Laatua ja vaikuttavuutta tutkijan työhön datanhallinnan avulla

Miten teet datastasi FAIR

Avoin tiede

Sisällys

Johdanto

Datanhallinta osana tutkimusta

Datan arvon ja laadun arvioiminen

Datan jakaminen ja tallentaminen

Mitä tutkimusdata on?

Tutkimusdatalla tarkoitamme tässä tietoaineistoa, joka syntyy ja/tai jota hyödynnetään tutkimuksessa.

Datan elinkaarella tarkoitetaan kaikkia digitaalisen tiedon olemassaolon vaiheita luomisesta pitkäaikaissäilyttämiseen tai tuhoamiseen.

Digitaalisten aineistojen hallinta vaatii suunnittelua ja osaamista. Hallinnoimaton aineisto voi korruptoitua itsestään tai se voidaan sotkea tai hävittää.

Vastuullinen tutkimus edellyttää läpinäkyvyyttä ja mahdollisuutta toistaa tai tarkistaa tehty työ. Data ja ohjelmistot ovat tärkeä osa tutkimusta.

Mitä FAIR-periaatteet tarkoittavat?

Löydettävyys (*Findable*), saatavuus (*Accessible*), yhteentoimivuus (*Interoperable*) ja uudelleenkäytettävyys (*Reusable*) ovat tavoitteita, joiden huomioiminen tukee hyvää datanhallintaa, sekä tutkimuksen laatua ja vaikuttavuutta

- Lisää Fair-periaatteista: https://www.go-fair.org/fair-principles/ ja https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/

FAIR-periaatteet on tarkoitettu kaikelle datalle, niin määrälliselle kuin laadulliselle, ja niiden tavoitteena on datan koneluettavuus ja yhteentoimivuus.

- Samat periaatteet pätevät myös metatietoihin, joilla data kuvataan ja tehdään löydettäväksi.

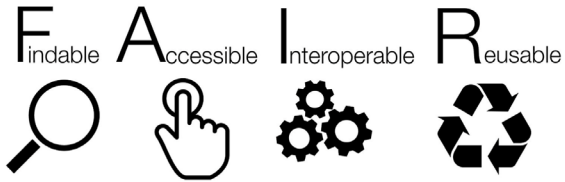
FAIR-data on ihmiselle ymmärrettävää ja sitä pystyy käsittelemään myös ohjelmallisesti. Esimerkiksi ihmiselle ymmärrettävä taulukko pdf-tiedostossa ei ole helposti muokattavissa, kun taas csv-tiedosto on. Hyvin rakenteistettua tietoa voi esimerkiksi yhdistää toisiin samanmuotoisiin tietoihin tai siihen voi kohdistaa hakuja.

FAIR-periaatteet ja tutkija

Tutkija voi lähestyä aihetta pohtimalla miten hyvin toinen tutkija pystyisi toistamaan tehdyn työn: löytyvätkö datat ja menetelmät, ovatko ne toisille ymmärrettäviä ja käytettäviä?

Tutkimusrahoittajat ja tieteelliset kustantajat voivat edellyttää FAIR-periaatteiden noudattamista ja datan hallinnointia.

Tämän dokumentin ohjeiden tarkoituksena on tukea tutkijaa oman työnsä suunnittelussa siten, että FAIR-periaatteiden noudattaminen on mahdollisimman vaivatonta. Jokaisen kappaleen lopussa on tutkijan tarkistuslista kappaleen aiheista.



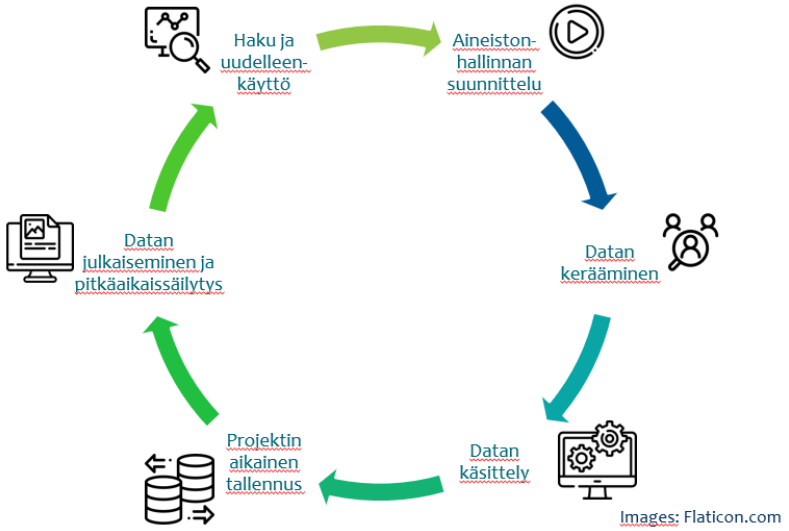
Datanhallinta osana tutkimusta

Koko tutkimuksesta FAIR

FAIR-periaatteet eivät saa olla tutkimusprosessista erillisiä. Niitä pitäisi noudattaa koko tutkimuksen ajan, ja niitä voi soveltaen käyttää kaikkiin tutkimuksen tuotoksiin.

