Tiedeinstituuttien avoin tutkimusdata

Digitaalisen aineistonhallinnan käsikirja 1.0

Harri Kiiskinen, Laura Nissin, Manna Satama

Digitaalisen aineistonhallinnan käsikirja 1.0
2020-10-21
Harri Kiiskinen, https://orcid.org/0000-0003-4187-5551
Laura Nissin, https://orcid.org/0000-0001-8050-3916
Manna Satama, https://orcid.org/0000-0003-3775-9363
Tiedeinstituuttien avoin tutkimusdata -hanke
Institutum Romanum Finlandiae sr
Suomen Ateenan-instituutin säätiö sr
Suomen Lähi-idän instituutin säätiö sr
Suomen Japanin instituutin säätiö sr
DOI: 10.5281/zenodo.4113901
Asiasanat (YSA): avoin tieto; metadata; tutkimusaineisto; humanistiset tieteet; kuvailu; projektit; pysyvät tunnisteet; tiedeinstituutit; tiedontuottajat; tieteellinen julkaisutoiminta; tutkimusprojektit; tutkimustyö

Tiivistelmä

Tämä asiakirja on tehty osana opetus- ja kulttuuriministeriön rahoittamaa Tiedeinstituuttien avoin tutkimusdata -hanketta. Hankkeen toteuttajina toimivat Suomen ulkomaisten tiedeinstituuttien taustayhteisöt. Hanketta hallinnoi Institututum Romanum Finlandiae sr.

Asiakirja on hankkeen tutkijoiden laatima ehdotus instituuttien työntekijöille ja tutkijoille suunnatusta ohjeistuksesta. Se koskee digitaalisten tutkimusaineistojen hallintaa, tuottamista, omistajuutta, säilyttämistä ja julkaisua. Asiakirja on tehty hankkeen tutkijoiden selvitysten pohjalta, ja sitä on käsitelty hankkeen ohjausryhmässä, joka on antanut sisältöön liittyviä ohjeita, kommentteja ja palautetta.

Asiakirjaan on kerätty palautetta erityisesti hankkeessa mukana olevilta instituuteilta (henkilöstö, säätiöiden hallitukset) kesällä ja syksyllä 2020. Palautteiden perusteella päivitetty versio laaditaan loppuvuonna 2020 ja julkaistaan versiona 1.1.

Asiakirja on tarkoitettu tiedeinstituuttien käyttöön näiden sopivaksi katsomallaan tavalla.

Sisällys

1. Johdanto

Tämän asiakirjan tarkoitus on ohjeistaa instituutissa työskentelevien tutkijoiden tutkimusaineiston hallinnan käytäntöä mahdollisimman konkreettisesti. Tutkija saa tuotettua:

- aineistonhallintasuunnitelman
- aineiston hallinnan perustan tutkimuksen aikana
- riittävän avoimen aineiston (tavoite: niin avoin kuin mahdollista, niin suljettu kuin pakollista)
- aineiston sellaiseen kuntoon, että sen voi toimittaa julkaisu- tai data-arkistoon (repositorioon) mahdollisimman vaivattomasti

Asiakirjassa esitellään avoimen tieteen pääperiaatteet ja perustellaan mitä hyötyä avoimen tieteen linjauksista on tutkijalle. Tämän lisäksi asiakirja antaa konkreettisia ohjeistuksia ja ohjaa lisätiedon äärelle.

Avoimen tieteen ja tutkimuksen hyöty tutkijalle

- **meritoituminen** – toimintakulttuuri muuttuu, vaikka hitaasti; pysyvät tunnisteet (ei vain julkaisuille) → viittausmahdollisuus → **tutkimuksen ja tutkijan näkyväksi tekeminen**
- **rahoittajan vaatimus** (rahoittaja, taustaorganisaatio)
- **tutkimuksen aikainen aineiston hallinta jäntevöityy**: aineiston löydettävyys (kuvailu), suunnitelmallisuus, tutkimusryhmien yhteiset toimintaohjeet, vastuut ja oikeudet.

Tutkijalla on oikeus ja vastuu suunnitella tutkimusaineistonsa hallinta siten, että tieteen eettiset periaatteet toteutuvat. Eettisistä kysymyksistä löytyy lisätietoja Tutkimuseettisen neuvottelukunnan sivuilta (https://tenk.fi/).

Mihin aineistonhallintasuunnitelmaa tarvitaan?

Hyvä aineistonhallinta edistää tutkimushanketta sen aikana ja sen jälkeen aineistojen ja tulosten jatkokäyttönä.

Asiakirjatasoinen, jäsennelty suunnitelma helpottaa tutkimusryhmän työskentelyä (yhteiset pelisäännöt), auttaa palaamaan tutkimuksen pariin tauon jälkeen, ohjaa ratkaisemaan aineiston turvaamisen tutkimuksen aikana ja sen jälkeen - mieti, mitä tapahtuu, jos tietokoneesi hajoaa.

