KANSALLISARKISTO

Digitaaliarkiston migraatio

Reko Etelävuori, kehittämispäällikkö

31.3.2022

KANSALLISARKISTO

Tavoitteet

- Digitaaliarkiston tallennusinfrastruktuurin sekä asiakaskäyttöliittymien alasajo kokonaisuudessaan (pl. siirtopalvelu)
- Massadigitoinnin infrastruktuurin tallennusinfrastruktuurin alasajo
- Digitaaliarkiston migraation yhteydessä tarkoitus harmonisoida digitoinnissa tuotettu ja säilytykseen saatetun aineiston tietorakenne
- Tiedostokonversiot ja datamassan tiivistäminen
- Tietoaineiston rikastuttaminen
- Tietoaineisto keskitetysti SAPA:n hallinnoin piirissä
- 1 Pt -> uusi tallennusratkaisu ja rakenne
- Edellytyksenä kontekstimetatietomigraatio (VAKKA -> AHAA)

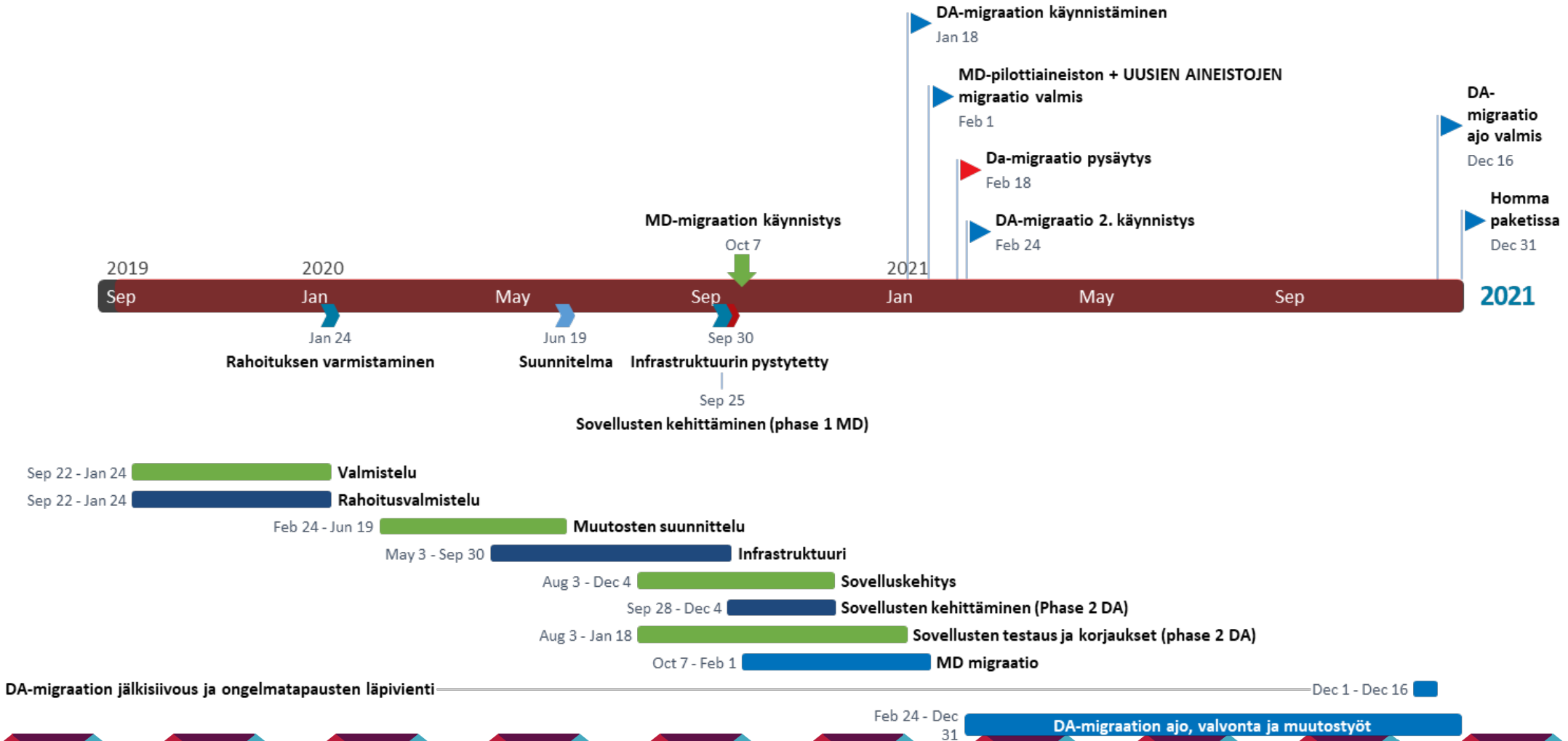
KANSALLISARKISTO

Suunnitteluvaihe

- Piirrettiin BPMN-prosessi
- Kuvattiin tarvittaessa lähdejärjestelmien tietorakenteet
- Määriteltiin käsittelysäännöt – jotka matkanvarrella muuttuivat useasti
- Päädyttiin rakentamaan oma infra, koska projekti mittava. Käytettiin mahdollisimman paljon vanhoja koodeja.
- Määriteltiin prosessi siten, että mikään ei etene, mikäli havaitaan mitään erilaisuutta suhteessa suunniteltu
