GeoForum ja DLF-Forum Prahassa

Anssi Jääskeläinen
Tutkimus-päällikkö
Xamk

Tämän kertainen DLM-Forum pidettiin Prahassa, jonka konferenssikeskuksessa puitteet olivat loistavat. Samaan aikaan järjestettiin myös joitakin ministeritason tapaamisia ja turvatoimet olivat sen mukaiset. Suomesta oli tällä kertaa mukaan kolme osallistujaa, kirjoittajan lisäksi Kati Saltiola Xamkista sekä Markus Merenmies Kansallisarkistosta. Ennen varsinaista konferenssia pidettiin myös päivän mittainen GeoForum Tšekin kansallisarkiston tiloissa kaupungin laitamilla.

Ennen tapahtumien sisältöjen esittelyä mainitsen muutamia asioita, joihin suomalaistenkin toimijoiden kannattaisi kiinnittää huomiota tekemisissään:

- Arkistointi ei ole päätarkoitus, vaan sen tulee luoda uutta (julkista) arvoa
- FAIR periaatteiden noudattaminen tiedonhallinnassa
- Annotointia ei välttämättä tarvita ollenkaan, jos teknisiä komponentteja osataan hyödyntää ristiin

GeoForum

GeoForum järjestettiin nyt toista kertaa, ja tapahtuma hakee selvästi vasta muotoaan ja osallistujiaan. Paikan päällä oli puhujien lisäksi noin 20 osallistujaa ja verkossa jonkin verran lisää. Puheenvuorot liittyivät pitkälti geoinformaation luomiseen, hallintaan, säilytykseen ja jakeluun.

Gregor Završnik Geoarhilta puhui geoinformaatioon liittyvien työnkulkujen mukauttamisesta sekä tiedon uudelleen käytöstä FAIR periaatteiden mukaisesti. Näistä olen kirjoittanut aiemminkin mutta kertauksen vuoksi FAIR tarkoittaa F=Findable (löydettävä) A=Accessible (saavutettava), I=Interoperable (yhteentoimiva) ja R=Reusable (uudelleenkäytettävä). Näihin neljään termiin sisältyy paljon asioita, ja haastankin kaikki tiedonhallintapalveluita tai arkistointia tarjoavat toimijat tutkailemaan omia tuotteitaan näiden periaatteiden mukaisesti. Jos koko kirjainsarja täyttyy, niin mahdollisuus hyvälle asiakaskokemukselle on olemassa, mutta epätäydellisten kirjainsarjojen haltijoiden kannattaa varautua yhä valveutuneimpien asiakkaiden yhteydenottoihin.

GIS:in (Geographic Information System) tapauksessa sen FAIR periaatteita pystytään parantamaan hyödyntämällä alalle sopivia määrityksiä. Hieman yllättävästi Gregor kuitenkin totesi, että Inspire-direktiivin mukaiset määritykset eivät sovellu kuin tietynlaiselle geoinformaatiolle. Kun tallennetaan geodataa, mukaan pitää tallentaa myös ohjeet siitä, kuinka kartta muodostetaan. Esimerkkinä hän käytti kakun leipomista, jossa pelkät ainekset eivät riitä, vaan lisäksi tarvitaan valmistusohjeet. Geodatan tapauksessa ilman ohjeita päästään helposti tilanteeseen, jossa esimerkiksi Suomeen sijoittuvat datapisteet ovat keskellä Pohjanmerta (kts. *Faili* 4/21).

Hieman ehdotuksen omaisesti Gregor esitteli Geospatial CITS (Content Information Type Specification) määrityksiä, jotka laajentavat E-ARK SIP (Submission Information Package) määrityksiä. Käytännössä siis SIP-pakettiin tulisi mukaan geoinformaatioon liittyviä rakenteita ja metatietoja.

Toisessa puheenvuorossa Jaroslav Nechyba käsitteli BIM:iä (Building Information Model) joka mahdollistaa digitaalisen rakennusdatan jouhevamman käsittelyn, siirrettävyyden ja yhteen toimivuuden GIS-järjestelmien kanssa. Aihe on pinnalla myös Suomessa, jossa esimerkiksi Senaatti-kiinteistöt alkoi vaatia BIM:n hyödyntämistä jo vuonna 2007.

