CSC

ICT Solutions for Brilliant Minds

# PAS-yhteistyöryhmä
# 15.6.2023

# Kokouksen avaus

- Todetaan läsnäolijat
- Esityslistan hyväksyminen
- Edellisen kokouksen muistio

# PAS-palveluiden rajapintojen uudistaminen

# PAS-palveluiden rajapinnat uudistuvat

- PAS-palveluiden rajapinnoista tulossa uusi pääversio (v3)
- Suurimmat muutokset:
  - SFTP-siirto korvataan HTTP-siirrolla
  - Siirtojen tilan seuranta ja tarkastusraporttien nouto HTTP:n kautta
  - Jakelupaketin muodostamiskutsu muuttuu
  - Uutena resurssina käsittelyhistoria
  - Uusia tilastotietoja?
- Vanhat rajapintaversiot toimivat vielä rinnakkain uuden kanssa

# Rajapintojen resurssit

- HTTP REST mallissa hallinnoidaan toimintoja resurssien kautta
- PAS-palveluiden rajapintojen resurssit:
  - **transfers** – Siirrot
  - **search** – Hakutyökalu
  - **preserved** – Säilytyspaketit
  - **disseminated** – Jakelupaketit
  - **statistics** – Tilastotiedot
  - **history** – Käsittelytiedot
  - **public-keys** – Julkiset avaimet

# Rajapintojen toiminnot

| Resurssi | Metodit | Kuvaus |
|---|---|---|
| /transfers | POST | Aineiston siirron alustaminen |
| /transfers/<siirron tunniste> | HEAD, PATCH | Aineiston siirto PAS-palveluun |
| /transfers | GET | Siirtojen listaus |
| /transfers/<siirron tunniste>/status | GET | Siirron tila PAS-palvelun vastaanotossa |
| /transfers/<siirron tunniste>/report | GET | Siirtopaketin tarkastusraportin nouto |
| /search | GET | Hakutoiminto |
| /preserved/<aip-id> | GET | Säilytyspaketin tiedot PAS-palvelussa |
| /preserved/disseminate | POST | Säilytettävän aineiston muodostaminen jakelupaketiksi |
| /disseminated | GET | Jakelupakettien listaus |
| /disseminated/<dip-id> | GET | Jakelupaketin tilan PAS-palvelussa |
| /disseminated/<dip-id>/download | GET | Jakelupaketin nouto |
| /history/transfers | GET | Siirtojen käsittelytietojen listaus |
| /history/preserved | GET | Säilytettävän aineiston käsittelytietojen listaus |
| /history/disseminated | GET | Aineiston jakelun käsittelytietojen listaus |
| /statistics/overview | GET | Tilastotiedot |
| /public-key/dip | GET | Jakelutoiminnon julkisen avaimen nouto |

# Aineiston siirto HTTP:n kautta

**Resurssi**

*transfers*

- Aineiston siirrossa, jokainen siirtotapahtuma on yksilöity resurssi
  - Siirroilla yksilöivä tunniste (transfer-id)
  - Siirto ei ole sama kuin siirtopaketti
  - Saman siirtopaketin voi siirtää monta kertaa, jokaisella kerralla luodaan uusi siirto
  - Jokaiselle siirrolle muodostetaan vastaanotossa tarkastusraportti
  - Siirrot ovat olemassa 20 päivää, jonka jälkeen ne poistetaan rajapinnasta
- HTTP-rajapinta ei hyödynnä organisaatioiden kotihakemistoa
  - Kun SFTP poistuu, organisaatioiden kotihakemistot poistuvat samalla ja kaikki hallinta tapahtuu rajapintakutsujen avulla
- Siirtäminen tapahtuu HTTP-pohjaisen TUS-protokollan kautta

# TUS-protokolla

- TUS-protokolla on luotu isoja siirtoja varten
  - https://tus.io/
  - Siirtäminen tapahtuu pilkkomalla tiedoston hyvin pieneen osiin, jolloin osat siirretään jonossa verkon yli
  - HTTP-pyyntöjen otsikkotiedoissa annetaan tietoa siirron etenemisestä ja tiedoston lopullisesta koosta
  - Siirron voi keskeyttää ja jatkaa myöhemmin, jolloin siirtäminen jatkuu siirtämällä jonon seuraavan tiedoston osan
  - Protokolla tietää milloin tiedosto on siirretty kokonaisuudessaan, hyödyntämällä tiedoston lopullista koko –tietoa
  - Jos siirto on vahinko, eikä tiedostoa haluta siirtää loppuun saakka, siirtämisen voi yksinkertaisesti vain keskeyttää eikä jatkaa, PAS-palvelun vastaanotto siivoaa pois keskeneräiset siirrot tietyn ajanjakson kuluttua
- TUS-protokollaa käytetään tyypillisesti valmiiden asiakasohjelmistojen avulla

