säilytettävään tietosisältöön ei tulisi muutoksia.
- Summa ei muutu tallennuslaitteen järjestelmän tietojen muuttuessa (esim. tiedostonimi).

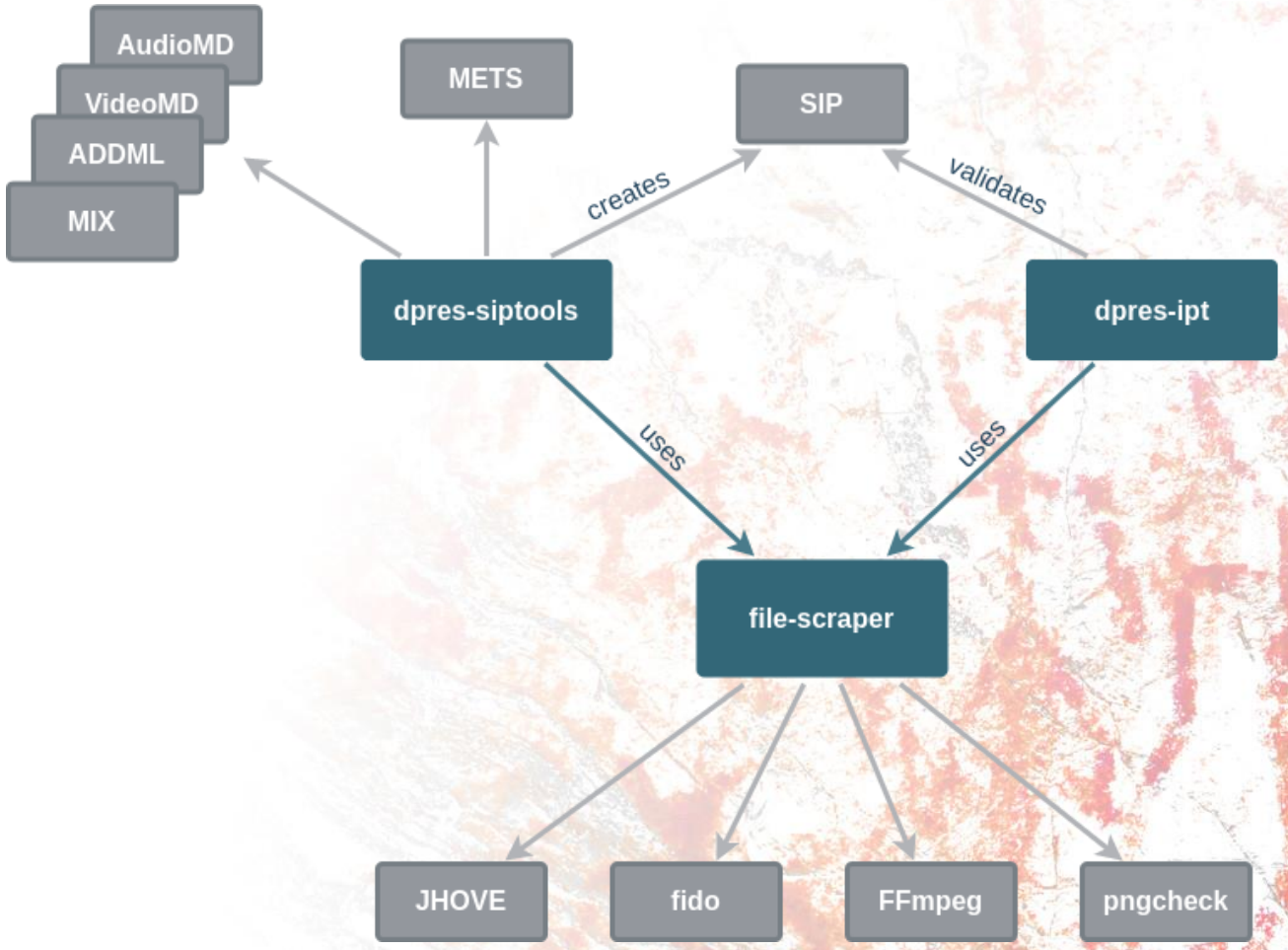
Looginen taso ja bittien säilyttäminen

Aineisto pitää säilyä käyttökelpoisena

- Valittu säilyttämisen taso käytännössä vaikuttaa vastuujakoon
  - Bittitason säilytykseen tulevan aineiston loogisen säilyttämisen vastuu on hyödyntävällä organisaatiolla
  - PAS-palveluiden arkkitehtuuriperiaate 7: *PAS-palvelut vastaavat säilytettävien aineistojen loogisesta ja bittitason säilyttämisestä sekä loogisen säilyttämisen tehtävien jaon sopimisesta hyödyntävien organisaatioiden kanssa.*
- Jos tiedosto on teknisesti rikki tai tuntematonta muotoa
  - Migraatiota ei välttämättä pystytä toteuttamaan normaalilla/tunnetulla migraatioprosessilla
  - Tiedoston kanssa voidaan joutua umpikujaan, jossa avaaminen on mahdotonta
  - Vaikka tiedostot saadaan tänä päivänä auki, näin ei välttämättä ole tulevaisuudessa
- Ohjelmistotuen kadotessa myöskään ehjiä tiedostoja ei välttämättä enää saa auki ilman loogista säilyttämistä

