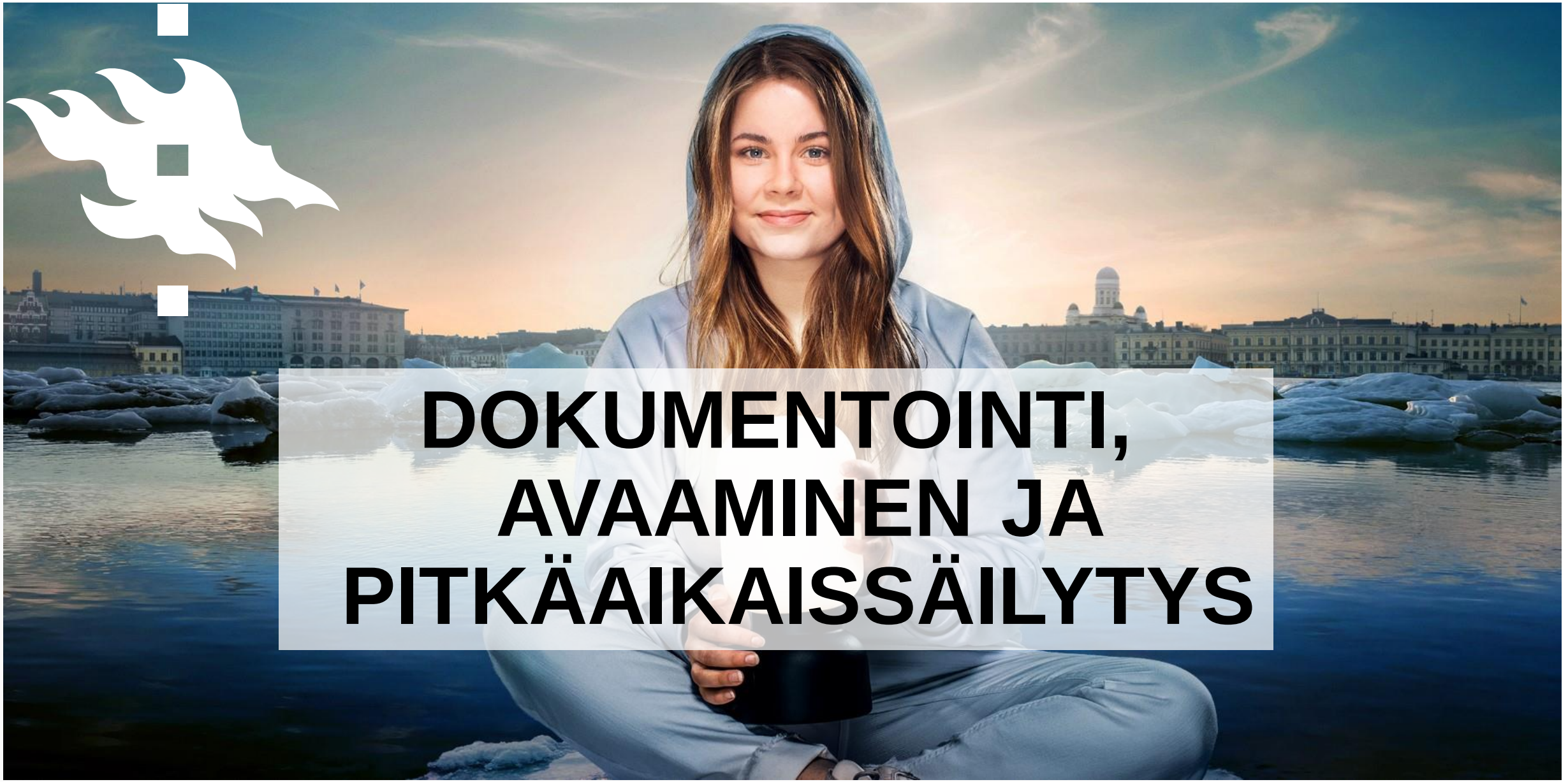


DOKUMENTOINTI, AVAAMINEN JA PITKÄAIKAISSÄILYTYS

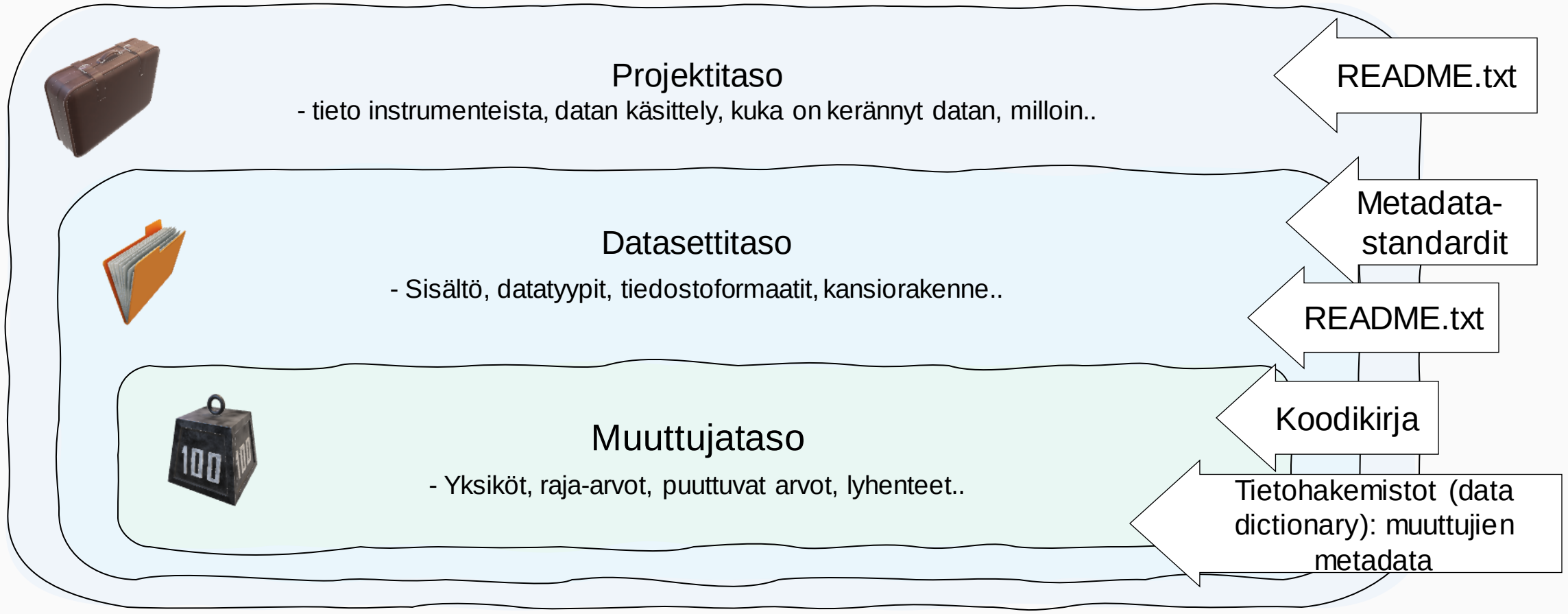
KUVAILU JA DOKUMENTOINTI

1. MITÄ KUVAILU JA DOKUMENTOINTI ON?

MITÄ DATAN KUVAILU ON?

- Tarkoittaa **datan kuvaamista** selkein ja yksinkertaisin termein.
- Antaa asiayhteyden ja selventää, **mitä** tiedot ovat ja **mistä** ne ovat peräisin.
- Avaa tiedostojen nimeämiskäytännöt ja kansioiden hakemistorakenteet.
- Auttaa seuraamaan versionhallintaa.
- **Metatiedot** (joita kutsutaan usein dataa datasta -tiedoiksi) ovat rakenteellinen datan dokumentoinnin muoto, jolla on suuri merkitys datan **löydettävyydessä**.
- **Metadataskeemat** noudattavat erityisiä **standardeja**, jotka tutkimusyhteisöt ja tieteenalat ovat määritelleet. Tieteenalakohtaiset standardit helpottavat **toistettavuutta**.
- Dokumentaatio voi kuvata dataa eri tasoilla.

DATAN KUVAILUN ERI TASOT



2. MITEN TUTKIMUSDATAA DOKUMENTOIDAAN?

DOKUMENTAATION ELEMENTIT

LÖYDETTÄVYYS

YMMÄRRETTÄVYYS

Datasetin “tuoteseloste"

Kuvailee datasetin sisällön. Pitäisi olla saatavilla, vaikka dataa ei voida avata.

