Kommentteja Rooman Instituutin Aineistonhallinnan konkretiaa humanisteille -webinaarin chat-keskusteluun / 24.11.2020

Viittaan 'Mitä on data'-kalvoon ja sen (mukaviin) kuviin, eli kysymys: Datan määrä tuskin pysyy vakiona tutkimuksen kuluessa --> miten tämä suhteutuu tietoon ja sen muuttuvuteen?

Aineistojen ja datan kasvu on syytä ottaa huomioon jo suunnitteluvaiheessa niiltä osin kuin mahdollista. Yksi keskeinen kysymys on selvittää itselleen, onko datani staattista vai muuttuvaa. Jos se on muuttuvaa, kuinka nopeasti se kasvaa (RDM:n näkökulmasta erityisesti kuinka paljon (lisää) levytilaa tarvitsen X ajanjakson aikana). Tietoon ja sen muuttuvuuteen tämä liittyy siten, että kannattaa kiinnittää huomiota datan kopioituvuuteen ja siihen, että raakadatan päälle/oheen alkaa syntyä uusia tasoja – eli käytännössä uusia tiedostoja. Raakadata (esim. kerätty data, haastatteluäänite tms.) kannattaa ”jäädyttää” ja jos sille on syytä tehdä jotain prosessointia, editointia, muokkausta tms., tämä tehdään sen kopiolle. Näin alkuperäinen säilyy koskemattomana ja dataa muokataan aina edelleen. Jäädyttää kannattaa aina sellainen versio, johon on syytä pystyä palaamaan syystä tai toisesta. Mikäli jäädytetyt versiot sisältävät sensitiivistä informaatiota, ne pitää säilyttää suojattuina.

Mitä jos on itse kerättyä aineistoa, jota käyttää myöhemmin uuteen tutkimukseen? Käytännössä haastattelumateriaali, jonka on itse kerännyt

Tähän on toki laajemmat käyttö- ja omistusoikeudet kuin muiden keräämään ja näitä oikeuksia pääsee myös paremmin määrittelemään jo keruuvaiheessa, mutta tämäkin pitää siis suunnitella ennen haastattelujen tekoa. Arkistoinnin (siivouksen, lisensoinnin) jälkeen tähän voi suhtautua kuin mihin tahansa muuhun uudelleen käytettyyn aineistoon, vaikka kyseessä olisi itse kerätty haastatteluaineisto.

Ovatko myös haastateltavien kanssa käydyt yksityisviestit, joissa esimerkiksi sovitaan haastattelun ajankohdasta, dataa? Tai pitääkö ne ottaa huomioon tässä?

Voivat olla, jos näin suunnitellaan ja viestittelijöiden kesken näin sovitaan – mielellään etukäteen. Menevät ehkä pikemminkin projektin manageriaalisen datan kategoriaan ja kannattaa miettiä, mikä näiden arvo lopulta on, eli onko näissä jotain sellaista informaatiota, johon olisi syytä pystyä palaamaan jatkossa. Itse etsin juuri sähköpostiviestejä parin kymmenen vuoden takaa, jotta pystyin kuvaamaan projektini aikajännettä ja siitä, milloin ja kenen kanssa olen ollut missäkin vaiheessa tekemisissä. Tämä kuuluu myös prosessin dokumentoimiseen, eli voi näistä olla apua.

Luokitellaanko tutkimuksessa lomakekyselyllä kerätty aineisto A vai C luokkaan?

Luokittelisin itse tuon luokkaan A, mutta sitten siitä syntyvän annotoidun, prosessoidun tai analysoidun kokonaisuuden luokkaan C. Mutta tämä on siis luovaa työskentelyä. Kategorioita ei ole tarkoitus ottaa annettuja ja jäykästi ohjaavina. Niiden on tarkoitus auttaa teitä ajattelemaan mm. minkälaisia käyttö- ja omistusoikeuksia tutkimuksen taustalla oleviin aineistokokonaisuuksiin kuuluu. Lomakekyselyn raakadata on tiukemmin kiinni tutkimukseen osallistujien informoinnissa kuin sitten se kokonaisuus, joka tuosta raakadatasta on prosessoitu – esim. annotoitu, koodattu tai anonymisoitu kokonaisuus, jossa ei ole enää suoria tunnisteita vastaajiin jne.

Tuleeko mainitsemasi suunnitteilla oleva aineistonhallinnan kurssi (5 op) myös avoimeen yliopistoon? Kun kurssi saadaan avoimeen, siitä ovat taatusti kiinnostuneita muutkin kuin tohtoriopiskelijat. Aihe on monelle muullekin paitsi tärkeä myös kiinnostava!