Suunnitelmallisuus ohjaa myös ratkaisemaan aineiston avaamisen mahdollisuuden. Tarkoitus ei ole avata kaikkea aineistoa, mutta aineiston julkaisemisella on hyötyä itsellesi (meritoituminen) ja muille (aineiston jatkokäyttö). Tutkijan ei kannata tehdä ennakkopäätöstä, keitä ja mitä tutkimusalaa oma tutkimus mahdollisesti kiinnostaa tai hyödyttää. ”Tämän päivän roska voi olla huomispäivän aarre” pätee tässäkin yhteydessä. Aineiston avaamisen kautta sekin osa aineistosta, joka jää tutkimushankkeen julkaisujen ulkopuolelle, tulee näkyväksi. Rahoittajilla on myös suoranaisia vaatimuksia aineistojen avaamiseen, Suomessa erityisesti julkisin varoin tuotetun tutkimuksen kohdalla.

Tämän käsikirjan ohella

- Tutustu oman yliopistosi aineistonhallintaohjeisiin.
- Tietoarkiston Aineistonhallinnan käsikirja, keskittyy erityisesti yhteiskuntatieteen aineistoihin (https://www.fsd.tuni.fi/aineistonhallinta/fi/).
- Seuraa Avoimen tieteen koordinaation viestintää tutkijoille (https://avointiede.fi/fi/tutkijalle).
- Kansainvälisiin avoimen tieteen verkostoihin voit tutustua esim. seuraavista: DARIAH-EU (https://www.dariah.eu/), Research Data Alliance (https://www.rd-alliance.org/) tai European Science Cloud (https://www.eosc-portal.eu/).

2. Perusasiaa aineiston hallintaan ja avoimeen tutkimukseen

7 perusaskelta aineiston hallintaan FAIR-periaatteet huomioiden

1. Suunnittele aineiston hallinta, palaa siihen ja päivitä suunnitelma.
2. Valitse data-arkisto, johon suunnittelet tallentavasi aineiston, tutustu sen ohjeisiin ja päivitä aineistonhallintasuunnitelma.
3. Pidä aineisto aineistonhallintasuunnitelmassa kuvatussa järjestyksessä; suosi yleisiä tiedostomuotoja.
4. Dokumentoi: README-tiedosto, joka kulkee aineiston mukana.
5. Kuvaile aineisto: metatiedot.
6. Säilytä aineisto tutkimuksen aikana.
7. Päätä valmiin tutkimusaineiston kohtalo: säilytys, hävittäminen, jakaminen.

Keskeiset käsitteet: aineisto/data ja metadata/metatieto

Aineistoa ovat esim.

- mittaukset
- paikkatiedot
- kuvat
- äänitteet
- tekstit
- luettelot
- muistiinpanot
- sähköpostit
- tutkijan luoma uusi aineisto toisen aineiston pohjalta

Aineisto voi olla analogisessa tai digitaalisessa muodossa. Usein digitaalista aineistoa kutsutaan dataksi.

Metatieto kuvailee datan sen koko elinkaaren ajan ja huomioi tiedon

- kontekstin
- sisällön
- rakenteen
- hallinnan
- käsittelyn

Tällaisia tietoja ovat datan nimi, tuottamisajankohta, tuottaja, muoto, aihe, käyttöoikeus.[1]

Metatiedolle suositellaan standardeja, joita on lukuisia eri tieteenaloja ja käyttötarkoituksia varten. Standardien käyttö mahdollistaa löydettävyyden, saavutettavuuden, yhteentoimivuuden ja uudelleenkäytön eli FAIR-periaatteiden toteutumisen.

Metatieto voidaan avata, vaikka itse aineistoa ei avattaisi.

[1]Ks. https://utuguides.fi/tutkimusdata/avaaminen#s-lg-box-14879918.

3. Miten toimitaan, kun tullaan instituuttiin tekemään tutkimusta

Onko jo olemassa aineistonhallintasuunnitelma?

- Kyllä → Käy se läpi suhteessa instituuttien malliasiakirjan kanssa ja vastaa mahdollisesti puuttuviin kohtiin. -> Toimita instituuttiin instituutin nimeämälle vastuuhenkilölle.
- Ei → Tee instituuttien malliasiakirjan pohjalta suunnitelma. → Toimita instituuttiin.

1. Aineistonhallintasuunnitelma

Ks. käsikirjan liite A Instituuttien aineistonhallintasuunnitelman malli.

Työkaluja ja ohjeita aineistonhallintasuunnitelmien tekemiseen:

- DMPTuuli (https://dmptuuli.fi/)
- easyDMP (https://easydmp.eudat.eu/)

2. Data-arkiston valinta

Data-arkiston alustava valinta on hyvä tehdä jo varhain. Suomessa esim. Tietoarkisto ottaa vastaan humanistiseen tutkimukseen liittyviä aineistoja, mutta aineisto pitää olla järjestetty ja muotoiltu arkiston ohjeiden mukaan. EU-tason peruspalvelu on Zenodo.org, joka ottaa vastaan kaikkea tutkimusaineistoa ilmaiseksi eikä aseta juurikaan vaatimuksia aineiston järjestämiselle.

Alakohtaisia data-arkistoja voi etsiä mm. Registry of Reserach Data Repositories -työkalun avulla osoitteessa https://www.re3data.org/search. Data-arkistojen sivuilta voi löytyä hyviä ohjeita aineiston järjestämiseen ja käsittelyyn. Monet arkistoista ovat kansallisia ja/tai maksullisia.