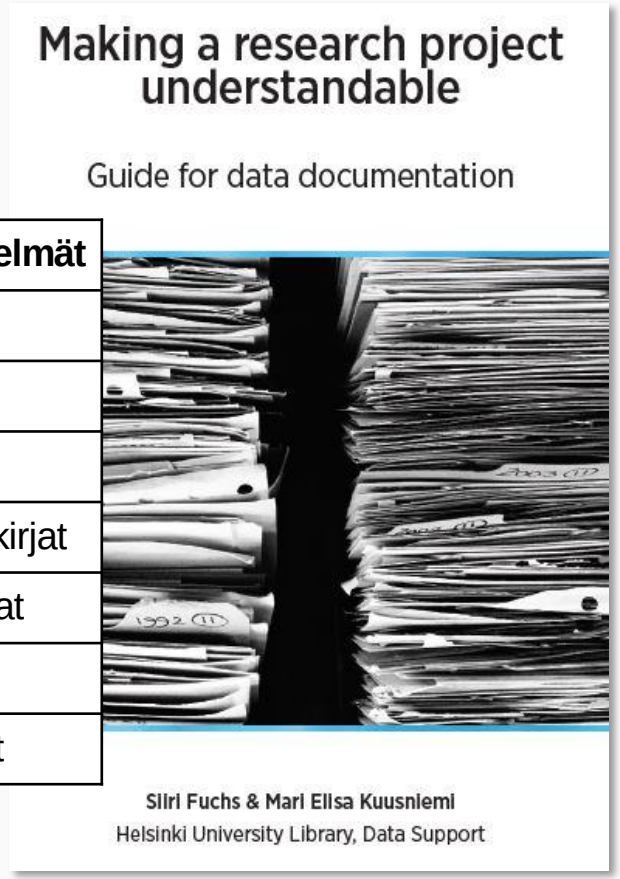
Nimeke, kuvaus, tekijä, pysyvä tunniste, hallinnolliset tiedot, lisenssit etc.

Datasetin “käyttöohje"

Datasetistä tehdään itsensä selittävä.

Tiedostojen nimeämiskäytännöt, tietohakemistot/koodikirjat (muuttujien selitykset), tägit, README.txt-tiedostot, hallinnolliset dokumentit etc.

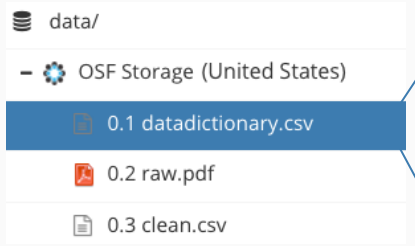
Dokumentaatiomenetelmät
Tiedostonimet
Kansiorakenne
Versiokontrolli
Tietohakemistot/koodikirjat
Laboratoriopäiväkirjat
Readme-tiedostot
Metadatastandardit



Zenodo. http://doi.org/10.5281/zenodo.1914401

TIETOHAKEMISTOT JA KOODIKIRJAT

- Taulukon muuttujien ja arvojen selitykset
- Yleensä tallennettu erilliseen tiedostoon (esim. xx.datadictionary.csv)
- Termejä koodikirja (*codebook*) ja tietohakemisto (*data dictionary*) käytetään päällekkäin. Koodikirjaa käytetään useammin kyselydatan kuvailuun.



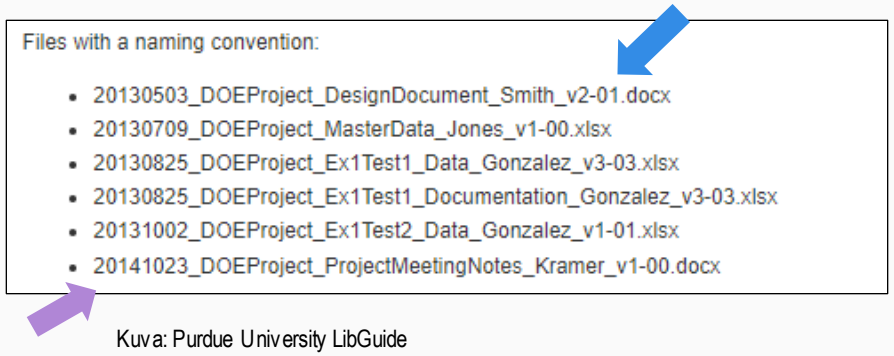
Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Lähde: How to Make a Data Dictionary, https://help.osf.io/article/217-how-to-make-a-data-dictionary; OSF Guides. Licensed under CC0.

TIEDOSTOJEN NIMEÄMINEN

- Suunnittele nimeämiskäytännöt mahdollisimman varhain
- Systemaattisuus, loogisuus, selkeys!
- Suosi deskriptiivisiä nimiä
- Nimeä versiot selkeästi (_v2-01.docx)
- Johdonmukainen päivämäärän muoto: yyyy-mm-dd ⇨ kronologisesti järjestyvä
- Yksilöivät tiedostonimet auttavat tunnistamaan tiedostot kansiorakenteesta riippumatta



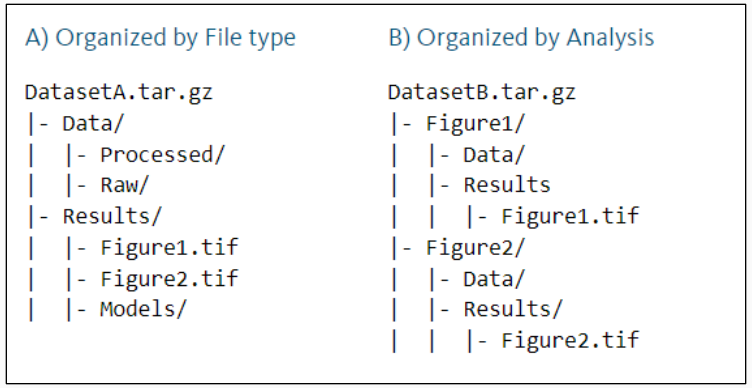
Kuva: Purdue University LibGuide
https://guides.lib.purdue.edu/c.php?g=353013&p=2378293

KANSIORAKENNE

- Riippuu projektin tarpeista
- Selkeä rakenne auttaa hallinnoimaan pääsyoikeuksia (esim. sensitiivinen data)
- Tiedostojen löytämistä auttaa:

  - Syvän ja matalan kansiorakenteen tasapainottaminen

  - Käytä avainsanoja ja tägejä



Kuva: Dryad
https://datadryad.org/stash/best_practices#organize

README-TIEDOSTOT

Tekstidokumentit (esim README.txt), joiden sisältämän informaation avulla varmistetaan, että tiedot tulkitaan oikein. Kuvailevat mm.

- tiedostojen nimet ja tiedostomuodot
- miten data on järjestetty (tiedosto- ja kansiorakenne)
- miten data on tuotettu sisältäen käytetyt laitteet, laitteistot ja ohjelmat
- miten dataa on muokattu tai prosessoitu/editoitu
- koodien, lyhenteiden ja muuttujien selitykset

Tallenna readme-tiedosto varsinaisen datan kanssa samaan paikkaan.

```
========== HEADER ==========
Readme.txt for ____[add name/title here]___ dataset
Documentation written on ____[add date as YYYYMMDD here]___
By ____[add Last name, First name here]___
Updated <YYYYMMDD>, by ____[add Last name, First name here]___


========== DATA DESCRIPTION ==========
ACKNOWLEDGEMENTS
Project title:
Funding agency/agencies:
Award Number:
Award Period:

Investigator Name:
Investigator Institution:
Investigator Address:
Investigator Email:
Investigator Role (related to this dataset): [e.g., data collection, data processing/cleaning, data
Investigator ID (if applicable): [e.g. ORCID]

Investigator Name:
Investigator Institution:
Investigator Address:
Investigator Email:
Investigator Role (related to this dataset): [e.g., data collection, data processing/cleaning, data
Investigator ID (if applicable): [e.g. ORCID]

...(repeat as needed)


DESCRIPTION
What this project or dataset is about, in terms of topical, geographic, or temporal coverage:

Other details about the content, formats, and internal relationships the dataset:


CITATION
Author(s)/Creator(s):
Title:
Year of Publication[when dataset was published/released, not data collection or coverage date]:
Publisher: [data center or repository]:
Identifier (DOI or any applicable identifier, including edition or version):
Availability and Access (URL or other location information for data): [e.g. https://www.datacite.org


DATE(S) of DATA COLLECTION
[single date, range, approximate date]
<suggested format YYYYMMDD>

GEOGRAPHIC LOCATION(S) of DATA COLLECTION
[where data was collected]
<suggested format: city, state, zip code, country, GPS>
```

Kuva: README.txt esimerkki, University of Texas LibGuide: https://guides.lib.utexas.edu/metadata-basics/key-concepts

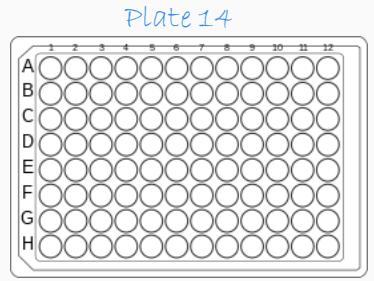
ELEKTRONISET LABORATORIOPÄIVÄKIRJAT

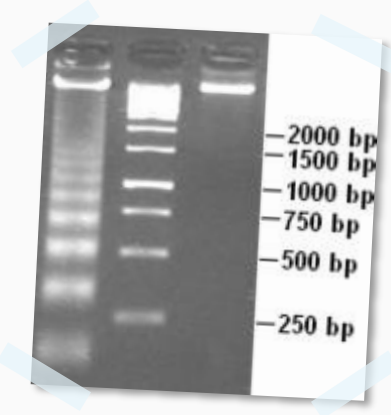
- Ohjelmissa usein dokumentointi ja metadatan luonti automaattisesti
- Laboratoriomuistiinpanot ovat paremmin tallessa ja mahdollistavat pääsyoikeuksien hallinnoinnin
- Tekee yhteistyöstä ja datan jakamisesta helpompaa
- Monikäyttöisempi verrattuna paperiseen laboratoriopäiväkirjaan, esim. hakutoiminnot
- Satoja ohjelmia saatavilla: linkki
- Splice-bio on listannut parhaimmat ELN-ohjelmat
- Valinnassa kannattaa kiinnittää huomiota hintaan, tietoturvallisuuteen, tarvittavaan perehdytykseen, mitä ohjelmaa tutkimusprojektin yhteistyökumppanit käyttävät ym..
- Ohjelmia voi etsiä haulla *electronic lab notebook; ELN*

