Digitoitujen vp-asiakirjojen palvelu

Tietohallintopäällikkö
Ari Apilo
ari.apilo@eduskunta.fi
28.9.2018

http://avoindata.eduskunta.fi/digitoidut/

Vuosien 1907-2000 vp -asiakirjat

- Suomen yksikamarinen eduskunta on toiminut toukokuusta 1907 alkaen ja tuottanut laajan aineiston
- N. 1830 kpl nidettä sidottuja vp-asiakirjoja
- Keskimäärin 1000 s/kirja
- Yhteensä n. 1,9 miljoonaa digitoitavaa sivua

Digitoinnin lähtökohdat ja tavoitteet – miksi?

- Eduskunnan verkkopalvelussa oli vp-asiakirjat vuodesta 2000 alkaen mutta vanhemmat puuttuivat
- Käsittelytietoja oli saatavilla digitaalisina vuodesta 1970
- Erityisesti vuosien 1991-2000 asiakirjojen puuttuminen aiheutti voimakasta kritiikkiä
  - Mm. Perustuslain esityöt puuttuivat palvelusta
  - 1990-luvulla tehtiin merkittäviä lainsäädännöllisiä uudistuksia esim. EU-jäsenyys

Digitoinnin lähtökohdat ja tavoitteet – miksi?

- Digitaalisten asiakirjojen käsittely on helpompaa kuin paksujen kirjojen
- Vanhoja vp-asiakirjoja tarvitaaan edelleen tutkimustarkoituksiin
- Painetun tekstin tunnistaminen mahdollistaa tekstihaut
  - hakemistokirjojen käyttö on hidasta eikä aina tuo oikeita tuloksia
- Digitoitu aineisto voidaan arkistoida digitaalisesti
  - Vain n. 40 täydellistä painettua asiakirjasarjaa jäljellä

Miten palvelun kehittäminen eteni?

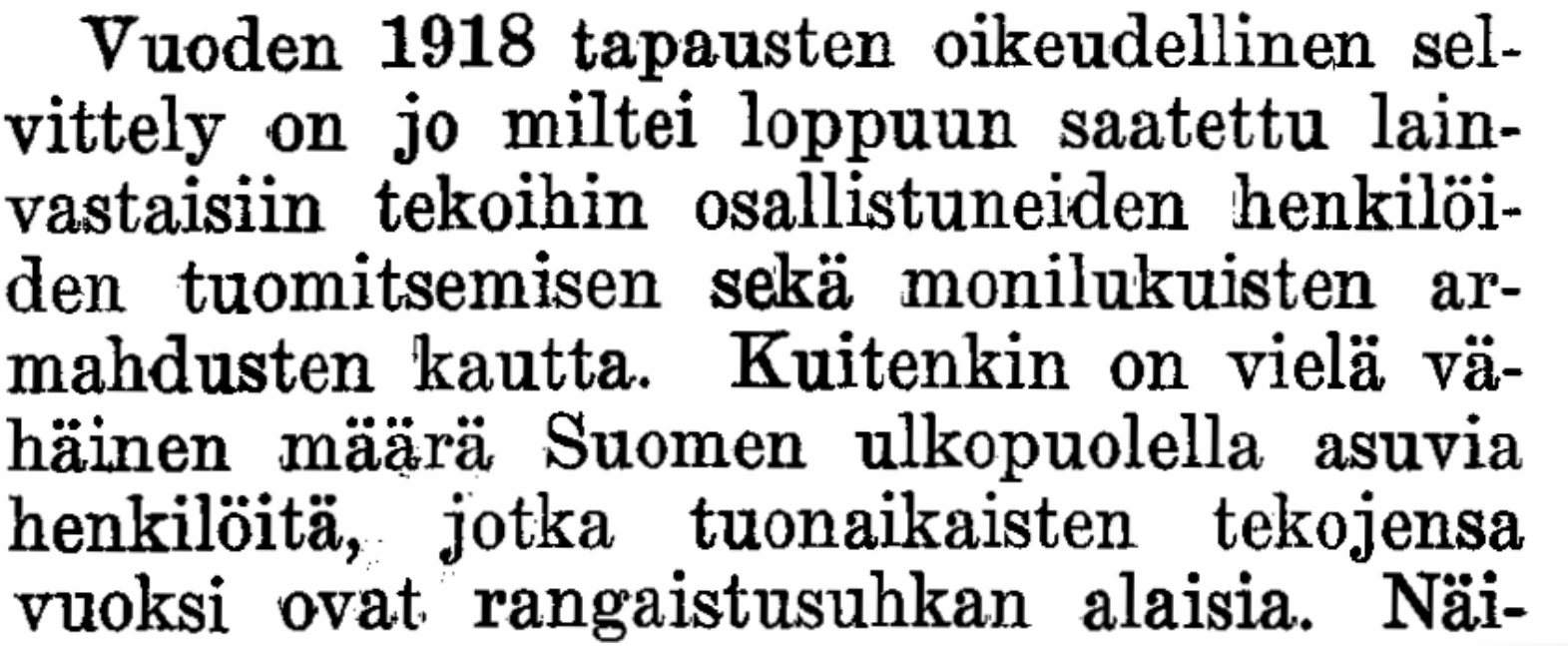
- ensimmäiset tulokset eduskunnan verkkopalvelussa 26.6.2015: vuosien 1998-2000 hallituksen esitykset
- Tammikuussa 2016 lisättiin eduskunnan verkkopalveluun vuosien 1991-2000 digitoidut vp-asiakirjat (erillisinä asiakirjoina)
- 1980-luvun suomenkielisiä valtiopäiväasiakirjat julkaistiin eduskunnan verkkopalvelussa kesällä 2017 (erillisinä asiakirjoina)
- Marraskuussa 2017 saatiin kirjojen digitointi vietyä vuoteen 1907 asti
  - n. 1800 kpl kirjoja, jossa keskimäärin vajaat 1000 sivua
  - OCR-teksti mukana - taso vaihtelee
- Erillisen jakelupalvelun kehittäminen 2018 keväällä; pilotti avattiin syksyllä 2018