Tarjolla olevien palveluiden kenttä on nopeasti muuttuva, joten pysyviä suosituksia on vaikea antaa. Esimerkiksi CSC kehittää uusia kansallisia ratkaisuja, jotka saattavat olla käyttökelpoisia ja suositeltavia joidenkin vuosien kuluessa.

3. Järjestele

Kiinnitä jo heti alkuun huomiota aineiston keskinäiseen järjestykseen ja perustele se (ks. kohta 4).

Suosi avoimen lähdekoodin tiedostomuotoja viimeistään siinä vaiheessa, kun aineisto siirretään data-arkistoon. Näin varmistat aineistosi käytettävyyden. Käsikirjan lopussa on taulukko, jossa on kerrottu eri vaihtoehtoja.

4. Dokumentoi → README malli.txt (käsikirjan liitteenä)

README-tiedosto turvaa tutkimusaineiston ymmärrettävyyden. Se kertoo keskeiset seikat. Ks. B. Liite: Malli README-tiedostolle. Tiedosto on txt-muotoinen (ei siis esim. word-dokumentti).

Varmista, että README-tiedosto kulkee aineiston mukana.

5. Kuvaile: metatiedot

Tärkeintä on noudattaa tutkimushankkeen kannalta toimivinta kuvailun tapaa. Tarkoitus ei ole pakottaa tutkimusta valmiisiin standardeihin, jos ne eivät ole tarkoituksenmukaisia. Kuvailun selittäminen on kuitenkin ehdottoman tärkeää, käytetään standardia tai ei. Tutkimushankkeessa kaikkien on ymmärrettävä, miten aineistoa kuvaillaan (termien sisällöstä alkaen). Jos käytössä on tietty standardi, muista mainita README-tiedostossa, mikä standardi on käytössä tai jos standardia ei käytetä, missä kuvailutietojen logiikka selitetään.

Standardit määrittelevät kolmea eri metatiedon tasoa: käytettävissä olevia tietokenttiä ("*schemes*"), kentissä käytettäviä sanastoja ("*taxonomies and vocabularies*") ja eri tietokenttien ja aineistoyksiköiden välisiä suhteita ("*ontologies*").

Esimerkkinä Dublin Core

Matalan kynnyksen skeema on Dublin Core (DC), jota käytetään laajasti ja jota on sillattu eli mäpätty muihin alakohtaisiin standardeihin ja malleihin. DC:n plussana on sen monikäyttöisyys: Se taipuu eri tieteenaloihin ja aineistoihin, sillä sen ns. peruskentät ovat valtavan yleisluonteisia ja helposti ymmärrettäviä ja niistä voi valita tarpeelliset kentät. Aineiston kuvailun voi aloittaa jo tutkimuksen alussa ja tarkentaa tutkimuksen aikana. DC:n kentät ohjaavat myös vastaamaan moniin aineistonhallintasuunnitelmissa esitettyihin kysymyksiin.

DC-standardia voikin pitää eräänlaisena kattotason jäsentelyn apuna tai aineiston julkaisemisen yhteydessä aineiston kokonaisuuden kuvailun työkaluna. Tätä varten avuksi on luotu käsikirjan liitteeksi taulukko "Kuvailu_DC_instituutti" (Liite C). Siinä on kerrottu Dublin Coren kentät (eng Elements), niiden tarkenteet englanniksi ja suomeksi sekä viittaukset käytettäviin sanasto-ym. standardeihin ja suosituksiin. Käytä siis standardoituja tapoja ja sanastoja ilmaisemaan kentissä kysyttyjä asioita (esim. tekijät, aikamääre, jne.).

Ks. myös Dublin Core -generaattori: https://nsteffel.github.io/dublin_core_generator/index.html

Data Documentation Initiative (DDI)

DDI-metatietostandardia käytetään erityisesti yhteiskunta- ja käyttäytymistieteissä. Se soveltuu hyvin kyselytutkimuksiin. Standardista on kaksi versioita, joista DDI-Lifecycle on tarkoitettu nimensä mukaisesti aineiston koko elinkaarelle ja DDI-Codebook jossain määrin kevyempään käyttöön. Tietoarkisto käyttää DDI Codebook -versiota. DDI:n versiot ovat rakenteista, koneluettavaa ja xml-pohjaista soveltuen siten pitkäaikaissäilytykseen.

DDI-standardissa on viisi pääkenttää ja lukuisa määrä näiden alla olevia lisäkenttiä. Se on sillattu moniin muihin standardeihin, ja esimerkiksi valtaosa Dublin Coren kentistä saa vastineen DDI:n kentistä (https://ddialliance.org/resources/ddi-profiles/dc). Kenttien ja niiden alakategorioiden kuvakset löytyvät DDI: n kotisivulta: https://ddialliance.org/Specification/DDI-Codebook/2.1/DTD/Documentation/version2-1-all.html.

Esimerkkinä Gettyn sanastot

Getty Research Institute ylläpitää taiteeseen ja kulttuuriperintöön liittyviä monikielisiä sanastoja, joiden kattamat terminologiat soveltuvat hyvin instituuttien piirissä tehtävän tutkimuksen aineistojen kuvailuun (https://www.getty.edu/research/tools/vocabularies/).