- Määriteltiin suhteet järjestelmiin ja miten järjestelmien tulisi toimia eri tavoin suhteessa toisiinsa
- Käytännössä SAPA-AHAA- Migraatioinfra. Tehtiin uusi pakettityyppi SAPA:lle, joka ohjasi toimintaa

KANSALLISARKISTO

Aikataulu



KANSALLISARKISTO

Toteutustapa ja tiimi

- Lähdettiin scrummilla ja viikon sprinteillä, muutettiin myöhemmin kerta viikossa tapaamisiin. Jatkuvaa kanssakäymistä.
- Tuoteomistaja + 1 koodari + 2 migraatiovalvojaa, jotka tekivät myös pienkehittämistä ja tiedostojen käsittelyä
- Kukaan ei täyspäiväinen

KANSALLISARKISTO

Liittyvät järjestelmät

Lähtöjärjestelmät

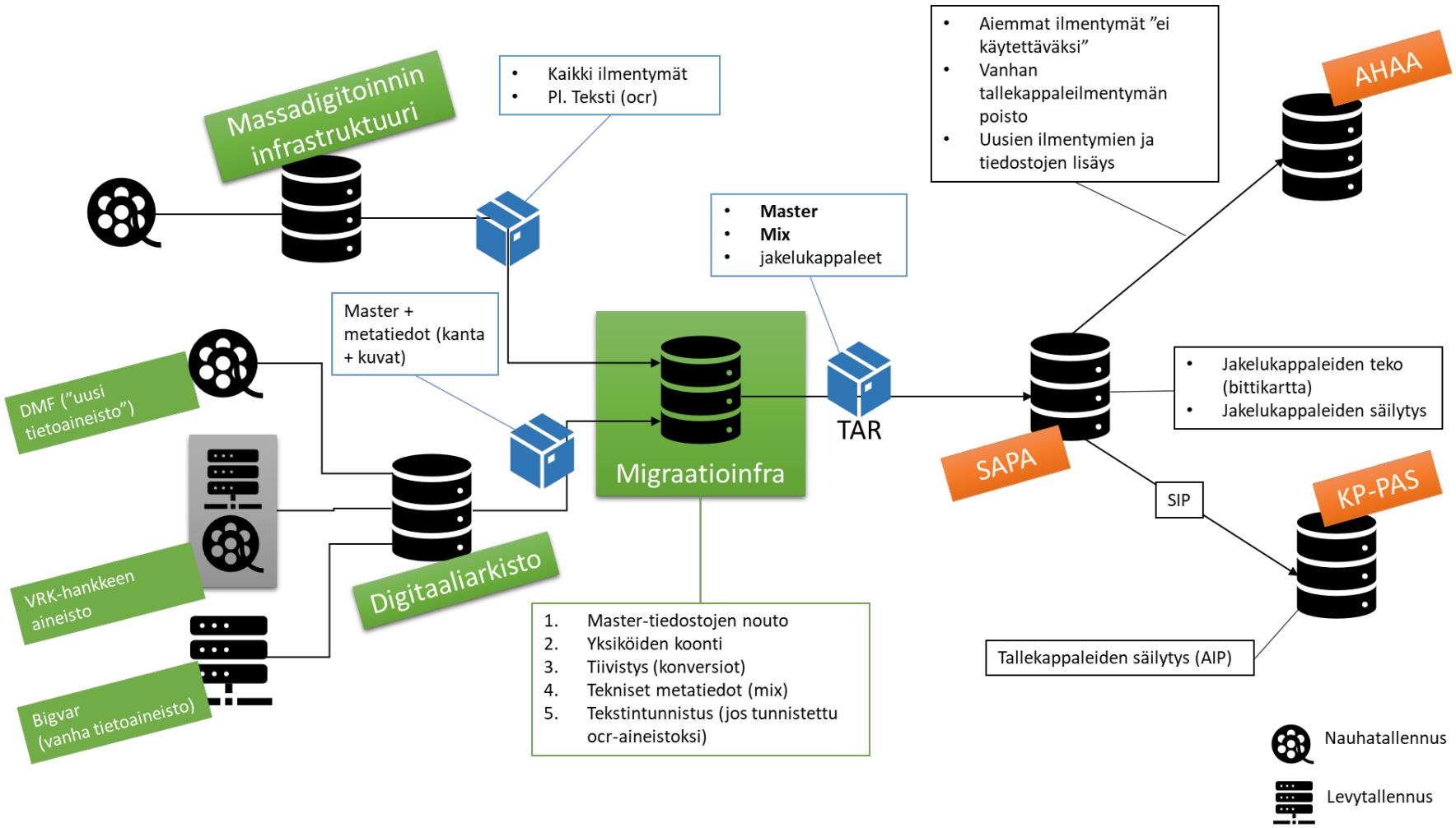
- **Digitaaliarkisto**
  - Kansallisarkiston vanha järjestelmä, jossa tallennusinfra, käyttökappaleiden tallennus, kontekstimetatietoa, työnohjaus ja prosessointisovellus, digitaalisen tietoaineistojen käyttörajoitusten hallinta sekä asiakaskäyttöliittymä.
- **Massadigitoinnin oma infrastruktuuri**
  - Kansallisarkiston digitaalisen tietoaineiston vastaanottoon, prosessointiin, tallennukseen ja saataville asettamiseen käytetty järjestelmä
- **(Vakka-arkistotietokanta)**
  - Entinen kontekstimetatietokanta