# Siirtäminen TUS-protokollan avulla

| Resurssi |
| --- |
| *transfers* |

- Ennen siirtämistä, siirtoa pitää alustaa:
  - `POST <base>/<contract>/transfers`
  - Kutsussa annetaan otsikkotiedoissa siirron tietoja, kuten lopullinen tiedostonkoko tavuissa
  - Alustaminen varaa työtilan palvelimelta ja palautusviestissä annetaan siirron yksilöivä tunniste (transfer-id)
- Tietoa siirron etenemisestä
  - `HEAD <base>/<contract>/transfers/{transfer-id}`
  - Kutsulla kysytään palvelimelta tietoja siirron etenemisestä, kuten montako tavua on jo siirretty ja mistä tavusta (offset) seuraava siirron osa alkaa
- Siirtäminen
  - `PATCH <base>/<contract>/transfers/{transfer-id}`
  - Kutsussa siirretään tiedoston osa ja palautetaan tieto seuraavan osan offsetistä
- Kun viimeinen tavu on siirretty, siirto on automaattisesti valmis

# Siirtojen tiedot

**Resurssi**

*transfers*

- Siirtojen listaaminen
  - `GET <base>/<contract>/transfers`
  - Palautuksessa annetaan lista kaikista aktiivisista siirroista (tunnisteet, päivämäärät …)
  - Yli 20 päivää vanhat siirrot eivät enää ole löydettävissä tämän toiminnon kautta
- Siirron tilan tarkastaminen
  - `GET <base>/<contract>/transfers/{transfer-id}/status`
  - Palautuksessa annetaan tietoa siirron tilasta PAS-palvelun vastaanotossa
  - Tilat ovat:
    - siirto kesken - tiedostoa ei ole vielä siirretty PAS-palveluun kokonaisuudessaan
    - käsittelyssä - vastaanoton tarkistukset ovat käynnissä
    - hyväksytty – aineisto on hyväksytty säilytykseen
    - hylätty – siirtopaketissa on virheitä

# Siirtojen tiedot jatkuu

**Resurssi**

*transfers*

- Vastaanoton tarkasturaporttien nouto
  - *GET <base>/<contract>/transfers/{transfer-id}/report*
  - Palautuksessa palautetaan vastaanoton tarkastusraportit XML- tai HTML-muodossa
  - Tämä toiminnoin voi suorittaa vain jos siirto on joko hyväksytty tai hylätty
  - Myös tarkastusraportit poistetaan PAS-palvelun rajapinnasta 20 päivän kuluttua, siirron poiston myötä

# Aineiston haku

**Resurssi**

*search*

- Säilytyksessä olevat aineistot ovat löydettävissä niiden aineistojen metatiedoilla METS-tietokannasta
  - *GET <base>/<contract>/search*
- Hakuehdossa käytetään Apache Lucene –syntaksia avain/arvo-pareilla ja vastauksessa palautetaan tulosjoukko, joissa on säilytyspakettien tulosalkioita
- Tämä resurssi tai sen toiminto ei muutu uudessa rajapinnassa

# Säilytyspaketteihin kohdistuvat toiminnot

| Resurssi |
| --- |
| *preserved* |

- Säilytyspaketin tiedot
  - *GET <base>/<contract>/preserved/{aip-id}*
  - Vastauksessa ilmoitetaan säilytyspaketin tiedot
  - Säilytyspaketista ilmoitetaan mm seuraavat tiedot:
    - Tietopaketin tunnisteet
    - Aikaleimat
    - Säilytyspaketin sisältämät tiedostot ja niiden tunnisteet
    - Säilytyspaketin rakenne
- Säilytyspakettien tietojen avulla, on mahdollista valita tietty aineisto tai aineiston osa aineiston jakeluun