Toistettavuus voi olla kriteeri, jonka mukaan päätetään mitä tallennetaan. Tällöin on ehdottomasti muistettava myös työnkulun dokumentointi ja tallennus.

“FAIR by design” -ajattelussa pyritään suunnittelemaan koko tutkimusprosessi siten, että esimerkiksi metatietoja karttuu automaattisesti ja tarvittavat tunnisteet on helppo luoda.



Datan hallinta alkaa jo tutkimuksen suunnitteluvaiheessa (DMP - Data Management Plan)

FAIR-periaatteiden toteuttamista tehdään koko tutkimuksen ajan. Suunnittele siis

- datan koko elinkaari ja huomioi se tutkimuksen suunnittelussa (esim. säilytettävät, pitkäaikaissäilytettävät ja tuhottavat versiot).
- dataan liittyvät tekniset tiedot sekä oikeuksiin liittyvät asiat.
- dokumentaation eri muodot (esim. lähdekoodi, koodisto, ohjeistukset) jotka yhdessä parantavat läpinäkyvyyttä.

Vastuulliseen tutkimukseen kuuluu prosessin läpinäkyvyys, eli sen avaaminen mitä datalle on tehty ja miten lopputulokseen päästy. Tämä on otettava huomioon aineistonhallintasuunnitelmassa, joka pitää laatia jo hankkeen suunnitteluvaiheessa ja jota on pidettävä yllä tutkimuksen edetessä.

Datan elinkaaresta (1): Datan kypsyystaso

Raakadatalla tarkoitetaan dataa alkuperäisessä muodossaan, niin kuin se on saatu suoraan lähteestä ennen kuin sitä on mitenkään käsitelty.

Primääridatalla tarkoitetaan tiettyä tarkoitusta varten kerättyä tai luotua ja ylläpidettyä dataa, esim. tekijän vastuulla olevaa “master-kopiota” datasta, jota tutkija käyttää tutkimuksessaan.

Julkaistu tutkimusdata on viitattavissa oleva datakokonaisuus, joka voi kasvaa hallinnoidun laatu-, dokumentaatio- ja versiointiprosessin kautta. Minimissään sen metadata on julkisesti saatavissa ja saatavuuden rajoitukset on määritelty koneluettavasti. Julkastu tutkimusdata voi koostua useammasta primärilähteestä tuodusta datasta.

Datan elinkaaresta (2): Luonti ja valmistelu

Datan syntyyn voi liittyä erilaisia instrumentteja tai keräysmenetelmä, jotka on tärkeä dokumentoida.

Raakadatasta kohti analysoitavaa dataa

- Tutkimusta varten kerättävä data on ensimmäisessä vaiheessa ns. raakadataa. Raakadata on datan prosessoinnin alkupiste, ja siihen olisi hyvä aina olla mahdollista palata.
- Analyysejä varten raakadataa laatutarkastetaan (eli **validoidaan**), **kuvaillaan**, ja joskus sitä rikastetaan ja luokitellaan, ennen kuin sitä voi analysoida.

Usein datan valmistelussa käytetään erilaisia komentoja tai muuta koodia

- **Lähdekoodi** (*program code*), on väline, jolla dataa tuotetaan. Esimerkiksi käytetyt mallit, analyysi- ja datan käsittelykoodi (esim. R-skriptit) on syytä säilyttää, dokumentoida ja julkaista mikäli perusteita julkaisemattomuudelle ei ole.

Datan elinkaaresta (3): Jalostaminen ja hallinta

Koodisto tarkoittaa muuttujien, entiteettien tai kategorioiden ilmaisemista datassa. Tällaiset aineistokohtaiset koodistot täytyy dokumentoida ja avata yhdessä datan kanssa. Pyri käyttämään olemassa olevia standardeja ja/tai koodistoja ymmärrettävyyden ja toistettavuuden parantamiseksi.

Työnkulku on sarja toimintoja, joiden avulla voidaan standardoida työn tekemisen vaiheet ja varmistaa johdonmukaiset tulokset. Työnkulut jäsentävät prosesseja ja parantavat tehokkuutta. Työnkulku mahdollistaa datan kehityshistorian (*lineage*) ja omistushistorian (*provenance*) tallettamisen ja julkaisemisen.

Kun tieto-objekti luodaan tai siitä tehdään muutettu versio, se merkitään versionumerolla. **Versionhallinta** pitää kirjaa tallennetuista versioista ja mahdollistaa palautuksen niihin, mikä on tärkeää datan jäljitettävyyden, muokkausten seurannan ja virheiden korjaamisen kannalta.