Jan Macura kertoi kuulijoille avoimeen lähdekoodiin perustuvasta Hub4Everybody-viitekehyksestä[1], jonka avulla erilaisten geodataa hyödyntävien palvelujen julkaisu on kuulemma niin helppoa, että kuka tahansa osaa sen tehdä. Lähtökohtaisesti tämä on hyvä ajatus, joka ammattisoftien kehittäjienkin kannattaisi pitää mielessä. Koko palvelun koodeja ei ilmeisesti jaeta avoimena, vaan hyödynnettävät oh-

1 https://hub4everybody.com/

Prahan konferenssikeskus. Kuva: Anssi Jääskeläinen.

jelmistot esitellään palvelusivuston Technologies välilehdellä. Ratkaisuna Hub4Everybody lienee suhteellisen tuore, koska Suomen alueelta palvelusta ei löydy yhtäkään karttakerrosta. Palveluun voi luoda karttoja myös kirjautumatta, mutta palvelu on kaupallinen, maksamalla saa luonnollisesti enemmän toiminnallisuuksia ja tukea.

Aamupäivän viimeinen puhuja Zuzana Syrova kertoi Tšekkiin tehdystä historiallisten paikkojen GIS-portaalista[2]. Portaalissa on osittainen kielituki englanniksi, mutta materiaalien ollessa tšekiksi on käyttö hieman haastavaa. Toiminnoiltaan ratkaisu vaikuttaa samanlaiselta kuin suomalainen Karttapaikka-palvelu, mutta se sisältää ilmeisesti myös ominaisuuksia joukkoistamiseen.

Michal Kepka esitteli ratkaisua, jossa geodataa hyödynnettiin modernissa multimediapohjaisessa turistioppaassa[3], joka keskittyi historialliseen Pilgrimin reittiin Bavarian metsissä. Opas sisältää muun muassa 3D-malleja, AR:ää ja karttoja. Tekniikasta kiinnostuneille voidaan mainita knoppitietona toteutuksessa käytetty CesiumJS-kirjasto, jolla 3D-geotietoa voidaan visualisoida www-selaimessa. Toteutuksessa oli kohdattu samoja ongelmia kuin Xamkissa meneillään olevassa King's Road Renascence -hankkeessa, jossa osaa historiallisista rakennuksista on vuosien saatossa entisöity ja tuhoutuneista on hyvin vähän tietoa olemassa. Näin ollen alkuperäisen ulkonäön selvittäminen mallinnusta varten on ollut hyvin aikaa vievää ja haasteellista. Lopputulos onkin yhdistelmä mittauksien, vanhojen dokumenttien, kuvien ja suunnitelmien perusteella tehtyjä 3D-malleja, joista oli toteutettu vielä soveltuvalla 3D-tulostimella pienoismallinäyttely.

DLM-Forum

Varsinainen DLM-Forum oli houkuttanut tällä kertaa paikalle noin 80 osallistujaa, joka on suurin osallistujamäärä pitkään aikaan. Verkossa osallistujia oli lisäksi noin 60. Krystyna Ohnesorge avasi tapahtuman DLM:n varapuheenjohtajan roolissa puheenjohtaja Anja Paulicin ollessa estynyt saapumasta paikalle.

Avauksessa oli mukana myös Tšekin sisäministeri Petr Vokác. Hän puhui EU-tason Digitaalinen kompassi 2030 -strategiasta, jota myös Tshekki pyrkii toteuttamaan. Hänen puheessaan oli luonnollisesti paljon poliittisella tasolla olevia lauseita, mutta hän mainitsi

2 https://geoportal.npu.cz/

3 https://peregrinus.online/current

myös tiedon kompleksisuuden ja määrän kasvamisen muodostavan arkistoille digitaalisen muutoksen haasteen. Tavoitteena on lisätä luottamusta tiedonjakopalveluihin sekä parantaa tiedon uudelleen käytettävyyttä. Prosessissa tulee huomioida tiedon koko elinkaari, ja arkistojen vastuulla on luonnollisesti syntyneen tiedon seulonta ja pitkäaikaissäilyttäminen. Hän totesi myös linkitetyn tiedon määrän ja sen mahdollisuuksien kasvavan. Vastauksena esitettyihin haasteisiin on kansainvälinen yhteistyö arkistojen ja alan toimijoiden keskuudessa. DLM-Forumilla on merkittävä rooli näiden toimenpiteiden toteuttajana ja alulle panijana.

Árpád Welker Euroopan komissiosta toi komission terveiset sähköiseen arkistointiin. Puhe sivusi digitaalisuuden vuosikymmentä sekä tavoitteita ja tarvetta osaavan työvoiman saatavuuteen. Tärkein asia, joka puheesta kannattaa muistaa, on toteamus, että arkistointi ei ole hostingpalvelua, kilpailua sovelluskehityksen kanssa eikä etenkään ”superpalvelua” kaikille mahdollisille tahoille.

Janet Anderson kertoi EARK-konsortion kuulumisia. CSP (Common Services Platform) -tarjouksesta ei ollut vielä kuultu mitään virallista, mutta jo Forumin seuraavan iltapäivän aikana saimme odotetun hyväksymistiedon komissiolta. Tässä konsortiossa on Suomesta mukana sekä Kansallisarkisto että Xamk. Janet totesi, että jatkumo EARK3-hankkeesta CSP:hen on ollut jouheva, koska jokaisessa neljästä rahoitetusta Generic Services -hankkeesta (joihin myös Xamkin koordinoima OneClick eArchiving kuuluu) on mukana alkuperäisen EARK3-konsortion jäsen.