RSpace COMMUNITY

labguru

SCINOTE



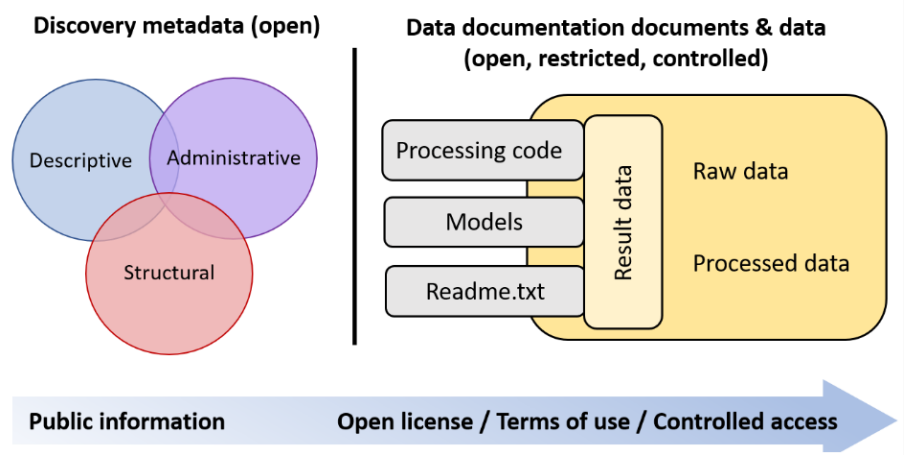


Kuvat: vasen, Nessa Carson, oikea Shinryuu.
Lähde: Wikimedia Commons, CC0

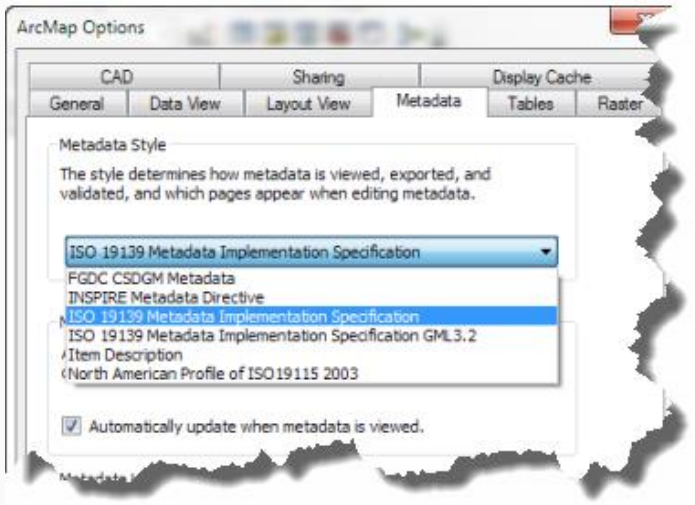
METADATA

METADATA

- Rakenteellisempi tapa kuvailla dataa
- ”Dataa datasta”
- Painotus datan löydettävyydessä
- Tietyt ohjelmistot tuottavat metadataa automaattisesti (REDCap, ArcGIS jne.)
- Standardisoidun metadatan käyttö edistää data-aineistojen vertailua ja uudelleenkäyttöä



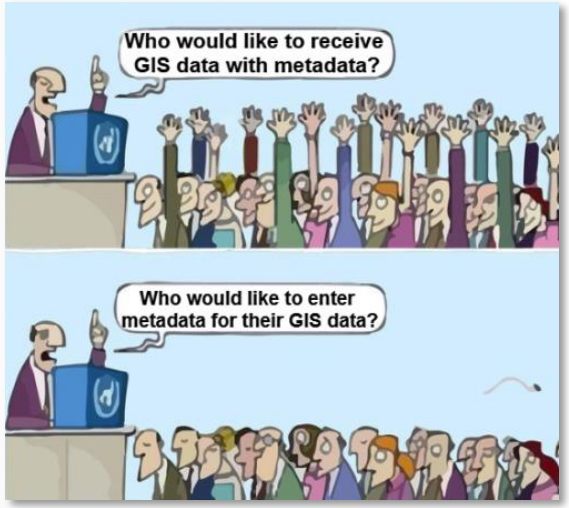
Kaavio: CSC Metadata and documentation, https://docs.csc.fi/data/datasets/metadata-and-documentation/ CC BY 4.0



Kuva: kuvakaappaus ArGIS-ohjelmistosta.
https://desktop.arcgis.com/en/arcmap/latest/manage-data/metadata/what-is-metadata.htm

METADATASTANDARDIT JA SKEEMAT

- Metadatastandardit voivat olla yleisiä tai tieteenalakohtaisia
- Standardeja ylläpitävät erilaiset yhteisöt, kuten FAANG.org
- Metadatastandardit noudattavat **skeemaa**, jossa hahmotellaan metatietojen rakenne
- Metatietoskeema = muoto, jossa tietokokonaisuus kuvataan hallitusti ja ennalta määritellysti, toisin kuin vapaamuotoinen kuvaus.
- Mahdollistavat koneluettavuuden
- Repositorion valinta tekee metadatastandardin/-skeeman valitsemisesta helpompaa



GIS=Geographic Information System.
Image: X, Matt Malone

Metadatastandardien katalogi: https://rd-alliance.github.io/metadata-directory/subjects/

FAANG
Functional Annotation of Animal Genomes

Functional Annotation of ANimal Genomes (FAANG) Project
— A coordinated international action to accelerate Genome to Phenome

Task Force: metaFAIR
— FAANG Metadata, Ontologies and FAIR Data Sharing Task Force

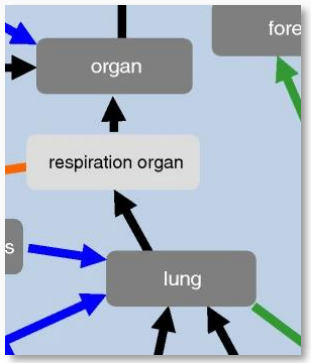
SANASTOT

- Tarjoavat johdonmukaisen tavan kuvata dataa ja lisäävät tietojen täsmällisyyttä
- Kontrolloidut sanastot: asiasanastot, thesaurukset yms.
- Tieteenalakohtaisia sanastoja, kuten CABI Thesaurus (vuodesta 1983, kattaa laajasti perus- ja soveltavat tieteet luonnonaloilta, teknologiasta ja sosiaalitieteistä). Suomalainen sanastopalvelu https://finto.fi/fi/

Kuva: https://www.cabi.org/cabithesaurus/

ONTOLOGIAT

- Datan kuvaamiseen käytettyjen termien suhteet ⇨ ontologia on jäsennelty sanasto
- Hierarkkinen luokittelu on yksi ontologioiden peruspiirteistä, esim. ”*keuhko* kuuluu *hengityselimet*-käsitteen alle”
- Rakennetaan ja muodostetaan eri lähteistä: artikkelit, sanakirjat, asiantuntijat
- Parantavat koneluettavuutta
- Erilaiset tutkimusalat: alakohtaiset ontologiat, kuten Uberon (monilajinen anatomian ontologia)



Kuva: Mungal et al (2012) Uberon, an integrative multi-species anatomy ontology

DOKUMENTOINNIN MUISTILISTA

- Suunnittele dokumentointi ennalta ja aloita tietojen kerääminen heti alussa: mitä aiemmin aloitat, sitä helpommin dokumentointi käy.
- Jos mahdollista, käytä metadatastandardeja ja kontrolloituja sanastoja.
- Datanhallintatyökalut (data management software) helpottavat dokumentointia.
- Minimissään tallenna datan kuvailu readme-tiedostoon.
- Jos et voi jakaa dataa vapaasti, voit kuitenkin avata metadatan.