Ohjeita datan elinkaareen ja sen dokumentointiin

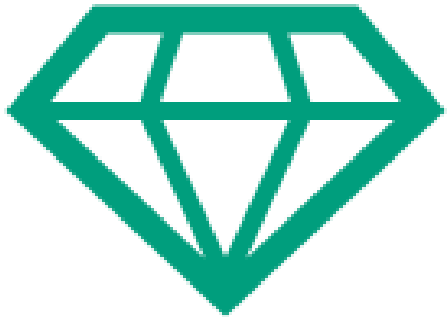
- Tallenna raakadata, mikäli sen säilyttäminen on mahdollista ja tarkoituksenmukaista.
- Dokumentoi datan lähteet. Käytä pysyviä tunnisteita mikäli mahdollista (esim. lähdedataseteille, infrastruktuureille, sensoreille, protokollille).
- Dokumentoi datan käsittely ja sen periaatteet tarkasti, kuten validointiprosessi, onko kirjoitusvirheitä korjattu, sekä konversiot jne.
- Suunnittele versiointi, pitkäaikaissäilyttäminen ja tuhoaminen, myös raakadatan osalta
- Varmista, että voit tarvittaessa palata versioissa taaksepäin huolehtimalla riittävästä määrästä versioita työn aikana. Varmista, että versiot on yksiselitteisesti tunnistettavissa.
- Huolehdi tiedostojen eheydestä laskemalla niille tarkistussummia ja muista varmuuskopiointi. Kirjaa myös nämä toimenpiteet.
- Mieti, mikä on tarpeen tutkimuksen toistettavuutta ajatellen.
- Varsinkin pitkäaikaissäilytettävän datan osalta tarkka dokumentaatio on tärkeää esimerkiksi käytettyjen laitteiden ja ohjelmistojen osalta

TARKISTUSLISTA: Onko tutkimuksesi toistettavaa?

1. Ohjaako aineistonhallintasuunnitelma (DMP) työskentelyäsi koko datan elinkaaren ajan siten, että datan koko käsittelyprosessi on läpinäkyvä ja riittävästi dokumentoitu?
2. Miten olet huomioinut datan avoimuuden ja käyttörajoitukset koko prosessin ajan?
3. Hyödynnätkö metatiedoissa ja itse datassa yhteisiä käytäntöjä, esim. standardeja ja sanastoja?
4. Oletko dokumentoinut tutkimusaineiston elinkaaren järjestelmällisesti ja vastaako kuvaus todellisuutta? Oletko automatisoinut mahdollisimman monen vaiheen datan prosessoinnissa ja onko koodi tallessa? Löytyykö käytettyjen ohjelmistojen ja asetuksien dokumentaatio (tekninen dokumentaatio)?
5. Oletko versioinut datan ja muut tuotokset?
6. Onko data ja sen dokumentaatio tallessa viitattavassa muodossa (pysyvät tunnisteet ja metatiedot)?

Toistettavuudesta huolehtimalla FAIR periaatteet toteutuvat luontevasti osana tutkimusprosessia, eikä tietoja tai dokumentaatiota tarvitse luoda erikseen artikkelin tai datan julkaisuvaiheessa

Datan arvon ja laadun arvioiminen



FAIR-periaatteet ja datan arvo

Datan uudelleenkäytettävyys lisää sen arvoa tiedeyhteisölle ja yhteiskunnalle. Tutkijalla on vastuu omalta osaltaan määritellä datan arvoa ja huolehtia arvon säilymisestä.

FAIR-periaatteiden mukainen aineistonhallinta mahdollistaa datan vastuullisen uudelleenkäytön ja lisää tutkimuksen vaikuttavuutta.

Mitä tarkemmin datan alkuperä on dokumentoitu, sitä helpommin sitä voi käyttää uudestaan. Siksi työnkulku, jolla raakadatasta saadaan julkaistava tutkimusdata, on tallennettava.

Tutkija saattaa kerätä datan määrittelemänsä oman tieteenalansa kysymyksen ratkaisemiseksi, mutta datalla voi olla arvoa myös osana laajempaa kokonaisuutta, esim. organisaation tietokannassa, vertailudatana, meta-analyysin osana tai jollain muulla tieteenalalla.

Datan ainutkertaisuus ja sen tuottamiseen käytettyjen resurssien määrä lisäävät yleensä datan arvoa. Datalla on usein myös arvo tutkimuksen läpinäkyvyyden ja (vertais)arvioinnin näkökulmasta.

Datan arvon muuttuminen

Datan arvo voi vähentyä tai kasvaa ajan myötä: ajankohtaisen ongelman ratkaisemiseen kerätty data voi jäädä kertakäyttöiseksi, mutta data voi saada uutta merkitystä esim. osana aikasarjaa tai muutoksen vertailukohtana tai osana kulttuuriperintöä. Ajan myötä tutkimusaineisto voi saada kulttuuriperintöarvoa

- Lue lisää datan kulttuuriperintöarvosta

On vaikeaa ennustaa mikä data on tulevaisuudessa arvokasta, siksi datan huolellinen valikointi ja säilytettävän datan ymmärrettävyyden varmistaminen on olennainen osa tutkimusta.

- Esimerkiksi Turun Aurajoen jäiden lähtemisen seuranta-aineisto, joka on kerätty niin arkistodatasta kuin vanhoista sanomalehtitiedoista (koko tutkimusartikkeli englanniksi). Sen avulla voidaan nyt 100-200 vuotta myöhemmin analysoida ilmastonmuutoksen vaikutuksia.
- Säilyttäminen voi lisätä myös raakadatan arvoa. Esimerkiksi näennäisesti poikkeavat havainnot voivat olla ensimmäisiä merkkejä alkavasta muutoksesta. Siksi raakadataa voi olla hyvä tallentaa, vaikka siinä vaikuttaisi olevan selittämättömiä virheitä keruuhetkellä.