# Säilytettävän aineiston jakelu

**Resurssi**

*preserved*

- Säilytyksessä olevan aineiston jakelu suoritetaan luomalla aineiston jakelupyyntö:
  - `POST <base>/<contract>/preserved/disseminate`
  - Pyynnössä annetaan JSON muotoisena tietona seuraavat tiedot:
    - catalog – Jakelupaketin skeemakatalogin version
    - format – jakelupaketin tiedostomuoto (zip tai tar)
    - metadata – jakelupaketti muodostetaan pelkistä metatiedoista
    - aips – lista säilytyspaketeista, joista jakelupaketti koostuu. Jakelupaketti voi muodostua säilytyspaketin osasta, jolloin on mahdollista antaa lista niistä säilytyspaketin tiedostoista tai rakenteesta, joista jakelupaketti muodostetaan. Nämä annetaan säilytyspaketin tunnisteen yhteydessä.
  - Vastauksessa toimitetaan jakelupaketin tunniste (dip-id)
- Jakelupakettia on mahdollista muodostaa aineistosta, joka ei ole riippuvainen yksittäisen säilytyspaketit muodostamasta kokonaisuudesta

# Jakelupakettien tiedot ja noutaminen

**Resurssi**

*disseminated*

- Jakelupakettien listaaminen
  - *GET <base>/<contract>/disseminated*
  - Vastauksessa annetaan lista kaikista jakelupaketeista (tunnisteet, päivämäärät …)
  - Yli 10 päivää vanhat jakelutpaketit eivät enää ole löydettävissä tämän toiminnon kautta
- Jakelupaketin tilan tarkastaminen
  - *GET <base>/<contract>/disseminated/{dip-id}*
  - Vastauksessa annetaan tietoa jakelupaketin valmistumisesta PAS-palvelun jakelutoiminnossa
- Jakelupaketin nouto
  - *GET <base>/<contract>/disseminated/{dip-id}/download*
  - Valmistuneen jakelupaketin voi noutaa tämän toiminnon avulla

# Käsittelytiedot

**Resurssi**

*history*

- Siirtojen käsittelytietojen listaus:
  - *GET <base>/<contract>/history/transfers*
  - Toiminto listaa lokitietomaisesti kaikki PAS-palveluun lähetetyt siirrot ja niiden tiedot
- Säilytypakettien käsittelytietojen listaus:
  - *GET <base>/<contract>/history/preserved*
  - Toiminto listaa lokitietomaisesti kaikki PAS-palvelussa olevat säilytyspaketit ja niiden tiedot
- Jakelupakettien käsittelytietojen listaus:
  - *GET <base>/<contract>/history/disseminated*
  - Toiminto listaa lokitietomaisesti kaikki PAS-palvelun muodostamat jakelupaketit ja niiden tiedot

# Tilastotiedot

**Resurssi**

*statistics*

- Yhteenveto sopimukselle kohdistuvista tilastoista:
  - *GET <base>/<contract>/statistics/overview*
  - Tämä toiminto ei muutu uudessa rajapinnassa
- Tilastotietoja on mahdollista laajentaa

# Julkiset avaimet

**Resurssi**

*public-key*

- Jakelutoiminnon julkisen avaimen nouto:
  - `GET <base>/<public-key>/dip`
- Tämä resurssi tai sen toiminto ei muutu uudessa rajapinnassa

# Rajapintauudistuksen eteneminen

- Rajapinnoista luodaan uusi määrittely, jonka on tarkoitus julkaista ensi vuoden alussa
- Rajapintojen toiminnallisuus toteutetaan vaiheittain:
  - Vuoden 2024 aikana toteutetaan HTTP siirto ja siirtojen tilan seuranta
  - Vuoden 2025 aikana toteutetaan muut toiminnallisuudet
- Vanhat rajapinnat (v1 ja v2) poistetaan kun rajapintojen v3 on toteutettu
  - Suunnitelman mukaan 2026 vuoden alussa
  - Täten hyödyntävät organisaatiot voivat kerralla siirtyä v1:stä ja v2:sta v3:seen mikäli niin haluavat

# Yhteistyöryhmän syksyn 2023 toiminnan suunnittelu

# Yhteistyöryhmän syksyn 2023 toiminnan suunnittelu

Ote joulukuun 2022 kokouksen pöytäkirjasta:

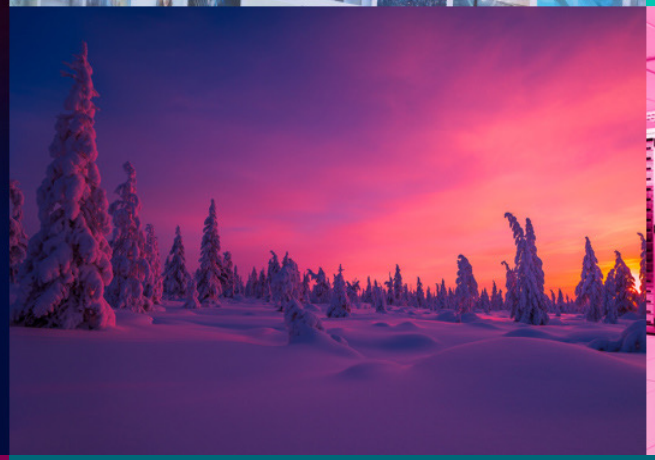
*Sovittiin että painopiste vuonna 2023 on seuraavissa asioissa, normaalien vuosikellon mukaisten asioiden lisäksi:*

- *PAS-palveluiden rajapinnan uudistaminen*
- *Hyödyntäville organisaatiolle tarjottavien työkalujen kehityksessä ja niihin liittyvissä koulutuksissa*
- *Aineiston poiston PAS-palvelusta suunnittelemisen aloittaminen*
- *PAS-kokonaisarkkitehtuurityön jatkaminen*

# Yhteistyöryhmän syksyn 2023 toiminnan suunnittelu

- Vuosikellon mukaiset asiat
- PAS-kokonaisarkkitehtuuri
- Kokemukset ensimmäisestä laajasta migraatiosta (case: arc/warc-migraatio)
- PAS-palveluiden rajapintojen uudistaminen
- Aineiston poistaminen PAS-palvelusta
- AlmaLinuxiin siirtyminen
- Paketointikomponentin uudistaminen
- Työpajat / Koulutukset
- Muuta?

# PAS-palveluiden ohjelmistokerros



# PAS-palveluiden ohjelmistokerros

- Viimeinen Python2 päivitys oli maaliskuussa
- RHEL 9 / AlmaLinux 9 -siirtymä
  - On tehty ja tehdään kesän aikana
  - Asennusohjeet syksyllä
- Paketointityökalun uudistus
  - Ensivilkaisu oli PAS-seminaarissa koulutuspäivänä
  - Viedään Githubiin saataville
    - Toivottiin PAS-seminaarissa tutustumista ja palautteen antamista varten
    - Ei sisällä kaikkia ominaisuuksia
    - Ei vielä suositella tuotantokäyttöön

# PAS-palveluiden uudet ominaisuudet Q2/2023

- ARC/WARC-migraation kehitystyö
  - Erilaisia viimeistelyjä migraatioon
  - Koeaineiston kerääminen
  - Koemigraation aloitus
- Nauhamigraatioon liittyvä kehitystyö on edistynyt
- Aputyökaluja tietokantojen ylläpitämiseen
- Tiedostomuotojen analysointityökalu
  - MPEG-PS (Program Stream) version tunnistaminen
  - file- ja libreoffice-ohjelmien polkujen käsittely
  - Mediainfo-bugi virallisessa jakelussa: selvitys ja korjauspyyntö sekä väliaikainen paikkaus
- Bugikorjauksia ja muita kehitystöitä
- SAPA-kehitystyö omassa projektiryhmässään
- Fairdata PAS
  - RHEL 9 / AlmaLinux 9 –työ on edistynyt
  - Uusi paketointityökalu on edistynyt
  - Hallintaliittymä
    - Käyttöliittymäsuunnittelutyö
    - Muita pienempiä muutoksia
  - Paketointi- ja validointiprosessin muutokset
  - Säilytyspäätöskohtainen autorisointiuudistus on aloitettu ja edistynyt
  - SEG-Y -tiedostomuodon tunnistus (bittitaso)
  - Bugikorjauksia ja muita kehitystöitä

# Sopimukset – KP-PAS (1/2)
(12.6.2023)

| Organisaatio | Kapasiteetti (Tt) | Aineistoa (Tt) | Täyttöaste | Säilytyspaketteja |
|---|---|---|---|---|
| Celia | 110 | 87.75 | 79.77 | 32374 |
| Kavi | 2 400 | 1511.0 | 62.96 | 3906 |
| Kansallisarkisto<br>Kansallisarkiston vastaanottamat alkujaan digitaaliset valtionhallinnon asiakirjalliset aineistot | 41 | 1.56 | 3.80 | 1947 |
| Kansallisarkisto<br>VAPA-järjestelmään siirretyt tietoaineistot | 1 | 0.14 | 14.00 | 389 |
| Kansallisarkisto<br>Kansallisarkiston massadigitointi-hankkeen aineistot | 114 | 87.89 | 77.10 | 73215 |
| Kansallisarkisto<br>Kansallisarkiston digitaaliarkistosta siirrettävät aineistot ja takautuvan digitoinnin aineistot | 805 | 296.55 | 36.84 | 579687 |
| Kansallisarkisto<br>Kansallisarkiston yksinomaan digitaalisessa muodossa olevat yksityisarkistoaineistot | 27 | 0 | 0.00 | 0 |

# Sopimukset – KP-PAS (2/2)

(12.6.2023)

| Organisaatio | Kapasiteetti (Tt) | Aineistoa (Tt) | Täyttöaste | Säilytyspaketteja |
|---|---|---|---|---|
| Kansallisgalleria | 20 | 6.32 | 31.60 | 478 |
| Kansalliskirjasto<br>Kulttuuriaineistolain nojalla kerätyt aineistot | 355 | 201.32 | 56.71 | 1982369 |
| Kansalliskirjasto<br>Kansalliskirjaston digitoimat kulttuuriperintöaineistot | 175 | 49.71 | 56.71 | 185429 |
| KOTUS | 60 | 9.83 | 16.38 | 407 |
| Museovirasto | 1 | 0.53 | 53.00 | 40186 |
| Musiikkiarkisto | 70 | 1.55 | 2.21 | 41 |
| Svenska litteratursällskapet i Finland | 50 | 1.14 | 2.28 | 166 |
| Yhteiskuntatieteellinen tietoarkisto | 1 | 0.06 | 6.00 | 8969 |
| **Yhteensä** | **4 230,00** | 2255.35 | 53.32 | 2909563 |

# Säilytyspäätökset – FD-PAS (1/2)
(12.6.2023)

| Organisaatio | Kapasiteetti (Tt) | Aineistoa (Tt) | Täyttöaste | Säilytyspaketteja |
|---|---|---|---|---|
| Geologian Tutkimuskeskus<br>GTK:n tomografialaitteen tuottamat tietoaineistot | 12 | 6.64 | 55.33 | 139 |
| Helsingin yliopisto<br>Helsingin yliopiston SMEAR-aineistojen valikoima meteorologisia - ja ilmanlaatumittauksia | 2 | 0.01 | 0.50 | 13 |
| Helsingin yliopisto<br>M. cinxia and C. melitaearum in the Åland metapopulation system | 2 | 0 | 0.00 | 1 |
| Helsingin yliopisto<br>FIRE (The Finnish Reflection Experiment) | 1 | 0 | 0.00 | 0 |
| Helsingin yliopisto<br>Luomuksen aineistot | 150 | 0 | 0.00 | 0 |
| Itä-Suomen yliopisto<br>SENSOTRA | 1 | 0 | 0.00 | 0 |

# Säilytyspäätökset – FD-PAS (2/2)
(12.6.2023)

| Organisaatio | Kapasiteetti (Tt) | Aineistoa (Tt) | Täyttöaste | Säilytyspaketteja |
|---|---|---|---|---|
| Jyväskylän yliopiston kiihdytinlaboratorio<br>250-Nobeliumin hajoamisspektroskopia | 1 | 0 | 0.00 | 0 |
| Oulun yliopisto, Sodankylän geofysikaalinen observatorio<br>Havaintoaineistot | 30 | 0 | 0.00 | 0 |
| Tampereen yliopisto<br>Yhteiskuntatieteiden tiedekunnan Kansanperinteen arkiston A-K-kokoelma | 2 | 1.08 | 54.00 | 5361 |
| Turun yliopisto<br>Historian, kulttuurin ja taiteiden tutkimuksen arkiston aineistot (HKT-arkisto) | 20 | 0.25 | 1.25 | 1 |
| Åbo Akademi<br>Samlingar vid Åbo Akademis bibliotek | 10 | 0 | 0.00 | 0 |
| **Yhteensä** | **231** | **7.98** | **3.45** | **5515** |

# Ilmoitusasiat

- PAS-määrittelyiden päivittäminen – kysely käynnissä 30.8.2023 asti
  - https://link.webropolsurveys.com/S/212D3C71B2F6BB29
- PAS-seminaari järjestettiin 24.-25.4.2023 CSC:llä
- CSC järjesti OPF:n vuosikokouksen (AGM) ja teknisen ohjausryhmän (OAG) kokoukset 24.-25.5.2023 Oodissa
- iPPRES 2023 -konferenssiin hyväksytty
  - *Calculating the Carbon Footprint of Digital Preservation - A Case Study*
- Aalto-yliopiston Tutkimuspalvelupäivät 21.-23.8.2023
  - *Calculating the Carbon Footprint of Digital Preservation*
  - *On Preserving Research Data – What to Consider Before Ingesting Data*

# Kokouksen päättäminen

- Muut asiat
- Seuraava kokous
  - Syyskuu: Tiistai 26.9. klo 12:30-15:00
  - Joulukuu: Tiistai 12.12. klo 12:30-15:00
- Kokouksen päättäminen