Datan jäljet -työpaja

Eri sidosryhmien roolit kansallisen dataviittausjärjestelmän kokonaisuudessa

Tuomas J. Alaterä

TIETOARKISTO

Kartoitus data-arkistonäkökulmasta

- Viittausmallit ovat olleet data-arkistoilla käytössä jo pitkään, viittausohjeet ~2010 alkaen
  - Tietoarkisto 2001 aineistoviite, ohjeet 2010

    Tietoarkisto: www.fsd.uta.fi/fi/aineistot/jatkokaytto/viittaaminen.html

    Gesis: www.gesis.org/angebot/daten-analysieren/weitere-sekundaerdaten/datenservice/forschungsdaten-zitieren/

    IASSIST IG: iassistdata.org/sites/default/files/quick_guide_to_data_citation_high-res_printer-ready.pdf
  - Tekijä mainitaan viitteessä nyt ensimmäisenä (meriitti)
    - Dataviite/avattu data lisää viittauksia julkaisuun
      - » Esim. Pivowar at al 2007, 2013, Pronk 2016
  - Viite kopioitavissa tutkijan ansioluettelomalliin
    - Tutkimustyön tieteellinen ja yhteiskunnallinen vaikuttavuus: ansiot tutkimus- ja tietoaineistojen tuottamisessa ja jakamisessa

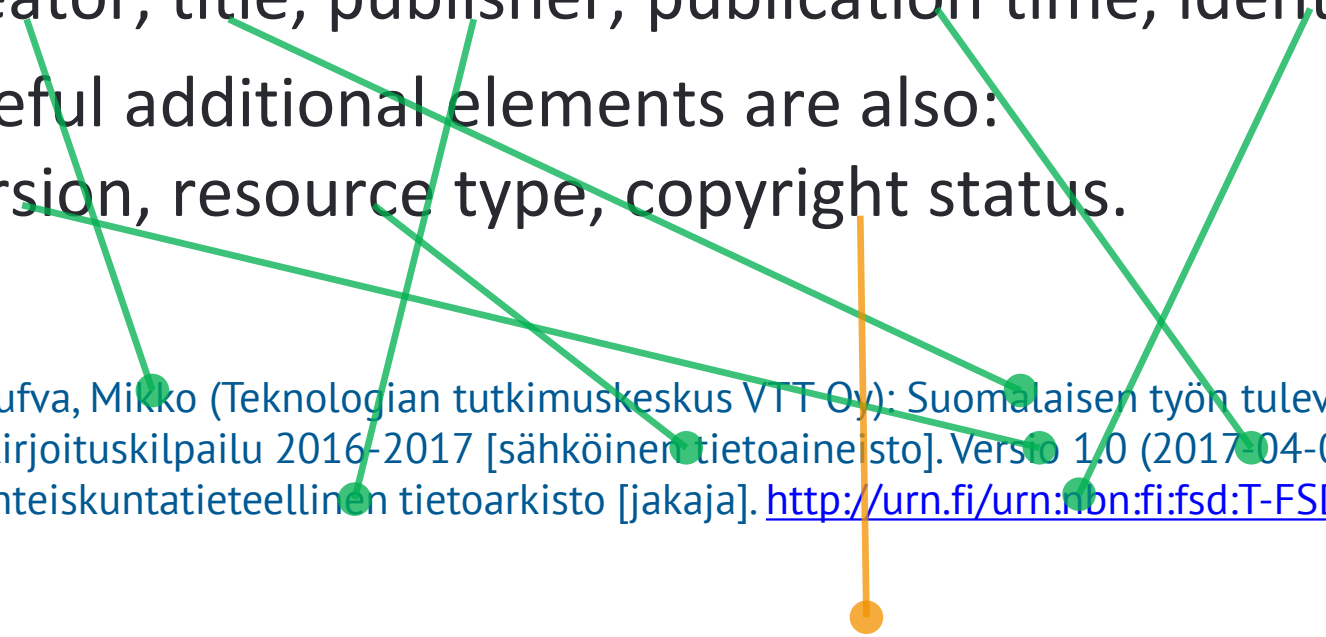
Dataan viittaaminen Tietoarkistossa

- Aineistojen jatkokäytön ehdot **vaativat** dataan viittaamista, ei vain hyvä tieteellinen käytäntö
  - Sanktiointi haasteellista, seuranta työlästä
- Tietoarkiston säilyttämät datat saavat viittaustiedot automaattisesti DDI-kuvailusta
- Datoilla(?) kuvailuilla(?) landing pageilla(?) on PID