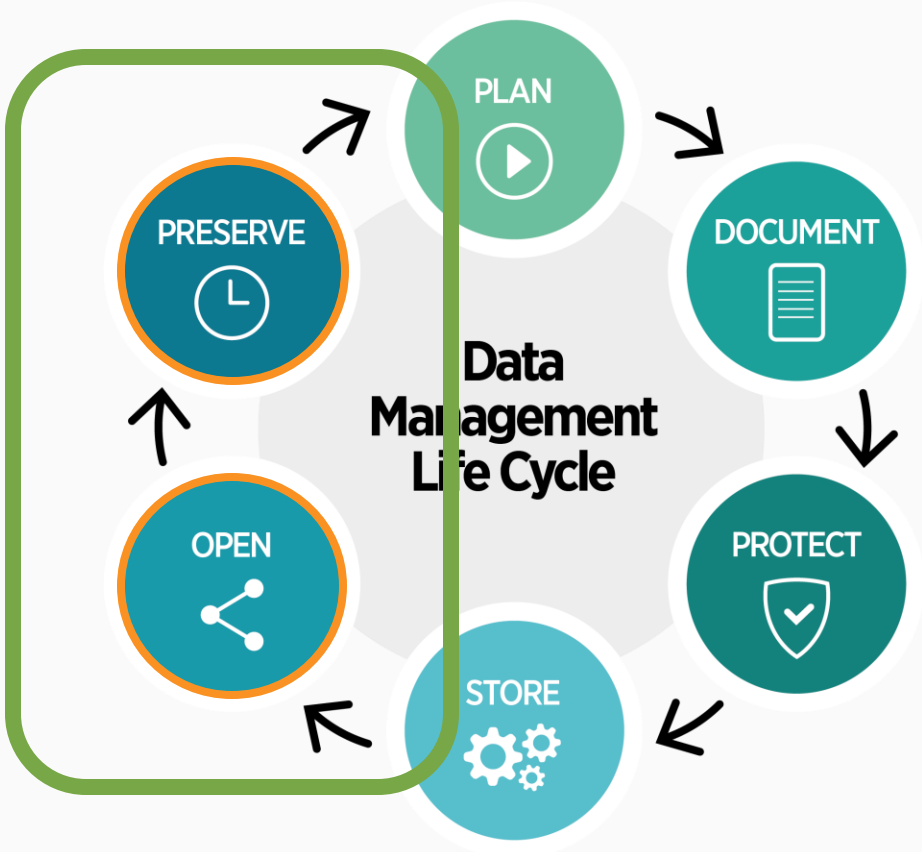
AVAAMINEN JA SÄILYTTÄMINEN

KUVA: JOEL GRANDELL

AVAAMINEN JA SÄILYTTÄMINEN

- *Mitä datalle tapahtuu projektin aktiivivaiheen jälkeen?*
- *Miten, milloin ja kenelle data tuodaan saataville?*
- *Miten ja missä tuodaan saataville data, joka on arvokasta pitkällä aikavälillä?*

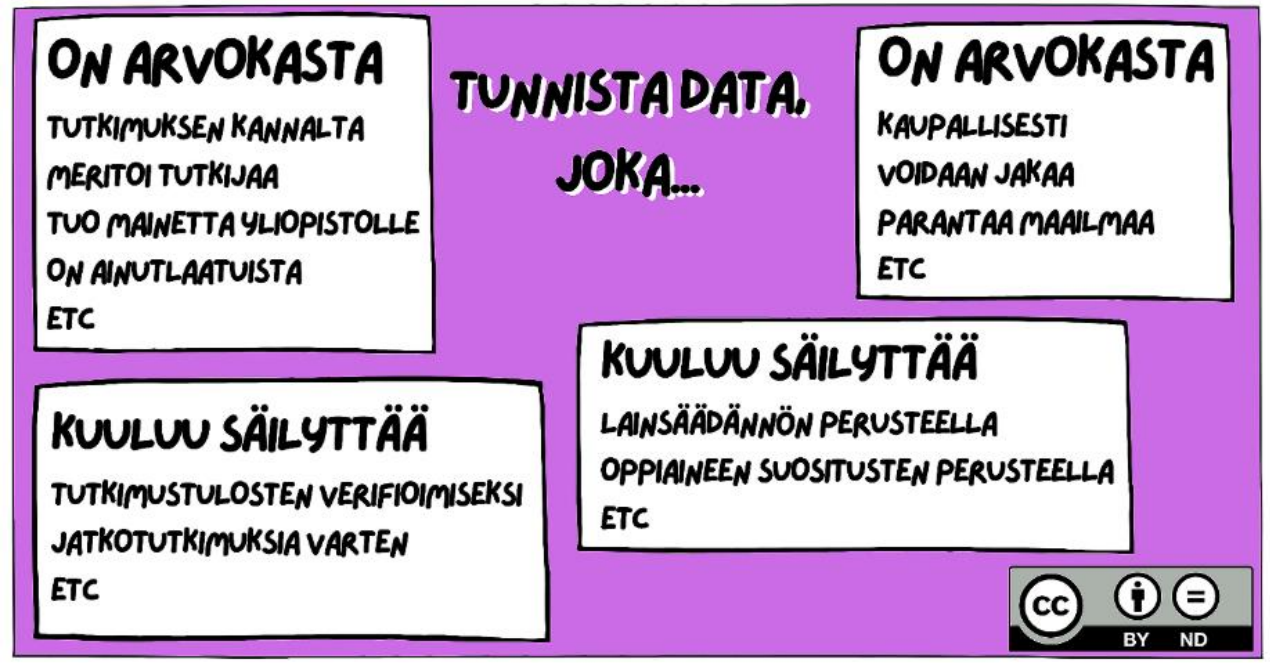
RISKIEN HALLINTA!



DATAN AVAAMISESTA HYÖTYVÄT MONET

- Tutkimuksen toistettavuus, avoimuus ja validointi
- Resurssien kestävä hyödyntäminen, turhien kokeiden vähentäminen
- Julkisilla varoilla tuotetun tutkitun tiedon saattaminen laajemmalle yleisölle
- Tutkimuksen ja innovaatioiden edistäminen ja yhteistyömahdollisuuksien lisääminen
- Lisää näkyvyyttä tutkimuksellesi (lisää viittauksia!)
- Investointien houkutteleminen tutkimukseen
- Laajojen, globaalien ongelmien tehokkaampi ratkaisu

MITEN VALITA SÄILYTETTÄVÄ DATA



Kuva: Lahtinen et al. 2023 How to become a data preserver: The official University of Helsinki guide to the responsible preservation of research data https://zenodo.org/records/10424017

Lisää aiheesta (englanniksi): Five steps to decide what data to keep https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep

DATAN ENSIKÄYTTÖ, JATKOKÄYTTÖ JA UUDELLEENKÄYTTÖ

- **Datan ensikäyttö**: Tutkija tai tutkimusryhmä kerää dataa tutkimusprojektia varten ja analysoi sitä vastatakseen tiettyyn tutkimuskysymykseen.
- **Datan jatkokäyttö**: Jos sama tutkija tai tutkimusryhmä palaa data-aineistoon myöhemmin, data-aineistoa analysoidaan edelleen datan tuottajien kontekstissa.
- **Datan uudelleenkäyttö**: Data-aineisto on tallennettu datarepositorioon ja joku muu lataa sen itselleen uutta projektia varten.

Lähde: Pasquetto et al. (2019) Uses and reuses of scientific data: the data creators advantage.
https://hdsr.mitpress.mit.edu/pub/jduhd7og/release/10

ESIMERKKEJÄ DATA-AINEISTOJEN KÄYTÖSTÄ

- Datan käyttötavat ovat moninaiset, mihin oheisista kategorioista nämä esimerkit putoaisivat?
  - Itse kerättyyn dataan palaaminen myöhempää vertailua varten
  - Julkisista tai yksityisistä lähteistä hankittujen data-aineistojen vertaaminen uutena kerättyyn dataan
  - Olemassa olevien data-aineistojen kartoittaminen uuden projektin taustoitusta varten
  - Uusien tutkimuskysymysten ratkaiseminen analysoimalla johonkin toiseen tarkoitukseen kerättyä dataa
- Usein on vaikea ennakoida tulevaisuuden relevantteja datan käyttökohteita

Lähde: Pasquetto et al. (2019) Uses and reuses of scientific data: the data creators advantage.
https://hdsr.mitpress.mit.edu/pub/jduhd7og/release/10

