Hankepäivityksiä DLM Forumissa Hollannissa

Mikkelin Digitalian sähköpostiarkistointi herätti kiinnostusta

Anssi Jääskeläinen
Miia Kosonen
TKI-asiantuntijat
Digitalia-tutkimuskeskus
Mikkelin ammattikorkeakoulu

Tämänkertainen DLM-Forumin jäsenkokous pidettiin Haagissa, noin 30 km Amsterdamista lounaaseen. Hollannin kansallisarkisto toimi tilaisuuden isäntänä. Varsinainen tapahtuma pidettiin 8.-9.6., mutta liikenneyhteyksien vuoksi oli järkevää mennä jo tiistaina ja palata perjantaina. Näin saa aina myös hieman aikaa tutustua kohteeseen sekä paikalliseen kulttuuriin.

Puheenjohtaja *Zoltán Szatucsek* kiitteli järjestäjiä ja avasi tapaamisen "coolissa ja calmissa" kaupungissa. Varsinaisen avauspuheen piti *Marens Engelhard*, Hollannin Kansallisarkiston johtaja. Puheessaan hän muistutti, että tieto on demokratian valuuttaa ja sivusi avoimen hallinnon lakia, joka tulee vaikuttamaan kaikkiin. Lopuksi hän totesi kentän olevan aivan liian laaja siihen, että kaikki tähtäisivät maaliin yksinään. Eri toimijoiden on tehtävä yhteistyötä, jotta lopputuloksesta tulee mahdollisimman hyvä.

Marensin jälkeen puheenvuoron sai *Joachim Jung* Open Preservation Foundationista (OPF). Puhe oli otsikoitu osuvasti "404 Page not found". Hän aloitti puheensa napakasti toteamalla, että kaikilla T&K -hankkeilla tulisi olla jatkuvuussuunnitelma (sustainability plan). Harvalla on. Tällä Jung viittasi myös siihen, että epäonnistuneiden kokeiden tulosten ja virheiden pitäisi olla muiden toimijoiden löydettävissä ja käytettävissä. Näin voitaisiin vähentää samojen virheiden toistamista T&K -kentällä. Tällä hetkellä OPF:llä on 23 jäsentä 13 eri maassa. He ylläpitävät mm. COPTR –työkalurekisteriä, jossa on saatavilla noin 350 digitaaliseen säilyttämiseen tavalla tai toisella liittyvää ohjelmistoa. Yksi kiinnostavimmista on verapdf, pdf-tiedostojen validointityökalu.

Nykyisellään verapdf on vasta kehitysvaiheessa, mutta Jungin mukaan se pystyy validoimaan kaikki tämänhetkiset pdf-versiot sekä niiden muunnelmat. Viimeisenä mainittakoon, että kaikki OPF:n ylläpitämät ja kehittämät työkalut ovat täysin avoimeen lähdekoodiin perustuvia, joten kuka tahansa voi osallistua kehitykseen ja testaukseen tai vain käyttää niitä.

Seuraavana oli vuorossa koko tämänkertaisen DLM-Forumin käytännönläheisin ja valmiin ratkaisun sisältävä esitys. Tämä oli Mikkelin ammattikorkeakoulun Digitalia -tutkimuskeskuksen puheenvuoro, jonka pitivät Anssi Jääskeläinen ja Miia Kosonen. Kosonen esitteli Digitalian toimintaa ja Kansalaisarkiston tähänastista kehitys- ja pilotointityötä, minkä jälkeen Jääskeläinen käänsi osallistujien huomion sähköpostien arkistoinnin dilemmaan erityisesti tavallisen käyttäjän näkökulmasta.

"Eri maiden arkistolaitoksilla on erinäisiä suosituksia säilytettäviksi sähköpostiformaateiksi, mutta kaikkiin ratkaisuihin liittyy tavalla tai toisella OAIS-mallin mukaiset SIP, AIP ja DIP -paketit. Tavallinen kadun kulkija ei ole kiinnostunut standardien mukaisista säilytyspaketeista: häntä kiinnostaa sähköpostiensa helppo luettavuus ja siirrettävyys eri laitteiden ja käyttöjärjestelmien välillä", Jääskeläinen arvioi.

Toimintavalmis sähköpostiratkaisu

Esityksessä nähtiinkin lähes toimintavalmis ratkaisu, jolla noin 10 000 sähköpostia sisältävä 1 Gt:n kokoinen Outlookin .pst tiedosto saadaan muutettua noin kahdeksassa minuutissa PDF/A-2b tiedostoiksi. Näissä muutetuissa sähköposteissa on alkuperäisen sähköpostin ulkonäkö ja tuntuma sekä kaikki alkuperäisen sähköpostitiedoston metatiedot, mukaan lukien normaalisti käyttäjälle näkymättömät välitys- ja otsikkotiedot.

