Soile Manninen ja Niina Nurmi

Läpinäkyvyys tuo luotettavuutta – IDCC 2024 -konferenssin antia

18. International Data Curation Conference (IDCC) järjestettiin 19.–21.2.2024. Konferenssin pääjärjestäjänä oli monille tuttu Edinburghin ja Glasgow'n yliopistojen yhteinen Data Curation Centre (DCC). Vuosien aikana konferenssi on kasvanut, ja tällä kertaa ensimmäistä kertaa hybridinä järjestetty tilaisuus keräsi osallistujia yli 270, joista noin 60 etänä.

Maanantaina ennen varsinaista konferenssia osallistuimme työpajaan, jonka aiheena olivat koneluettavat aineistonhallintasuunnitelmat (maDMP) sekä niiden yhteentoimivuus. Samanaikaisesti järjestettiin muitakin työpajoja muun muassa pysyvistä tunnisteista (PIDit), repositorioiden luotettavuuden ohjeistuksista, tutkimusdatan jakamisen linjauksista Isossa-Britanniassa sekä eri FAIR-projektien tarjoamista yhteiskäyttöisistä resursseista.

Työpajailua DMP:stä

Työpaja alkoi DMP-työkalujen sekä aineistonhallintapalveluiden edustajien puheenvuoroilla, mutta varsinainen työskentelymme rakentui "Ten principles for machine-actionable data management plans" (2019) -artikkelin ympärille. Artikkeli koneluettavien DMP:iden kymmnestä periaattesta sai alkunsa IDCC:ssä vuonna 2017, ja tarkoituksena on toteuttaa kahdeksan vuoden kuluttua vastaava tarkastelu.

Edinburghin kuninkaallisen lääkäriseuran kokoelmiin kuuluu yli 70 000 teosta, mm. ensipainos ensimmäisestä painetusta lääketieteellisestä kirjasta, Celsuksen ***De Medicinasta*** (1478). Kuva: Soile Manninen CC BY 4.0

Pienryhmissä mietimme, mitä periaatteista pitäisi muotoilla uusiksi, mitä säilytetään ja mitä poistetaan. Yhteisten keskustelujen perusteella suurin osa periaatteista säilyy, mutta uudelleenmuotoiluja ehdotettiin runsaasti. Työpajan tuloksia lähdetään työstämään muun muassa Research Data Alliancen (RDA) eri työryhmissä.

Iltapäivän osuudessa tuotettiin erilaisia käyttäjätarinoita ja koottiin käyttäjien toiveita koneluettaville aineistonhallintasuunnitelmille, jos kaikki toteutuisi ”unelmien maailmassa”. Kehittämiskohteina nähtiin muun muassa tutkimustietojärjestelmistä saatavien tietojen parempi hyödyntäminen. Työpajan päätteeksi esiteltiin Salzburgin julistus, jossa eri DMP-työkaluja kehittävät ja ylläpitävät tahot sitoutuvat yhteistyössä edistämään koneluettavia aineistonhallintasuunnitelmia.

Maanantai-iltana nautimme tervetuliaismaljat Edinburghin kuninkaallisen lääkäriseuran (Royal College of Physicians of Edinburgh) tiloissa. Viktoriaaniset ja yrjönaikaiset puitteet olivat hienot, ja juhlasalin lisäksi pääsimme ihastelemaan Skotlannin vanhinta ja laajinta lääketieteellisen kirjallisuuden kokoelmaa, joka on perustettu jo vuonna 1682.

Kuratointi näkyväksi

IDCC-konferenssin tämän vuoden teema oli ”Trust through transparency”. Kun kerromme avoimen tieteen periaatteiden mukaisesti, miten tutkimustietoa kerätään, todennetaan ja käytetään, voimme luottaa myös tutkimuksen tuloksiin.

Varsinainen konferenssi käynnistyi tiistaina Ingrid Dillon (Data Archiving & Network Services, DANS) avauspuheenvuorolla, joka johdatteli yleisön sujuvasti konferenssin teemaan. Esityksen runkona toimi allegoria joesta, joka alkaa luottamuksen määritelmästä ja kiemurtelee luottamuksen eri elementtien, avoimen datan, tieteellisen petoksen ja datan todenperäisyyden kautta kahvitauolle. Puheenvuorossa nostettiin muun muassa esimerkkejä Alankomaissa tapahtuneesta tutkimusdatan väärentämisestä sekä väärin perustein tehdystä syytöksestä koskien tekoälyn käyttöä artikkelien kirjoittamisessa.

