OHJE TUTKIMUSAINEISTOJEN KUVAILUUN

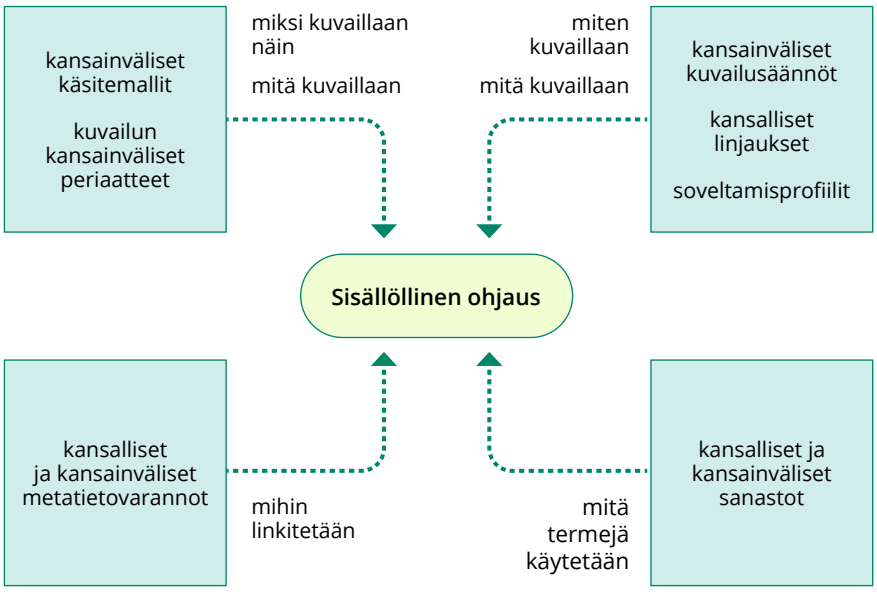
SISÄLLYS

OHJE TUTKIMUSAINEISTOJEN KUVAILUUN

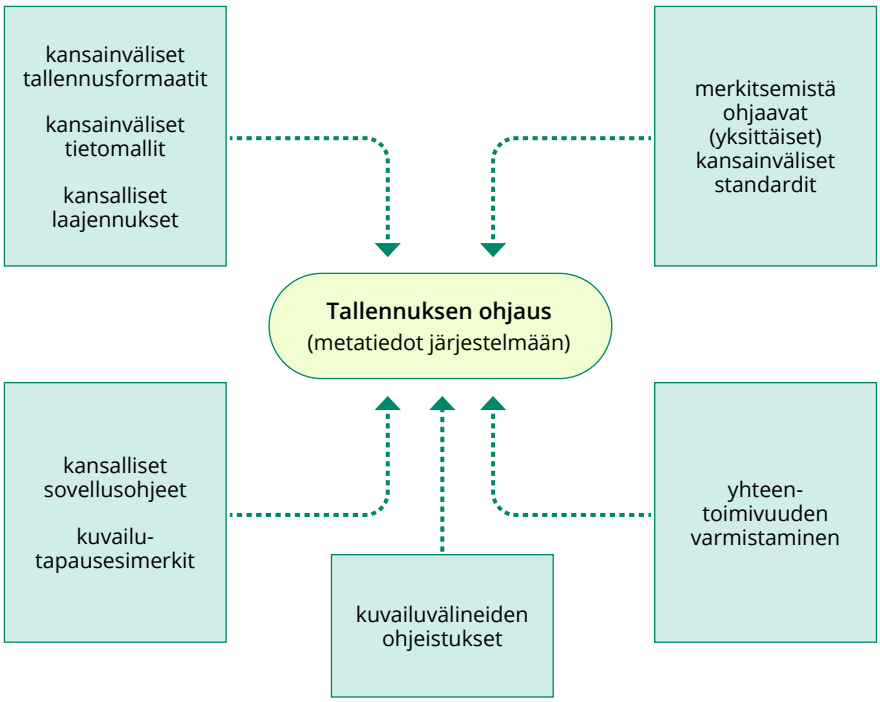
10.5281/zenodo.13847137

OHJE TUTKIMUSAINEISTOJEN KUVAILUUN

Tässä Avoimen tieteen ja tutkimuksen koordinaation piirissä laaditussa kokonaisuudessa luodaan yleiskatsaus aineiston-kuvailun keskeisimpiin elementteihin. Tarkoitus on helpottaa tutkijalukijaa orientoitumaan kuvailutyöhön ja avata yleisellä tasolla siihen liittyviä prosesseja. Keskeisellä sijalla ovat kuvailevaan metadataan (*descriptive metadata*) liittyvät teemat. Moni organisaatio ja tieteenala on laatinut tutkijoidensa avuksi tätä yleisesitystä yksityiskohtaisempia ohjeita, joihin tutkijan on hyvä tutustua mahdollisimman varhaisessa vaiheessa. Niistä löytyy muun muassa neuvoja tutkimusaineiston kannalta parhaiten soveltuvan repositorion valintaan.



Kuva 1. Miten ja miksi kuvailua tehdään? Miten kuvailutyötä ohjataan? Kuva: Marja-Liisa Seppälä 2023 (CC BY).



Kuva 2. Tallennuksen ohjaamisen periaatteet. Kuva: Marja-Liisa Seppälä 2023 (CC BY).

JOHDANTO: TUTKIMUSAINEISTON KUVAILU ON OSA KAIKKEA TUTKIMUSTA

Tutkimusaineiston kuvailu on osa kaikkea tutkimusta riippumatta siitä, voidaanko aineisto avata. Tutkimusaineiston kuvailuprosessissa syntyy kokonaiskuva tutkimuksen lähtökohdista ja toteutuksesta sekä siinä kerätystä aineistosta kontekstitietoineen. Aineiston huolellinen kuvailu parantaa osaltaan tutkimuksen laatua ja tekee sen arvioimisen ja rakentavan kritiikin mahdolliseksi. Parhaimmillaan tutkijalle avautuu uusia yhteistyömahdollisuuksia, kun samojen teemojen kanssa työskenteleville kollegoille tarjotaan näkymä aineistoon.

