CSC

ICT Solutions for Brilliant Minds

Aineistonhallinta

PAS-palveluiden hyödyntäminen käytännössä -koulutus

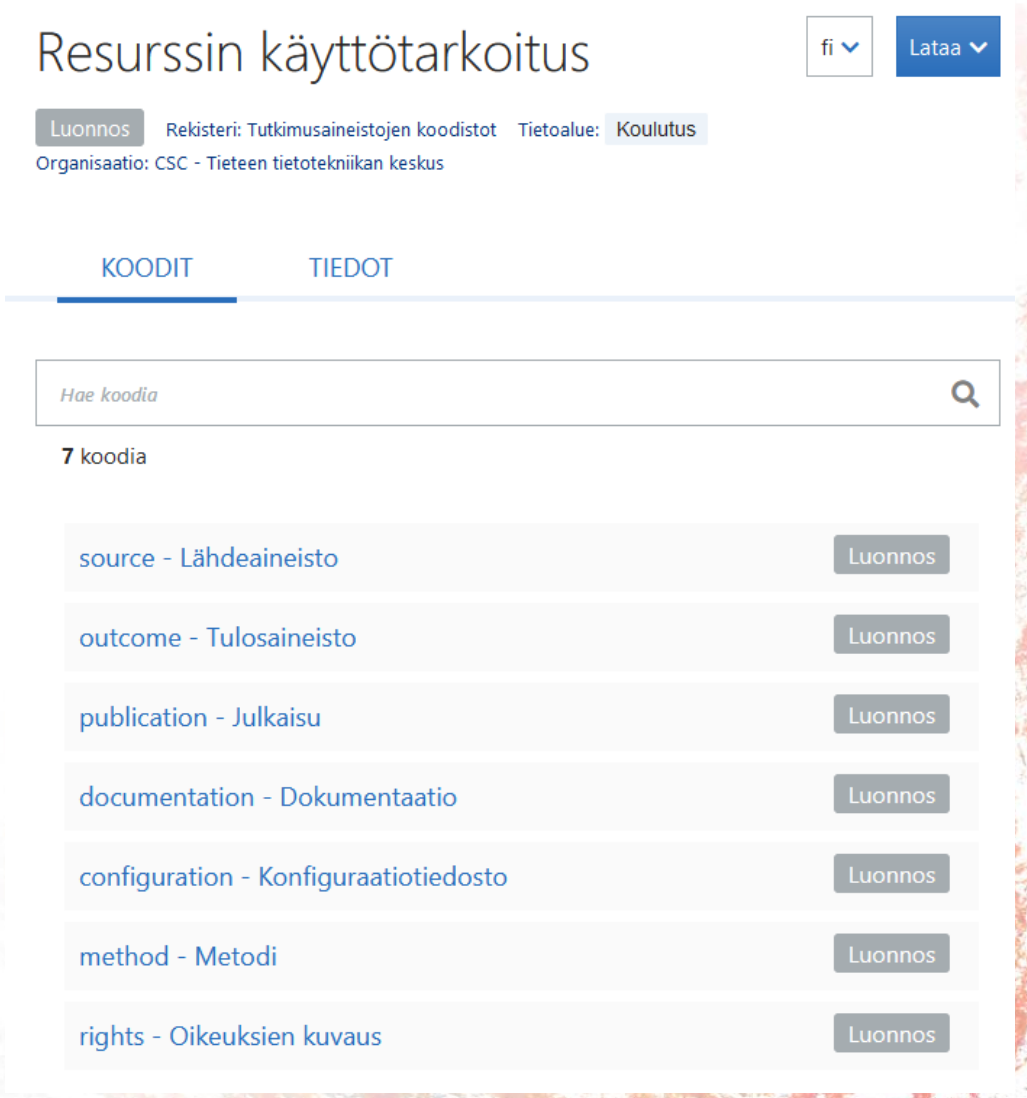
25.4.2023

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

Hyödyntäviä organisaatioita ohjaavat PAS-määrittely

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

Yleiset rajoitukset tiedostomuodoille

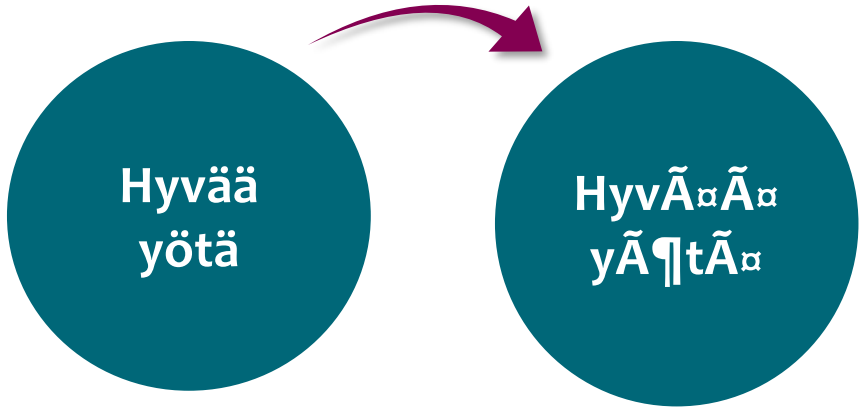
- Tiedostoissa ei saa käyttää salasanasuojauksia eikä mitään muita salaustekniikoita
- Tiedostoissa ei saa käyttää DRM (Digital Rights Management) -tekniikoita
- Tiedostoja ei saa allekirjoittaa digitaalisesti, jos se estää tiedoston käsittelyn
- Tiedostoja ei saa (tarpeettomasti) pakata
- Tiedostosta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