Esimerkkinä Finto eli suomalaiset asiasanastot

”Finto on suomalainen sanasto- ja ontologiapalvelu, joka mahdollistaa sanastojen julkaisun ja selailun. Palvelu tarjoaa myös rajapinnat sanastojen ja ontologioiden hyödyntämiseen muissa ohjelmistoissa” (http://finto.fi/fi/).

Esimerkkinä CIDOC-CRM

CIDOC-CRM on erityisesti kulttuuriperintöaineiston kuvailuun luotu ontologia. Sen avulla kuvataan tallennettavan tiedon sisäisiä suhteita, ja laajennuksineen ontologia soveltuu esim. arkeologisen kaivausprojektin kaikkien vaiheiden kuvailuun.

CIDOC-CRM:n tarkoitus on mahdollistaa konkreettisen tai abstraktin kohteen koko elinkaaren kuvailu. CIDOC-CRM:n soveltaminen aineistojen kuvailuun ei ole aivan yksinkertaista, mutta potentiaalisesti se hyvin hyödyllistä aineiston pitkäaikaiskäytön kannalta (http://www.cidoc-crm.org/).

Esimerkkinä Wikidata

Wikidata on avoin tietoarkisto, jonka tarkoitus on toimia sekä ihmisen että koneen luettavissa ja muokattavissa olevan tiedon avoimena tallennusalustana. Wikidata ei suoraan sovellut tutkimusaineiston kuvailuun, mutta siitä voi olla hyötyä tutkimusten tulosten julkaisussa tai tutkimusdatan kokonaisuuden kontekstualisoinnissa (https://www.wikidata.org/).

Kuvailun työkaluja

Kenttätöihin soveltuvat työkalut tai dokumentointialustat tuottavat kuvailutietoa, kuten esim. alempana kohdassa 6 mainitut Arches ja iDAI.field,

Käytännössä kuvailuun on kolme vaihtoehtoa:

1. Metadata sijoitetaan tiedoston sisään, esim. toimisto-ohjelmatiedostoissa tai kuvatiedostoissa. Tällöin metadata pysyy tiedon mukana mutta voi olla vaikeasti löydettävissä
2. Metadata sijoitetaan ulkoiseen tiedostoon, joko datatiedostokohtaisesti tai esim. koko kansiota koskevaan CSV[2]-tiedostoon. Tällöin metadata ei välttämättä pysy itse datan mukana, mutta sen saavutettavuus ja löydettävyys on hyvä.
3. Aineiston hallintaan käytetään järjestelmää, joka linkittää datan ja metadatan automaattisesti toisiinsa ja pitää huolen sekä datan eheydestä että metadatan löydettävyydestä. Tällaisissa ratkaisuissa datan käytettävyys osana tutkijan päivittäistä työntekoa on hankalaa koska aineisto ei ole suoraan omalla koneella hyödynnettävissä.

Mikään näistä ei ole täysin tyydyttävä. Mikä tahansa näistä ratkaisuista on parempi kuin ei mikään näistä ratkaisuista.

[2] *Comma Separated Values*, ks. https://fi.wikipedia.org/wiki/CSV

6. Säilytä tutkimuksen aikana

Muista versionhallinta ja varmuuskopiointi ja tee niistä rutiini tutkimushankkeessasi. Jos kyseessä on useamman henkilön työllistävä hanke, sopikaa rutiinit ja kirjatkaa ylös aineistonhallintasuunnitelmaan, jotta ne ovat kaikkien tutkimukseen osallistuvien tarkistettavissa ja käytettävissä.

Instituutilla voi olla omat pilvipalveluratkaisunsa, kuten tämän hankkeen aikana käytössä ollut TeamDrive, joidenkin suomalaisten yliopistojen käyttämä Seafile, tai esim. ownCloud. Osa näistä on asennettavissa myös instituutin omille palvelimille.

EUDAT-palveluun kuuluva B2DROP on yksi mahdollinen vaihtoehto, jonka käyttö on yksittäisillekin tutkijoille tai tutkimusryhmille mahdollinen ja EU-kansalaisille maksuton. Sen käyttö ei siis edellytä tutkijan kuulumista johonkin palvelussa olevaan tutkimusorganisaation (https://www.eudat.eu/services/b2drop). Tähän palveluun on mahdollista liittää myös B2SHARE, jossa tutkimusaineistoja voi julkaista ja jossa niille voi saada mm. DOI-tyyppisen pysyvän tunnisteen.

Toimiva vaihtoehto saattaa myös olla jonkinlainen aineistonhallintajärjestelmä, jonka avulla on mahdollista hallinnoida suuriakin määriä digitaalisia objekteja sekä mahdollisesti myös kuvata itse tutkimuskohteita. Tällaisia järjestelmiä ovat mm. Arches, Islandora, eXist-db ja iDAI.field

Aineistonhallintajärjestelmiä

iDAI.field on erityisesti arkeologiseen kenttätyöhön suunniteltu ohjelmisto, joka asennetaan pilveen ja jota käytetään mobiililaitteilla suoraan kentällä. iDAI.field on osa Saksan arkeologisen instituutin (DAI) aineistonhallinnan työkaluja (https://idai.world/).