Laatuvaatimukset

- asiakirjojen tiedostostandardi pysyvään arkistointiin sopiva PDF/A-1b
- riittävä skannaustarkkuus, jotta asiakirjat on helppo tulostaa ja lukea
  - 300 dpi mustavalkoskannaus
- tehokas mustavalkoinen skannaus, jotta tiedostot pystytään jakamaan tehokkaasti verkossa (n. 52 Kb/sivu)
  - harmaasävy ja värikuvat: JPEG-formaatti
  - häviöllisestä JPIG2-pakkauksesta luovuttiin, koska se saattaa aiheuttaa sisältövirheitä
- tekstintunnistuksen (ocr) laatu tulee olla riittävä hakuja ja tekstikopiointia varten
- asiakirjan nimi ja muut metatiedot ovat oikein
  - esim. he_1+1999.pdf, pevm_20+2000.pdf jne
  - asiakirjan otsikko (Title) on pitkä tunniste esim. ”Hallituksen esitys HE 1/1998 vp”

Skannaustarkkuus

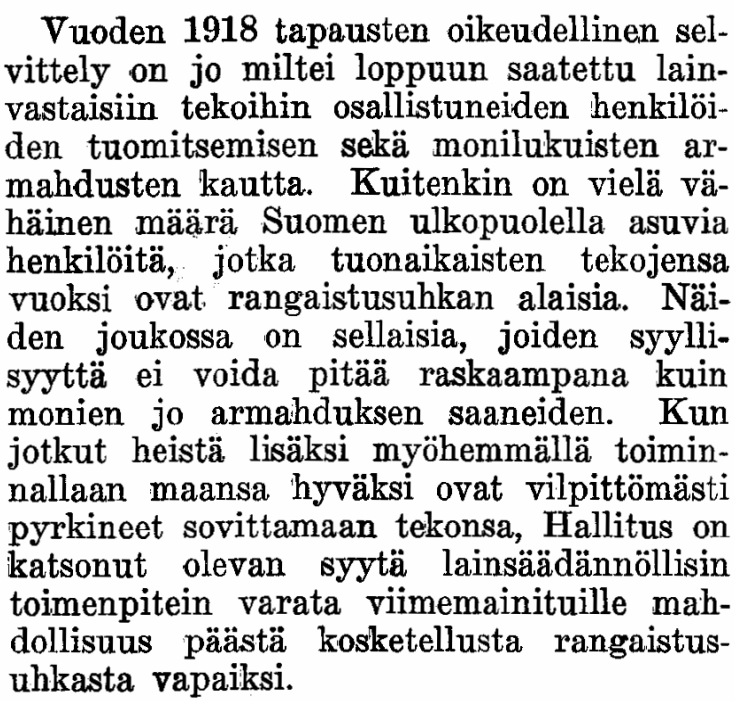
-> painolaatu ja tekstisisältö näkyvät helposti:
300 dpi tarkkuus riittää!