Aamupäivä jatkui rinnakkaissessioilla, joissa esiteltiin tuoreita tutkimusartikkeleita tai muita laajempia selvityksiä. Aiheena olivat aineistonhallinnan suunnittelu ja data-ammattilaisuus, kuratointiteknologiat ja -tekniikat, aineistojen jakaminen ja hävittäminen, standardit ja kestävyys.

Niina valitsi sessiot, joissa käsiteltiin luottamuksen ja läpinäkyvyyden rakentamista tutkimukseen, kuratointiin ja työnkulkuihin sekä kuratoinnin kustannuksia. Yhdysvalloissa toimivan datan kuratoinnin verkoston (DCN) esityksessä esiteltiin verkostossa tehtyjen haastattelujen (35 kpl) tuloksia. Kuratointi näyttäytyi suhteellisen yksinäisenä ja eristäytyneenä työnä. Haastateltavat korostivat kuratointi-

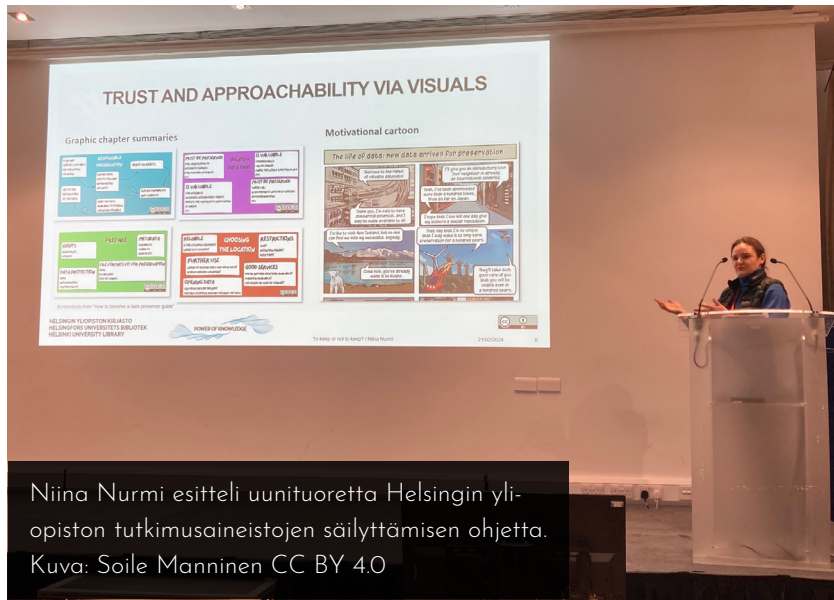
Kun tutkijat tekivät datanhallintaa itse, nousivat datanhallinnan kustannukset korkeammiksi kuin tutkijoilla, jotka hyödynsivät tukipalveluita.

työn näkyväksi tekemisen tärkeyttä ja painottivat, että kyseessä on ihmisten tekemää työtä ihmisille. Haastatteluissa ilmeni myös, että DCN-yhteisö on kuraattoreille korvaamaton tukiverkosto. Kuratoijat voivat ottaa yhteyttä verkostossa oleviin kollegoihin, kun he tarvitsevat neuvoja vaikkapa jonkin tieteenalan tiedostomuodoista. Myös organisaatiot hyötyvät, sillä kuratoinnin asiantuntijuuden kapasiteetti laajenee verkoston myötä.

Mitä kuratointi maksaa?

Wendy Kozslowski Cornellin yliopiston kirjastosta kertoi organisaatioiden tutkimusdatanhallinnan kustannuksia kartoittavasta mallista. Vaikka organisaatiot ovat investoineet laajasti datanhallinnan ja datan avaamisen tukipalveluihin ja infrastruktuureihin, kustannuksista on saatavilla hyvin vähän vertailevaa tietoa. National Science Foundationin (NSF) rahoittamassa tutkimuksessa saatiin yhteensä 69 vastausta kuuden tutkimusorganisaation palveluyksiköstä. Tulosten perusteella keskimääräiset datanhallinnan ja avaamisen vuosikustannukset (henkilöstö ja infra) olivat kaikki yksiköt mukaan lukien yhteensä noin 750 000 dollaria. Tarkemmassa tarkastelussa suurimmat vuosittaiset kulut kaatuivat kirjastoille, toiseksi suurimmat puolestaan tietotekniikkayksikölle. Tutkimuksessa kartoitettiin myös tutkijoiden datanhallinnan kuluja. Vastauksia saatiin yli 200 vastuulliselta tutkijalta (PI). Ilmeni, että rahoituskauden keskimääräiset datanhallinnan palveluihin ja infraan käytetyt kulut olivat noin 30 000 dollaria tai kuusi prosenttia kokonaisrahoituksesta.