Arches on erityisesti muistitieto-organisaatioiden käyttöön kehitetty järjestelmä, jossa on vahva tuki CIDOC-CRM-ontologialle ja sen tieteenalakohtaisille laajennuksille. CIDOC-CRM mahdollistaa niin esineiden, niitä kuvaavan dokumentaation kuin itse tutkimusprosessinkin semanttisen mallinnuksen sellaisella tasolla, joka soveltuu myös monimutkaisten data-aineistojen tallentamiseen. The Getty Conservation Institute ja World Monuments Fund ovat Arches-järjestelmän kehittäjäorganisaatiot (https://www.archesproject.org/).

Islandora on erityisesti digitaalisen aineiston arkistointiin soveltuva järjestelmä, joka on parhaimmillaan suurten aineistojen hallinnassa ja julkaisussa (https://islandora.ca/).

eXist-db on xml-pohjainen tietokanta ja julkaisujärjestelmä, joka soveltuu erityisen hyvin TEI-pohjaisen (esim. tekstin rakennetta kuvaavan) tekstiaineiston julkaisemiseen.

7. Valmis tutkimusaineisto: data-arkistot, lisenssit, käyttöluvat, pysyvät tunnisteet

Repositoriot

Tässä vaiheessa (toukokuu 2020) suositellaan Zenodo-palvelua, joka on luotettava, laajasti eri tieteenaloilla käytetty, tuottaa pysyvän doi-tunnisteen ja hyväksyy eri tyyppistä aineistoa postereista datasetteihin. Kirjautuminen Zenodo.org -palveluun onnistuu ORCID-tunnuksella. Ks. ohjeet esim. https://instruct-eric.eu/help/other/zenodo-upload-guidelines.

Yksilöimätön lisenssi

Aineistoa julkaistaessa on sille valittava lisenssi – ilman näkyvää lisenssiä aineisto ei ole periaatteessa kenenkään muun käytettävissä ilman erillisiä käyttösopimuksia. Tutkimusaineistojen julkaisussa suositellaan usein käytettävän Creative Commons eli CC-lisenssejä. CC-lisenssien käytöllä on myös laaja kansallinen ja EU:n laajuinen tuki. Näiden CC-lisenssitermien lyhyet selitykset on koottu tämän käsikirjan loppuun, ja lisää tietoa CC-lisensseistä löytyy sivulta https://creativecommons.fi/.

Aineistojen julkaisussa on muistettava, että lisenssin voi määritellä vain itse omistamalleen aineistolle. Jos hankkeen käyttöön on saatu kuvia esim. paikallisilta museoviranomaisilta, ei niitä voi omavaltaisesti jakaa eteenpäin. Tällaisen aineiston kohdalla on siis muistettava, että ne voivat olla julkaistavan tutkimusaineiston osana vain silloin, kun tähän on aineiston omistajan lupa (joko alun perin tai julkaisua varten erikseen saatu lupa tai julkaisemisen mahdollistava lisensointi).

Tietyillä erityisaloilla on omia lisenssikäytänteitään. Esim. mikäli hankkeessa syntyy ohjelmakoodia, on sen julkaisemiseen suositeltavampaa käyttää ohjelmistoille tarkoitettuja lisenssejä. Näistä löytyy lisätietoja mm. https://www.gnu.org/licenses/license-list.html.

Yksilöity käyttölupa

Käyttölupa ei ole sama asia kuin yllä mainitut lisenssit. Jos hankkeelle tai sen tutkijalle on annettu lupa käyttää ja julkaista kuvaa haluamallaan tavalla, on kyseessä osapuolten väliseen sopimukseen perustuva käyttölupa; CC-lisenssit taas määrittelevät, miten kuka tahansa asiaan voi suhtautua.

Aineistojen käyttöoikeuksia koskeva tieto on pidettävä tallessa ja aineistojen yhteydessä koko hankkeen ajan siten, että myös aineistojen julkaisuvaiheessa on mahdollista erottaa aineisto, jota ei voi julkaista.

Tunnisteet

Pysyvä tunniste eli *PID* (Persistent IDentifier) on yksiselitteinen ja ainutkertainen julkinen nimi, joka annetaan tutkimusaineistolle tai sen osalle. Tunnisteet tulevat julkaisemisen yhteydessä, eikä niistä huolehtiminen ole yleensä tutkijan omalla vastuulla.[3] Erilaisia PID-tyyppejä:

- **URI**: Uniform Resource Identifier, joka ilmaisee
  - tiedon paikan (URL eli Uniform Resource Locator, joka esim. kertoo www-sivun http-protokollan avulla)
  - nimen (URN eli Uniform Resource Name, kuten julkaisuissa käytetty ISBN, joka itsessään ei sisällä ko. resurssin sijaintia)

[3]Tunnnisteista esim. Hakala J. 2018. "URN:NBN ja muut toiminnalliset tunnistejärjestelmät". Tietolinja 2. http://urn.fi/URN:NBN:fi-fe2018093036991. Avoimen tieteen koordinaatiossa tuotettu kansallinen suositus pysyvien tunnisteiden käytöstä: https://doi.org/10.5281/zenodo.3560738.