Kiitos ehdotuksesta! Aivan ehdottomasti kartoitamme mahdollisuuksia tarjota tätä mahdollisimman laajalle.

Suomen Ateenan-instituutti/TATD-hanke: Kyllä. Ja koulutuksen tulisi sisältyä yliopistolliseen perusopetukseen, mistä on paljon puhuttu TSV:n avoimen tieteen koordinaatiossakin.

Mikä siinä [AVOIMESSA TIETEESSÄ] maksaa?

Tyhjentävää listaa on mahdotonta antaa, eli erityisesti kannustan ennakoimaan – nämä voivat olla hyvinkin yllättäviä kuluja! Esimerkkeinä voidaan mainita esim. tietyt repositoriot, eli aineistokokonaisuuksien avaaminen ja arkistointi saattaa jossain tapauksissa maksaa. Kaikilla aloilla ja kaikille aineistotyypeille ei ole olemassa ilmaisia ratkaisuja. Tutkimusaineistojen hankinta tai avoimeksi ostaminen saattaa aiheuttaa kustannuksia. Ja ylipäänsä se, että aikoo seurata avoimen tieteen periaatteita teettää jonkin verran työtä, johon kannattaa varata omaa työaikaansa, eli palkkakustannuksia. Isommissa projekteissa on mahdollista palkata työvoimaa, jolla tätä voi omalta osaltaan helpottaa.

Arkeologisen aineiston julkaiseminen niin että se on uudelleenkäytettävää saattaa vaatia ulkoisen ammattilaisen palkkaamista!

Avoimen julkaisun teemaan liittyen kysymys maksullisuudesta: osa avoimen julkaisemisen journaaleista toimii kaupallisesti (ymmärtääkseni) siten, että tutkijat joutuvat maksajiksi julkaisemisesta (eivätkä vertaisarviointiprosessit ole niin vahvoja -- voiko vertaisarviointi jopa puuttua?). Näihin liittyy toki tutkimuksenkin kannalta iso huoli. Miten tunnistaa tällaiset julkaisijat niistä avoimen julkaisun journaaleista, jotka toimivat tieteen avoimuuden puolesta eivätkä markkinaehtoisesti (jopa ei-vertaisarvioiden)?

Tutkimusartikkelien ja -kirjojen avoin julkaiseminen on hieman oma maailmansa verrattuna datan ja aineiston avaamiseen, mutta toki näissä on myös yhtymäkohtia. Tässä puhutaan niin sanotuista saalistajalehdistä. Näistä kannattaa olla yhteydessä oman yliopiston julkaisun tukeen. Esim. Helsingin yliopistossa Avoimen tieteen -ryhmä auttaa tutkijoita tällaisissa kysymyksissä – niin näissä artikkelimaksuissa (APC) kuin saalistajajulkaisujen tunnistamisessa. Yksin ei kannata jäädä ja missään tapauksessa itse ei saa maksaa kenellekään yhtään mitään ennen kuin on pyytänyt asiantuntijoita tarkistamaan julkaisukanavan. Tässäkin toki kannustan tutkijoita itse harjaantumaan näiden tunnistamisessa. Ovelia ovat, mutta julkaisukanavan kriittinen arviointi on taito, jossa voi kehittyä ja kuuluu myös tutkijan perustaitoihin – siinä missä oman alan vakuuttavien julkaisukanavien tunteminen.

Zenodosta olisi mielenkiintoista kuulla hieman lisää. Eli tämä Zenodo olisi suositeltava paikka ladata kaikki artikkelit ja datasetit, jolloin viimeistään ne saavat tunnisteet?

Voisiko ajatella että esim. instituutti perustaa Zenodoon "yhteisön" jonka kuratoinnista se vastaa. Eli ajattelisi Zenodoa työkaluna ja alustana?

Zenodo on hyvä matalan kynnyksen ilmainen erilaisten tutkimustuotteiden (data, aineistot, julkaisut, esitelmät) julkaisualusta. Jo aiemmin muualla julkaistujen artikkelin osalta ottaisin ensin yhteyttä ko. julkaisijaan ja yrittäisin selvittää, onko heidän mahdollista takautuvasti tarjota julkaisuille pysyviä tunnisteita – jos ei niin sitten Zenodo voisi olla hyvä paikka.