Tutkimusaineistoa kuvaillaan ja kuvailua täydennetään koko tutkimuksen ajan. Prosessin aikana tutkija dokumentoi muun muassa miten tutkimus on toteutettu, miten tutkimusaineistot on kerätty ja mitä ovat tutkimuksessa käytetyt keskeiset muuttujat. Kun aineiston kuvailu täydentyy tutkimusta tehtäessä, aineiston ja sen kuvailun avaaminen tutkimuksen valmistuttua ei aiheuta lisätyötä. Välittömästi tehty kuvailu myös parantaa kuvailun laatua ja tarkkuutta.

AINEISTON KUVAILU JA FAIR-PERIAATTEET

Tutkimusmaailmassa ja erityisesti tutkimusaineistoihin liittyvissä keskusteluissa viitataan tätä nykyä tiheästi alun perin vuonna 2014 muotoiltuihin FAIR-periaatteisiin. FAIR-lyhenne tulee sanoista **Findable, Accessible, Interoperable ja Re-usable**. Se tarkoittaa, että tutkimuksessa käytettyjen lähdeaineistojen, menetelmien ja tietotuotteiden tulee olla löydettävissä, saatavilla, siirrettävissä tai yhdistettävissä ja uudelleenkäytettävissä. (https://www.opendata.fi/fi/opas/tietoaineiston-fair-periaatteet)

Jos aineistot kuvaillaan huolellisesti, edellä mainitut FAIR-periaatteet toteutuvat. Huolellisesti, ymmärrettävästi, tieteenalan käytäntöjen mukaisesti kuvailtu ja tallennettu aineisto, jonka teknisestä uudelleenkäytettävyydestä on huolehdittu, on FAIR (ks. esim. Keiski ym. 2023). Voi olla, että itse aineistoa ei esimerkiksi sen arkaluonteisuuden takia voida avata, mutta lähes aina aineistoa kuvaileva metatieto voidaan asettaa avoimesti saataville. Avoimet metatiedot lisäävät tutkijan ja tutkimusaineiston näkyvyyttä myös silloin, kun aineistot ovat suljettuja. Yhteentoimivuuden ja uudelleenkäytettävyyden kannalta on tärkeää, että kuvailtu aineisto tallennetaan koneluettavassa muodossa.

KUVAILUN PERUSELEMENTTEJÄ – YHTEISTEN KÄYTÄNTÖJEN MERKITYS AINEISTONKUVAILUSSA

Eri tieteenaloilla on omat vakiintuneet käytänteensä esimerkiksi viittaamisessa ja lähteiden merkitsemisessä. Sama pätee myös aineiston kuvailuun: yhden tieteenalan tapa kuvailla aineistoa ei välttämättä sovellu sellaisenaan toisen tieteenalan käyttöön. Aineiston löydettävyyden, saatavuuden, yhteentoimivuuden ja uudelleenkäytettävyyden kannalta on kuitenkin tärkeää, että tieteen- tai oppialasta riippumatta aineistonkuvailussa noudatetaan joitain yhteisesti sovittuja käytäntöjä.

Oleelliset kaikkia tieteenaloja koskevat aineistonkuvailun metatietoelementit esitellään taulukossa 1. Suurin osa elementeistä on johdettu datan kuvailussa käytetyistä metadatastandardeista (DCAT, Dublin Core). Enimmän osan kuvailutiedosta luo useimmiten aineiston tuottamiseen osallistunut tutkija. Julkaisupäivän ja pysyvän tunnisteen tarjoaa useimmiten tutkimusaineiston julkaiseva arkisto tai repositorio aineiston julkaisuvaiheessa. Arkisto tai repositorio saattaa myös määritellä, mitä tieteenalaluokituksia tai sanastoja sisällönkuvailussa käytetään.

Taulukko 1. Tutkimusaineistojen kuvailussa keskeisiä metatietoelementtejä

Elementti	Kuvaus/Ohje	Mihin FAIR-periaatteeseen liittyy*	Tiedon tuottaja
Aineiston nimi	Anna aineistolle yksilöivä ja kuvaava nimi 1) aineiston kielellä ja 2) englanniksi.	Löydettävä	Tutkija
Tekijä & ORCID	Nimeä henkilö tai organisaatio, joka on tuottanut aineiston. Kirjaa nimet muodossa Sukunimi, Etunimi; Sukunimi2, Etunimi2. Nimien lisäksi on hyvä käyttää yksilöllistä pysyvää tutkijatunnistetta (esim. ORCID), jonka avulla samannimiset tutkijat voi erottaa toisistaan.	Löydettävä, saavutettava, uudelleenkäytettävä, yhteentoimiva	Tutkija
Tekijän affiliaatio	Ilmoita tekijän organisaatiokytkös, jonka piirissä ja resursseilla aineisto on tuotettu. Käytä mahdollisuuksien mukaan aina organisaation nimen lisäksi organisaation koodia, esim. Tiedejatutkimus.fi -portaalin käyttämät organisaatiokoodit.	Löydettävä	Tutkija
Aineiston omistaja	Nimeä aineiston omistajat. Mikäli aineiston oikeuksia ei ole työsopimuksella tai muilla sopimuksilla (esimerkiksi hankesopimus) siirretty organisaatiolle, ovat omistajia aineiston tekijät. Henkilöiden nimien lisäksi on hyvä käyttää yksilöllistä pysyvää tutkijatunnistetta (esim. ORCID). Organisaation nimen lisäksi on hyvä käyttää organisaation koodia.	Uudelleenkäytettävä	Tutkija
Julkaisija	Nimeä toimija, jolla on oikeus julkaista pelkät metatiedot TAI metatiedot ja data TAi metatiedot ja datan dokumentaatio: esimerkiksi repositorio, arkisto tai tutkimusorganisaatio. Organisaation nimen lisäksi on hyvä käyttää organisaation koodia.	Löydettävä, saavutettava	Tutkija
Rahoittaja	Kuvaile taho, jonka rahoituksella ja resursseilla aineisto on tuotettu. Organisaation nimen lisäksi on hyvä käyttää organisaation koodia.	Löydettävä	Tutkija
Tutkimus-infrastruktuuri	Kerro, onko aineisto tuotettu jonkin tutkimusinfrastruktuurin piirissä?	Yhteentoimiva, uudelleenkäytettävä	Tutkija
Kuvaus	Laadi aineiston sisällön tiivis (1-2 kappaletta) kuvaus. Mitä aineistolla tutkitaan tai on tutkittu? Mistä aineisto koostuu? Mikä on tutkimuskohde ja havaintoyksikkö? Kun kuvailutiedot ja/tai kuvailtava aineisto julkaistaan, kuvauksen tulee olla niin informatiivinen, että ulkopuolinen voi ymmärtää ja käyttää aineistoa.	Löydettävä, uudelleenkäytettävä	Tutkija
Menetelmä	Laadi lyhyt kuvaus aineiston keruussa, muodostamisessa ja analyysissa käytetystä menetelmästä tai menetelmistä sekä ohjelmistoista. Aineistotyypistä riippuen anna lyhyt kuvaus keruuinstrumentista ja otantamenetelmästä (esimerkiksi mittalaite, puolistrukturoitu haastattelu jne).	Yhteentoimiva, uudelleenkäytettävä	Tutkija
Avainsanat	Valitse vapaasti valittavat, aihetta kuvailevat sanat, jotka edesauttavat aineiston löydettävyyttä, kun kuvailutiedot ja/tai aineisto julkaistaan.	Löydettävä	Tutkija

Elementti	Kuvaus/Ohje	Mihin FAIR-periaatteeseen liittyy*	Tiedon tuottaja
Asiasanat	Valitse kontrolloiduista sanastoista tai ontologioista (yleisistä tai tieteenalakohtaisista) haetut aihetta kuvailevat sanat, jotka parantavat aineiston löydettävyyttä, kun kuvailutiedot ja/tai aineisto julkaistaan. Esimerkkejä sanastoista: KOKO-ontologia, YSO, MeSH.	Löydettävä	Tutkija
Tieteenala	Määritä aineiston tieteenala. Eri repositoriot/arkistot käyttävät erilaisia tieteenalaluokituksia. Esimerkki: OKM:n tieteenalaluokitus.	Löydettävä	Arkisto/ Repositorio/ Julkaisija
Aineistotyyppi	Kuvaa millaisesta tutkimusaineistosta on kysymys, esimerkiksi haastattelu, mittausdata, fyysiset näytteet, teksti, kuva, video, koodi, tilasto. Pyri käyttämään kontrolloitua sanastoa, esim. YSO.	Löydettävä	Tutkija
Kieli	Määrittele kieli, jota käytetään aineistossa. Käytä ISO 639-2/3 -standardin mukaista kolmimerkkistä kielikoodia, esim. eng, fin, swe.	Löydettävä, saavutettava, yhteentoimiva, uudelleenkäytettävä	Tutkija
Ajallinen kattavuus	Kuvaa ajanjakso, jonka aineisto kattaa. Käytä ISO 8601 -standardin mukaista muotoa YYYY-MM-DD. Dokumentoi tarvittaessa kellonaika muodossa HH:MM:SS.	Löydettävä, yhteentoimiva	Tutkija
Maantieteellinen kohdealue	Määrittele maantieteellinen alue, jonka aineisto kattaa. Jos mahdollista, käytä ISO-standardin tai jonkin muun koneluettavan koodiston mukaista paikannimeä tai koordinaatteja, esim. YSO-paikat.	Löydettävä, yhteentoimiva	Tutkija
Versio	Jos aineistosta on useita versioita, ilmoita tiedosto- ja kansiotasolla selkeästi, mikä versio on kyseessä, esim. v01.1.	Uudelleenkäytettävä	Tutkija
Tiedostomuoto/ -formaatti	Ilmoita datan tallennusformaatti. Käytä mahdollisuuden mukaan yleisesti käytössä olevia tiedostoformaatteja, jotka ovat riippumattomia (kaupallisista) ohjelmistoista kuten MS Office -ohjelmistoista (esim. .txt, .mp4). Pyri käyttämään kontrolloitua sanastoa, esim. MIME typejä. Tietoa pitkäaikaissäiltyspalvelujen säilytys- ja siirtokelpoisista tiedostomuodoista.	Saavutettava, yhteentoimiva, uudelleenkäytettävä	Tutkija
Aineiston saatavuus (avoin, embargo, kirjautuminen, rajattu, vaatii luvan)	Kuvaile valmiin aineiston saatavuus. Jos aineisto pysyy perustellusta syystä rajatun pääsyn takana tai kokonaan suljettuna, kuvaile tämä. Jos aineistoa ei ole mahdollista avata, perustele, miksi aineisto ei ole avoimesti saatavilla. Esimerkkejä avoimuuden asteista: ”Saatavilla avoimesti verkosta”; ”Saatavilla pyynnöstä” sekä perustelu rajatulle pääsylle; ”Saatavilla metatietojen, dokumentaation tai aineiston tasolla yhteydenotolla aineiston omistajaan” sekä perustelu rajatulle pääsylle; ”Ei saatavilla” sekä perustelu aineiston pitämiselle poikkeuksellisesti kokonaan suljettuna. Jos aineisto on kaupallisesti saatavilla, kuvaile, miten aineiston voi ostaa käyttöönsä.	Saavutettava, uudelleenkäytettävä	Tutkija

Elementti	Kuvaus/Ohje	Mihin FAIR-periaatteeseen liittyy*	Tiedon tuottaja
Perustelu mahdolliselle pääsyn rajoittamiselle	Jos aineistoa ei ole mahdollista julkaista avoimesti, kuvaile perustelu saatavuuden rajoittamiselle. Tutkimusaineistojen koodistojen rajoiteperusteluokissa on kuvattu perusteltuja syitä saatavuuden rajoittamiselle.	Saavutettava, uudelleenkäytettävä	Tutkija
Lisenssi	Lisenssi määrittelee ehdot, joilla aineistoa saa käyttää uudelleen. Yleisesti käytettyjä lisenssejä teksti-, kuva- ja taulukkoaineistoille ovat Creative Commons 4.0 -lisenssit. Skripteille, koodille ja ohjelmistoille yleisiä ovat esimerkiksi GNU- ja MIT-lisenssit.	Uudelleenkäytettävä	Tutkija
Aineistoon liittyvät muut tutkimusaineistot	Avaa aineiston suhde toisiin tutkimusaineistoihin: 1) Aineisto on johdettu jostain toisesta aineistosta (IsBasedOn), 2) Aineisto on toisen aineiston osa (IsPartOf), 3) Aineistoon liittyy toisia aineistoja (HasPart). Esimerkkejä: 1) seuranta-aineisto, jossa aineistonkeruut toistuvat esimerkiksi vuosittain, 2) kuva-aineisto, joka sisältyy laajempaan moniosaiseen aineistoon, 3) terveystutkimus, joka jakaantuu kysely- ja haastatteluaineistoihin.	Löydettävä, uudelleenkäytettävä	Tutkija
Aineistoon liittyvät muut tutkimus-tuotokset	Ilmoita mahdolliset julkaisut tai muut tuotokset, joilla on yhteys tähän tutkimusaineistoon.	Löydettävä, uudelleenkäytettävä	Tutkija
Henkilötiedot	Jos aineisto sisältää henkilötietoja, kuvaa seuraavat tiedot: 1) rekisterinpitäjä, 2) sisältääkö aineisto erityisiin henkilötietoryhmiin kuuluvia tietoja tutkimukseen osallistujista ja/tai aineistossa esiintyvistä henkilöistä, ja jos sisältää, mitä.	Saavutettava, uudelleenkäytettävä	Tutkija
Salassa pidettävät tiedot	Sisältääkö aineisto salassa pidettäviä tietoja (esim. liikesalaisuus, sensitiivinen lajitieto)?	Saavutettava, uudelleenkäytettävä	Tutkija
Julkaisupäivä	Ilmoita päivä, jolloin aineisto on julkaistu data-arkistossa tai aineistorepositoriossa. Ilmoita aika ISO 8601 -standardin mukaisessa muodossa YYYY-MM-DD. Tarvittaessa kellonajan voit ilmoittaa muodossa HH:MM:SS.	Löydettävä, yhteentoimiva	Arkisto/ Repositorio/ Julkaisija
Säilytyspolitiikka	Kirjaa päätös ja suunnitelman kuvaus aineiston säilyttämisestä pysyvästi tai määrätyn ajan. Perustele miksi juuri tämä päätös aineiston säilyttämisestä on tehty, erityisesti jos aineistoa ei säilytetä pysyvästi. Tutkija on vastuussa aineistonsa tuhoamisesta määräajan päätyttyä.	Saavutettava, uudelleenkäytettävä	Tutkija
Pysyvä tunniste	Anna aineistolle yksilöllinen, pysyvä tunniste. Repositoriot ja arkistot, joihin tutkimusaineistoja tallennetaan, tarjoavat usein pysyvän tunnisteen kuten DOI, URN tai arkistotunnus (accession number).	Löydettävä, saavutettava, uudelleenkäytettävä	Arkisto/ Repositorio/ Julkaisija

Erityisesti yhteensopivuuden ja uudelleenkäytettävyyden takaamiseksi kuvailussa on hyvä tallentaa myös aineistoa koskevat paratiedot eli kuvaus tutkimusaineiston keräämisestä ja käsittelystä. Hyvän tieteellisen käytännön mukaisesti laadukkaaseen aineistonkuvailuun sisältyy myös tekijätietojen ilmaiseminen (TENK 2023). Aineiston laajuus vaikuttaa siihen, millä tasoilla aineistoa kuvaillaan. Tavallisesti käytettyjä tasoja ovat

- koko aineiston taso (kansiotaso),
- kansiorakenteen taso (alakansiot),
- tiedostojen taso,
- tiedoston sisäisten rakenteiden taso (esim. muuttujat) sekä
- datan sisältöjen taso (esim. koodaukset, selitteet).

Joillakin tieteenaloilla on tärkeää, että tutkimus voidaan tarvittaessa toistaa. Huolellisesti oman tieteenalan käytänteiden mukaisesti tehty kuvailu edesauttaa myös tässä.

REPOSITORION VALINTA

Aineistonkuvailun taustalla on oletus siitä, että aineisto tai vähintään aineistoa koskevat metatiedot asetetaan saataville tuonnempaa käyttöä varten. Saatavilla pitämisestä huolehtivat tavallisesti erilaiset repositoriot (esimerkkejä taulukossa 2). Aineistolleen sopivan repositorion valinnassa tutkijan kannattaa turvautua oman tieteenalansa ja organisaationsa ohjeisiin. Niitä noudattamalla hoituu yleensä myös aineiston yhteentoimivuuden ja uudelleenkäytettävyyden kannalta tärkeä koneluettavuus. Luotettavan repositorion tunnistaa helpoimmin siitä, että sillä on CoreTrustSeal sertifikaatti.

Taulukko 2. Esimerkkejä datarepositorioista

Tieteenala	Repositorio	Standardi	Huomiota	Linkki
Monitieteinen	Zenodo	DataCite	Pakolliset kentät: Publication date, title, authors, description, access right, license	https://www.re3data.org/repository/r3d100010468
Monitieteinen	Figshare	DataCite	Pakolliset kentät: Item title, item type, authors, categories, keywords, description, license	https://www.re3data.org/repository/r3d100010066
Monitieteinen	FAIRdata IDA Fairdata Qvain	IDA:ssa olevien datojen kuvailu tapahtuu Qvaimessa, ks. ylempi rivi. Fairdata Metax tietomalli Perustuu DCAT 3.0 Konvertoituu myös DataCite -formaattiin Fairdata Metax tietosisältö selkokielellä	Säilyttäminen IDAssa vaatii aktiivisen projektin. Käyttöoikeus myönnetään organisaatiolle. Ei sensitiiviselle datalle. Kuvailu tapahtuu Qvaimessa. Qvain on kuvailutyökalu IDAssa olevalle tai muualle arkistoidulle datalle. Qvaimen pakolliset kentät: Lisenssi, aineiston kuvaus ja otsikko, julkaisupvm, avainsanat, tekijä (henkilö tai organisaatio) ja julkaisija (henkilö tai organisaatio). Hakupalvelu julkaistuille metadatoille: Fairdata Etsin.	https://www.fairdata.fi/ida/ https://www.re3data.org/repository/r3d100012157 https://qvain.fairdata.fi/
Monitieteinen	Dryad	Dublin Core, DataCite, OAI-ORE, RDF DataCube	Pakolliset kentät: Journal name; Title; Author(s); Abstract; Research domain; Keyword(s)	https://datadryad.org/stash/requirements
Bio- ja ympäristötieteet	Pangaea	Darwin Core, Dublin Core, ISO 19115, DIFF	Pakolliset kentät: Event; Expedition; PI; Author(s); Data set title; Reference(s);-Method; Abstract;	https://wiki.pangaea.de/wiki/Metadata
Biotieteet	BOLD System	Useita	BOLD = Barcode of life data system. Esimerkiksi valokuvan pakolliset kentät ovat: Image file; Original specimen; View metadata; Sample ID; License; License year; License contact.	http://www.boldsystems.org/index.php/Public_BINSearch?searchtype=records
Lääketieteet	Useita, ks. esim. NIH-Supported Data Sharing Resources	Useita		https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
Kielitieteet	Kielipankki	CLARIN Standards Information System	Ilmoitus kielivarasta – pakolliset kentät: Information provider, email address of information provider, organization, name of the language resource in Finnish and in English, type Tallettaminen | Kielipankki	https://www.kielipankki.fi/ https://www.re3data.org/repository/r3d100011807

Tieteenala	Repositorio	Standardi	Huomiota	Linkki
Kaikki, erityisesti yhteiskunta-, humanistiset ja kasvatustieteet	Tietoarkisto	DDI, Data Documentation Initiative	Tietoarkiston hakuportaali on Aila Aineistoa ehdotetaan Tietoarkistoon, joka tarkistaa sen sopivuuden. Pakolliset kentät ilmoituksessa: Aineiston tekijä tai kerääjän nimi, vastaus tutkittavien informoinnista, aineiston nimi ja lyhyt kuvaus, aineiston koko, ilmoittajan nimi, taustaorganisaatio ja email	https://www.fsd.tuni.fi/fi/ https://www.re3data.org/repository/r3d100010490
Monitieteinen ja/tai yhteisökohtainen	EUDAT CDI B2SHARE / EUDAT B2SHARE Premium	EUDAT Core ja Extended schema (perustuvat DataCite schemaan), sekä yhteisökohtaiset laajennukset. https://schema.eudat.eu/eudatextended/ https://schema.eudat.eu/communityextensions/	EUDAT CDI B2SHARE on ilmaispalvelu, jossa kaikille yhteiset käyttöehdot. Yhteisökohtaiset räätälöidyt skeema-laajennukset mahdollisia myös ilmaispalvelussa, esimerkiksi sopivissa yhteishankkeissa. EUDAT B2SHARE Premium -palvelut ovat organisaatioille maksullisia räätälöityjä palveluita, joissa mm. tallennuskapasiteettia, metatietomallia ja palvelun toiminnallisuuksia voidaan räätälöidä asiakkaan kanssa sovittavalla tavalla. Pakolliset kentät (Premium-palveluissa räätälöitävissä): https://schema.eudat.eu/eudatcore_metadataelements/	EUDAT CDI B2SHARE: https://b2share.eudat.eu https://www.re3data.org/repository/r3d100011394 https://www.eudat.eu/catalogue https://docs.eudat.eu/b2share/overview/ Esimerkki B2SHARE Premium palvelusta – Ilmatieteen laitoksen METIS-palvelu: https://www.re3data.org/repository/r3d100013582 https://fmi.b2share.csc.fi EUDAT CDI B2FIND -palvelu koostaa metatietoja kaikista B2SHARE-palveluista ja muista yhteisöjen repositorioista: https://b2find.eudat.eu

ESIMERKKEJÄ AINEISTONKUVAILUSTA

Alla olevassa luettelossa mainitut esimerkit on haettu Tieto-arkistosta. Niille on tarjolla vastaava kuvailu myös englanniksi. (Alaterä, 2024.)