Julkaistuja ohjeita datan arvon ja laadun arvioimiseen

Erilaisia tapoja ja kriteeristöjä datan arvon määrittämiseksi on monia. Laadun arviointi voi perustua esimerkiksi datan eheyteen, kattavuuteen, luotettavuuteen ja yhdenmukaisuuteen. Kun kaikkea dataa ei voi säilyttää, arvon ja laadun määrittely tukee erityisesti pitkäaikaissäilytettävien aineistojen valintaa.

Seuraavaksi esittelemme pääkohdat kahdesta julkaistusta ohjeesta, joita tutkija voi hyödyntää ja soveltaa arvioidessaan datansa arvoa ja laatua:

- Helsingin yliopiston ohje datan arvon määrittämiseen pitkäaikaissäilytystä ajatellen: Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.
- Tilastokeskuksen tietoaineistojen laadunvarmistuskehikko

Myös Kansallisarkiston ohjeet sisältävät luvun, joka käsittelee tutkimusaineistoja ja arvon määritystä erityisesti arkistonäkökulmasta.

Näkökulmia datan odotetun uudelleen käytön arvioimiseen

- Onko datasi kokonaisuutena riittävän kattavaa, jotta sitä voi tulevaisuudessa käyttää eri tavoin? Kattavuus voi tarkoittaa tiettyä ilmiötä, ajanjaksoa tms.
- Onko datasi teknisesti niin hyvää, että sitä voi käyttää tulevaisuudessa eri tavoin? Vai pitääkö sitä käsitellä tai valmistella tavalla, joka aiheuttaa kohtuuttomia kustannuksia?
- Voiko datasi laajentaa olemassa olevaa aineistoa?
- Voiko datasi toimia jonkin muun aineiston vertailukohtana?
- Onko datasi analysoitu vain osittain?
- Onko kohtuullista olettaa, että dataasi voi tulevaisuudessa hyödyntää vielä uusin tavoin?

Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

Näkökulmia datan tulevan arvon arvioimiseen

- Onko datallasi tulevaisuudessa merkitystä jonkin tieteenalan tai tutkimusaiheen kehitykselle?
- Voiko datasi johtaa huomattaviin tieteellisiin löydöksiin tai julkaisuihin?
- Onko datasi tieteellisesti tai kulttuurillisesti ainutlaatuista?
- Voiko datansi uusiokäyttö johtaa kaupallisiin sovelluksiin, yritysyhteistyöhön tai patentteihin?
- Onko datallasi huomattavaa arvoa opetuksessa, esimerkiksi tutkijakoulutuksessa?

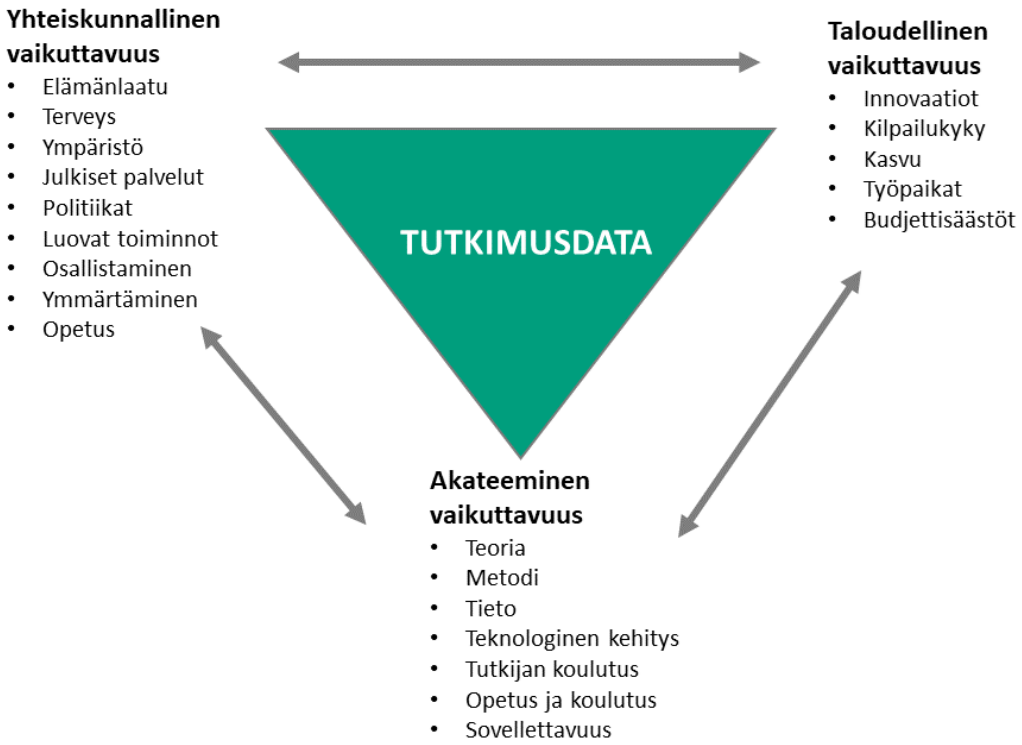
Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo. https://doi.org/10.5281/zenodo.5785971