Paketointikomponentti, tiedostojen anaalyysityökalu, IPT

- Paketointikomponentti luo siirtopaketteja.
- IPT (dpres-ipt) validoi siirtopaketit.
- Tiedostojen käsittely on delegoitu anaalyysityökalulle (file-scraper) :
  - Tiedostomuotojen tunnistaminen, teknisten metatietojen keruu tiedostoista, tiedostojen eheyden tunnistaminen, tuki (PAS-palvelussa)
  - Työkalu on käytössä paketointikomponentissa ja IPT:ssä.



Tiedostojen analyysityökalun toiminta

- Työkalulle syötetään tiedosta ja mahdollisia argumentteja (kuten tiedostomuoto ja –versio, validoinnin ohittaminen)
- Jos tiedostomuotoa ei syötetä työkalulle, muoto tunnistetaan työkalussa.
- Annetun tai tunnistetun muodon avulla valitaan sopivat komponentit ja metatietomallit
  - Jos validointi ohitetaan, vain ne komponentit, jotka tunnistavat tiedostomuodon ja keräävät metatietoja, ajetaan
  - Analyysityökalu normalisoi yksittäisten työkalujen metatiedot ja yhdistävät ne.
- Jos tiedostoa validoidaan, työkalu ilmoittaa tiedoston eheydestä sekä kaikki mahdolliset virheet jos tiedosto on rikki.
  - Jos yksikään komponentit ilmoittaa virheistä, tiedosto on rikki
  - Jos ei löydy sopivaa komponenttia, tiedosto ei ole tuettu

Analyysityökalun tuloste


```
{
    "path": "tests/data/audio_x-wav/valid_2_bwf.wav",
    "MIME type": "audio/x-wav",
    "version": "2",
    "metadata": {
        "0": {
            "index": 0,
            "mimetype": "audio/x-wav",
            "stream_type": "audio",
            "version": "2",
            "audio_data_encoding": "PCM",
            "bits_per_sample": "8",
            "codec_creator_app": "Lavf56.40.101",
            "codec_creator_app_version": "56.40.101",
            "codec_name": "PCM",
            "codec_quality": "lossless",
            "data_rate": "705.6",
            "data_rate_mode": "Fixed",
            "duration": "PT0.86S",
            "num_channels": "2",
            "sampling_frequency": "44.1"
        }
    },
    "grade": "fi-dpres-recommended-file-format",
    "well-formed": true
}
```


- Tiedostomuoto ja –versio
- Kerätyt metatiedot
- Tuki PAS-palvelussa
- Tiedoston eheys
- Mahdolliset virheet