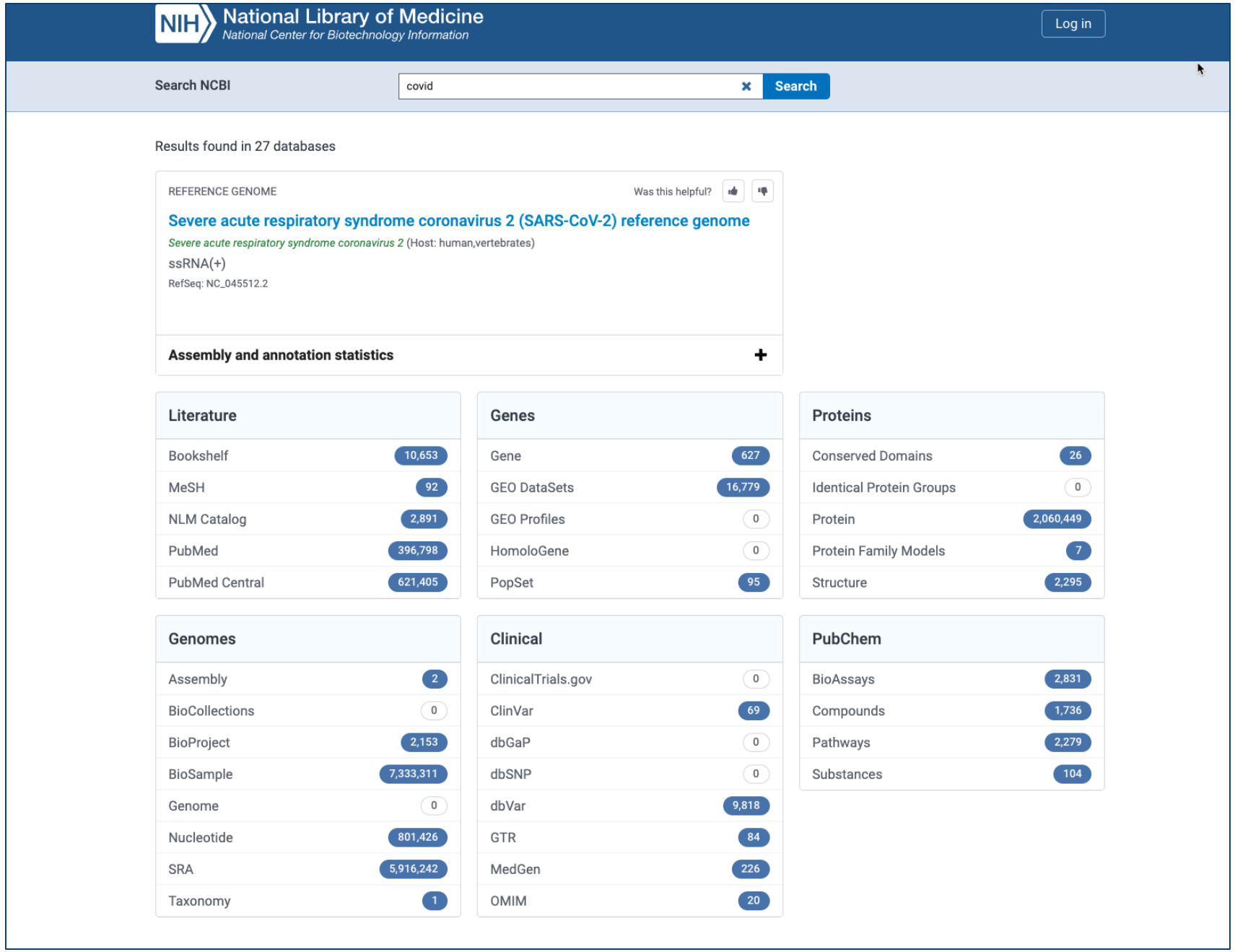
KAIKKI LÄHTEE HYVÄSTÄ SUUNNITTELUSTA

- Aineistonhallintasuunnitelma (DMP):
  - Datan avaaminen yleisessä repositoriossa tai tieteenalakohtaisessa repositoriossa
  - Data-artikkelin kirjoittaminen (Data in Brief, Data)?
  - Riittääkö datan avaaminen artikkelin lisäosiona tai tutkimusprojektin verkkosivuilla?
  - Yhteistyöprojektissa tulee sopia, kenen tulisi avata data?
  - Miten paljon datan valmistelu avaamista varten vie aikaa?
  - Datan avaamiseen liittyvät maksut? Esimerkiksi Dryad-repositoriolla on $150 dollarin kuratointimaksu

DATAREPOSITORIOT

Virtuaalinen, yleensä tieteenalakohtainen arkisto tai tietokanta, johon tutkijat voivat siirtää tutkimusdatansa jakamista, raportointia ja jatkokäyttöä varten. Data-arkisto säilyttää tutkimusdataa, asettaa sen käytettäväksi ja järjestää sen loogisella tavalla. Datarepositoriot helpottavat myös tutkimusdataan viittaamista, kun käytetään pysyviä tunnisteita.

(Lähde: Helsingin yliopiston tutkimusdatapolitiikka)


NIH National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
covid
Search
Results found in 27 databases
REFERENCE GENOME
Was this helpful?
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) reference genome
Severe acute respiratory syndrome coronavirus 2 (Host: human,vertebrates)
ssRNA(+)
RefSeq: NC_045512.2
Assembly and annotation statistics
Literature
Bookshelf 10,653
MeSH 92
NLM Catalog 2,891
PubMed 396,798
PubMed Central 621,405
Genes
Gene 627
GEO DataSets 16,779
GEO Profiles 0
HomoloGene 0
PopSet 95
Proteins
Conserved Domains 26
Identical Protein Groups 0
Protein 2,060,449
Protein Family Models 7
Structure 2,295
Genomes
Assembly 2
BioCollections 0
BioProject 2,153
BioSample 7,333,311
Genome 0
Nucleotide 801,426
SRA 5,916,242
Taxonomy 1
Clinical
ClinicalTrials.gov 0
ClinVar 69
dbGaP 0
dbSNP 0
dbVar 9,818
GTR 84
MedGen 226
OMIM 20
PubChem
BioAssays 2,831
Compounds 1,736
Pathways 2,279
Substances 104

HYVÄN DATAREPOSITORION KRITEERIT

- Mahdollistaa data-aineistojen kuvailun monipuolisella metadatalla
- Antaa data-aineistolle pysyvän tunnisteen (DOI, accession number..)
- Nojaa kestävään yritysmalliin tai taustalla on laajalti tunnettu organisaatio

Työkaluja dataresitorioiden arviointiin:

- TRUSTworthy repositoriot (TRAC)
- FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) periaatteet
- F-UJI: https://www.f-uji.net/ (automatisoitu FAIR arviointityökalu)

Digitaalisten repositorioiden TRUST periaatteet

Periaate	Ohjeistus repositorioille
Transparency Läpinäkyvyys	Läpinäkyvyys repositorion tuottamista palveluista sekä repositoriossa säilytettävästä datasta
Responsibility Vastuullisuus	Vastuu data-aineistojen eheyden ja aitouden varmistamisesta ja repositorion palvelun luotettavuudesta ja pysyvyydestä
User Focus/ Käyttäjälähtöisyys	Dataa käyttävien yhteisöiden datan hallinnan käytänteiden ja odotusten täyttäminen
Sustainability Kestävyys	Palveluiden kestävyys ja data-aineistojen säilyttäminen pitkällä aikavälillä
Technology Teknologia	Infrastruktuurin ja kyvykkyyksien tarjoaminen turvallisten, pysyvien ja luotettavien palveluiden tueksi

KUINKA LÖYTÄÄ SOPIVA REPOSITORIO?

- Repositorioiden katalogi **Re3data**, https://www.re3data.org/
- Julkaisijoilla ja lehdillä on omia oppaita ja suosituksia verkossa:
  - Scientific Data: https://www.nature.com/sdata/policies/repositories
  - PlosOne: https://journals.plos.org/plosone/s/recommended-repositories
- Muista tarkistaa valitsemasi lehden suositukset ajoissa!
- Parhaat tieteenalakohtaiset datarepositoriot ottavat vastaan vain tietyn tyyppistä data → Pyri löytämään kaikille tutkimuksesi tuottamille datatyypeille sopiva repositorio.
- Kuratoitu vs. kuratoimaton repositorio
  - Onko datalla laadunvarmistusta?

RAJOITETTU DATAN AVAAMINEN

- Tutkimuksessa julkaisuviive eli **embargo** on rajoitus joka asetetaan lopputyölle, artikkelille tai tutkimusdatalle. Julkaisuviiveen aikana vain otsikko, abstrakti ja viittaustiedot näkyvät yleisölle. Varsinaisen sisällön ollessa piilotettuna määritetyn ajan.
- Monet datarepositoriot tarjoavat julkaisuviiveen mahdollisuuden.
- Luottamuksselliselle ja arkaluontoiselle tutkimusdatalle on tarjolla rajallinen määrä sopivia repositorioita.
- Arvioi voisiko data-aineiston avata anonymisoituna (vaatii aikaa ja osaamista)
- Todennäköisesti voit avata datan karkeistettuna jossain muodossa (anonymisointi a turvallisempi vaihtoehto)

METADATAN AVAAMINEN

- Vaikka et voisikaan avata tutkimusdataa, tulisi löytämistä tukeva metadata kuitenkin avata.
  - Löytävyyttä tukeva metadata: data-aineiston otsikko, abstrakti, viittaustiedot, jne.
- Julkisen kuvailun voi tehdä CSC:n Qvain-työkalulla, joka tukee kontrolloitujen sanastojen käyttöä, antaa kuvailulle lisenssin ja luo metadatalle pysyvän tunnisteen.
  - Qvaimen avulla voit julkaista metadatasi suoraan Etsin-hakupalvelussa.
- Kuvailutietoja voi julkaista myös esimerkiksi Zenodossa

DATA ARTIKKELI JA DATA LEHDET

Data Article

Skin, gut, and sand metagenomic data c
placebo-controlled sandbox biodiversity
intervention study

Marja I. Roslund[a,b], Anirudra Parajuli[b,1], Nan Hui[b,2]
Riikka Puhakka[b], Mira Grönroos[b], Laura Soininen[b],

of movement sensor dataset for
r classification

*, Sanni Somppi[b], Heini Törnqvist[b,e],
la Cardó[b], Pekka Kumpulainen[a], Heli Väätäjä[c,d]
[c], Veikko Surakka[c], Miiamaaria V. Kujala[b,e],