Huomionarvoista tuloksissa oli, että pienemmän rahoituksen saaneissa projekteissa datanhallinnan kulut olivat suhteessa korkeammat, eli noin 15 prosenttia kokonaisrahoituksesta. Tutkimuksessa kävi myös



Niina Nurmi esitteli uunituoretta Helsingin yliopiston tutkimusaineistojen säilyttämisen ohjetta. Kuva: Soile Manninen CC BY 4.0

ilmi, että kun tutkijat tekivät datanhallintaa itse hyödyntämättä organisaation tarjoamaa tukea, nousivat datanhallinnan kustannukset korkeammiksi kuin tutkijoilla, jotka hyödynsivät tukipalveluita. Tärkeänä viestinä tutkimuksesta nousi tiedotuksen rooli organisaation omista datanhallinnan palveluista.

Postereita ja valaisevia pikaesityksiä

Ensimmäisenä konferenssipäivänä suoritettiin myös postereiden esittelyä Minute Madness -sessiossa. Neljänkymmen posterin esittelykierroksen käynnisti Soile Helsingin yliopiston DMP-prosessia esittelevällä posterilla, ja hetken päästä oli vuorossa Jari Friman Tampereen yliopistosta, aiheena koneluettavan DMP-pohjan kehittäminen ja hyödyntäminen. CSC:n voitokkaan posteriesittelyn tutkimuksen dokumentoinnin haasteista piti Pinja Immonen.

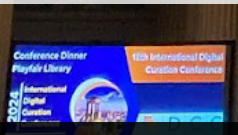
Posterit olivat nähtävillä molempina konferenssipäivinä, ja niitä esitellessä sai jutella äänensä käheäksi. Parhaita olivat tietysti hankalat kysymykset posterin aiheesta, mikä antoi lisäpontta kehitystyölle. Posterin äärellä tuli puhuttua myös DMP-työkaluista ja Suomen DMP-konsortiomallista. Edellisen päivän työpajan esityksissä näytettiin DMPTuulin uusia toimintoja, ja kehuttiin Suomea erinomaisena uusien toiminnallisuuksien testiympäristönä. Pienessä maassa toimijat tuntevat toisensa, joten yhteistyötä on helppo tehdä.

Päivän lopuksi pidettiin rinnakkaissessioina lyhyitä pikaesityksiä (lightning talks), joissa kymmenen minuutin esitystä seurasi viiden minuutin kysymys-vastaus -osio. Esityksissä tuotiin esille käytännön kokeiluja ja kokemuksia datanhallintaan liittyen. Uutena tietona tuli se, että Espanjassa on lähes Suomen DMP-konsortiota vastaava järjestely. Mari Elisa Kuusniemen esityksen aiheena oli Helsingin yliopiston kirjaston datan kuratoinnin lukupiiri, ja sen hyödyntäminen perehdytyksessä sekä osaamisen ja luottamuksen rakentamisessa. Esitys herätti paljon mielenkiintoa ja kysymyksiä, ja keskustelua jatkettiin vielä konferenssi-illallisella. William Playfairin suunnittelemassa kirjastossa (Playfair Library) säkkipillin ja kelttiharpun soidessa nautimme tuhdin skotlantilaisillallisen, johon sisältyi paikallisia herkkuja, kuten haggista.

Töitä riittää

Konferenssin päätöspäivä käynnistyi jälleen pidemmillä esityksillä, joiden aiheena olivat muun muassa sensitiivisen datan palveluiden kehitys ja terveysdataa välittävien palveluiden prosessien läpinäkyvyys. Iltapäivällä jatkettiin lightning talk -esityksillä, jolloin Niina kertoi Helsingin yliopiston alkuvuodesta 2024 julkaisemasta datan säilyttämisen ohjeesta "How to become a data preserver".

Konferenssin päätöspuheenvuoro ei ollut varsinaisesti mikään tunnelman nostattaja, mutta hyvin tarpeellinen.



Uusklassismia edustava Playfair-kirjasto toimi nimensä mukaisesti kirjastona vielä 1960-luvulle saakka.
Kuva: Soile Manninen CC BY 4.0

Materiaalitutkimukseen keskittyvän Royce-instituutin datakuratoija Stavrina Dimosthenous kertoi, miten datanhallinta on materiaalitutkimuksessa vielä alkutekijöissään ja jäljessä muita lähitieteitä. Datan jakamisessa on ongelmia erityisesti silloin, kun rahoitus on tullut teollisuudelta ja tutkimuksessa tuotettu data palautuu yrityksille eikä jää tutkijalle. Datan avointa saatavuutta ei voi tuoda julkaisuissakaan esille, joten FAIR-periaatteet jäävät toteuttamatta. Tarve datanhallinnalle on kuitenkin valtava, koska materiaalitutkimusta on tehty käytännössä jo pronssikaudelta ja dataa tuotetaan paljon.

Konferenssin loppuyhteenvedon on perinteisesti pitänyt Cliff Lynch (Coalition for Networked Information). Tänäkin vuonna hän kertoi omat havaintonsa konferenssin kuumista aiheista ja nosti esille teemoja, joista pitäisi keskustella enemmän. Useimmissa esityksissä käsiteltiin resilienssiä eri näkökulmista, hallinnon roolia dataympäristössä sekä arkistojen ja erikoiskokoelmien roolia luottamuksen rakentajina yhteistyössä. Yksi kiinnostava aihe oli synteettinen data ja sen osuus datahuijauksissa. Kyberturvallisuudesta sekä ohjelmistojen kuratoinnista ja testaamisesta Lynch olisi toivonut kuulevansa useammankin esityksen.

Mitä jäi mieleen?

Esityksistä ja keskusteluista jäi mielikuva haasteiden samankaltaisuudesta, olipa puhuja sitten sitten Floridasta tai Suomesta. Siksi myös ratkaisuja on hyvä miettiä yhdessä. Tieteelliset julkaisut ja niihin kuuluva data liikkuvat eri palveluissa, eikä pysyviä tunnisteitakaan aina löydy, mikä helpottaisi tuotosten löytymistä ja yhdistämistä. Arkistopainotteisista puheenvuoroista kuului kirjastolaisillekin tuttu ylipalvelemisen eetos, mikä voi johtaa siihen, että otetaan lisää tehtäviä ja vastuita, vaikka resurssit eivät riitä.

Hienointa oli tietysti tavata kollegoita eri puolilta maailmaa. IDCC-konferenssia on viime vuodet järjestetty pelkästään virtuaalisena, jo-

Posteriesittelyssä riitti nähtävää ja keskusteltavaa 38 posterin verran. Joillakin oli jopa kaksi posteria esiteltävänä. Kuva: Soile Manninen CC BY 4.0

ten ilo kollegoiden tapaamisesta pitkä tauon jälkeen oli valtava. Suurimmalla osalla osallistujista oli posterin esittely tai puheenvuoron pitäminen luvassa, joten uusien tuttavuuksien kanssa juttu käynnistyi luonnollisesti yhteisen jännityksen jakamisella.

Yksi konferenssin mieleenpainuvimmista hetkistä koettiin Ingrid Dillon avajaispuheenvuoron päätteeksi. Dillo päätti puheensa muistelemalla vuoden 2023 joulukuussa yllättäen menehtynyttä, monille osallistujille läheistä kollegaa ja ystävää Sarah Jonesia. Dillo muisteli yhteistyön iloa, ja siteerasi Sarahin kuvailua infrastruktuurien infrastruktuurista, EOSCista: Visio yhteentoimivan ekosysteemin rakentamisesta on kaunis, mutta sen toteuttaminen on uskomattoman monimutkaista. Konsensuksen saavuttaminen ja yritys palvella merkittävää osaa enemmistön tarpeista on mutkikasta, aikaa vievää, sekä vaatii suurta sitoutumista, hyvää tahtoa, kollegiaalisuutta ja luottamusta.

Dillon mielestä toiveet kollegiaalisuudesta ja luottamuksesta sopivat hyvin konferenssin teemaan. Sarah ajoi vahvasti avoimuuden asiaa, uskoi yhteisöllisyyden voimaan, ja näitä periaatteita hän myös käytännössä toteutti. Sarahia muistettiin monissa puheenvuoroissa ja yhteisissä juttutuokioissa konferenssin aikana. Hänen merkityksensä avoimen tieteen kansainväliselle yhteistyölle on mittaamaton. Kesällä 2024 julkistettiinkin Sarahin muistoksi perustettu palkinto ja rahasto, jolla tuetaan avoimen tieteen yhteistyön lisäämistä.

Tilastoja ja yhteenvetoa

IDCC24-konferenssista: https://dcc.ac.uk/events/idcc24/summary

IDCC 2024 konferenssimateriaalit Digital Curation Centren Zenodo-yhteisön sivuilla.

Seuraava, 19. IDCC-konferenssi järjestetään 17.–19. 2. 2025 Haagissa, Alankomaissa. https://dcc.ac.uk/events/idcc25. Teemana ”Twenty years back, twenty years forward: lessons and directions in digital curation.”

Kirjoittajat

Soile Manninen
Helsingin yliopiston kirjasto
soile.manninen@helsinki.fi
https://orcid.org/0000-0003-1009-1180

Niina Nurmi
Helsingin yliopiston kirjasto
niina.nurmi@helsinki.fi
https://orcid.org/0000-0003-2036-3346