FSD3166 Suomalaisen työn tulevaisuus -kirjoituskilpailu 2016-2017
Dufva, Mikko (Teknologian tutkimuskeskus VTT Oy): Suomalaisen työn tulevaisuus - kirjoituskilpailu 2016-2017 [sähköinen tietoaineisto]. Versio 1.0 (2017-04-04). Yhteiskuntatieteellinen tietoarkisto [jakaja]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3166

Vertaillaanpa tiekarttaan

- Data reference should consist of following elements: creator, title, publisher, publication time, identifier.
- Useful additional elements are also: version, resource type, copyright status.



Dufva, Mikko (Teknologian tutkimuskeskus VTT Oy): Suomalaisen työn tulevaisuus -kirjoituskilpailu 2016-2017 [sähköinen tietoaineisto]. Versio 1.0 (2017-04-04). Yhteiskuntatieteellinen tietoarkisto [jakaja]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3166

CESSDA ERIC PID-policy (valmistelussa)

- Velvoittaa data-arkistoja (Suomessa Tietoarkistoa)
- Ei ristiriitaa tiekarttaan nähden.
  - Taustalla samoja lähteitä, mm. Force11
- Suurilta osin samat tavoitteet sekä keinot
  - PIDit, viittausmallit ja riittävä metadata viitettä varten…
- landing pagen sisällössä ehkä hieman eroa
  - ei myöskään vaadi viittausta dataan, ohi landing pagen
  - how to access the data, licensing rules, different versions and provenance, tombstone page

CESSDA PID-policyn periaatteet

- "…ability to easily locate and access digital resources and then associate them with the related metadata is essential to allow for the citation, retrieval and preservation of those data…"
  - 1 – Identifying (yksiselitteisyys)
  - 2 – Locating (perustiedot, saatavuus)
  - 3 – Resolving (24/7)
  - 4 – Referencing and Citation (viittauksen karkeustaso!)
  - 5 – Visibility (data ja sen ristiviittausten maksimointi)
  - 6 – Flexibility (muutokset, esim. organisaatiotunnisteet)

Dataviittaus data-arkistossa/lle

- Osoitus siitä, että data on arkistoitu (ja saatavilla tietyin ehdoin)
- Omaa sellaisenaan sangen vähän toiminnallisuutta
- PID mainittu suosituksen tärkeimpänä elementtinä
  - Keskeinen rakennuspalikka
  - Huomioitava, ettei pelkästää toimi ihmisluettavana, innostavana viitteenä jatkotiedusteluihin
- Julkaisijan ja data-arkiston tiedonkulun suunta
  - Kumpi haravoi? Haravoiko jokin kolmas osapuoli?

Miksi viittausmalli data-arkistolle tärkeä?

- Valttikortti, jolla vakuuttaa tutkija arkiston palvelusta
  - A Trusted Digital Repository simply has one, sertifiointi
- Voidaan yhdistää ja rikastaa tarjottavaa kuvailutietoa
  - Julkaisut, koodi, tutkimusorganisaatiot ja -rahoittajat, aineistoon liittyvät datasetit, versiot, lisenssit…
- Voidaan jakaa (meta)dataa ja seurata datan käyttöä
- Mahdollistaa tiedon aineiston dearkistointinista
- Duplikaattien/versioiden tunnistaminen

Ajatelkaamme nyt arkiston sidosryhmiä

- Viittaus edellyttää, että on dataa johon viitata
  - Sidosryhmistä tärkein on siis tutkija/tutkijaYHTEISÖ
  - Datan jakamisen palveluita ja porkkana(keppejä) on edelleen kehitettävä ja tuettava
  - Rahoittajien ja tutkimusorganisaatioiden suositukset, koulutus ja niihin vaikuttaminen
  - Viitattu data vaatii pysyvän hoivaketjun (päättäjät)
    - Rahoitus, jatkuvuus, osaaminen, infrastruktuurivaateet
  - **Lainsäädännöllinen kehys**
    - Toimiedellytykset, tietosuojan varmistaminen yms.