Hyviä kvantitatiivisia esimerkkejä voisivat olla esim.

- Kansallinen vaalitutkimus (https://urn.fi/urn:nbn:fi:fsd:T-FSD3467)
- Kansalaispulssi (https://urn.fi/urn:nbn:fi:fsd:T-FSD3826)
- Hyvinvointi ja eriarvoisuus Suomessa (https://urn.fi/urn:nbn:fi:fsd:T-FSD3539)

Laadullisia erimerkkejä voisi tarjota

- Yksinäinen yhdessä ja erikseen (https://urn.fi/urn:nbn:fi:fsd:T-FSD3609)
- Kirjoituksia koirani elämästä (https://urn.fi/urn:nbn:fi:fsd:T-FSD3724)

LOPUKSI: MUISTA AINAKIN NÄMÄ

Tutkijalle voi olla vaikeaa kuvata tutkimusaineistoa erillään tutkimustuloksista. Toivottavasti tämä ohje on osaltaan auttanut lukijaa hahmottamaan, mitä aineistonkuvailu on ja miksi sen tekeminen on tärkeää.

Tutkimusaineiston kuvailua tehdään tutkimuksen aikana ja tutkimusaineistojen julkaisemisen yhteydessä laaditaan julkiset kuvailutiedot. Kuvailutiedot on hyvä julkaista, vaikka varsinaista tutkimusaineistoa ei julkaistaisikaan.

Lukijan on hyvä muistaa ainakin nämä asiat:

1. Tutkimusaineiston kuvailun kohteena on se, mitä aineisto sisältää. Tutkimuksen ja tutkimustulosten sijaan kuvaillaan tutkimusaineistoa – sitä, miten tutkimusaineisto on koottu, mikä on perusjoukko, miten aineisto on järjestetty, kuinka paljon aineistoa on, mitä rajoituksia aineistoon liittyy ja niin edelleen.
2. Ellet tiedä, kysy! Omasta organisaatiostasi löytyy varmasti asiantuntijoita jotka vähintään osaavat ohjata avun pyytäjää eteenpäin.
3. Aineistonkuvailu edistää omalta osaltaan tieteellisen tiedon leviämistä ja saatavuutta kansalaisyhteiskunnassa; tutkittu tieto kuuluu kaikille.

MUISTILISTA KÄYTÄNNÖN TUEKSI

a. Suunnittele ja laadi kuvailun käytänteet tutkimuksen eri vaiheisiin.
b. Valitse julkaisuarkisto tai metatietotietokanta ja tutustu sen vaatimuksiin.
c. Hyödynnä standardeja ja oman alasi kuvailuohjeita. Kuvailun perus-elementit -taulukosta löydät keskeisimmät tietokentät.
d. Mieti millaista kuvailutietoa tarvitaan, jotta aineistosi on ymmärrettävää ja sitä on mahdollista hyödyntää tutkimuksen jälkeenkin.
e. Ole järjestelmällinen kuvailussa ja dokumentoinnissa koko tutkimuksen ajan.
f. Laadi julkiset kuvailutiedot ja julkaise ne (mahdollisesti yhdessä aineistosi kanssa).

SANASTO

Sana	Selitys	Englannin-kielinen termi
Asiasanat	Asiasanat ovat toisiinsa linkitettyjä termejä, joita voidaan valita ylläpidetystä, ennalta määritellystä sanastosta kuvaamaan aineistoa. Kontrolloitujen asiasanojen käyttö parantaa aineiston löydettävyyttä, sillä sanastot on pyritty laatimaan yksiselitteisiksi (synonyymit viittaavat samaan termiin) ja ne sisältävät myös tietoa asiayhteydestä. Koska kontrolloidut asiasanastot ovat rajallisia, kuvailua on hyvä täydentää vapaamuotoisemmilla avainsanoilla.	Subject Heading
Avainsanat	Avainsanat ovat vapaamuotoisia aineiston sisältöön liittyviä sanoja, joilla aineistoa voidaan kuvailla. Avainsanoilla voidaan laajentaa ja tarkentaa kontrolloiduin asiasanoin tehtyä kuvailua kyseisen tutkimusalan tarkalla käsitteistöllä. Kontrolloimattomat avainsanat eivät välttämättä ole yksiselitteisiä laajemmassa kontekstissa, vaan niiden merkityksissä voi olla tieteenalakohtaista vaihtelua.	Keyword
Data-repositorio	Datarepositorio on tutkimusaineistojen tallentamiseen ja etsimiseen tarkoitettu digitaalinen arkisto. Arkistoon tallennetaan varsinaisen tutkimusaineiston tiedostojen lisäksi myös kuvailutietoja. Datarepositorion valintaan on hyvä perehtyä huolella (ks. ohjeen kohta “datarepositorion valinta”).	Data Repository
FAIR-periaatteet	FAIR-periaatteet muodostuvat tutkimusaineiston löydettävyydestä (F), saavutettavuudesta (A), yhteentoimivuudesta (I) ja uudelleenkäytettävyydestä (R). Periaatteita noudattamalla tutkimuksen laatu ja vaikuttavuus todennäköisesti kasvavat, koska tällöin aineiston kuvailutiedot löytyvät helpommin eri tietokannoista. Aineiston sisältöä on kuvattu siten, että myös aineiston kerääjän on helpompi palata siihen myöhemmin. On tärkeää huomioida, että FAIR-periaatteisiin sisältyy myös mahdollisuus rajata tutkimusaineistoon pääsyä, mikäli se on välttämätöntä. https://tieteentermipankki.fi/wiki/Avoin_tiede:FAIR-periaatteet	FAIR Principles
Koneluettavuus	Koneluettavuus tarkoittaa, että tieto on rakenteistettu siten, että kone pystyy käsittelemään tietoja. Esim. PDF-muoto ei ole koneluettava, sillä vaikka ihmisen on helppo lukea sitä, niin koneellisesti sitä ei pysty tulkitsemaan.	Machine Readability
Kuratointi	Kuratointi tarkoittaa tutkimusaineistosta huolehtimista. Tähän kuuluu mm. aineiston järjestäminen siten, että tiedostot on nimetty systemaattisesti ja ymmärrettävästi, tiedostokansiot ovat järkevästi järjestettyjä ja niiden sijainti on varmuuskopioitu ja aineistoon pääsy voidaan tarvittaessa rajata. Kuratointiin kuuluu myös aineiston sisällöllisen laadun tarkastaminen (esim. mahdollisten virheiden tai puuttuvien tietojen asianmukainen käsittely) sekä aineistosta kertovien kuvailutietojen lisääminen aineiston yhteyteen. https://tieteentermipankki.fi/wiki/Avoin_tiede:tietoaineistojen_kuratointi	Curation
Kuvailutiedot (metadata, metatiedot)	Kuvailu- eli metatiedot ovat tietoja tutkimusaineistosta. Niissä esitetään tavallisesti aineiston nimen ja yleiskuvauksen lisäksi avain- ja asiasanat, keräämiseen osallistuneet henkilöt tai organisaatiot, aineiston julkaisuajankohta sekä versiohistoria, tiedot aineiston avoimuudesta ja käyttöoikeudesta. Myös viittaukset toiseen tutkimusaineistoon tai aineistoon liittyvä julkaisu voidaan sisällyttää kuvailutietoihin. https://tieteentermipankki.fi/wiki/Avoin_tiede:metatieto	
Lisenssi	Lisenssillä kuvataan, millaiset oikeudet aineiston hyödyntämiseen annetaan. Vaikka aineisto olisi vapaasti saatavilla, lisenssi voi silti rajoittaa aineiston jatkokäyttöä. https://tieteentermipankki.fi/wiki/Avoin_tiede:lisenssi	License
Metadata-standardi	Metadatastandardin avulla kuvailutietojen rakenne yhtenäistetään esimerkiksi tietokantaa varten ja määritellään, missä muodossa tiedot tulee kerätä ja tallentaa. Erilaisten standardien avulla kuvailutietoja voidaan muotoilla myös eri tieteenaloille paremmin soveltuviksi.	Metadata Standard

Sana	Selitys	Englannin-kielinen termi
Ontologia	Ontologialla tarkoitetaan tutkimusaineistojen kuvailutietojen yhteydessä sitä, että aineistoon liittyvät termit sisältävät tiedon termien yksiselitteisestä merkityksestä sekä yhteyksistä toisiin termeihin (mahdolliset ylä- ja alakäsitteet). Ontologioita hyödyntämällä pyritään siihen, että sanoihin voidaan liittää kontekstia myös tietojärjestelmien ymmärtämään muotoon.	Ontology
Paradata	Paradata on tietoa tutkimusaineiston keräämisestä ja käsittelystä, josta voi olla hyötyä tutkimusaineistoa tulkittaessa.	
Pitkäaikais-säilytys	Pitkäaikaissäilytys tarkoittaa aineiston ylläpitämistä käytettävänä ja löydettävänä vuosikymmenten tai vuosisatojen ajan. Tällöin tutkimusaineisto on kuvattava tarkasti ja aineisto talletettava sellaiseen muotoon, että sitä on mahdollista käyttää myös tulevaisuudessa. https://tieteentermipankki.fi/wiki/Avoin_tiede:pitk%C3%A4aikais-s%C3%A4ilytys	Digital Preservation
Pysyvä tunniste	Pysyvä tunniste on viite, jonka avulla tutkimusaineisto voidaan löytää, vaikka sen sijainti muutuisikin. Pysyvä tunniste on hyvä sisällyttää tutkimusaineiston kuvailutietoihin, jotta kuvailutiedot yhdistyvät yksiselitteisesti aineistoon. Mikäli tutkimusaineiston sisältö muuttuu, myös pysyvä tunniste vaihtuu. Pysyviä tunnisteita on siten ylläpidettävä niin, että ne viittaavat tietyn tutkimusaineistoon tiettyyn versioon. https://tieteentermipankki.fi/wiki/Avoin_tiede:pysyv%C3%A4_tunniste	Persistent Identifier, PID
Readme-tiedosto	Readme-tiedosto on vapaamuotoinen tekstitiedosto, johon voi dokumentoida tutkimusaineiston kuvailutietoja erityisesti silloin, kun käytettäväksi valittu metadatastandardi ei riitä. Readme-tiedostoon voi myös kerätä kuvailutietoja myöhemmin datarepositorioon syötettäväksi. Vaikka tiedosto on vapaamuotoinen, se kannattaa tehdä selkeäksi ja ymmärrettäväksi, jotta siitä olisi hyötyä aineiston ymmärtämisen kannalta.	Readme File
Repositorio	Repositorio on kokonaisuus, joka koostuu tieteellisten aineistojen tallentamiseen ja saatavilla pitämiseen soveltuvasta teknisestä järjestelmästä sekä teknisen järjestelmän ympärille rakennetuista palveluista.	Repository
Skeema	vrt. metadatastandardi	Schema
(Yleis)kuvaus	Yleiskuvaus on datarepositoriossa esitettävä kokotekstimuotoinen ihmisluettava sepite aineiston sisällöstä. Kuvaus on tärkeä aineistojen löydettävyyden ja selailtavuuden kannalta, sillä se mahdollistaa datarepositorioihin tallennettujen aineistojen kartoittamisen nopealla tutustumisella. Kuvaustekstissä kerrotaan rakenteisessa metatiedossa esitettyjä aineistoa kuvaavia keskeisimpiä tietoja mahdollisimman helposti omaksuttavassa muodossa. Kuvaus voidaan mieltää aineiston käyntikortiksi tai hissipuheeksi.	