Näkökulmia todistetun merkityksen arvioimiseen

- Onko dataasi käytetty jonkun erityisen tärkeän julkaisun tai löydöksen tekemiseen?
- Onko datasi ratkaisevan tärkeää kansalliselle tai globaalille infrastruktuurille?
- Onko dataasi hyödynnetty merkittävässä tutkimusyhteistyössä eri organisaatioiden välillä?
- Onko datasi tuottaminen vaatinut merkittäviä investointeja ja resursseja?
- Onko esim. eettinen neuvosto aiemmin arvioinut datasi? Eettisen neuvoston arviointiraportit voivat olla hyödyllisiä päätöksenteossa sillä, joskus datan säilyttäminen voi olla perusteltua eettisistä syistä.
    - Eettisen ennakkoarvioinnin ohje

Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

Vaikuttavuuden ulottuvuudet



Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

Laadukas data kuvaa todellisuutta

Virheettömyys
kuvaa sitä, miten data vastaa todellisuutta. Datan virheettömyyttä tarkastelemalla voidaan saada kiinni myös systemaattisia vääristymiä tietoaineistossa.

Tarkkuus
kuvaa sitä, miten hyvin datan tiedot vastaavat sitä mitä tavoitellaan. Tarkkuus kuvaa sitä, kuinka hyvin data osuu oikeaan.

(Ajantasaisuus)
mikäli merkityksellinen

Johdonmukaisuus
kertoo siitä, että data on yhtenäinen ja ristiriidaton. Johdonmukaisuudella voidaan kuvata myös eri datojen keskinäistä johdonmukaisuutta.

Kattavuus
kuvaa datan tavoitellun ajallisen ja alueellisen kattavuuden sekä tavoitellut kohdeyksiköt ja ominaisuustiedot. Toisaalta kattavuus kertoo miltä osin data sisältää tavoiteltuja tietoja.

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

Laadukas data on kuvattu hyvin

Alkuperäisyys
kertoo siitä, että dataan ja sen tietoihin tehdyt muutokset voidaan jäljittää. Datan alkuperä tunnetaan.

Metatietojen ymmärrettävyys
tarkoittaa sitä, miten kattavasti metadata kuvaa dataa ja auttaa sisällön ymmärtämisessä.

Suositustenmukaisuus
kertoo siitä, että data ja sen ominaisuustiedot noudattavat tunnettuja standardeja, käytäntöjä ja säädöksiä ja ne on kerrottu datan yhteydessä.

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

Laadukasta dataa on mahdollista käyttää uudelleen

Koneluettavuus
kuvaa, onko data rakenteistettu siten, että sitä voidaan käsitellä koneellisesti ja käsittely on mahdollista eri tietojärjestelmissä.

Käyttöoikeudet
kuvaa sitä, miten datan käyttöoikeus on määritelty ja mitä datalla voi tehdä eli mihin käyttötarkoituksiin tietoaineistoa voi hyödyntää.

Täsmällisyys
tarkoittaa sitä, että data on käytettävissä ilmoitettuna ajankohtana ja riittävän tiheästi täydennettynä.

Niin avointa kuin mahdollista, niin rajoitettua kuin välttämätöntä

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

Kuinka arvokasta datasi on?

Jokainen tutkimusdata on omassa kontekstissaan ainutkertainen. Datan ja sen pohjalta tehdyn tutkimuksen kuvaaminen ja datan avaaminen kasvattaa sen arvoa tieteelle. Kaikkea dataa ei kuitenkaan voi säilyttää ikuisesti ja säilytettävästä datasta tulee jollain olla pitkäaikainen vastuu.

Arvonmäärityksessä sinä tutkijana ja tutkimusdatasta vastaava organisaatio voi harkita:

- Datan uudelleen hyödyntämisen mahdollisuudet
  - merkitys tieteelle nyt ja lähitulevaisuudessa tai pitkällä aikajänteellä
  - käyttö omalla tieteenalalla vs. laajempi käyttö täydentävänä/vertailun mahdollistavana datana
  - Datan sisältämän analysoimattoman tiedon määrä
- Datan ja dokumentaation laatu
  - kuinka tukee tutkimuksen toistettavuutta tai laadunvarmistusta
- Datan rahallinen arvo
  - panostetut varat ja työresurssit, kaupallinen arvo
  - tutkimuksen rahoittajan intressi
- Datan merkitys tutkimuksen vaikuttavuuden kautta yhteiskunnalle
  - esimerkiksi sen merkitys ihmiskunnan viheliäisten ongelmien ratkaisussa.
- Datan historiallinen ja kulttuurillinen merkitys tai sen tieteellinen ja kulttuurinen ainutkertaisuus
  - osana pitkiä aikasarjoja sekä tutkijoille että yhteiskunnalle.