Vastaanottavat järjestelmät

- **SAPA**
  - Kansallisarkiston digitaalisen tietoaineiston vastaanottoon, prosessointiin, tallennukseen ja saataville asettamiseen käytetty järjestelmä
- **AHAA**
  - Kansallisarkiston kontekstimetatietojen tallennusjärjestelmä = metatietovaranto
- **KP-PAS**
  - Tietoaineistojen pitkäaikaissäilytysjärjestelmä
- **Astia**
  - Tietoaineistojen esitysjärjestelmä

KANSALLISARKISTO

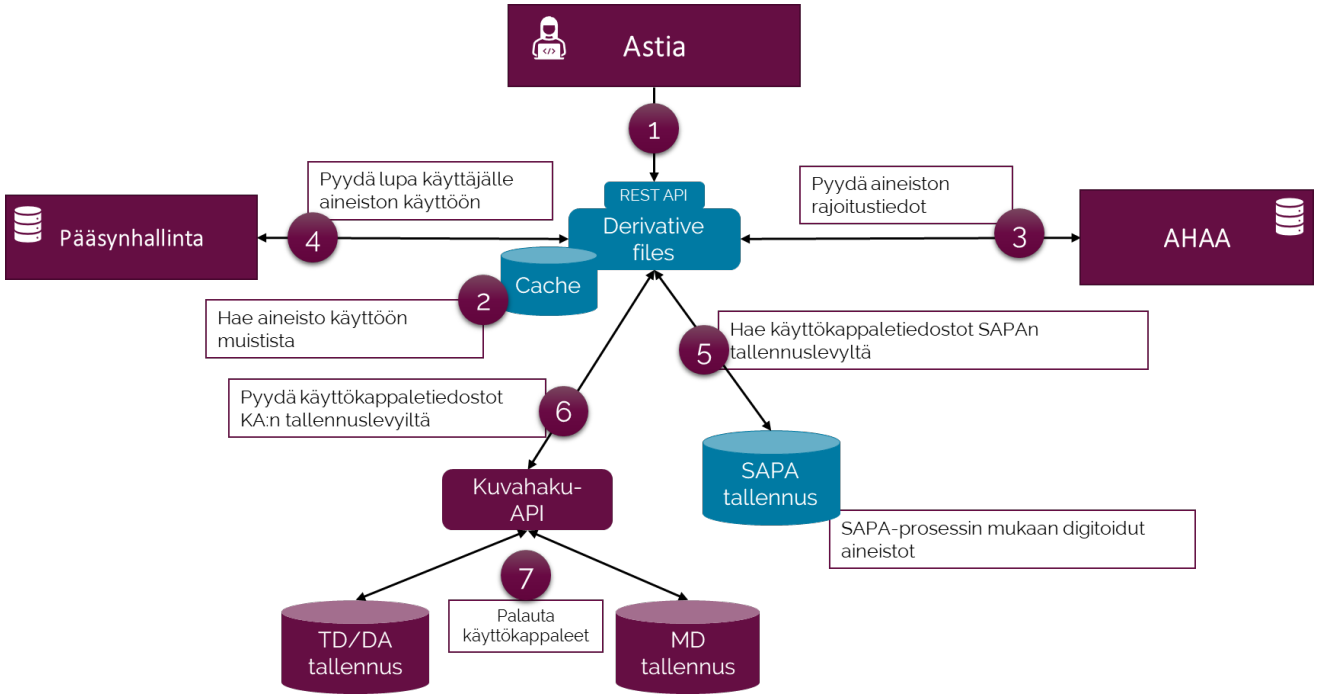
Kokonaisuus



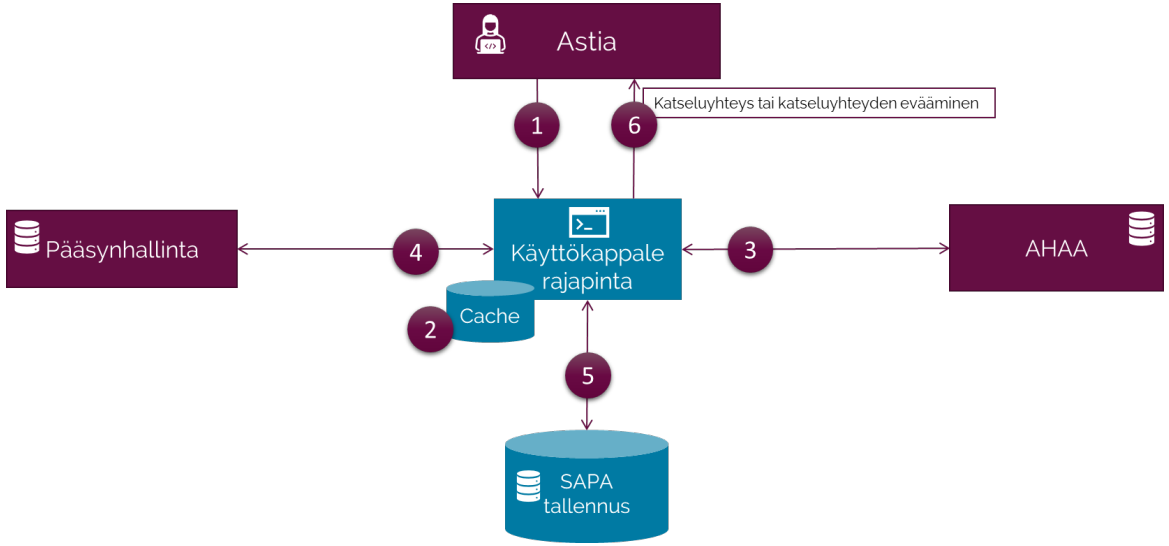
KANSALLISARKISTO

Käyttöönsaatto

Ennen



Jälkeen



KANSALLISARKISTO

Tietoaineistoblokit

- Massadigitoidun tietoaineiston migraatio
  - Massadigitoitu aineisto, joka on digitoitu ennen kuin massadigitoinnin prosessointisovellus kytkettiin SAPAan. Kaikki uudelleen OCR
- Digitaaliarkiston ”uuden” tietoaineiston migraatio
  - Uusi tietoaineisto on aineistoa, joka on digitoitu 06/2015 jälkeen. Tuolloin Digitaaliarkiston ulkoistaminen CSC:lle valmistui ja takautuvan digitoinnin prosessointi oli sovelluksen ja infran puolesta kokonaisuudessaan CSC:llä
- VRK-hankkeen tuottaman tietoaineiston migraatio