- **DOI**: Digital Object Identifier
- Handle

Kuvailutiedoissa voit aineiston sisällä käyttää omaa tunnistesysteemiä, joka pitää aineiston järjestyksessä tutkimuksen aikana. Muista dokumentointi, jotta systeemi on ymmärrettävä muillekin.

Muista tekijän tunniste!

On tärkeää huolehtia myös omasta näkyvyydestään aineiston julkaisun yhteydessä. Vahvassa nousussa näyttäisi olevan nk. ORCID-tunniste, jonka tutkija voi luoda itselleen affiliaatioistaan riippumatta. Lisätietoja tunnisteesta löytyy sivulta https://tutkijatunniste.fi/

- **ORCID**-tunniste tutkijoille (nopea ja helppo rekisteröinti): https://orcid.org/
- **ISNI**-tunniste muut (luovan alan tekijät) kuin tutkijat: http://www.isni.org/

4. Käsikirjan ymmärtämisen avuksi

Tiedostomuodoista

Tiedostomuodot (erit. pysyvä säilytys) – suosi avoimen lähdekoodin ratkaisuja, vältä omisteisia ohjelmistoja.

Ks. https://www.fsd.tuni.fi/aineistonhallinta/fi/tiedostoformaatit-ja-ohjelmistot.html

Tyyppi	Suositus	Vältä
Teksti	plaintext/markdown, xml/html, odt, pdf/a	word
Taulukko	CSV, TSV, SPSS portable	excel
Media	Container: mp4, ogg Codec: Theora, Dirac, FLAC	Quicktime, HZ64
Kuva	DNG, TIFF, JPG 2000, PNG	GIF, JPG
Rakenteinen data (structured)	xml, rdf, json, yaml	RDBMS, Access

FAIR (ks. Fairdata.fi -palvelu) – Viittaa dataan, sekä omaan että muiden

Aineiston kuvailutiedot (ks. myös metatiedot) ovat keskeiset. Ne kertovat tutkijan käsityksen siitä, mistä on kyse ja miten tutkimustulokset perustellaan. Hyvät kuvailutiedot ovat avainasemassa FAIR-periaatteiden toteutumiseen: aineisto on löydettävissä, siihen päästään käsiksi, se toimii yhdessä eri järjestelmien ja hakukoneiden kanssa sekä on käytettävissä uudelleen.

Fairdata.fi -palvelu tarjoaa päivitettyä ja luotettavaa tietoa datan hallinnasta. Palvelun järjestää opetus- ja kulttuuriministeriö ja toteuttaa CSC – Tieteen tietotekniikan keskus Oy (https://www.fairdata.fi/).

Fairdatan perusominaisuudet

F – Findable: Aineisto on löydettävissä.

- pysyvä ja yksilöivä tunniste (PID eli persistent identifier)
- kattavat metatiedot (kuvaileva, hallinnollinen, rakenteellinen)
- kuvailu ja rekisteröinti hakupalveluun

A – Accessible: Aineisto on saavutettavissa. Aineisto tai sen metatiedot ovat noudettavissa standardisoidun yhteyskäytännön kautta

- aineisto on niin avoin kuin mahdollista ja niin suljettu kuin tarpeellinen
- yhteyskäytäntö on avoin, maksuton ja yleisesti käytettävissä

I – Interoperable: Aineisto on yhteen toimivaa.

- sekä ihmis- että koneluettavaa ja sisällöt siirreltävissä järjestelmien välillä
- avoimet, koneluettavat sanastot, ontologiat ja koodistot

R – Reusable: Aineisto on uudelleen käytettävissä.

- kattavat metatiedot
- käyttölisenssit ilmaistu selkeästi
- aineiston syntyprosessi

Yhteenveto Creative Commons -lisensseistä

Apuna lisenssin valinnassa voi käyttää mm. lisenssiavainta: https://creativecommons.org/choose/?lang=fi. Lisenssiavaimen avulla voi myös selvittää lisenssitekstien muodot ja niissä käytetyn terminologian useilla eri kielillä.

CC-termejä voi myös yhdistellä: esim. CC BY-NC tai CC BY-NC-ND ovat tavallisia.

CC0

"Ei käyttörajoituksia"

Tekijä ei aseta käyttörajoituksia aineistolle. Tämä soveltuu esim. tutkimushankkeita kuvaavan ylätason metadatan jakeluun, jolloin näiden kuvailutietojen levittäminen ja hyödyntäminen on mahdollisimman helppoa. (Tämä ei tarkoita sitä, että tutkimus itsessään olisi CC0-lisensoitu.)

CC BY (Nimeä / ByAttribution)

"Saa käyttää, kunhan nimi mainitaan."

Käytössä esim. metadatan lisenssivaihtoehtona useissa palveluissa. Esim. Helsingin yliopisto kannustaa jakamaan myös tutkimusaineistot tällä lisenssillä tieteen avoimuuden edistämiseksi. Mikäli tutkija haluaa, että muut tutkijat voivat hyödyntää hänen keräämiään aineistoja omissa tutkimuksissaan, on tämä käytännössä ainoa vaihtoehto; mutta tutkimusjulkaisujen kohdalla tämä ei välttämättä ole paras vaihtoehto.

CC NC (Ei kaupallinen / NonCommercial)

"Lupa annetaan ei-kaupalliseen käyttöön."