Received: 14 January 2022 | Revised: 23 March 2022 | Accepted: 29 March 2022
DOI: 10.1002/ecy.3740

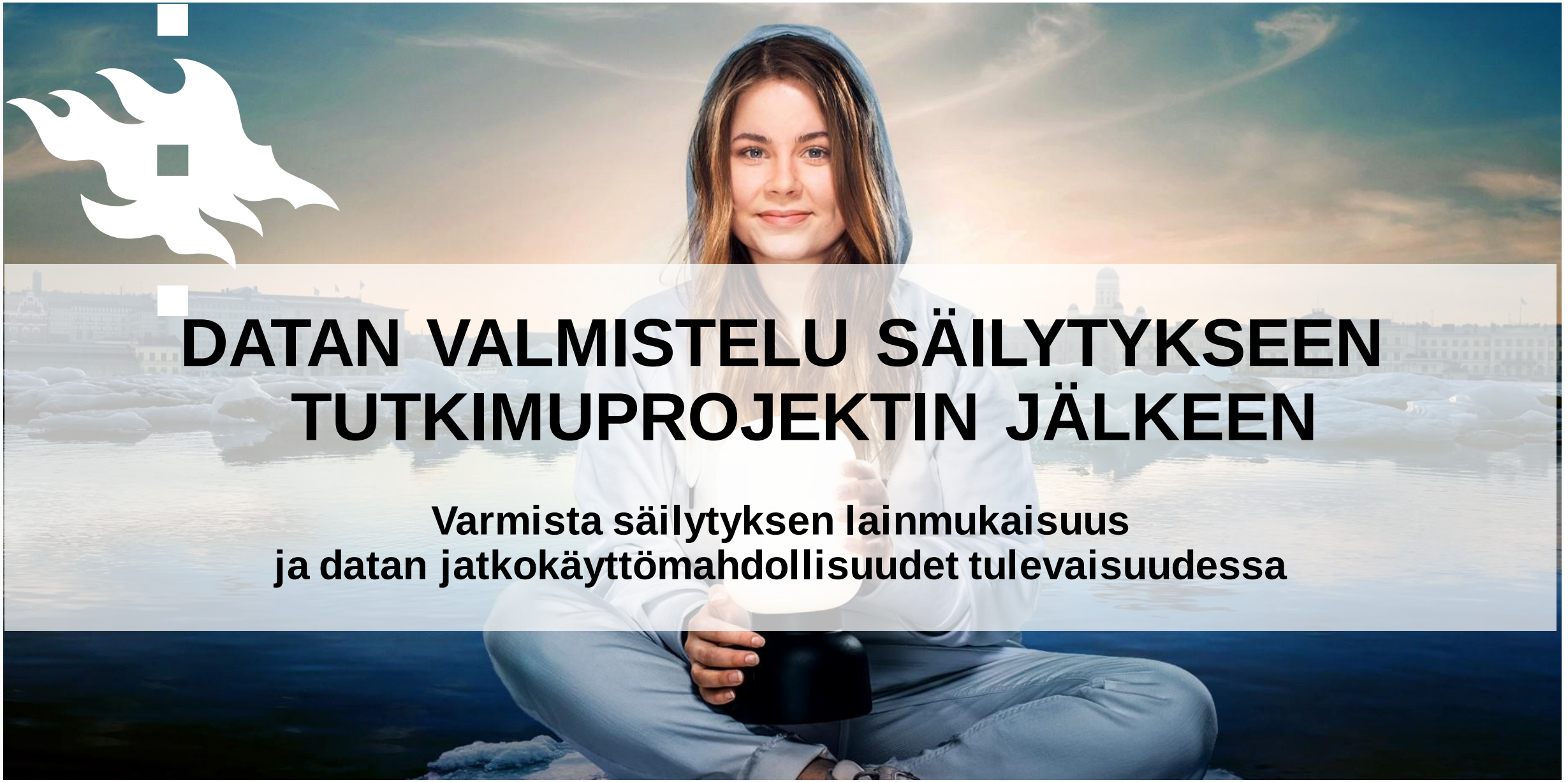
DATA PAPER

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Wild bee larval food composition in five European cities

Joan Casanelles-Abella[1,2] | **Alexander Keller**[3] | **Stefanie Müller**[1,4] | **Cristiana Aleixo**[5] | **Marta Alós-Orti**[6] | **François Chiron**[7] | **Lauri Laanisto**[6] | **Łukasz Myczko**[8] | **Pedro Pinho**[5] | **Roeland Samson**[9] | **Piotr Tryjanowski**[8] | **Anskje Van Mensel**[9] | **Lucía Villarroya-Villalba**[1] | **Loïc Pellissier**[2,10] | **Marco Moretti**[1]

[1]Biodiversity and Conservation Biology, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland
[2]ETH Zurich, Institute of Terrestrial Ecosystems, Zurich, Switzerland
[3]Organismic and Cellular Interactions, Biocenter, Faculty of Biology, Ludwig-Maximilians-Universität München, Martinsried, Germany
[4]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[5]Centre for Ecology, Evolution and Environmental Changes (cE3c), Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
[6]Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Tartu, Estonia
[7]Université Paris Saclay, CNRS, AgroParisTech, Ecologie Systématique Evolution, Orsay, France



DATAN VALMISTELU SÄILYTYKSEEN TUTKIMUPROJEKTIN JÄLKEEN

Varmista säilytyksen lainmukaisuus ja datan jatkokäyttömahdollisuudet tulevaisuudessa

Kuva: Dorothy Whitaker töissä National Oceanographic Data Center (NODC) magneettinauha kirjastossa. Lähde Wikimedia Commons. contributor NODC. Lisenssi Public domain

Kuva: Dunhuang käsikirjan digitointi IDP UK Studiolla. Lähde Wikimedia Commons, International Dunhuang Project. Lisenssi CC Attribution-Share Alike 3.0

Kuva: Automaattinen magneettinauha kammio CERN tietokonekeskuksessa vuonna 2008. Lähde https://cds.cern.ch/record/1138232, Claudia Marcelloni; Maximilien Brice. Lisenssi © 2008-2023 CERN

TUTKIMUSDATAN SÄILYTTÄMINEN

- Digitaalinen materiaali on laiteriippuvaista ja haurasta huolimatta sen helposta kopioinnista ja jakamisesta
- Tiedon ja asiayhteyden katoaminen ovat uhkat datan säilyvyydelle ja eheydelle
- Digitaaliset alustat muuttuvat ja ne ovat usein riippuvaisia pitkistä ja monimutkaisista riippuvuuksien ketjuista
- **Valinta, arviointi ja hallittu hävittäminen** ovat säilyttämisen keskeisiä komponentteja
- Kaikkea tutkimusdataa ei edes tulisi pyrkiä säilyttämään

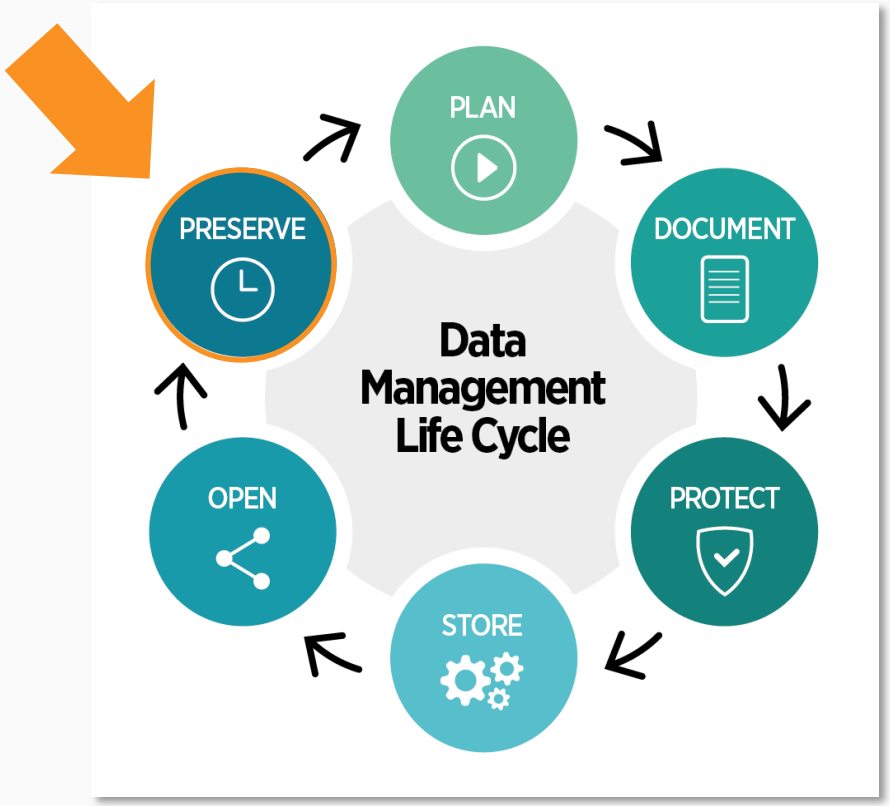
Kuva: Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Lähde: https://www.dpconline.org/handbook/digital-preservation/why-digital-preservation-matters