DLM Forumin näyttämönä oli Hollannin kansallisarkisto National Archief. Kuva: Miia Kosonen

Puheenvuorot herättivät runsaasti keskustelua sekä positiivisia kommentteja. Muun muassa OPF:n Jung pyysi testaamaan ratkaisua heidän työkaluillaan ja maailman meteorologisen instituutin edustaja *Susana Rodriguez* totesi: "Vihdoin käytännöllinen ratkaisu todelliseen ongelmaan". Hän myös totesi olevansa myöhemmin yhteydessä ratkaisuumme liittyen. DLM-Forumin jälkeen toimivuutta on kehitetty ja tällä hetkellä ratkaisu tuottaa PDF/A-3b validoituja tiedostoja jotka sisältävät kaikki alkuperäisen sähköpostin metatiedot sekä sen mukana mahdollisesti tulleet liitetiedostot. Ratkaisusta voi kysellä lisää Digitalia tutkimuskeskukselta.

Seuraavaksi palattiin takaisin ruotuun teoreettisempien puheenvuorojen muodossa. Vuorossa oli Leeuwardenin kaupungista *Rienk Jonkerin* esitys tiedosta konseptuaalisena labyrinttina. Hän esitteli kehittämäänsä tietomallia, josta kiinnostuneet voivat itse tutkia asiaa tarkemmin jutun lopussa olevan kalvolinkin kautta.

Tämän jälkeen esiteltiin automaattista mallia, jolla päätetään, otetaanko sähköinen asiakirja pysyvään säilytykseen vai hävitetäänkö se. Tämän puheenvuoron piti *Marcel van Vlaanderen* (Ministry of the Interior and Kingdom Relations). Automaattinen seulontapäätös vaatii mahdollisimman tarkan optisen merkintunnistuksen (OCR), mahdollisesti jälkikorjausta ja jopa mahdollisesti skannattujen kuvien korjaamista ennen OCR:ää. Mitään tuloksia puheenvuorossa ei esitelty, joten oletamme, että tämäkin oli suunnitteilla tai kehitteillä oleva projekti.

Myös seuraava puheenvuoro Hollannin kansallisarkiston *Mette van Esseniltä* käsitteli tietojen automaattista luokittelua. Heillä oli kehitteillä Electronic Discovery Reference Model v. 3.0. Työnsä rakenteistamattoman tiedon lähteenä he olivat käyttäneet sähköposteja. Tätä perusteltiin sillä, että yhden sekunnin aikana lähetetään 2,5 miljoonaa sähköpostia, mikä ylittää selvästi esimerkiksi tweettien tai Youtubeen lähettävien videoiden määrän (väite perustuu http://www.internetlivestats.com/one-second/ sivuston tietoihin).

Automaattinen seulontapäätös vaatii mahdollisimman tarkan optisen merkintunnistuksen (OCR), mahdollisesti jälkikorjausta ja jopa mahdollisesti skannattujen kuvien korjaamista ennen OCR:ää.

Tällä hetkellä työssä on menossa ensimmäinen vaihe, jossa tunnistetaan käyttötapaukset sekä kerätään dataa. Seuraavana vaiheena on teknologioiden ja työkalujen valinta, ja tässä vaiheessa ainakin Mamk on kuulolla. Loppukaneettinaan hän kehotti luottamaan teknologiaan ja totesi, että ihmiset luulevat edelleen tekevänsä asiat paremmin kuin koneet ja teknologiset ratkaisut.

Päivän viimeinen tapahtuma oli jäsenkokous, jossa esiteltiin DLM:n toimintaa, taloutta sekä tulevaisuutta. Riikan jäsenkokouksen jälkeen talous on korjaantunut ja uusi verkkosivusto toiminnassa. Lisäksi on palkattu pysyvä sihteeri. Juuri tehdyn tarkastuksen mukaan DLM-Forumilla on tällä hetkellä 43 jäsenmaksunsa maksanutta jäsentä. Suomestakin on kolme jäsentä, mutta vain yhdestä oli jälleen edustus paikalla, Mamkista. Esimerkiksi Tanskan kansallisarkiston *Jan Dalsten Sørensen* uteli, missä kollegat Suomen Arkistolaitokselta ovat, koska heitä ei ole hetkeen näkynyt jäsenkokouksissa. DLM-Forumin toimeenpanevaan komiteaan (executive committee) on tulossa kuusi avointa paikkaa, joten tässä olisi hyvä mahdollisuus päästä vaikuttamaan Forumin asioihin paremmin.

Varsinaisen MoReq-speksin tilanne on pitkälti ennallaan. IST/INESC-ID Portugalin johdolla speksistä on tehty täysin nettipohjainen versio, joka pitäisi julkaista muutaman viikon sisällä. *Jon Garde* puolestaan kertoi samat asiat MoReq:n laajettavuudesta ja myönsi tehneensä virheen siinä, että MoReq 2010 julkaistiin 500-sivuisena jättiläisenä eikä eroteltuna spesifikaationa ja teknisenä lisäosana.

Ennen seuraavaa jäsenkokousta toimenpidelistalla on E-arkin seu-

raajaprojektin suunnittelua, EU:n tietosuoja-asetuksen (GDPR, general data protection regulation) käyttöönottoon liittyvän ohjeistuksen tekeminen sekä seuraavan Brightonissa pidettävän Triennial-konferenssin järjestelyjä. Lopuksi jäsenkokouksessa julkistettiin seuraavan kokouksen pitopaikka, joka ei ole Slovakia, koska siellä ei ole yhtäkään jäsentä. Seuraava jäsenkokous pidetään Oslossa 15.-17.11.2016. Näistä kaksi ensimmäistä päivää on DLM-Forumin kokous ja viimeinen päivä E-Arkiin liittyvä lisäpäivä. Vuoden 2017 ensimmäisen jäsenkokouksen pitopaikka huhtitoukokuussa varmistettiin Maltaksi, jolloin isäntänä toimii pieni Maltan Kansallisarkisto.

Torstaiaamun avasi Susana Rodriguez (World Meteorological Organization), joka kertoi uuden EDMS (Electronic Document Management System) -järjestelmän kehittämisestä standardien mukaan. Tähän asti kaikki kuulostaa helpolta, mutta Rodriguezin espanjalainen veri kuohahti hänen kertoessaan suosta, joka syntyy, kun huomioon pitää ottaa dokumenttienhallinnan standardit, EU:n tietosuoja-asetus ja Sveitsin (päämajan) omat standardit. Hän esittikin aiheellisen kysymyksen, onko standardi apu vai hienosti ilmaistuna kipu takapuolessa. Ongelmaan on hänen mukaansa ratkaisuna se, että yhtäkään standardia ei noudateta orjallisesti, vaan niitä räätälöidään omaan käyttötarkoitukseen sopivaksi. Toiseksi, kaikkien kehittämiseen osallistuvien tahojen on oltava tietoisia, siitä mitä standardia seurataan ja miksi sitä muovataan.

Seuraavaksi *Erik Saaman* kertoi DUTO:sta. Se on lyhennelmä hollanninkielisistä sanoista, jotka vapaasti suomentaen tarkoittavat vaatimuksia, joilla taataan esteetön pääsy hallinnolliseen tietoon. Hollannissa pysyvästi säilytettävät asiakirjat siirretään vasta 20 vuoden päästä arkistoivalle instituutille. Saamanin mukaan arkistoinnin kannalta tämä on huono asia, jos asiakirjojen (joko sähköisten tai paperisten) hallinta ei ole ollut kunnossa koko niiden elinkaaren ajan. Hän myös totesi, että arkistointi tai säilyttäminen ei saa olla päämäärä, vaan esteetön pääsy tietoon. Tiedon sijainnilla tai järjestelmällä ei ole väliä, kunhan pääsy on taattu. Ongelmaksi hän näki edelleen paperimaailman säännöillä toimivan asiakirjahallinnan.

MoReq:n isä Jon Garde esitteli katsauksen "lapsensa" nykyisestä tilanteesta. Suurimpana haasteena Jon kertoi olevan EU:n tulevan tietosuojadirektiivin, jonka mukaan henkilötietoja sisältävien järjestelmien tulee pystyä pyynnöstä etsimään kaikki henkilöön liittyvät tiedot. Lisäksi kyseisen määräykseen liittyy oikeus tulla unohdetuksi. Tätä Garde esitti vertauskuvalla. Suuri osa lukijoista tietänee, että selaimet tarjoutuvat muistamaan luottokorttitiedot niitä syötettäessä. Tässä aukeavassa valikossa on myös vaihtoehto "ei koskaan tällä sivulla". Kysymys kuuluukin: miten selain voi muistaa, että korttia ei saa muistaa, jos siitä ei saa tallentaa mitään tietoja? Garde totesi, että MoReqiin tarvitaan jonkinlainen identiteetin hallintapalvelu.