Niin hyvä kuin Zenodo onkin, on syytä muistaa, että se on julkaisualustojen karvalakkimalli. Eli se ei ole kuratoitu. Tämä tarkoittaa erityisesti sitä, että kukaan ei valvo, mitä Zenodossa julkaistaan. Kukaan ei konsultoi, onko julkaistava aineisto/data avaamiseen sopivaa – esim. onko siitä tunnisteellista (erityisesti sensitiivistä) tietoa. Ja tämä on tärkeää: TUTKIJA ON ITSE VASTUUSSA SIITÄ, ETTÄ AINEISTO ON AVAAMISEEN SOPIVAA! Toisekseen Zenodossa kukaan ei katso aineistojen ja datan perään jälkikäteen, eikä turvaa niiden pitkäaikaista käytettävyyttä, eli seuraa tiedostojen toimivuutta, tee migraatioita tms. Tämäkin jää siis alkuperäisen avaajan vastuulle. Pitkäaikaisella käytettävyydellä tarkoitetaan hieman tilanteesta riippuen ajanjaksoja yli viidestä vuodesta ikuisuuteen. Niin ikään kukaan ei seuraa säilyykö Zenodossa julkaistun aineiston anonymisointi – sekin voi nimittäin rikkoutua ajansaatossa. Kuratoitu ja sertifioitu arkisto huolehtii tällaisten asioiden hoidosta pitkään myös julkaisemisen/avaamisen jälkeen.

Instituutin perustama Zenodo-yhteisö kuulostaa hyvältä ratkaisulta, mutta kuratoinnin lupaamista kannattaa kyllä tarkkaan harkita, miettiä ja suunnitella. Se on aikamoinen hanke.

From Laura Nissin / IRF : Esim. brittiläinen https://archaeologydataservice.ac.uk/ sopii arkeologiselle materiaalille. On maksullinen. Mahdollisesti jotain rajoituksia?

> Tosin: ADS "The ADS is the leading accredited digital repository for heritage data generated by UK-based fieldwork and research" eli käsittääkseni ei kv datoille.

Ruotsissa on kanssa kansallinen repositorio, missä työskentelee arkeologeja: https://snd.gu.se/en

Saako tunnisteet liitettyä vanhoihin julkaisuihin? / Entä vanhat artikkelit, joilla ei ole tunnisteita. Miten niille voisi sellaisen saada?

On mahdollista, jos julkaistaan uudestaan alustalla, joka tarjoaa pysyvän tunnisteen. Mutta toki mieluummin niin, että pyrkii jo lähtökohtaisesti valitsemaan julkaisijan tai alustan, joka tarjoaa tunnisteen tai painostaa julkaisijaa muuttamaan toimintaansa tunnisteelliseen suuntaan – esim. journal.fi tarjoaa mahdollisuuden pysyviin tunnisteisiin, mutta vieläkin palvelussa on julkaisuja, jotka eivät ole ottaneet tätä mahdollisuutta käyttöön. Suoraan sanoen ihmettelen suuresti, miksi ei. Ehkäpä näiden tunnisteiden tärkeyttä ei vain yksinkertaisesti ymmärretä.

Julkaisijat voivat hankkia tunnisteet vanhoihin artikkeleihin, jos ovat alkaneet julkaisemisen jälkeen DOI-tunnisteita antamaan. Joten kannattaa ottaa yhteyttä julkaisijaan.

Juuri näin!!

Ymmärsinkö oikein (avoimen julkaisemisen kohtaan ja DOI:n käyttöön liittyen): myöskään Academia.edu-sivustolla ei tulisi julkaista julkaisuaan, vaan ainoastaan DOI, joka ohjaa avoimesti julkaistuun alkuperäisjulkaisuun? (Mutta entä niiden julkaisujen kohdalla, joita ei ole julkaistu avoimesti?)

Paikallisissa datatuissa tms. tukipalveluissa emme tietenkään voi kieltää tutkijoita käyttämästä somealustoja myös julkaisualustoina, mutta millään muotoa suositeltavaa se ei ole. Eikä missään tapauksessa linjassa minkään avoimen tai vastuullisen tieteen periaatteen kanssa. Tutkijan oma etu on, että hän ohjaa kaiken latausliikenteen yhteen tai kahteen vastuulliseen arkistoon – esim. oman yliopiston repositorioon tai esim. omaan Zenodo-yhteisöönsä, jos tällaisen on perustanut. Myös latausluvut ja arvioin siitä, kuinka paljon omat aineistot ovat lähteneet liikkeelle on helpompi ottaa talteen.

Nyt esittää vastakysymys, että millä perusteilla julkaisun, jota aiemmin ei ole avoimesti julkaistu, saa laittaa sosiaaliseen mediaan? Suurella todennäköisyydellä näin toimiessaan tutkija sortuu sopimusrikkomukseen.