Vuoden 1918 tapausten oikeudellinen selvittely on jo miltei loppuun saatettu lainvastaisiin tekoihin osallistuneiden henkilöiden tuomitsemisen sekä monilukuisten armahdusten kautta. Kuitenkin on vielä vähäinen määrä Suomen ulkopuolella asuvia henkilöitä, jotka tuonaikaisten tekojensa vuoksi ovat rangaistusuhkan alaisia. Näi-

Tekstintunnistus ja skannaustarkkuus

Teksti kuvana



Vuoden 1918 tapausten oikeudellinen selvittely on jo miltei loppuun saatettu lainvastaisiin tekoihin osallistuneiden henkilöiden tuomitsemisen sekä monilukuisten armahdusten kautta. Kuitenkin on vielä vähäinen määrä Suomen ulkopuolella asuvia henkilöitä, jotka tuonaikaisten tekojensa vuoksi ovat rangaistusuhkan alaisia. Näiden joukossa on sellaisia, joiden syyllisyyttä ei voida pitää raskaampana kuin monien jo armahduksen saaneiden. Kun jotkut heistä lisäksi myöhemmällä toiminnallaan maansa hyväksi ovat vilpittömästi pyrkineet sovittamaan tekonsa, Hallitus on katsonut olevan syytä lainsäädännöllisin toimenpitein varata viimemainituille mahdollisuus päästä kosketellusta rangaistusuhkasta vapaiksi.

Tunnistettu teksti (OCR)

- Vuoden 1918 tapausten oikeudellinen selvittely on jo miltei loppuun saatettu lainvastaisiin tekoihin osallistuneiden henkilöiden tuomitsemisen sekä monilukuisten armahdusten kautta. Kuitenkin on vielä vähäinen määrä Suomen ulkopuolella asuvia henkilöitä, jotka tuonaikaisten tekojensa vuoksi ovat. rangaistusuhkan alaisia. Näiden joukossa on sellaisia, joiden syyllisyyttä ei voida pitää raskaampana kuin monien jo armahduksen saaneiden. Kun jotkut heistä lisäksi myöhemmällä toiminnallaan maansa hyväksi ovat vilpittömästi pyrkineet sovittamaan tekonsa, Hallitus on katsonut olevan syytä lainsäädännönisin toimenpitein varata viimemainituilla mahdollisuus päästä kosketellusta rangaistusuhkasta vapaiksi.

-> 3 virhettä 695 kirjaimessa = 0,43 % merkeistä on virheellisiä

OCR tulos ei ole koskaan täydellinen – mutta toivottavasti riittävä!

Tekstintunnistuksen virheiden merkitys palvelun käyttäjälle

- Oletettavaa, että 0,3-0,8 % tunnistetuista merkeistä poikkeaa alkuperäisestä painetusta tekstistä
  - OCR tulos ei ole koskaan täydellinen
  - Painolaatu huonoin 1920-1930 luvun aineistossa – eniten virheitä
- Jotkut sanat ovat jakautuneet kahteen osaan tavutuksen vuoksi
  - OCR ohjelma on pystynyt yhdistämään suurimman osan näistä sanoista – joitain irrallisia tavuja voi silti olla
- Merkitys käyttäjälle: tekstihaun tulos ei ole täydellinen vaan suuntaa-antava
- Rinnalla kannattaa edelleen käyttää hakemistoja