André van der Meer SAS-instituutista oli otsikoinut puheensa "Big Dataa ei enää ole!". Esitys pohjautui pitkälti siihen, että todellinen ongelma on business datan louhinta big datasta ja että prosessijohtoisesta tiedosta pitäisi siirtyä tietojohtoisiin prosesseihin. Tietomassoihin liittyen hän esittikin aiheellisen kysymyksen: suorittaako tiedoille seulonta etukäteen vai tallentaa kaikki mahdollinen ja pyrkiä tekemään syväluotaavaa analyysiä koko massalle, jolloin asioiden suhteet, seuraukset ja syyt voisivat tulla paremmin esiin?

Asiakirjat vaihtuvat Virossa X-roadilla

DLM Forumin vakiovierailijoihin kuuluva *Liivi Karpistsenko* Viron talous- ja viestintäministeriöstä kysyi puheessaan "Missä asiakirjat elävät?". Hän kertoi, kuinka muutos Virossa vuoden 2013 viimeiseltä neljännekseltä vuoden 2014 viimeiselle neljännekselle oli ollut huomattava. Noin neljästä miljoonasta EDRMS-järjestelmien välillä vaihdetuista asiakirjoista päästiin noin 50 miljoonaan muiden tietojärjestelmien välillä vaihdettuun asiakirjaan. Osansa tässä kehityksessä on luonnollisesti X-Roadilla, mutta suurempi vaikutus on ollut muutosjohtamisella ja kanssakäymisellä eri osapuolten kanssa. Karpistsenko totesi, että erinäisillä pakottavilla toimenpiteillä ei saada tuloksia aikaiseksi. Muutosvastarinta on pienintä, jos muutoksen kohteet ovat olleet mukana muutoksen toteuttamisessa. Tähän huomautamme, että muutosvastarinnan pienentämiseen voi riittää jopa mielikuva siitä, että mukanaolo on vaikuttanut muutoksen lopputulokseen.

Elena Cortes Ruiz Espanjan koulutus-, kulttuuri- ja urheiluministeriöstä kertoi, kuinka taataan kansalaisten laillinen pääsy julkiseen sähköiseen tietoon. Pyrkimyksenä on saada kaikki hallintoon liittyvä tieto sähköisenä yhden portaalin kautta saataville. Sen lisäksi, että hallinnollinen tieto on oltava saatavilla, kansalaisten tulee päästä käsiksi myös itseään koskevaan hallinnol-

liseen tietoon sekä Espanjan kulttuuriperintöaineistoon.

Kaksi viimeistä sessiota esittelivät E-Ark -projektia niin ylä- kuin alatasoiltakin. Tässä käsitellään yleisesti kyseisessä projektissa meneillään olevia asioita. E-Ark -projektin hallinto on siirtynyt Portsmouthin yliopistosta Brightonin yliopistolle. Projekti on edennyt muutenkin kaikin puolin hyvin ja EU:n review-prosessissa oli saatu hyvä arvostelut. E-Arkin projektipäällikön (Janet Delve) mukaan erinomaista arviointia ei voi saada kuin viimeisenä toimintavuotena, mutta he olisivat sen ansainneet jo nyt, koska kaikki tehtävät on tehty ajallaan, talous on kunnossa jne. Lisäksi puhetta oli myös E-Arkin pakettien yleismäärittelystä (common specification), joista on tehty Einsteinin mukaan yksinkertaisia "everything should be made as simple as possible but not simpler", sekä erinäisistä E-Arkiin liittyvistä prosesseista.

Myös relaatiotietokantojen arkistoinnista, siirtämisestä sekä pääsystä E-Arkissa kehitetyillä tai muokatuilla työkaluilla kerrottiin. Näissä työkaluissa avainasemassa ovat SIARD-formaatti sekä täysin avoimeen lähdekoodin perustuva Roda (Repository of authentic digital objects) -arkistointiratkaisu. Päivän viimeisessä sessiossa jatkettiin E-Ark linjalla ja puhuttiin mm. EARK-Web-alustan prototyypistä, joka toteuttaa E-Arkin työprosessin (workflow) aina SIP-paketista arkistoon asti.

Haagin DLM-Forumin kaikki esitykset löytyvät http://goo.gl/RLNWaE linkin takaa.