Tulevatko diat jälkikäteen saataville? Paljon hyödyllistä infoa

From Laura Nissin / IRF : Tulevat kyllä. Laitetaan Suomen Rooman-instituutin sivuille

Onko aineistonhallinnassa kyse ainoastaan digitaalisista aineistoista? Miten ei-digitaalisten aineistojen määrä arvioidaan?

Aineistonhallinnassa on toki kyse kaikesta aineistosta, johon tutkimusprojekti pohjautuu. Aineistonhallintasuunnitelmassa on syytä rajautua keskittymään siihen osaan aineistosta, josta on itse vastuussa. Jos olet vastuussa fyysisestä tai analogisesta aineistosta, pitää tämän käsittelyä toki kuvata, vaikkei se digitaalista olekaan. Yleisen tason vastausta siihen, miten ei-digitaalisten aineistojen määrää arvioidaan on tosi vaikea antaa – tämä riippuu täysin, mistä aineistosta puhutaan. Arkijärki ja luova kuvailu varmaankin auttaa tässä: esim. 350 kelanauhaa, noin 7 hyllymetriä tai 700 keraamista objektia, paino 1-2 kg / objekti ja tilantarve… tms. Te olette oman aineistonne asiantuntijoita – myös sen kuvailussa!!

Miksei tiedonkeruumenetelmää voi muuttaa kesken prosessin? Esim. jos kentällä huomaa, että joku muu menetelmä kannattaa lisätä repertuaariin?

Toki voi muuttaa, jos esim. virheen huomaa, mutta RDM:n datan laatukysymykset ohjaavat tutkijaa ennakoimaan tähän liittyviä ongelmia. Kysymys ohjaa perustelemaan, miten tällaisessa tilanteessa pidät aineistosi eheänä ja analysointikelpoisena. Jos kyselypatteriston/tiedonkeruumenetelmän muuttaminen kesken aineistonkeruun EI aiheuta laatuun ja konsistenssiin liittyviä ongelmia, niin hyvä niin, mutta tämä pitää osata perustella.

Entäpä jos henkilöt, joita tutkii elivät 1500-1600-luvuilla?

From Laura Nissin / IRF : "Tietosuoja-asetusta (EU 679/2016) ei sovelleta kuolleiden henkilöiden henkilötietojen käsittelyyn". Eettisiä kysymyksiä on silti, etenkin jos on elossaolevia sukulaisia (vaikka 1900-luvun Lähi-idän tutkimus)

https://www.jyu.fi/fi/yliopisto/tietosuoja/ukk/usein-kysyttya-tietosuojasta

Mihin kuvailutietoja kannattaa arkistoida?

Tutkimuksen aikana omaan systemaattisesti dokumentoituun systeemiinsä – suosittelen systemaattista kansiojärjestelmää tietokoneella, jossa nämä tarkemmat tiedostot taulukkotiedostossa (esim. csv tai muuta taulukko-ohjelma; taulukko siitä syystä, että siihen kuvailu on helpompi tuottaa rakenteisena).

Jos saan Zenodosta pysyvän tunnisteen, tuleeko sama tunniste käyttöön myös jos myöhemmin annan aineiston jollekin arkistolle?

Jokainen julkaisu saa oman pysyvän tunnisteensa, eli tuo arkisto antaa ko. setille uuden tunnisteen, vaikka kokonaisuus olisi sama kuin alkuperäinen Zenodo-julkaisusi. Tämä ei kuitenkaan ole ongelma – esim. Zenodo-julkaisutietojen yhteyteen voi käydä lisäämässä tiedon rinnakkaisesta tunnisteesta, joka voidaan määritellä esim. tismalleen sama kuin tuo toinen, sille jollain tavalla hierakinen tms.

Kiitos [resurssi]taulukosta! Omassa hankkeessa metatietojen, aineistonhallinnan ja arkistoinnin parissa kuluu yllättävän paljon aikaa, ja sen näkyväksi tekeminen tuntuu haastavalta kun iso osa on päivittäistä "siinä sivussa" -tekemistä..

Tämä on juuri aineistonhallintasuunnitelman ja RDM-taitojen kehittämisen tarkoitus, että nämä ymmärrettäisiin projektiin kuuluvina hoidettavina työtehtävinä, ei yllättävinä resurssisyöppöinä. Niin ikään sitten kun nämä ymmärretään osana geneerisiä tutkijataitoja tilanne toivottavasti helpottuu ja näitä opitaan myös resursoimaan projekteihin.