Seuraavaksi käydyssä paneelikeskustelussa nostettiin esiin mielestäni aiheellinen kysymys. Mikä on kolmen samalla toiminta-alueella olevan organisaation Nestor, DLM-Forum ja OPF olemassaolon tarve erillisinä toimijoina? Tämä oli mielenkiintoinen kysymys siinäkin mielessä, että kolmesta keskustelijasta John Sheridan edusti DLM Forumia, Christian Keitel Nestoria ja Remco van Veenendaal OPF:ää. Lyhyesti esitettynä Nestor on pääasiassa Saksassa toimiva osaajaverkosto, joka keskittyy tiedon pitkäaikaissäilyttämiseen ja saavutettavuuteen, kun taas OPF on kansainvälinen organisaatio, joka työskentelee pitkäaikaissäilytyksen standardien ja työkalujen parissa. Muun muassa Digitalian työnkulkujenkin hyödyntämä VeraPDF-työkalu on OPF:n kehitystyön tulosta. Vastaukset esitettyyn kysymykseen pyörivät pitkälti yhteistyön ympärillä ja siinä, että jokaisella organisaatiolla on oma spesifinen toiminta-alueensa. On parempi tehdä yhteisyötä kuin yrittää tehdä kaikkea yhden suuremman toimijan alla. Tämä on verrattavissa muun muassa Linuxin sovellusohjelmien toimintalogiikkaan: ”Tee vain yksi asia ja tee se hyvin”.

Päivän toisessa sessiossa kuultiin muun muassa LINDAS-projektista[4], jossa on tehty konversioita olemassa olevista CMS ja RMS -järjestelmistä LInked DAta Service -järjestelmään. Kyseessä on RDF triplet -muotoinen tietovarasto, jossa jokaisella datasetillä on SPARQL-päätepisteet. Jos ja todennäköisesti kun nämä termistöt ovat lukijalle hepreaa, niin kyseessä on linkitetyn tiedon peruskäsitteitä, joita hyödynnetään muun muassa Finto sanasto- ja ontologiapalveluiden taustalla.

4 https://lindas.admin.ch

Toisessa puheenvuorossa käsiteltiin tietovarantojen, rekisterien, tietojärjestelmien ja palveluiden suureen määrään ja hallittavuuteen liittyvää ongelmaa. Tšekissä ongelmaa on lähdetty ratkaisemaan muutamalla perusrekisterillä, jotka toimivat katalogina muille tietoa sisältäville järjestelmille ja rekistereille.

Seuraavaksi esiteltiin DUTO:a (Hollannissa kehitelty toimintamalli, jolla digitaalinen tieto pyritään tuomaan saataville kestävällä tavalla) ja sen uudistettavaa viitekehystä. Vanhojen DUTO-määritysten havaittiin käytännön kokeiluissa olevan liian abstraktilla tasolla, joten uudessa painotetaan joustavuutta Mooren julkisen arvon kolmion mukaisesti.

Session viimeinen puheenvuoro olikin vaihtelun vuoksi tiedonhallinta- ja arkistoalan ulkopuolelta. Tšekit Jiráková ja Pavlinec puhuivat EU:n laajuisesta potilastietojen vaihdosta ja sen nykytilanteesta Tšekissä. Suomi on ollut mukana jo vuodesta 2018 lähtien. Toiminta perustuu Euroopan komission eHealth-määrityksiin, joka on terveyspuolella eArchiving-toimintaa vastaava komission hallinnoima kokonaisuus.

Ryhmäkeskustelujen (Roundtable) jälkeen oli päivän viimeisen session vuoro, jossa olikin Digitalian toiminnan kannalta kiinnostavimmat teknisemmät esitykset. Pavel Ircing West Bohemian yliopistolta kertoi heidän AI-laboratorionsa kehittelemästä audiovisuaalisen aineiston automaattisesta indeksoinnista, jossa hyödynnetään puheen- ja kuvantunnistusta. Toteutuksessa oli hyödynnetty Facebookin NLP-ryhmän kehittämää wav2vec-työkaluja, mutta se on kuitenkin riippuvainen annotoidun opetusdatan määrästä. Annotoidun opetusdatan puutteen vuoksi virheitä kompensoitiin käyt-

tämällä T5 (Text-To-Text Transfer Transformer) -työkalua yhdessä BERT (Bidirectional Encoder Representations from Transformer) -kielimallin kanssa. Tulokset ovat kuulemamme mukaan olleet vähintäänkin kelvollisia.