Datan jakaminen ja tallentaminen

Datan käsittelyvaiheista

Datan jakamista *(data sharing)* voi tehdä jo tutkimushankkeen aikana. Tällöin datasetti voi olla aktiivinen. Datanhallinta ja kuratointi tapahtuu sovitusti datanhallinnan suunnittelun pohjalta siten, että kaikki osalliset tietävät dataan liittyvät oikeudet ja vastuut, kuten kuka saa editoida dataa, miten versiointia tehdään ja miten dataa yleensä saa ja voi käyttää. *Jakamista voit tehdä esimerkiksi verkkolevyltä, tietoaltaasta tai sähköpostin avulla tai luottamuksellisen datan jakamiseen tarkoitetussa palvelussa.*

Dataa tallennettaessa *(data storing)* data dokumentoidaan huolellisesti ja siihen liitetään käytön ja hallinnan kannalta oleelliset metatiedot. *Datan tallentamiseen voi käyttää sopivaa palvelua (data repository tai data service).*

Datan julkaiseminen *(publishing data)* tekee datasta viitattavan aineiston, jolla pitää olla pysyvä tunniste ja viittausohje. Julkaistu data on dokumentoitu huolellisesti ja ainakin sen löytämisen mahdollistava metatieto on julkaistu. Metatiedoissa kerrotaan, miten aineiston saa käyttöönsä ja minkä laatuista se on, jotta sitä voi käyttää tutkimuksen toistamiseen tai uuteen tutkimukseen. *Datan julkaisemiseen käytetään sopivaa palvelua (data repository tai data service).*

Datan pitkäaikaissäilyttäminen *(digital preservation)* on aineiston ymmärrettävyyden ja eheyden säilyttämistä pitkällä tähtäimellä, yli vuosikymmenten tai jopa vuosisatojen. Silloin otetaan huomioon esimerkiksi eri teknologioiden, medioiden ja formaattien muuttuminen. Myös luottamuksellista aineistoa on mahdollista pitkäaikaissäilyttää sertifioiduissa palveluissa. *Voit pitkäaikassäilyttää dataa tietyissä palveluissa, joista esittelemme muutamia myöhemmin näissä ohjeissa.*

Huom! Kaikissa vaiheissa aineistoon pääsy voi olla rajoitettua. Riittävästä tietosuojasta ja -turvasta on huolehdittava!

Datasetin kokoaminen ajatellen jatkokäyttöä

Datasetti on datasta koottu kokonaisuus. Dataa julkaistaessa on otettava kantaa paitsi sen eri jalostusasteeseen (level) ja dokumentaatioon, myös esimerkiksi

- granulariteettiin ja resoluutioon (rakeisuus) ja tarvittava viittaustarkkuus (tunnisteet)
- tiedostoformaattiin, miten data on jaettu tiedostoiksi ja miten ne on dokumentoitu (rakenne, tunnisteet)
- miten julkaistua aineistoa mahdollisesti versioidaan tai kartutetaan tulevaisuudessa
- aikasarjojen ja eri muuttujien ilmaiseminen metatiedoissa ja niiden tekeminen haettaviksi yli datasettien.

Huomioi tehdessäsi ratkaisuja FAIR-periaatteet - ajattele aineistojen viitattavuutta, löydettävyyttä, yhteentoimivuutta ja jatkokäyttöä.

- Dataa on usein helpompi jakaa osiin kuin yhdistellä uudestaan erillisistä tiedostoista
- Suuria tiedostoja voi toisaalta olla raskas siirtää ja käsitellä
- Yritä miettiä tulevia käyttäjiä ja uusia käyttötapoja
- Muista, että datastasi voi tehdä myös sekundaarijulkaisuja (jotka eivät ole primaarijulkaisun uusia versioita)
- Seuraa alasi hyviä käytäntöjä

Tallennus- ja julkaisupalveluiden valinta

- FAIR-periaatteet edellyttävät, että tutkimuksen tuotoksilla on pysyvä tunniste ja niiden metatiedot ovat löydettävissä.
  - Käytännössä tämä edellyttää tutkimusdatan julkaisemiseen tarkoitettujen yhteisten palveluiden hyödyntämistä
- Määrittele vähimmäisvaatimukset, joita julkaisupalvelulla tulisi olla.
  - Tietoturva, käytettävyys, saavutettavuus, työkalutuki, tietomallit, pysyvät tunnisteet
- Hyödynnä tieteenalakohtaisia palveluita mikäli mahdollista.
  - Ota huomioon kuka dataasi voisi myöhemmin hyödyntää - mistä data löytyy helpoimmin?
    - metadatan koneluettavuus, hakumahdollisuudet
  - Mitkä ovat käytettävissä olevat resurssit?
  - Tarkista eri formaatit, sisällön yhteentoimivuus, sisällön rakenne
  - Muista, että samasta aineistosta voi luoda eri tarkoituksiin erilaisia kopioita
- Määrittele datalle käyttöoikeudet.
  - Sopimukselliset perusteet
  - Rahoittajan ja organisaation linjaukset

Luotettavien palveluiden tunnistaminen

Palveluiden luotettavuuden arvioinnissa auttavat sertifikaatit kuten CoreTrustSeal.

Sopivia palveluita voi hakea esimerkiksi re3data.org -palvelusta. Valitse alakohtainen palvelu jos mahdollista.

Konsultoi oman organisaation ohjeita ja tukipalveluita.

Suomessa luotettavia, tutkijoille sopivia palveluita tarjoaa tutkimusta tekevien organisaatioiden lisäksi kulttuuriperintöorganisaatiot ja CSC.

Oikeuksien siirtäminen tai jakaminen organisaation tai palvelun kanssa ei poista tekijänoikeutta, mutta mahdollistaa aineistojen hallintaa pitkäaikaisesti.

Staattisen datan julkaiseminen

Staattinen data on muuttumaton kokonaisuus, esimerkiksi tutkimuksen tausta-aineisto tai data, johon julkaistu artikkeli perustuu. Se voi olla myös otos muista datalähteistä. Artikkeliin liitetään viite ja tieto datan saatavuudesta (*Data & Code availability statement*). **Datasetin ja sen versioiden pitää olla yksiselitteisesti tunnistettavissa.**

Tutkimuksen evidenssiksi riittää usein hyvin dokumentoitu tiedosto tai kokoelma tiedostoja tai aineistoja, mikäli tutkimuksen yhteydessä on syntynyt uusia aineistoja. Huomioi tällöin myös ohjelmistot ja koodit ja niiden jakaminen.

Valitse avoimesti dokumentoitu, yleinen ja avoin tiedostomuoto, koska se luo edellytykset datan jatkokäytölle.

Hyvä dokumentointi edellyttää useimmiten sekä ihmisen tuottamaa, sisältöä kuvaavaa kuvailumetatietoa että automaattisesti tuotettua metatietoa datan teknisistä piirteistä.

Usein tarvitaan myös muuta dokumentaatiota ja viittauksia eri lähteisiin, kuten standardeihin, sanastoihin, menetelmiin, koodistoihin tms.

Muuttuvan datan julkaiseminen

Muuttuvaa dataa on erilaista. **Karttuvassa datassa** aineisto kumuloituu eikä jo kerättyä dataa muuteta. **Dynaamisessa datassa** myös takautuvat muutokset ovat mahdollisia.

Karttuvassa datassa uuden aineiston voi lisätä joko uusina tiedostoina tai vanhojen tiedostojen päivityksinä. Karttuvaa dataa voi julkaista joissakin palveluissa.

Tietokannat vaativat tieteellisinä aineistoina erityistä dokumentaatiota, suunnittelua ja ylläpitoa. Kuvailuun pätee kuitenkin myös samat suositukset kuin staattisen datan julkaisuun.

Pysyäkseen toiminnallisina tietokannat vaativat ylläpitoa joka voi tulla pitkällä aikavälillä kalliiksi. Siksi niiden elinkaari ja ylläpitojärjestelyt tutkimushankkeiden rahoituksen loputtua pitää suunnitella etukäteen.

Yhteentoimivuus ja pitkäaikaissäilytyksessä sovellettavien standardien hyödyntäminen helpottavat datan siirtämistä tarvittaessa uusiin ympäristöihin.

Tutkimuksen toistettavuuden tähden on syytä kiinnittää huomiota täsmällisen ja yksiselitteisen viittaamisen tarpeeseen. **Pysyvän tunnisteen pitäisi aina ohjata käyttäjä yksiselitteisesti oikean datan äärelle.**

Esimerkkejä viittaamisesta muuttuviin aineistoihin

A. Viittaa tiettyyn datan osajoukkoon mainitsemalla lähde, tarkat rajaukset, sekä pysyvä tunniste
   - Kuvitteellinen esimerkki: Data Request T.Jansen; SAHFOS; Work published 2014 via SAHFOS ; Area Def: 54-65°N, 0-45°W. Temporal Def: 1980-2012 (April-August) Taxonomic Def: All zooplankton; (dataset). https://doi.org/10.7487/2014.15.1.1

A. Viittaa datan kopioon tiettynä ajan hetkenä.
   - Kuvitteellinen esimerkki: König-Langlo, G., & Sieger, R. (2010). BSRN snapshot 2010-01 as ISO image file (3.75 GB) [Data set]. PANGAEA - Data Publisher for Earth & Environmental Science. (dataset). https://doi.org/10.1594/pangaea.833424

A. Viittaa jatkuvasti päivittyvään dataan lisäämällä tarkka viittausajankohta

HUOM: Tämä tapa ei välttämä mahdollista toistettavuutta, mutta on joskus ainoa mahdollinen tapa.
   - Kuvitteellinen esimerkki: Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center. (dataset). https://doi.org/10.xxxx/notfoo.547983 . Accessed 1 May 2011.