LÄHTEET

CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023–2025 (V01.00). Zenodo. https://doi.org/10.5281/zenodo.7051012

DC (Dublin Core) https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (katsottu 23.1.2024)

DCAT (Data Catalog Vocabulary): https://www.w3.org/TR/vocab-dcat-3/ (katsottu 23.1.2024)

Digi- ja väestötietovirasto. Tiedon jakamisen toimintamalli. https://www.avoindata.fi/fi/toimintamalli (katsottu 28.3.2024)

Digital Curation Center: http://www.dcc.ac.uk/resources/metadata-standards/list (katsottu 30.1.2024)

FAIRSharing: https://fairsharing.org/ (katsottu 30.1.2024)

Lehtisalo Anneli, Ari Asmi, 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, & Hanna Koivula. (2023). Improve the quality and impact of your research through data management - A guide for making your data FAIR. Zenodo. https://doi.org/10.5281/zenodo.8012377

Metadata Standards Catalog: https://rdamsc.bath.ac.uk/ (katsottu 30.1.2024)

Rosti Tomi & Manna Satama (2023). Tutkimusdata talteen: kuka, miten ja missä? *Tietolinja*, 2023(2). Pysyvä osoite: https://urn.fi/URN:NBN:fi-fe20231218155442

TENK Tutkimuseettinen neuvottelukunta (2023). Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa. Toimittaneet Riitta Keiski ym. Tutkimuseettisen neuvottelukunnan julkaisuja 2/2023. Saatavilla https://tenk.fi/sites/default/files/2023-03/HTK-ohje_2023.pdf (katsottu 24.1.2024).

Kiitokset

Ohjetta laatiessaan työryhmä on tukeutunut oman asiantuntemuksensa lisäksi muutaman ryhmän ulkopuolisen henkilön tietämykseen. Tietoarkiston erityisasiantuntija Tuomas Alaterän kanssa käydyt keskustelut esimerkkikuvailuista ovat olleet hedelmällisiä. CSC:n kehittämispäällikkö Jessica Parland von Esseniltä saadut kommentit auttoivat kirkastamaan tekstiä vielä viimeistelyvaiheessa. Lääkärilehden tieteellinen päätoimittaja, neurologi Pertti Saloheimo auttoi löytämään lääketieteen eri aloilla luotettavaksi koetellun listauksen repositorioista. Kiitokset teille!

LAATIJAT

Nimi	Organisaatio	ORCiD
Nina Hyvönen	HY/Kansalliskirjasto	https://orcid.org/0000-0001-6467-5961
Riitta Koikkalainen	HY/Kansalliskirjasto	https://orcid.org/0000-0003-3289-1832
Tanja Vienonen	Turun ammattikorkeakoulu	https://orcid.org/0009-0000-8965-4577
Virpi Turkulainen	Haaga-Helia AMK	
Manna Satama	UEF	https://orcid.org/0000-0003-3775-9363
Raija Järvimäki	SAMK	https://orcid.org/0009-0009-0800-9767
Katja Ikonen	LUKE	
Anna Suorsa	Oulun yliopisto	https://orcid.org/0000-0003-1472-3675
Erja Kortelainen	CSC	https://orcid.org/0000-0003-4482-6704
Mari Elisa Kuusniemi	HY	https://orcid.org/0000-0002-7675-287X
Liisa Siipilehto	HY	
Kaisa Kylmälä	Tampereen yliopisto	https://orcid.org/0000-0001-8227-6817
Hannele Keckman-Koivuniemi	Tietoarkisto	https://orcid.org/0000-0002-9223-7261
Tiina Sipola	Oulun yliopisto	https://orcid.org/0000-0002-9242-8463
Liisa Näpärä	HY/Kansalliskirjasto	https://orcid.org/0000-0002-6473-5896
Marja-Liisa Seppälä	HY/Kansalliskirjasto	
Anna Salmi	Jyväskylän yliopisto	https://ordic.org/0000-0002-7599-9870
Essi Lempiäinen	Turun yliopisto	https://orcid.org/0000-0002-3675-0108
Seliina Päällysaho	SeAMK	https://orcid.org/0000-0002-3554-7773
Eeva Savolainen	Aalto-yliopisto	https://orcid.org/0000-0002-5435-0300
Niina Tiainen	Ruokavirasto	https://orcid.org/0000-0001-7947-5620
Anne Holappa	Humanistinen ammattikorkeakoulu	https://orcid.org/0000-0003-3044-0173
Jari Parkkinen	Jyväskylän ammattikorkeakoulu	https://orcid.org/0000-0003-1562-9381
Pekka Nygren	Suomen Metsätieteellinen Seura	https://orcid.org/0000-0002-3730-9840
Niina Nurmi	HY	https://orcid.org/0000-0003-2036-3346
Ulriikka Puura	Suomalaisen Kirjallisuuden Seura SKS	https://orcid.org/0000-0002-1846-516X
Päivi Leinonen	Oulun yliopisto	https://orcid.org/0000-0003-0210-5589
Minna Rahnasto	Itä-Suomen yliopisto	https://orcid.org/0000-0002-8882-747X
Udayanto Dwi Atmojo	Aalto-yliopisto	https://orcid.org/0000-0002-6865-0806
Kaisa Hartikainen	Itä-Suomen yliopisto	https://orcid.org/0000-0002-1626-7653