Päivän viimeinen esitys oli Digitalian meneillään olevien hankkeiden kannalta merkittävin. Michal Hradiš Brnon teknisestä yliopistosta kertoi keinoälyyn perustuvasta PeroOCR -työkalusta[5], joka esityksen mukaan toimii todella hyvin sekä koneettä käsinkirjoitetulle materiaalille. Kollegani Tuomo Räisänen Xamkilta on tämän työkalun ottanutkin jo käyttöön Digitalian palvelimella. Haittapuolena todettakoon, että Githubissa olevat koodit ja ohjeet ovat vielä suhteellisen vajavaisia. Helposta käyttöönotosta ei siis kannata vielä haaveilla.

Loistavan gaalaillallisen ja hyvien keskustelujen jälkeen seuraavan aamun avasi sessio, jossa myös allekirjoittanut kertoi OneClick eArchiving -hankkeessa toteutetusta helppokäyttöisestä Dockeroidusta SIP-paketin muodostajasta, joka on *Failin* julkaisuhetkellä jo avoimesti ladattavissa osoitteesta https://gitlab.com/jaaskela79/oneclick-full. Ennen latailuita ja testejä kannattaa kuitenkin huomioida, että hanke on meneillään 23.1.2023 loppuun saakka, joten työkalussa tapahtuu vielä potentiaalisia korjauksia ja toiminnan paranteluja, ja myös ohjeistusta korjataan testien ja palautteiden perusteella. Esitys herätti mielenkiintoa, koska validin SIP-paketin luominen ei ole kuitenkaan aivan yksinkertaisimpia asioita toteuttaa. Tällä hetkellä ratkaisu on kokeilukäytössä ainakin Latvian kansallisarkistossa. Kirjoittajaan voi olla yhteydessä digitalia.fi-sivustolta löytyvien yhteystietojen avulla, jos ratkaisusta haluaa lisätietoja.

Vincent Hooltin esitys samassa sessiossa liittyi tärkeiden asioiden löytämiseen sähköpostimassoista. Puheessaan hän viittasi monesti NARAn capstone-malliin, joka hänen mielestään ei ratkaise ongelmaa. Hollannissa oli päädytty siihen, että julkisten toimijoiden sähköpostiosoitteet on linkitetty julkiseen tehtävään, jota saa myös omalla vastuullaan käyttää henkilökohtaisten asioiden hoitamiseen. Lisäksi ennen sähköpostien arkistointia työntekijöille annetaan tietty aika tuhota henkilökohtaiset sähköpostit, mutta tästäkin huolimatta arkistoinnin yhteydessä on löytynyt muun muassa avioeron liittyviä henkilökohtaisia asiakirjoja.

Session viimeisessä puheessa Paul Young UK:n kansallisarkistosta puhui keinoälyn hyödyntämisestä seulontapäätösten tekemisessä. He olivat testanneet viittä kaupallista toimijaa noin 100 000 tiedoston testiaineistolla. Tärkeimpänä kriteerinä oli ollut: tärkeän tiedon häviämättömyys vs. ei tallenneta liikaa tavaraa. Heidän kokemuksiensa mukaan keinoälyn opetusvaihe oli ollut aikaa vievä ja opetusdatan pitää vastata todella hyvin sitä, mitä tullaan lopultakin seulomaan. Tulokset olivat lupaavia, mutta AI ei voi tässä kontekstissa ainakaan vielä korvata ihmisen tekemää työtä.

Seuraavan ryhmäkeskustelun jälkeen päästiinkin lounaan kautta DLM-Forumin viimeiseen sessioon, jossa puhumassa oli myös Xamkin Kati Saltiola aiheenaan Memory Campus ja Memory Lab. Molemmat aiheet ovat todennäköisesti Failin lukijakunnalle ainakin jollakin tasolla tuttuja. Viimeisen sessio muut esitykset käsittelivät Euroopan juutalaisten yhteisöarkistoa sekä lohkoketjutekniikan hyödyntämistä arkistoissa. Kaikkien esitysten videotaltionnit ovat nähtävissä DLM-Forumin jäsenille sivustolle kirjautumisen jälkeen.

Seuraava DLM-Forum järjestetään Ljubljanassa Sloveniassa 10–11.5.2023 ja GeoForum 8–9.5.2023. Lisäksi Archiving by Design -ryhmä kokoontuu 12.5.2023. Jos DLM-Forumiin liittyminen kiinnostaa, niin Forumin pääsivun alareunasta löytyy ohjeistusta ja jäsenmaksuhinnasto.

5 https://pero-ocr.fit.vutbr.cz/

Anssi Jääskeläinen aloittamassa esitystään naurettavan helposta SIP-paketoijasta. Kuva: Kati Saltiola.