Annat muiden kopioida, välittää, levittää ja esittää sinun tekijänoikeuksiisi kuuluvaa teosta sekä sen pohjalta tehtyjä muokattuja versioita teoksestasi vain epäkaupallisessa käytössä. Mikä on kaupallista ja mikä epäkaupallista on välillä epäselvää.

CC ND (Ei muutoksia / NoDerivatives)

”Aineistoa saa käyttää ja levittää ilman muokkausta.”

Annat muiden kopioida, välittää, levittää ja esittää sinun tekijänoikeuksiisi kuuluvaa alkuperäistä teosta, mutta et salli muokattujen versioiden tekemistä teoksesta.

CC SA (Jaa samoin / ShareAlike)

”Muokatut teokset jaettava samalla lisenssillä.”

Annat muiden julkistaa omasta teoksestasi muokattuja teoksia vain samalla lisenssillä, jolla oma teoksesi on julkaistu.

A. LIITE: Instituuttien aineistonhallintasuunnitelman malli

Tähdellä * merkityt kohdat vastaavat pitkälti tutkimusaineiston README-tiedoston tietoja.

Tutkimusprojektin yleiset tiedot

- *Projektin nimi:
- *Tekijä eli vastuullinen tutkija:
- *Edellisen ORCID-tunniste:
- *Organisaatio(t) (instituutti + muut mahdolliset organisaatiot, jotka mukana projektissa):
- Rahoittaja(t) (+ rahoituspäätöksen numero, jos jo tiedossa):
- Projektin tiivis kuvaus (abstrakti):
- Tekijän yhteystiedot (sähköposti, puhelin):

Tutkimusaineiston yleiset tiedot

- Aineiston alkuperä (onko kyseessä tutkimusta varten kerättävä aineisto ja/tai tutkimuksessa syntyvä aineisto ja/tai uudelleen käytettävä aineisto):
- Aineistotyyppi (esim. analoginen, digitaalinen/numeerinen data, taulukko, kuva, teksti, mittaus, paikkatieto, muistiinpanot):
- Tiedostomuodot ja käytettävät ohjelmat:
- Arvio aineiston koosta:
- *Alueellinen kattavuus (voi olla tekstimuotoinen kuvaus tai standardi):
- *Ajallinen kattavuus (voi olla tekstimuotoinen kuvaus tai standardi):
- Jatkokäytön tarpeen arviointi eli kenelle tai millaiseen tutkimukseen aineistosta voi olla hyötyä tämän tutkimusprojektin jälkeen:
- Täydentyykö aineisto tämän tutkimusprojektin jälkeen:

Eettiset ja oikeudelliset kysymykset – eli mahdolliset rajoitukset, jotka liittyvät aineiston jatkokäyttöön

A) Henkilötiedot

Sisältääkö aineisto henkilötietoja?

Ei

Kyllä, nimi tai yhteystiedot.

Kyllä, henkilötunnus, syntymäaika tai ikä.

Kyllä, sukupuoli.

Kyllä, ammatti, koulutus, työpaikka, opiskelupaikka tai koulu.

Kyllä, kotitalouden koostumus.

Kyllä, siviilisääty.

Kyllä, ajoneuvon rekisterinumero.

Kyllä, ainutkertaiset elämäntapahtumat, arkaluonteiset henkilötiedot (terveydentila, uskonnollinen tai poliittinen vakaumus, seksuaalinen suuntautuminen).

Jos vastasit mihinkään yllä olevista kysymyksistä ”kyllä”, aineistosta pitää laatia tieteellisen tutkimuksen tietosuoja-/rekisteriseloste: Onko sellainen tehty? (Kyllä/Ei -> Laadi tietosuojaseloste oman tutkimusalasi käytäntöjen mukaan).

Jos aineistossa on henkilötietoja, onko siitä saatavissa

- anonymisoitu aineisto (tunnisteeton aineisto eli aineistoa ei voida ”palauttaa” henkilötasolle)
- pseudonymisoitu aineisto (aineistosta on olemassa koodiavain, jonka avulla henkilöt voidaan tunnistaa -> henkilötietosuojan piirissä oleva aineisto)
- ei, aineistoa ei ole anonymisoitu eikä pseudonymisoitu (-> henkilötietosuojan piirissä oleva aineisto)

B) Sopimukset ja lisenssit (Tarkista, että aineiston README-tiedostossa on samat tiedot)

- Käytätkö tutkimuksessa aineistoa, jonka käyttöä ja/tai myöhempää julkaisemista rajoittaa tekijänoikeus, lisenssit tai muut syyt? Jos vastaat kyllä, kerro millaisista rajoituksista on kysymys.
- Millä tavoin aineiston tekijänoikeudet, julkaisuoikeudet ja omistajuus sovitaan (esim. kirjallinen sopimus, suullinen sopimus, kenen kanssa sopimus tehdään, missä kirjalliset sopimukset säilytetään)?
- Millä lisenssillä arvioit siirtäväsi aineiston data-arkistoon? (Ks. Creative Commons -lisenssiapu: https://creativecommons.org/choose/?lang=fi):

Tutkimusaineiston dokumentaatio ja kuvailu (metadata/metatieto)