  - VRK aineisto on aineistoa, joka digitoitiin DVV:n ja Kansallisarkiston yhteisessä hankkeessa. Hanke päättyi 2014 eikä aineistoja koskaan viety DA:n tallennusjärjestelmiin.
- Digitaaliarkiston ”vanhan” tietoaineiston migraatio
  - Vanhempi tietoaineisto, jonka prosessointiin tehtiin muutoksia, sillä rakenne oli erilainen.

KANSALLISARKISTO

Laadunvarmistus

- Mikäli kaikki ei ollut ”normaalisti”, käsiteltiin arkistoyksikkö tiimin toimesta
- Luotiin suuria määriä eri virhetyyppejä, osallistettiin myös muita laitoksen asiantuntijoita
- Kaikki:
  - Tiivistesummat
  - Kuvamäärät
  - Mikäli näissä eroja, ei ay liikkunut eteenpäin

KANSALLISARKISTO

Lukemat

Aineistotyyppi	Määrä Digitaaliarkisto	Datamäärä SAPA	Datamäärä KP-PAS	Kpl-määrät arkistoyksikkö	Kpl-määrät tiedostot
Tallekappaleet	~1 PB (1000 TB)	x	222 TB	505 628	78 189 766
Käyttökappaleet	~120 TB	142 TB	x	x	172 106 927
Arkistoyksiköiden määrä	502 828 kpl			497 967	
OCR-tietuiden määrä (AltoXML)	Ei lainkaan	Ei saatavilla	x		15 727 395

[1] KP-PAS sisältää myös epäonnistuneesti tai vaillinaisesti siirretyt yksiköt, jotka korjattiin myöhemmin ja siirrettiin uudelleen. SAPA on hyväksynyt vastaanotossaan saman määrän. SAPA:n luvusta puuttuu ne yksiköt, jotka on siirretty uudelleen.