VALMISTELU ON VÄLTTÄMÄTÖNTÄ

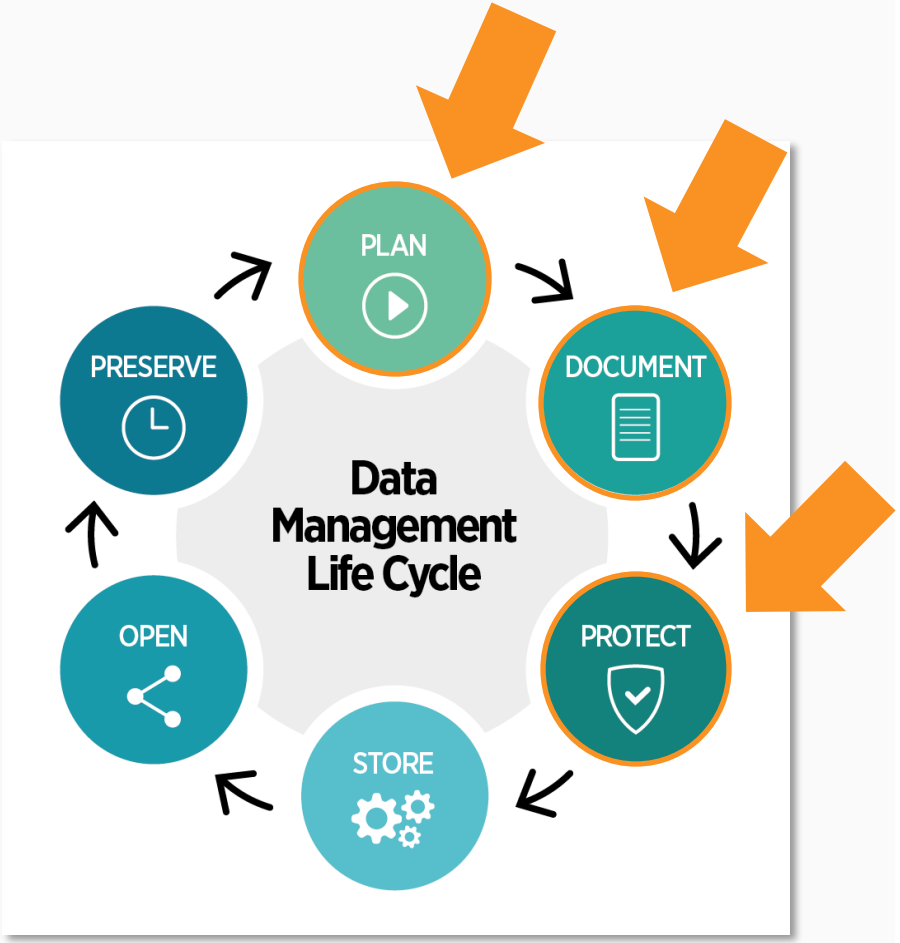
jotta voidaan varmistaa:

- Turvallinen säilytys
- Datan katoamattomuus
- Säilytyksen lainmukaisuus
- Datan tehokas uudelleenkäyttö tulevaisuudessa



VALMISTELU KUULUU AINEISTONHALLINNAN ERI VAIHESIIN

- Valinta, arviointi ja kontrolloitu hävittämien
- Säilytykseen soveltuvat tiedostomuodot
- Käyttöoikeudet, sopimukset ja lisenssit
- Tietoturva
- Datan kuvailu ja metadata
- Koodien ja ohjelmistojen valmistelu



SÄILYTTÄÄKÖ VAIKO EI?

- Data säilytetään
  - tutkimuksen tulosten verifioimiseksi
  - uudelleenkäyttöä varten
  - jos säilytysvelvoite on kirjattu lakiin tai sopimuksiin
- Tunnista pitkäaikaissäilytykseen soveltuva data
  - yhteiskunnallinen, historiallisen ja/tai tieteellisen arvo
  - ainutlaatuinen tai poikkeuksellisen pitkäaikaisen havaintoaineisto
- Hävitä data
  - joka tulee tuhota lainsäädännöllisiin syihin nojaten
  - joka on luvanvarainen säilytysaika on päättynyt
  - joka on hankittu kolmannelta osapuolelta ja jolle ei ole säilytysoikeutta

Kuva: Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo
https://zenodo.org/records/5785971

SÄILYTYKSEEN SOVELTUVAT TIEDOSTOMUODOT

- Suosi avoimia tai tieteenalallasi vakiintuneita tiedostomuotoja
- Tarkista tallennuspaikastasi mitä tiedostomuotoja he suosittelevat
- Erikoisemmat tiedostomuodot voivat soveltua säilytykseen, kunhan niille löytyy riittävä dokumentaatio.
- Esimerkiksi, suosi tiedostomuodon.xlsx sijaan muotoa .csv
  - Yksinkertaisuus vähentää alttiutta datan korruptiolle
  - Yhteensopivuus mahdollistaa käytettävyyden useiden ohjelmistojen kanssa
  - Keveys vähentää tallennustilalle asetettuja kokovaatimuksia

Lista avoimista tiedostomuodoista (englanniksi):
https://en.wikipedia.org/wiki/List_of_open_file_formats

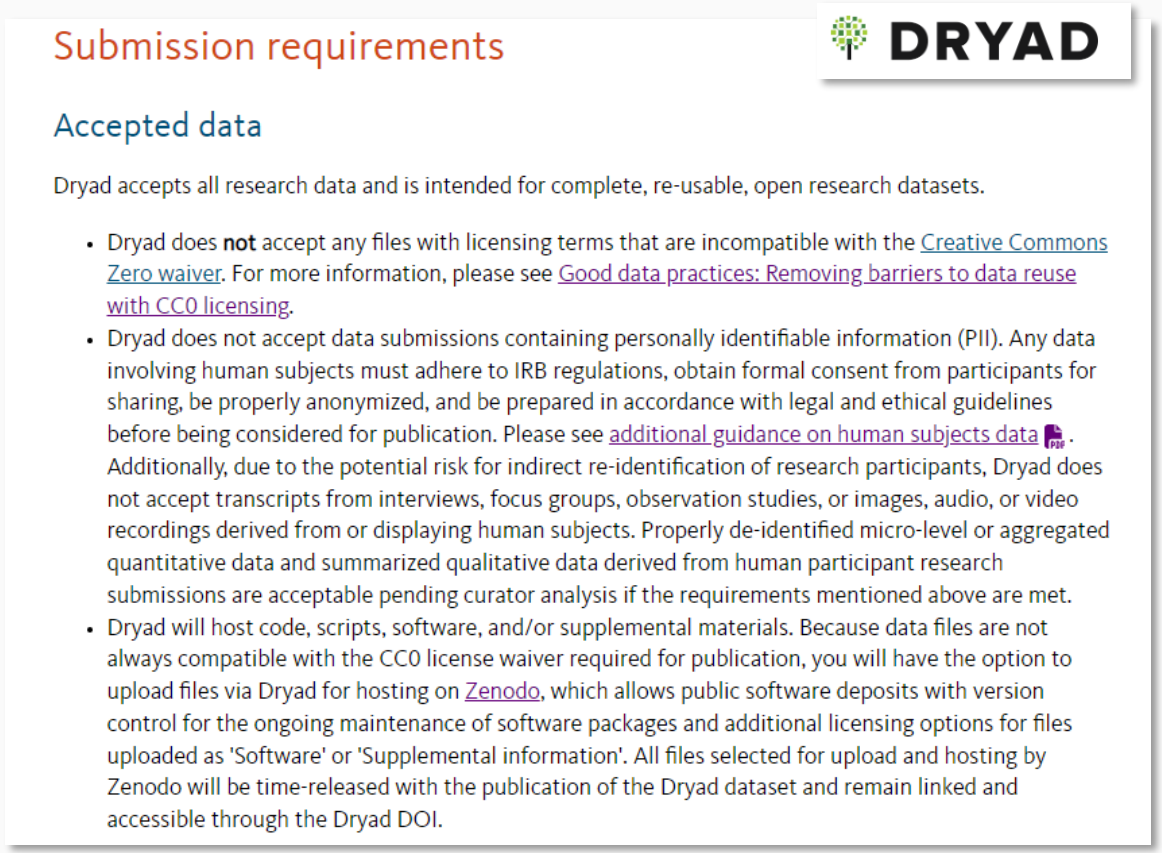
Kuva: Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

KÄYTTÖOIKEUDET, SOPIMUKSET JA LISENSSIT

- Käyttöoikeuksista on sovittava kaikkien datan keräämiseen osallistuneiden osapuolten kesken
- Sopimukset on syytä laatia jo ennen datan keräämisen aloittamista
- Jos data on saatu muualta, tutkijan oikeudet datan käsittelyyn tulee varmistaa
- Eri tallennuspalveluilla on omat käyttöehtonsa, joihin tulisi tutustua etukäteen
- Tallennuspalvelut eivät yleensä ota vastaan henkilötietoja sisältävää dataa tai kaupallisten toimijoiden omistamaa dataa
- Datalle asetettu lisenssi kertoo muille millä ehdoilla dataa saa uudelleenkäyttää

Kuva: Jørgen Stamp digitalbevaring.dk
CC BY 2.5 Denmark

KÄYTTÖOIKEUDET, SOPIMUKSET JA LISENSSIT



Submission requirements

DRYAD

Accepted data

Dryad accepts all research data and is intended for complete, re-usable, open research datasets.