Analyysityökalu hyödyntää kolmannen osapuolten komponentteja

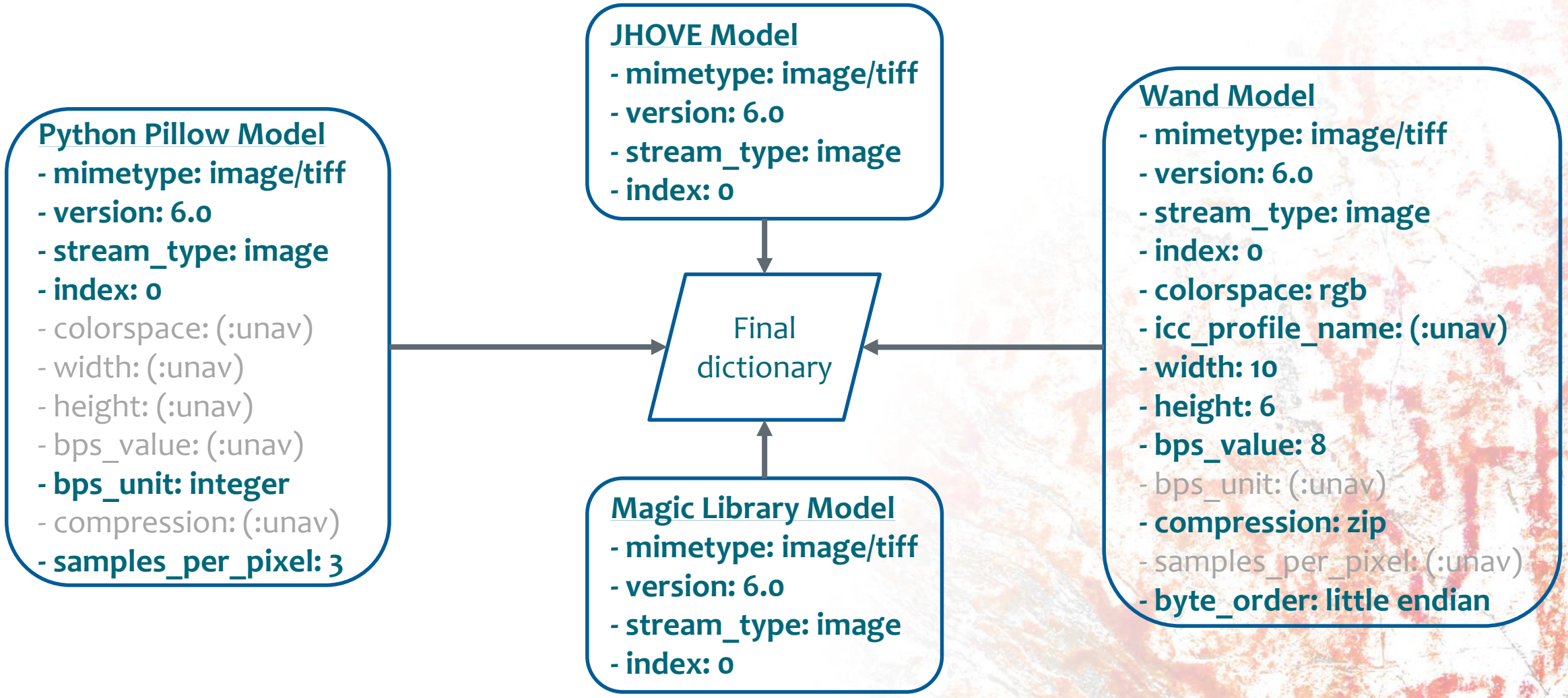
Lista ei ole täydellinen …

- PAS-palvelujen tuottamat kirjastot:
  - addml
  - audiomd
  - dpres-signature
  - **dpx-validator**
  - mets
  - premis
  - nisomix
  - videomd
  - xml-helpers
- Kolmannen osapuolten komponentit:
  - click
  - **FFMpeg**
  - ffmpeg-python (patched by DPS)
  - **file**
  - file-magic
  - **Ghostscript**
  - **ImageMagick**
  - **iso-schematron-xslt1** (patched by DPS)
  - **JHove**
  - **LibreOffice**
  - lxml
  - M2Crypto
  - **Mediainfo**
  - olefile
  - opf-fido (patched by DPS)
  - Pillow
  - **pngcheck**
  - **pspp** (patched by DPS)
  - pymediainfo
  - python-mimeparse
  - six
  - wand
  - **v.Nu**
  - **veraPDF**
  - **warc-tools**
  - **xmllint**

Komponentit vaikuttavat validointiin

- Koska käytössä on runsaasti eri kolmannen osapuolten komponentteja, tiedostomuotojen validoinnin tapa vaihtelee
  - Joitakin muotoja analysoidaan syntaksiltaan hyvin tarkasti
  - Joillekin varmistetaan, että tiedostomuoto on prosessoitavissa tietyllä ohjelmistolla virheittä
- Joillekin tiedostomuodoille ajetaan useampia validaattoreita
  - Esim. PDF/A: Jhove, Ghostscraper, VeraPDF
- Merkistön tunnistaminen perustuu tilastolliseen analyysiin
- Kaikille muodoille ei löydy valmista työkalua
  - PAS-palvelut joutuvat soveltamaan tai kehittämään omaa …

TIFF-tiedoston metatietojen yhdistäminen



(:unav) on arvo, joka tarkoitta *ei tiedossa*. Se ohitetaan jos toinen komponentti tuottaa metatiedolle oikean arvon.

Kaksi tapaa käsitellä komponenttien konflikteja

Case 1: Vain yksi ääni
Jos mahdollista, vain yksi komponentti palauttaa tietylle metatiedolle arvon. Kaikki muut palauttavat tarkoituksella (:unav).

Python Pillow Metadata	Wand Metadata
width: (:unav)	width: 1920
height: (:unav)	height: 1080
samples_per_pixel: 3	samples_per_pixel: (:unav)

Valinta on komponentti- tai metatietokenttäkohtainen.

Case 2: Yhden ääni on painottunut
Joskus komponentit tuottavat ristiriitaista tietoa. Silloin painottuneemman komponentin ääni ratkaisee.

File format	Ghostscript Metadata	VeraPDF Metadata
Valid PDF 1.7 (not PDF/A)	version: 1.7	Not run
Valid PDF/A-2b (also PDF 1.7)	version: 1.7	version: A-2b

PDF/A –tiedostojen kohdalla, VeraPDF-komponentilla on painottuneempi ääni tiedostomuodon version tunnistamisessa.

Validoinnin tarkoitus

- Tiedostomuotojen validoinnilla on useampia tarkoituksia:
  - Rikkinäisten tiedostojen havaitseminen
  - Väärintunnistettujen tiedostojen havaitseminen
  - Toteaminen, että tiedostoa pystyy käsittelemään tietyllä nykyhetken ohjelmistolla
- Vain eheä tiedosto ilman teknisiä virheitä voidaan migroida tunnetulla migraatioprosessilla
- Ainoastaan eheässä tiedostossa voidaan todeta, että tietosisältö säilyy luotettavalla tavalla tunnetuilla säilytysmenetelmillä


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