A. Viittaa aikaleimattuun tietokantakyselyyn versioidussa tietokannassa.
   - Kuvitteellinen esimerkki: R. Roe. 2017. "The Moo Data Query" created at 2017-07-21 10:25:30 PID https://doi.org/10.xxxx/notmoo.857988 Subset of Moo Database (dataset) PID https://doi.org/10.xxxx/bigmoo.360873

Perustuu DataCiten ohjeisiin

Käyttörajoitettu aineisto

Aineistojen käyttöön ja tallentamiseen voi liittyä eettisiä, lainsäädännöllisiä tai sopimuksellisia rajoituksia.

Käyttöön vaadittavat sopimukselliset, sekä prosesseihin ja tietoturvaan liittyvät toimet riippuvat aineistokohtaisista rajoituksista.

Erityistä huolellisuutta vaativat henkilötietoja sisältävät tai muiden tahojen omistamat aineistot, joihin liittyy sopimuksia ja ehtoja.

Ole yhteydessä organisaatiosi datatukeen, jos olet epävarma käyttöoikeuksiin liittyvistä asioista. Myös organisaatiosi datapolitiikka voi tarjota ohjeet, joiden nojalla voit toimia rauhallisin mielin.

Jotta käyttörajoitettuun aineistoon voi viitata, siitä pitää olla riittävästi kuvaavaa metatietoa, tunniste ja laskeutumissivu. Metatiedoissa kerrotaan, mihin rajoitukset perustuvat ja miten aineiston voi saada käyttöönsä.

- Esim. CLARIN ACA linsenssi, jossa aineisto on rajoitettu tutkimuskäyttöön (sisältää henkilötietoja) esim. loppukäyttäjän lisenssisopimus +NC +PRIV +DEP +OTHER v2.1

Datan pitkäaikaissäilyttäminen Suomessa

Tutkimusdatan pitkäaikaissäilyttäminen *(digital preservation)* tarkoittaa, että aineistojen ymmärrettävyys säilytetään periaatteessa pysyvästi. Laitteisto-, ohjelmisto- ja tutkijasukupolvien vaihtuminen ja tutkimusparadigmojen muuttuminen otetaan huomioon muokkaamalla dataa tarvittaessa.

Arvokkaaksi määritettyä dataa pitää säilyttää mahdollisimman pitkään. Datan arvon määrityksessä on monia ulottuvuuksia, joita käsiteltiin edellisessä luvussa.

- Datan arvo, juridiset ja omistajuuteen liittyvät kysymykset pitää ottaa huomioon jo tutkimusta suunniteltaessa, jotta pitkäaikaissäilytykseen siirtämisen prosessi on mahdollisimman kivuton.

Suomessa digitaalista pitkäaikaissäilytystä tarjoaa organisaatioille esimerkiksi CSC:n PAS-palvelu. Fairdata PAS-palvelu on kehitetty tutkimusdatan pitkäaikaissäilytykseen. Aineistojen siirto Fairdata-PAS palveluun edellyttää aina organisaation ja Opetus- ja kulttuuriministeriön välistä sopimusta, eli polku palvelun hyödyntämiseen alkaa tutkijan osalta aina neuvotteluista oman organisaation kanssa.

TARKISTUSLISTA: Ovatko data-asiasi hallussa?

1. Suunnittele huolella: Mitkä ovat tallennettavaksi sopivat arvokkaat tuotokset, jotka ovat tärkeitä toistettavuuden ja vastuullisuuden kannalta?
2. Ovatko sopimus-, ylläpito- ja käyttöoikeusasiat kunnossa myös tulevaisuutta ajatellen?
3. Miten pitkään eri tuotoksia kannattaa säilyttää ja milloin aineistoa mahdollisesti tuhotaan, esimerkiksi turhat versiot? Rahoittajan ja taustaorganisaatioiden vaatimukset vaikuttavat.
4. Ovatko valitsemasi julkaisu- ja tallennuspalvelut luotettavia, aineistoillesi sopivia, FAIR-periaatteita tukevia ja mielellään tieteenalakohtaisia?
5. Oletko ajatellut tulevia käyttäjiä, kun koostat ja dokumentoit aineistosi?
6. Oletko ottanut huomioon, että data tulisi mahdollisesti pitkäaikaissäilyttää uudelleenkäytön mahdollistamiseksi, sen kulttuuriperintöarvon tai rahoittajien vaatimusten mukaisesti?
7. Oletko kertonut taustaorganisaatiollesi pitkäaikaissäilyttämisen arvoisesta datastasi?
8. Oletko informoinut tutkittavia siitä, miten aineistoa jaetaan, tallennetaan ja julkaistaan?
9. Oletko muistanut viitata dataan julkaisuissa, joissa olet sitä käyttänyt?

Tekijät

Anneli Lehtisalo
Ari Asmi
Hanna Koivula
Heidi Troberg
Jessica Parland-von Essen
Juha Hakala
Katja Laine
Maria Söderholm
Marjut Vuorinen
Mika Virtanen
Nina-Mari Salminen
Pekka Nygren
Saila Huuskonen
Sonja Sipponen
Tanja Lindholm
Tarja Mäkinen
Timo Taskinen
Tomi Rosti
Tuomas Alaterä
Tuula Pääkkönen

Avoin tiede

FAIR-periaatteiden soveltaminen -ryhmä