- Dryad does **not** accept any files with licensing terms that are incompatible with the Creative Commons Zero waiver. For more information, please see Good data practices: Removing barriers to data reuse with CC0 licensing.
- Dryad does not accept data submissions containing personally identifiable information (PII). Any data involving human subjects must adhere to IRB regulations, obtain formal consent from participants for sharing, be properly anonymized, and be prepared in accordance with legal and ethical guidelines before being considered for publication. Please see additional guidance on human subjects data. Additionally, due to the potential risk for indirect re-identification of research participants, Dryad does not accept transcripts from interviews, focus groups, observation studies, or images, audio, or video recordings derived from or displaying human subjects. Properly de-identified micro-level or aggregated quantitative data and summarized qualitative data derived from human participant research submissions are acceptable pending curator analysis if the requirements mentioned above are met.
- Dryad will host code, scripts, software, and/or supplemental materials. Because data files are not always compatible with the CC0 license waiver required for publication, you will have the option to upload files via Dryad for hosting on Zenodo, which allows public software deposits with version control for the ongoing maintenance of software packages and additional licensing options for files uploaded as 'Software' or 'Supplemental information'. All files selected for upload and hosting by Zenodo will be time-released with the publication of the Dryad dataset and remain linked and accessible through the Dryad DOI.

CC0 vaatimuksena

Ei henkilötietoja

Koodeille Zenodo-integraatio

Lähde: https://datadryad.org/stash/requirements

TIETOTURVA

- Varmista sisältääkö data henkilötietoja, tai GDPR:n määrittelemiä erityisiä henkilötietoryhmiä
- Selvitä, onko data mahdollista anonymisoida tai pseudonymisoida
- Palauta mieleesi, miten tutkittuja on informoitu ennen datan keräämistä
  - Tutkittavien informoinnissa on linjattu mm. mihin data tallennetaan projektin jälkeen, millä ehdoin dataa saa uudelleenkäyttää ja mille tahoilla datan saa tulevaisuudessa luovuttaa
- Tarkista myös, sisältääkö data muuta salassa pidettävää informaatiota, kuten liikesalaisuuksia tai uhanalaisten eliöiden esiintymistietoa

DATAN KUVAILU JA DOKUMENTOINTI

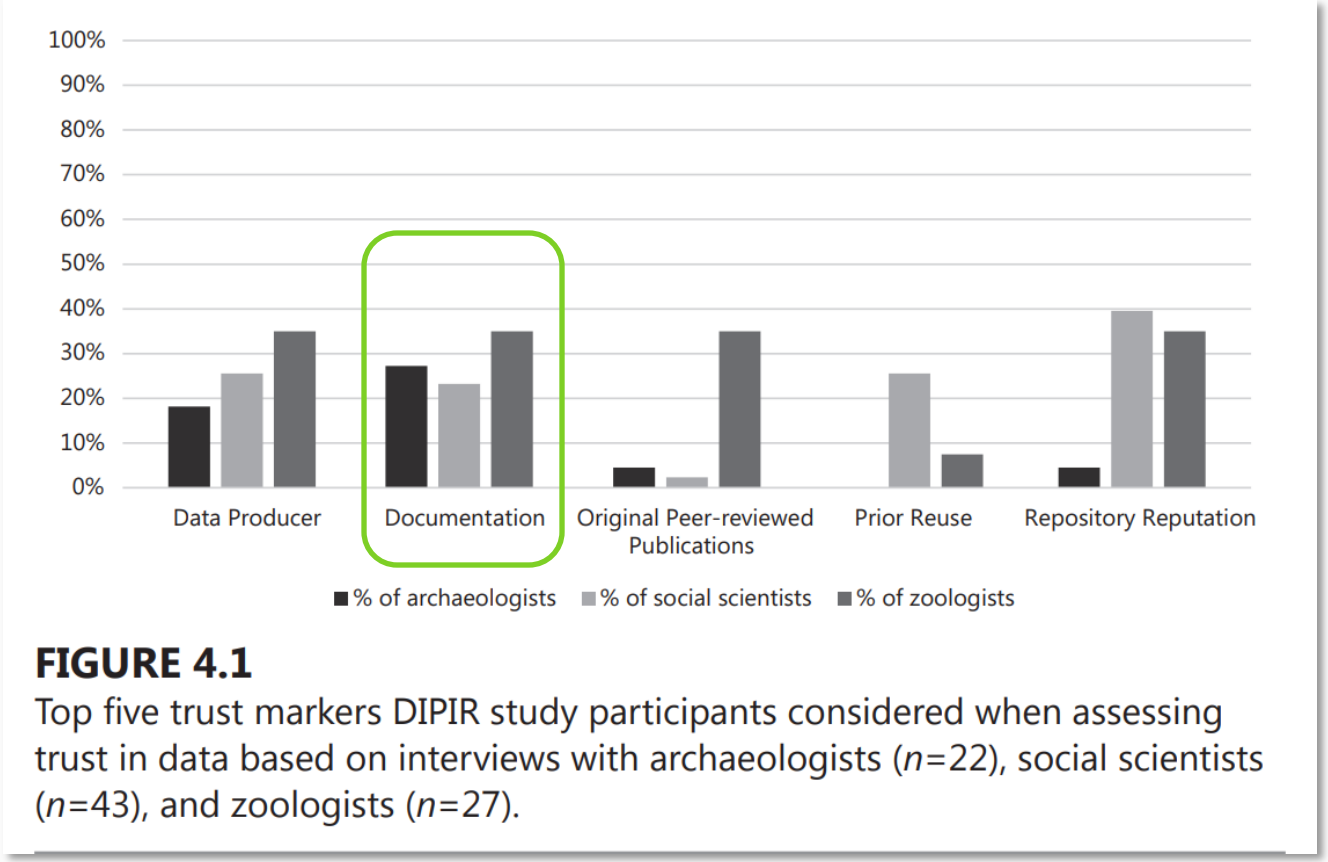


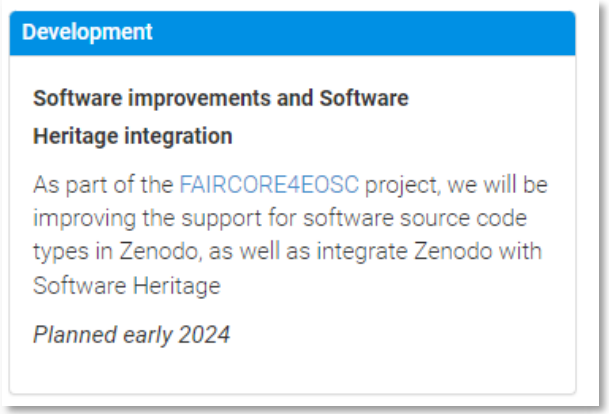
FIGURE 4.1
Top five trust markers DIPIR study participants considered when assessing trust in data based on interviews with archaeologists (*n*=22), social scientists (*n*=43), and zoologists (*n*=27).

Luottamus

Kuvankaappaus: Faniel & Yakel 2017. Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation.
DIPIR= Dissemination Information Packages for Information Reuse
https://www.oclc.org/research/publications/2017/practices-do-not-make-perfect.html

KOODIT JA OHJELMISTOT

- Eivät sovellu kaikkiin tallennuspalveluihin
- Zenodo-GitHub integraatio: Zenodo antaa DOI-tunnisteen ja säilyttää koodin
- Software Heritage on pitkäaikaissäilytyspalvelu koodeille, integroitu myös Zenodoon
- Kirjaa kaikki ohjelmistojesi riippuvuudet, kuten selaimet, käyttöjärjestelmät, ohjelmistokehityspaketit (SDK) yms.
- Kirjaa ja seuraa ohjelmistorajapintoja sekä avointen tai lisensoitujen standardien käyttöä, sillä ne voivat muuttua tai tulla korvatuksi uusilla rajapinnoilla tai standardeilla.
- Tarkista onko ohjelmistosi riippuvainen pääsystä johonkin julkiseen verkkopalveluun, infrastruktuuriin tai tietokantaan, joka saattaa muuttua tai kadota.
- Koodeille ja ohjelmistoille on omia metadatastandardeja, kuten CodeMeta
- Valitse koodillesi lisenssi, joka mahdollistaa uudelleenkäytön ja/tai lisenssi, joka on yhteensopivia riippuvuuksien lisensseistä.



Kuvakappaus: https://about.zenodo.org/roadmap/

Lisätietoja: https://www.software.ac.uk/ ; https://archive.softwareheritage.org/ ; https://the-turing-way.netlify.app/reproducible-research/code-reuse/code-reuse-overview

SÄILYTYSPAKETTI VALMIINA!

VALMISTELUT PÄHKINÄNKUORESSA

Kuva: Lahtinen et al. 2023 How to become a data preserver: The official University of Helsinki guide to the responsible preservation of research data https://zenodo.org/records/10424017