- Millä tavoin varmistat, että aineisto on ymmärrettävää? (Kirjoita esim. näin "README-tiedostossa annetaan aineiston perustiedot, joiden perusteella aineisto on ymmärrettävää.")
- Tässä yhteydessä voi mainita myös aineiston ymmärrettävyyttä varmistavien muistiinpanojen tallennus ja mahdollinen julkaisu.
- *Kuvailussa käytetty metadatastandardi (esim. Dublin Core):
- Jos kuvailussa ei käytetä metadatastandardia, miten aineisto kuvaillaan:
- *Käytetyt sanastot (esim. Finto https://finto.fi/fi/ + ne sanastot, joita käytät, esim. MAO/TAO tai PTO – Paikkatieto-ontologia jne; Getty Vocabulary, Art & Architecture Thesaurus https://www.getty.edu/research/tools/vocabularies/aat/):
- Jos kuvailussa ei käytetä valmiita sanastoja, miten käytettyjen termien sisällön ymmärrettävyys varmistetaan:

Tutkimusaineiston säilytys tutkimuksen aikana

- Missä ja miten aineisto säilytetään (digitaalinen aineisto: esim. oman tietokoneen kiintolevy, instituutin tietokone, verkkolevy, muistitikku, ulkoinen kiintolevy, instituutin pilvipalvelu, julkinen pilvipalvelu (Google Drive, Dropbox jne), data-arkisto, jne.; analoginen aineisto: esim. instituutin tilat, instituutin arkisto, tutkijan koti jne.):
- Kenellä on pääsy aineistoon ja miten pääsyä hallinnoidaan (koskee erit. projekteja ja arkaluonteista tietoa sisältävää tutkimusta):
- Miten digitaalisen aineiston luettavuus turvataan, kun käytetyt ohjelmistot vanhenevat:

Tutkimusaineiston säilytys ja julkaiseminen tutkimuksen päätyttyä

- Minkä osan aineistosta voi julkaista ja/tai avata?
- Milloin aineisto ja/tai sen kuvailutiedot avataan (eli sis. mahdollisen embargon)?
- Missä aineisto ja/tai sen kuvailutiedot avataan (esim. sama kuin pitkäaikaissäilytyksen data-arkisto)?
- Arvio aineiston pitkäaikaissäilytyksen kestosta (x vuotta)?
- Esim. Zenodo, EUDAT (tutkimuksen aikainen säilytys)

Tutkimusaineiston hallinnan vastuut ja resursointi

- Kuka vastaa tutkimusaineiston hallinnasta tutkimushankkeen ajan (esim. erityiset tehtävät, jotka eivät kuulu tutkimuksen arkipäivän rutiineihin)?
- Arvio ajallisen, tiedollisen/taidollisen ja rahallisen resursoinnin tarpeesta:
- Onko tämä huomioitu tutkimushankkeen suunnitelmassa ja rahoitushauissa/rahoituksessa

B. LIITE: README-tiedosto

Tämän [AINEISTON NIMI]readme.txt tiedoston on luonut [VVVVKKPV] [Sukunimi, Etunimi]

YLEISET TIEDOT – Aineiston:
---
*Nimeke: [esim. tutkimusaineisto artikkeliin/projektiin/julkaisuun "x"]
*Tekijä: [Sukunimi, Etunimi]
*Tekijän ORCID-tunniste:
Muu(t) tekijä(t): [Sukunimi, Etunimi; Sukunimi, Etunimi]
Muiden tekijöiden ORCID-tunnisteet:
Organisaatio(t): [esim. Suomen xxx-instituutti]
Alueellinen kattavuus:[voi olla tekstimuotoinen kuvaus tai standardi]
Ajallinen kattavuus: [voi olla tekstimuotoinen kuvaus tai standardi]

JAKAMINEN/SAAVUTETTAVUUS/PÄÄSY

Lisenssit (creative commons):
Rajoitukset (erit. sisältääkö arkaluonteista aineistoa):
Aineistoon viittamisen suositus (nimi, jolla halutaan aineistoon viitattavan):
Linkki tähän aineistoon:
Linkki muuhun sijaintiin, jossa on avoin pääsy tähän ainestoon (kokonaan tai osaan):

AINEISTO & TIEDOSTOT: YLEISTÄ

Tiedostolista (tiedostojen nimet, rakenne, lyhyt kuvaus tiedostoista):
Tiedostojen suhde toisiinsa, jos se on olennaista:
Lista muista lähteistä, jos aineisto on peräisin niistä:

METODIT

Kuvaus metodeista, joiden avulla aineisto on koottu/kerätty linkkeineen, jos sellaisia on:
Kuvaus metodeista, joilla aineisto on analysoitu:
Ohjelmistot ja/tai laitteet, joita tarvitaan aineiston tulkintaan (muista ohjelmistojen versionumerot):
Standardit ja kalibroinnit, jos tarpeen mainita:

AINEISTOON LIITTYVÄ INFORMAATIO (Luo oma osio jokaiselle tiedostolle tai datasetille, jos se on tarpeen)

Muuttujat, lyhenteet, mittayksiköt, symbolit, koodit, jne. (eli sellainen informaatio, jota tarvitaan aineiston perusymmärrykseen)
Metatietostandardit ja sanastot, joita on käytetty aineiston kuvailussa: