



# PAS-paketointimäärittelyt

PAS-seminaari 25.4.2023 / Juha Lehtonen



# Yhteistyö mahdollistaa aineistojen pitkäkestoisen hyödyntämisen



# Hyödyntäviä organisaatioita ohjaavat PAS-määrittelyt



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

# Määrittelyt päivitetään vuosittain yhteistyössä



## Palautteen kerääminen hyödyntäviltä organisaatioilta

- Epäviralliset palautteet ja virallisempi palautekysely



## Päätökset muutoksista

- Ovatko muutokset välttämättömiä juuri nyt?
- Perustelut



## Määrittelyiden päivittäminen yhteistyössä hyödyntävien organisaatioiden kanssa

- Hyväksyntä yhteistyöryhmässä



## Uusien versioiden julkaiseminen

- XML-skeemat
- Muu dokumentaatio

# Määrittelyt saatavilla

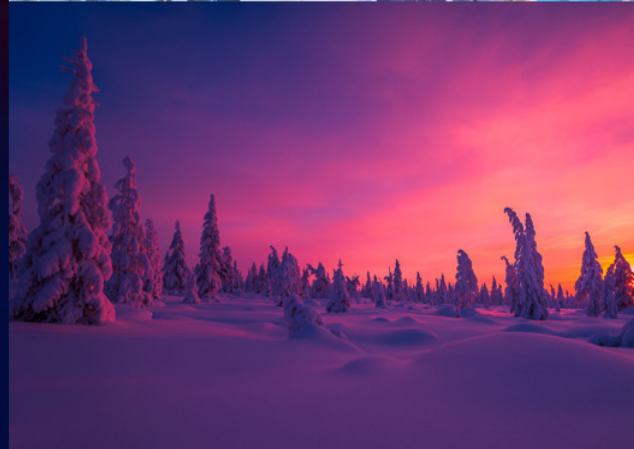
Myös englanniksi



- <http://digitalpreservation.fi/specifications>
  - Aineistojen ja niiden metatietojen paketointi pitkäaikaissäilytykseen
    - URN:NBN:fi-fe2020100578093
  - Säilytys- ja siirtokelpoiset tiedostomuodot
    - URN:NBN:fi-fe2020100578095
  - PAS-palveluiden rajapinnat
    - URN:NBN:fi-fe2020100578097
- Vuosittaiset laaturaportit ja muita julkaisuja
- Työkalut @GitHub
  - <https://github.com/Digital-Preservation-Finland/>

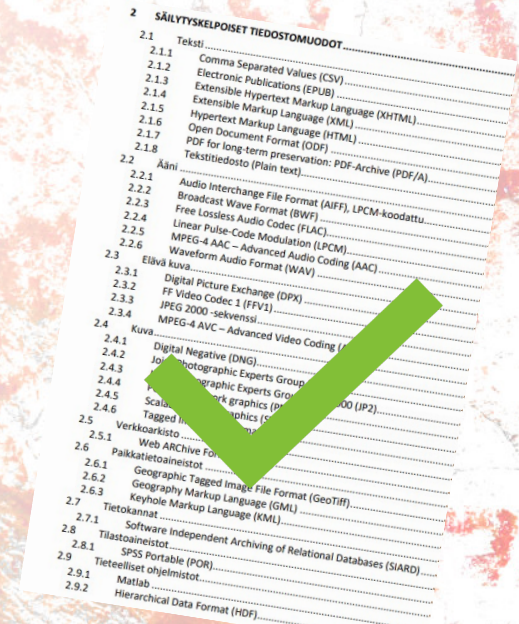


# Säilytys- ja siirtokelpoiset tiedostomuodot



# Säilytyskelpoiset tiedostomuodot

- Säilytyskelpoinen tiedostomuoto on sellainen, jonka tietosisällön säilyminen ja ymmärrettävyys voidaan taata pidemmällä aikavälillä
- PAS-palvelu vastaanottaa säilytyskelpoisia tiedostomuotoja siirtopaketeissa
- PAS-palvelu lupaa voivansa migroida hallitusti ja suunnitellusti
- Tiedostomuodot on jaettu aineistotyyppittäin, esimerkiksi
  - Kuva: JPEG, PNG, SVG, TIFF ...
  - Teksti: plain text, CSV, ODF, PDF ...
  - Tietokannat: SIARD, POR
  - Ääni: BWF, FLAC ...
  - Elävä kuva: FFV<sub>1</sub>, AVC ...



2	SÄILYTYSKELPOISET TIEDOSTOMUODOT
2.1	Teksti
2.1.1	Comma Separated Values (CSV)
2.1.2	Electronic Publications (EPUB)
2.1.3	Extensible Hypertext Markup Language (XHTML)
2.1.4	Extensible Markup Language (XML)
2.1.5	Hypertext Markup Language (HTML)
2.1.6	Open Document Format (ODF)
2.1.7	PDF for long-term preservation: PDF-Archive (PDF/A)
2.1.8	Tekstitiedosto (Plain text)
2.2	Ääni
2.2.1	Audio Interchange File Format (AIFF)
2.2.2	Broadcast Wave Format (BWF)
2.2.3	Free Lossless Audio Codec (FLAC)
2.2.4	Linear Pulse-Code Modulation (LPCM)
2.2.5	MPEG-4 AAC - Advanced Audio Coding (AAC)
2.2.6	Waveform Audio Format (WAV)
2.3	Elävä kuva
2.3.1	Digital Picture Exchange (DPX)
2.3.2	FF Video Codec 1 (FFV1)
2.3.3	JPEG 2000 -sekvenssi
2.3.4	MPEG-4 AVC - Advanced Video Coding (AVC)
2.4	Kuva
2.4.1	Digital Negative (DNG)
2.4.2	Joint Photographic Experts Group
2.4.3	Joint Photographic Experts Group
2.4.4	Joint Photographic Experts Group
2.4.5	Scalable Vector Graphics (SVG)
2.4.6	Tagged Image File Format (TIFF)
2.5	Verkoarkisto
2.5.1	Web ARChive For
2.6	Paikkatietoaineistot
2.6.1	Geographic Tagged Image File Format (GeoTIFF)
2.6.2	Geography Markup Language (GML)
2.6.3	Keyhole Markup Language (KML)
2.7	Tietokannat
2.7.1	Software Independent Archiving of Relational Databases (SIARD)
2.8	Tilastoaineistot
2.8.1	SPSS Portable (POR)
2.9	Tieteelliset ohjelmistot
2.9.1	Matlab
2.9.2	Hierarchical Data Format (HDF)

# 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 aineistotyyppittäin kuin säilytyskelpoisissa muodoissa
  - Kaikilla aineistotyyppiryhmillä ei kuitenkaan ole siirtokelpoisia muotoja

3	SIIRTOKELPOISET TIEDOSTOMUODOT.....
3.1	Teksti.....
3.1.1	Microsoft Office Suite.....
3.1.2	Portable Document Format (PDF).....
3.2	Ääni.....
3.2.1	Audio Interchange File Format (AIFF-C).....
3.2.2	Moving Pictures Expert Group (MPEG) MPEG-1 layer-3, MPEG-2 layer-3 (MP3).....
3.2.3	Windows Media Audio (WMA).....
3.3	Elävä kuva.....
3.3.1	Digital Video ja sen variantit (DV).....
3.3.2	Moving Pictures Expert Group (MPEG-1, MPEG-2).....
3.3.3	Windows Media Video (WMV).....
3.4	Kuva.....
3.4.1	Encapsulated postscript (EPS).....
3.4.2	Graphics Interchange Format (GIF).....



# Tiedostomuotojen arviointikriteeristö

<b>Avoimuus</b>	Kuinka helppoa tiedostomuodosta on saada tietoja?
<b>Käyttö PAS-standardina</b>	Missä määrin tiedostomuoto on muodollisesti hyväksytty pitkäaikaissäilytyksen välineeksi?
<b>Vakaus/ yhteensopivuus</b>	(a) Missä määrin tiedostomuoto on eteen- ja taaksepäin yhteensopiva? (b) Missä määrin tiedostomuoto on suojattu tiedoston korruptoitumista vastaan? (c) Kuinka usein tiedostomuodosta julkaistaan korvaavia versioita?
<b>Riippuvuudet / yhteentoimivuus</b>	Missä määrin tiedostomuoto on sidottu esimerkiksi tiettyyn laitteistoon tai ohjelmistoon?
<b>Standardisuus</b>	Missä määrin tiedostomuoto on käynyt läpi perusteellisen standardointiprosessin?

Arviointi: A, B tai C



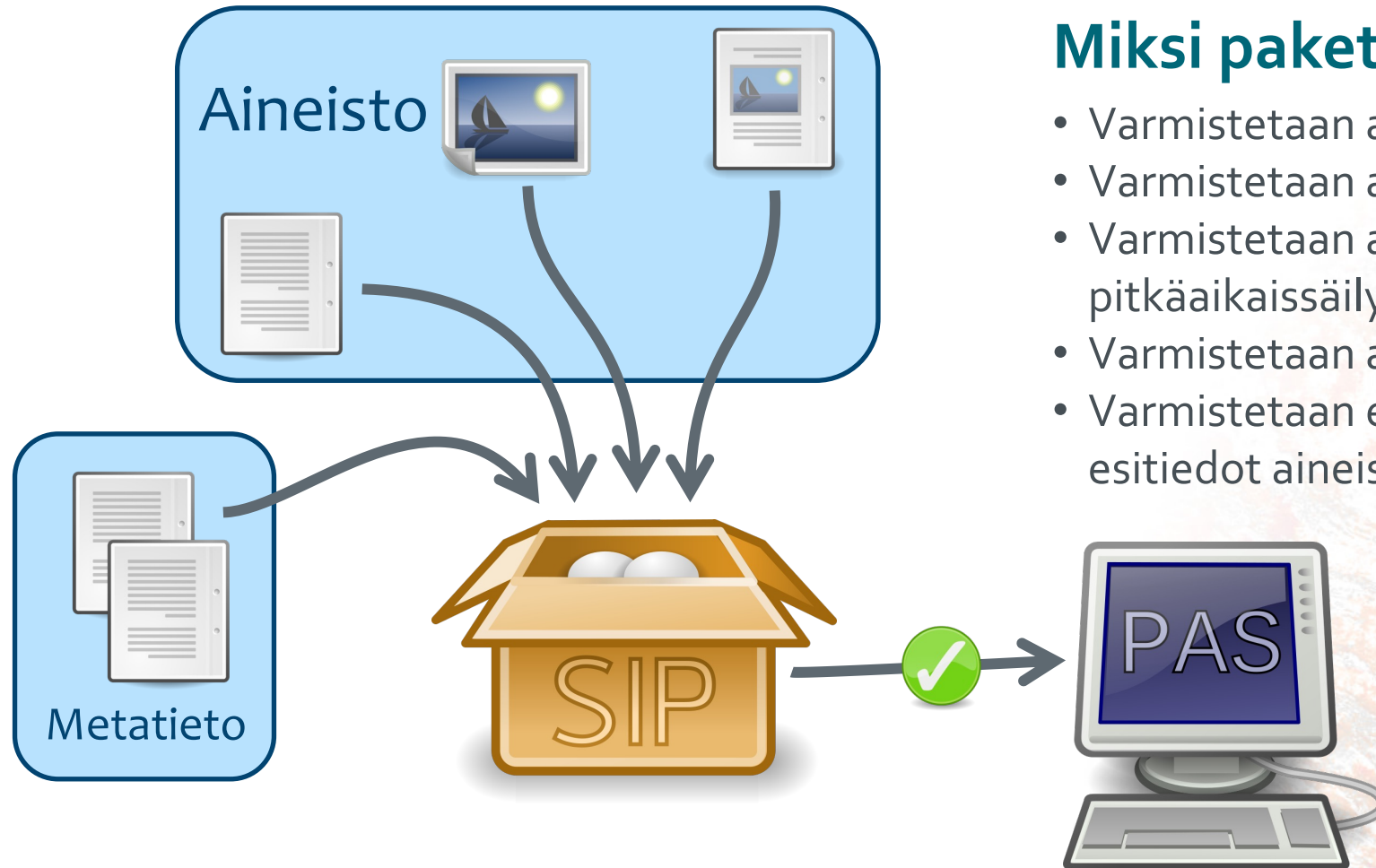
# Aineistojen ja niiden metatietojen paketointi pitkäaikaissäilytykseen



# Aineistojen ja niiden metatietojen paketointi pitkäaikaissäilytykseen

- Määrittelee pakolliset ja ehdollisesti pakolliset metatiedot
  - Hallinnolliset metatiedot
  - Rakenteelliset metatiedot
  - Tuettujen kuvailevan metatietomuotojen lista
- Määrittelee siirtopaketin (SIP) ja jakelupaketin (DIP) teknisen rakenteen
- Perustuu METS ja PREMIS XML-standardeihin
- Vaatimukset mahdollistavat aineistojen automaattisen validoinnin ja vastaanoton

# Aineistojen paketointi pitkäaikaissäilytykseen



## Miksi paketoidaan?

- Varmistetaan aineiston eheys
- Varmistetaan aineiston aitous
- Varmistetaan aineiston sopivuus pitkäaikaissäilytykseen
- Varmistetaan aineiston ymmärrettävyys
- Varmistetaan että PAS-palvelulla on riittävät esitiedot aineiston säilyttämiseksi

Aineistoja ei kannata paketoida etukäteen “varastoon”.

# Paketoinnin vaihtoehdot

## PAKETOINTIPALVELUN HYÖDYNTÄMINEN

Vain tutkimusaineistoille

### A. IDA-säilytyspalvelu

✓ Aineistot IDA:ssa

- + Paketointi automaattista
- + Ei vaadi omaa koodausosaamista
- + Ei vaadi XML-tuntemusta
- Joustamaton

### B. Tiedostojen koontipalvelu

✓ Aineistoa maltillisesti; yksinkertainen rakenne

- + Paketointi automaattista
- + Ei vaadi omaa koodausosaamista
- + Ei vaadi XML-tuntemusta
- Joustamaton

## ORGANISAATIO PAKETOI ITSE

### C. Paketointityökalun hyödyntäminen

✓ Paljon aineistoa

- + Melko joustava
- + Ylläpidettävyys määrittelyiden muuttuessa
- + Vähentää tarvetta XML formaattien osaamiselle
- Jonkin verran omaa koodausosaamista
- Määrittelyiden ymmärtäminen

### D. Organisaation räätälöimä paketointi

✓ Paljon aineistoa; monimutkaisia tarpeita

- + Joustava ja skaalautuva
- (Runsaasti) omaa koodausosaamista
- Määrittelyiden syvälinen ymmärtäminen
- Ylläpidettävyys määrittelyiden muuttuessa

# METS

- Metadata Encoding and Transmission Standard
- Kääremuoto
  - XML-standardi tiedostojen ja metatietojen niputtamiseksi yhteen
  - Kuvaileva, hallinnollinen, rakenteellinen metatieto
- METS XML -rakenteen määrittelee skeematiedosto, jota vastaan XML-dokumentin voi validoida
- Laajalti pitkäaikaissäilytyspiireissä käytössä
- METS Editorial Boardin ylläpitämä
- <https://www.loc.gov/standards/mets/>

# PREMIS

- PREMIS Preservation Metadata
- Metatietomalli säilytysmetatiedolle
  - Määrittelee tietokentät (sanastot, sallitut arvot, kardinaliteetit) ja rakenteen
  - PREMIS XML-skeema on virallinen tuettu implementaatio tietomallista
- Laajalti pitkäaikaissäilytyspiireissä käytössä
- PREMIS Editorial Committeeen ylläpitämä
- <https://www.loc.gov/standards/premis/>

# METS-dokumentin rakenne

## METS DOCUMENT

### METS HEADER

### DESCRIPTIVE METADATA

*Descriptive metadata standards approved*

### ADMINISTRATIVE METADATA

Provenance information

Technical metadata

Access rights

Source metadata

Preservation plan

### STRUCTURAL METADATA

#### Logical perspective

Title page  
Chapters  
Sections  
....

#### Physical perspective

Pages  
Columns  
Page sections  
...

Linkage of objects in different sections

### FILE REFERENCES

#### Group 1

e.g TIFF files (originals)

#### Group 2

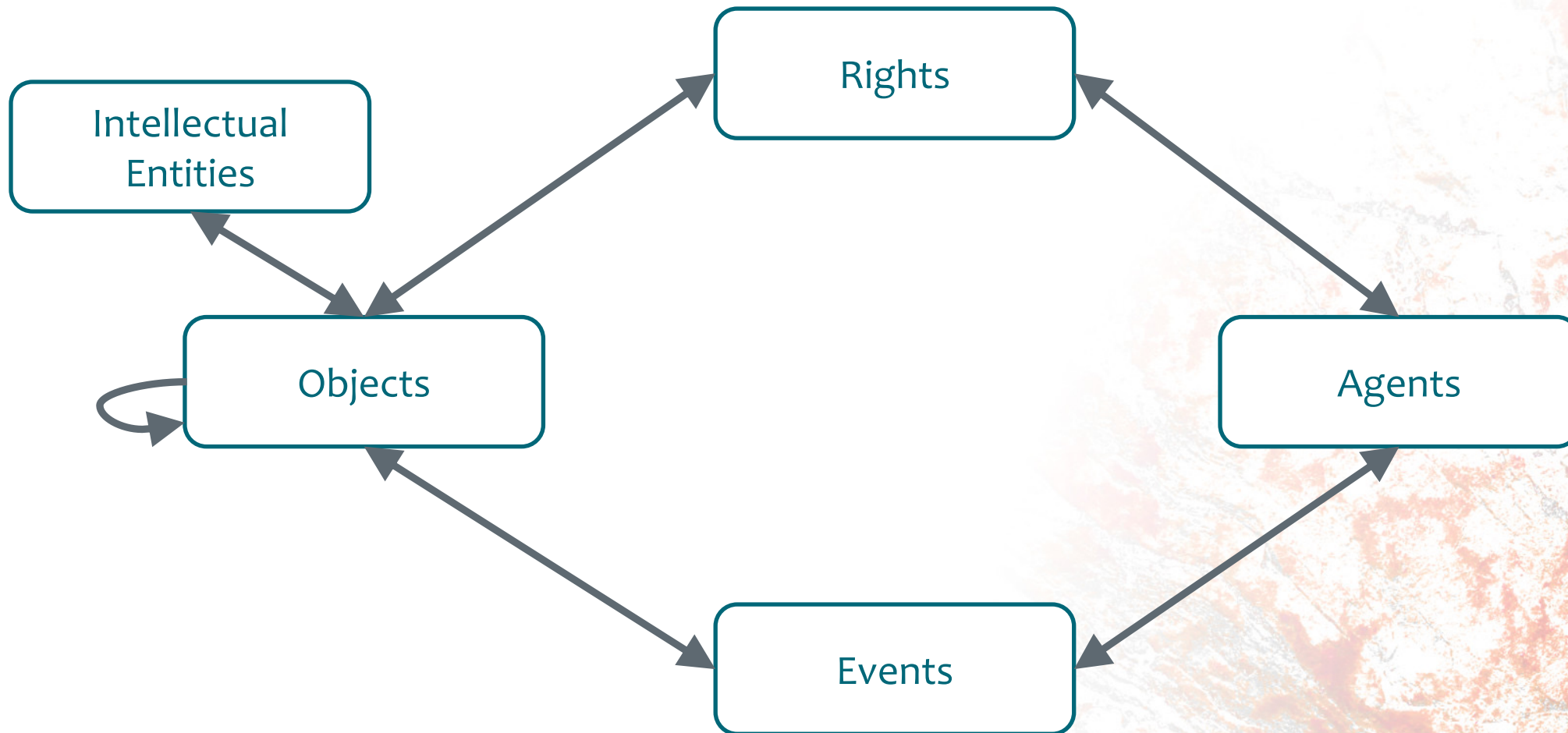
e.g. JPEG files (thumbnails)

#### Group 3

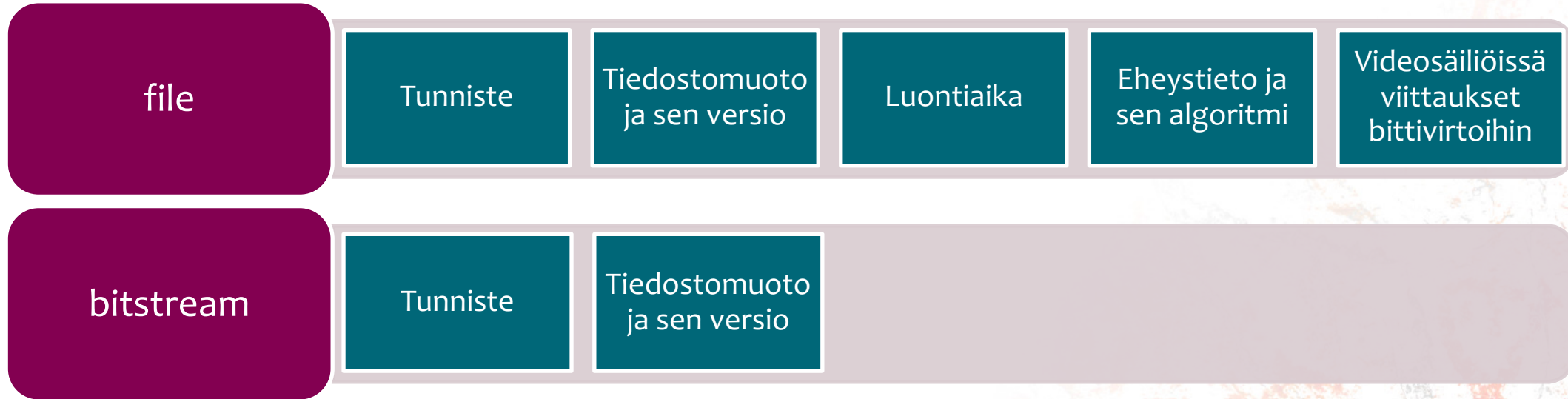
...



# PREMIS tietomalli



# Pakolliset tekniset metatiedot – PREMIS Object



- Muut ehdollisesti pakolliset tekniset metatietomuodot:
  - MIX, VideoMD, AudioMD, ADDML
- Paketointityökalu osaa tuottaa digitalisten objektien pakolliset tekniset metatiedot automaattisesti

# Yhteistyötä organisaatioiden kanssa nyt ja tulevaisuudessa

(myös muiden kuin asiakkaiden)

- PAS-palveluiden sähköpostilista
  - <https://www.digitalpreservation.fi/2021-liity-pas-palveluiden-sahkopostilistalle>
- Kuukausittaiset #PASKaffet
  - Joka kuukauden ensimmäinen perjantai klo 14:00-14:45
- #PASKlinikat
  - Säännöllisen epäsäännöllisesti
- Koulutustilaisuuksia enemmän tai vähemmän säännöllisesti
- [twitter.com/dpres\\_FI](https://twitter.com/dpres_FI)
- Virallinen tukiosoite: [pas-support@csc.fi](mailto:pas-support@csc.fi)
- <https://www.digitalpreservation.fi>
- <https://www.fairdata.fi>



**YOU'RE NOT  
ALONE...**

[pas-support@csc.fi](mailto:pas-support@csc.fi)