Data-arkisto ja tutkija

- Tutkija on arkiston tärkein tiedonlähde
- Arkisto laatii dataviitteet
- Arkisto ohjeistaa (aineistonhallintasuunnitelmissa, tietopalveluna, koulutuksena) tuottamaan riittävää metadataa aineistosta
- Dataa hyödyntävä tutkija sitoutuu käyttöehtoihin/lisensseihin, ml. viittaamiseen
- Palvelut osana kansallista tutkimusinfrastruktuuria

I would share my data but…

- Someone could use it inappropriately
- No one has made a case that it will affect my academic career
- It is too much work to compile various parts of the data into a coherent dataset suitable for reuse
- There are no clear definitions of ownership or usage rights
- It would be a copyright infringement
- Someone misinterprets our data and it has a negative impact our group
- Data security issues regarding identifiers prevent it
- Data cannot be anonymized or it is too expensive to do, or it becomes useless when anonymized
- Legal reasons or research ethics code prevents sharing
- File formats are old or damaged, and there is not enough documentation (metadata) to be sure what to share
- No one else can understand the data, it is too personal
- It is old, it doesn't answer to the questions researchers ask today
- There is no funding to produce a reusable copy of the data
- No rules or recommendations on which data should be shared
- Data is classified or a NDA was required by the funder/partner
- People will criticize my methods.
- I don't have documentation - only I understand the data
- My data is something special I can offer my own students.
- I just don't have the time to fill in all the metadata again
- I spent a lot of money on this research, and it's too valuable to just give away.
- What if another researcher uses my research for commercial purposes? That's not how my funder wanted this used.
- Sharing might lead to need to give advice/guidance to those who reuse of the data
- I promised my IRB/participants I would destroy it when I am done
- I promised IRB/participants only the research team would see it
- I promised I would keep data locked in my office
- I don't want it to be used to reverse social policies my research has been used to support
- My university holds ownership and won't let me
- My PI/Collaborators won't share
- I'm not done with it yet
- I want to be able to co-author all publications
- My students need it to complete their theses before release
- The rights owner is dead so he can't give permission for deposit
- Data got lost in the mail
- There is no sharing culture in my field
- It is not the optimal thing for me as a researcher to do
- I do not know how or why
- Repository systems are too complicated to deposit
- I don't understand the licenses you require

As described on IASSIST listserv

Data-arkisto ja yliopisto

- Tekijyyden määrittely – kuka vastaa arkistoinnista?
- Arkistonti on sopimus mm. säilyttämisvastuusta ja datan välittämisestä --> kuka on sopimusosapuoli?
- Luvanvaraiset aineistot vaativat luvanantajan
  - Jatkuvuudesta huolehdittava
- Yliopiston tunnettava laajasti oman alansa viitekehys ja datanhallinnan toimijat

Data-arkisto ja rahoittaja

- Rahoittajan vaatimusten tulisi sisältää ehdottomia viittausvaatimuksia
- Sisällytettävä aineistonhallintasuunnitelmaan
- Saatavuudesta sovittava aineistohallintasuunnitelmassa
- Rahoitus ja tuotos yhdistettävissä tunnistein?

- Kustantajat myös mukana viittausohjeissa, esim. Elsevier

Data-arkisto ja julkaisija

- Missä vaiheessa data saa tunnisteen?
  - Mahdollista jo hyvin aikaisessa vaiheessa ”ennen dataa”
- Edistettävä datan ja julkaisun yhdistämistä
- Datan merkittävyys ja vertaisarviointi?
  - Mikä taho vastaisi tästä? Data-arkisto, tiedeyhteisö?
  - Data tutkimustuotoksena nostaa sen arvoa, mutta arkisto tai rahoittaja ei yksinään voine arvioida datan tieteellistä laatua. Siihen tarvitaan tiedeyhteisöä.
  - Datanjulkaisualustojen jufo-luokitus?!?
    - I publish in repository X because it is better than Y

Data-arkisto ja tieteellinen kirjasto

- Työnjako: onko kirjasto osa tutkimuslaitosta, jolloin vastuulla olisi enemmän koulutus ja käyttö, kuin arkistointi?
- Missä on aineistojen jatkokäytön osaaminen ja tuki?
  - Perinteinen data librarianship
  - Erityisesti aineistot, jolla ei ole temaattista palvelukeskittymää

Landing page

- Luonnollinen arkiston ylläpidettäväksi
- Voi vaatia sopimuksen rekisterinpitäjän (tutkija) kanssa, jos sisältää henkilötietoja
- Lisenssi, käyttölupaehdot, erityisehdot
- Keskeiset tietolähteet:
  - Tutkija
  - Mahdollisesti rahoittaja
  - Mahdollisesti organisaatiorekisteri
  - Sanastot (esim. Finto)

Lyhyt katsaus suosituksiin

- All datasets intended for citation must have a globally unique persistent identifier
- Finnish data repositories should use either DOI or URN as their PID of choice
  - DOI usein nähdään olevan “paras ja kaunein”, tuottaa jonkin verran työtä olla yhteentoimiva
- This persistent identifier must resolve to a landing page that supports access to the actual data set.
  - Pääsy voi olla ehdollinen, lisenssi/käyttöehtopohjainen, käyttötarkoitussidonnainen tms.
- Assigning PIDs and creating landing pages is the responsibility of the data repository
- Landing page should facilitate access to metadata, either by holding metadata or a link to metadata.
  - http://urn.fi/urn:nbn:fi:fsd:T-FSD3166
- The landing page should include reference model for citation, and ideally also metadata helping with discovery, in human-readable and machine-readable format
- National data centers, libraries and archives should agree on the required metadata content of a data landing page.
  - Kuinka yksityiskohtainen tämän pitäisi olla, mihin sen tulisi pohjautua, ja kuinka sitova sopimuksen tulisi olla?
  - Tietoarkisto linkittää aineistokuvailuun. Eri aineistotyypit eroavat kuvailevan metadatan osalta
  - Entä kv-standardit/suositukset ja yhteentoimivuus
- Data that no longer exist should have a persistent landing page
- License all metadata with a CC0 license or equivalent.
  - CC BY käytössä Tietoarkistossa
- Make metadata freely harvestable through open APIs
- The persistent identifier must be embedded in the landing page in machine-readable format
- Pilot the RDA Data Citation model for dynamic data in one or several national data centers.
  - Ei suoraan sovellu Tietoarkiston dataan toistaiseksi. Onko muita malleja?
- Release all data citation related content intended for broad audiences, in open format, i.e. CC-BY, or equivalent.

Lopuksi on sanottava…

- Ehdotetut viittausohjeet ovat hyvin kattavat ja yhdenmukaiset aihepiirin muiden ohjeiden kanssa
- Syntyykö kansainvälisiä, rinnakkaisia, malleja hyvin pienillä eroilla?
- Miten tarkkaan määritellään landing page?
- Kuinka hienojakoisia ovat tutkijan tarpeet dataviittauksissa? Generoituihin osiin viittaaminen?
- Periaatteista käytäntöön ja yhteentoimivuuden haasteet.

Vielä lopun jälkeen todettava…

- Sharing benefits the scientific community, while it maybe is not the most optimal strategy for the individual researcher. (Pronk 2016)
  - Citation benefit may be an answer.

- Kuinka varmistaa, että viittaukset tutkimustuotoksiin tulevat näkyviksi ja yhdistyvät toisiinsa ja kaikki osapuolet hyötyvät?
    - Useita avauksia (esim. Scholix) mutta ei määräävää

Kiitos

Alaterä, Tuomas J. (Tietoarkisto) : Eri sidosryhmien roolit kansallisen dataviittausjärjestelmän kokonaisuudessa. Datan jäljet -työpaja. [sähköinen tietoaineisto]. Versio 1.0 (2017-10-20). Data-asian kansalliskomitea [jakaja]. CC BY. (PID puuttuu, ehkä hyvä niin)

Tuomas J. Alaterä
tuomas.alatera@uta.fi
orcid.org/0000-0002-3448-3448

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Mainitut lähteet

- Pivowar, Heather A. & Day, Roger S. & Fridsma, Douglas B. (2007): Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE2(3): e308. https://doi.org/10.1371/journal.pone.0000308
- Piwowar, Heather A & Vision, Todd J. (2013): Data Reuse and the Open Data Citation Advantage. PeerJ 1:e175 https://doi.org/10.7717/peerj.175
- Pronk, Tessa (2016): A Game Theoretic Analysis of Research Data Sharing. Konferenssiesitys IASSIST 2016, Bergen, Norway. http://iassistdata.org/conferences/2016/presentation/7426

- "I would share my data but" -editoitujen sitaattien lähde IASSISTin keskustelulista lokakuu 2017.