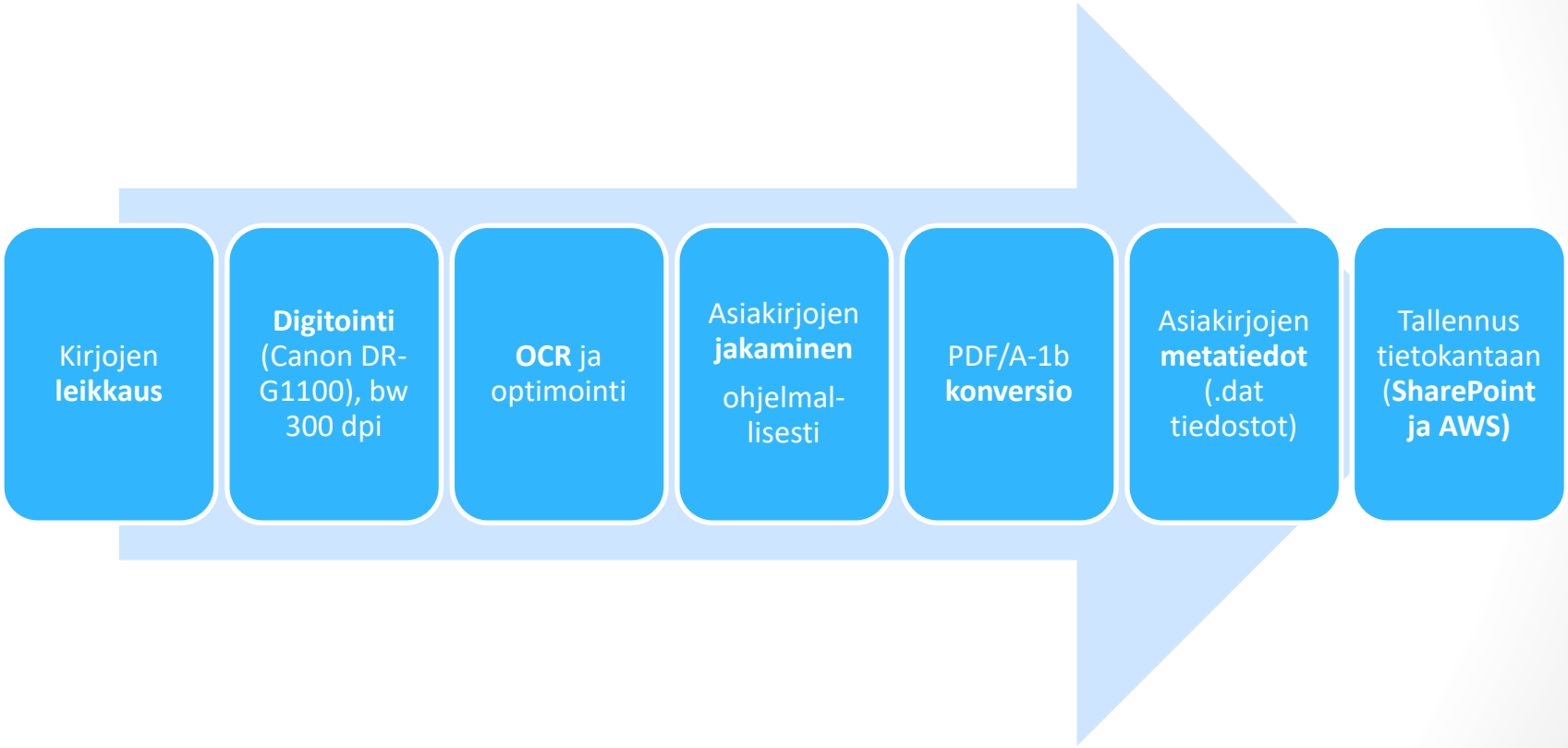
Miten digitointi tehtiin?

- tuotantoprosessi:
  - Kirja-aineiston valinta ja haku digitoitavaksi
  - Esikäsittely: sidosten purkaminen leikkaamalla
  - Digitointi kuvamuotoon skannaamalla (300 dpi, mustavalkoinen, Canon DR-G1100); sivujen kääntö pystyyn

Miten digitointi tehtiin?

- tuotantoprosessi:
  - Kuva-aineiston ohjelmallinen korjaus mm. sivujen suoristus ja ohjelmallinen tekstintunnistus (OCR) (Adobe Acrobat X)
  - Asiakirjoiksi jakaminen ja metatietojen lisääminen räätälöidyllä ohjelmalla
  - Ohjelmallinen muuntaminen julkaistavaan muotoon (Adobe Acrobat X Preflight PDF/A-1b)
  - Tallennus julkaisualustalle sekä indeksointi (SharePoint / AWS)
  - Laadunvarmistus/validointi (ohjelmallinen ja manuaalinen) eri vaiheissa

Digitointiprosessi kaaviona



Digitoidut valtiopäiväasiakirjat -jakelupalvelu

- Jakelupalvelun suunnittelu ja toteutus keväällä 2018; eteneminen iteroiden; pieni kehitystiimi
- Palvelun pilotti valmistui 5/2018; palvelu käyttöön 9/2018
- Pilvipalvelu: tehty Amazon Web Servicen (AWS) -teknologian avulla
  - Elastic Search-käytössä
- Eduskunnan kumppanina Gofore Oy

Digitoidut valtiopäiväasiakirjat -jakelupalvelu

- Palvelusta löytyvät suomen- ja ruotsinkieliset valtiopäiväasiakirjat vuosilta 1907-2000
  - Kaikki käsiteltyjen asioiden asiakirjat, täysistuntopöytäkirjat eli puheet sekä asiahakemistot
- Vapaatekstihaku ocr-tekstistä
  - Haun rajaus vp-vuodella ja asiakirjatyypillä
  - Tekstit tuodaan luettavaksi 100 sivun ”nippuina” (tiedostokoko n. 5 MB)
- Latausmahdollisuus kirjoittain (digitaalinen kirjahylly)
  - Kirjojen tiedostokoko 50-100 MB/kirja – latausaika riippuu yhteyden nopeudesta
- Palvelua voi käyttää:
  - http://avoindata.eduskunta.fi/digitoidut/
- Toimii parhaimmin Google Chrome-selaimella