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



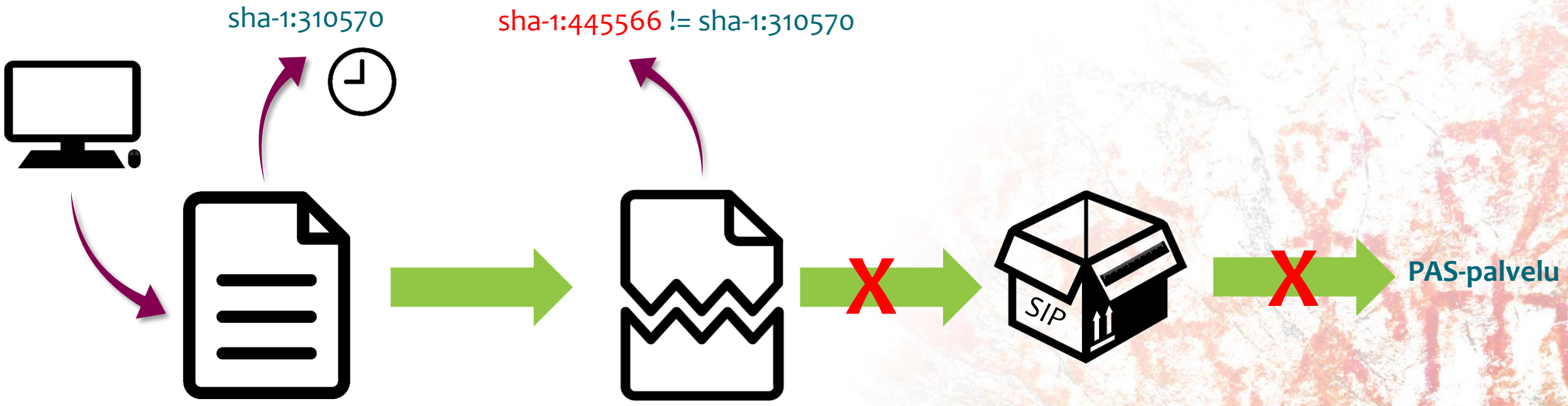
Tarkistussummat

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

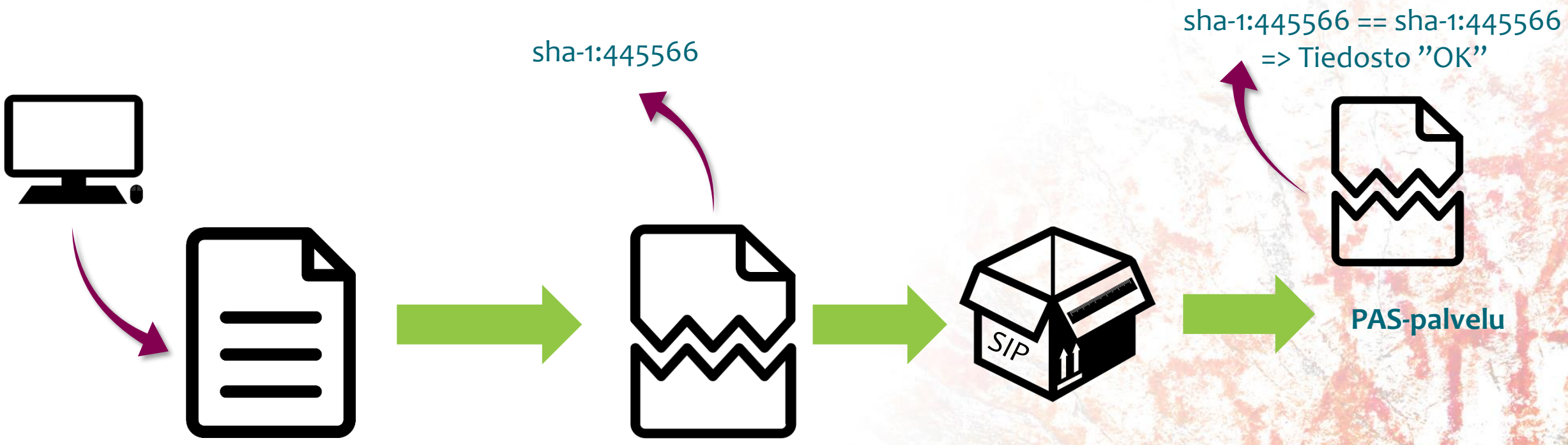
Tarkistussummat

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

Synty- ja tapahtumahistoria tapahtumina

Tapahtumatyyppi	• Tapahtuman tyyppi (kontrolloitu lista)
Tunniste	• Tunnisteen tyyppi ja arvo
Aikaleima	• Voi olla myös epätarkka aika
Selitys	• Kuvaus tai muu oleellinen ihmisluettava tieto tapahtumasta
Tapahtuman tulos	• “success”, “failure”, “unknown”
Lopputuloksen lisätiedot	• Esim. ajetun ohjelman koko tuloste
Viittaukset agentteihin	• suositellaan myös agenttien roolit tapahtumassa
Viittaukset objekteihin	• suositellaan myös objektien roolit tapahtumassa

Mitä metatietoja agentille?

Tyyppi	• *person, organization, software, hardware*
Tunniste	• tunnisteen tyyppi ja arvo
Nimi	• ohjelman tapauksessa nimi ja versiotieto
Lisätiedot	• esimerkiksi kuvaus konfiguraatiosta, jota ohjelmalle käytetään
(Lisäosio)	• Vapaamuotoinen paikallisesti koneluettava lisäosio
(Viittaukset)	• Paketointikomponentti luo viitaukset tapahtumaan, ei agenttiin

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-seminaari-2023-04-25-aineistonhallinta-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