CSC

ICT Solutions for Brilliant Minds

Johdanto yhteentoimivuuteen - tutkimusdatan maisema Suomessa

Jessica Parland-von Essen

Tutkimuksen digitalistoituminen

By Brian Snelson from Hockley, Essex, England - link to the photo, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=18441433

- Digitalisoituminen tarkoittaa prosessien radikaalia uudelleen muokkaamista
- Ensimmäinen vaihe mediamurroksessa on kuitenkin useimmiten vanhojen prosessien kopioiminen tietokoneympäristöön
- Mahdollisuuksia jää silloin hyödyntämättä ja syntyy yllättäviä haasteita, vrt. avoimuus
- Tutkimuksen prosessien miettiminen uudestaan on vielä kesken, keskeistä olisi kuitenkin huomioida *toistettavuuden vaatimus*
- Kaupalliset toimijat pyrkivät joskus viemään kehitystä tieteen kannalta väärään suuntaan

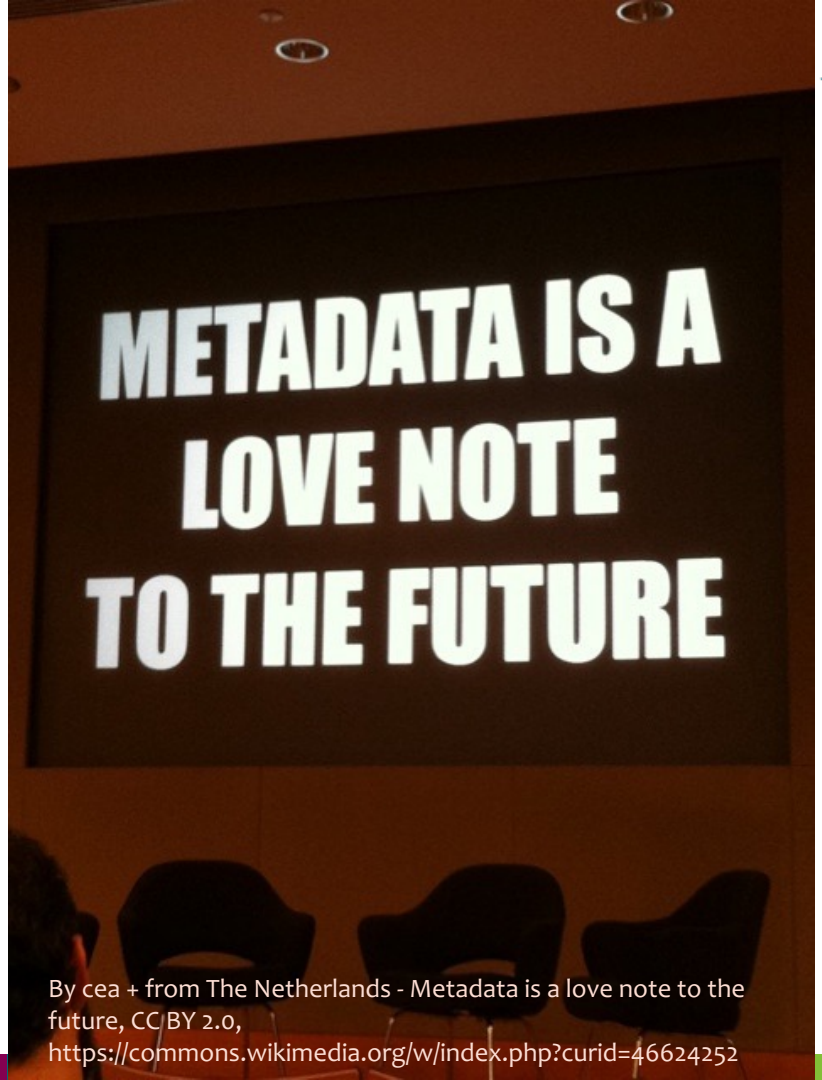
Metatiedot tutkimuksessa

- Tutkimuksen toistettavuus
- Viittaaminen ja hyvä tieteellinen käytäntö
- Meritoituminen
- Harvoin riittävä dokumentointi pitkäaikaiseen säilyttämiseen

TTA TUTKIMUKSEN TIETOAINEIST

Avoin tiede

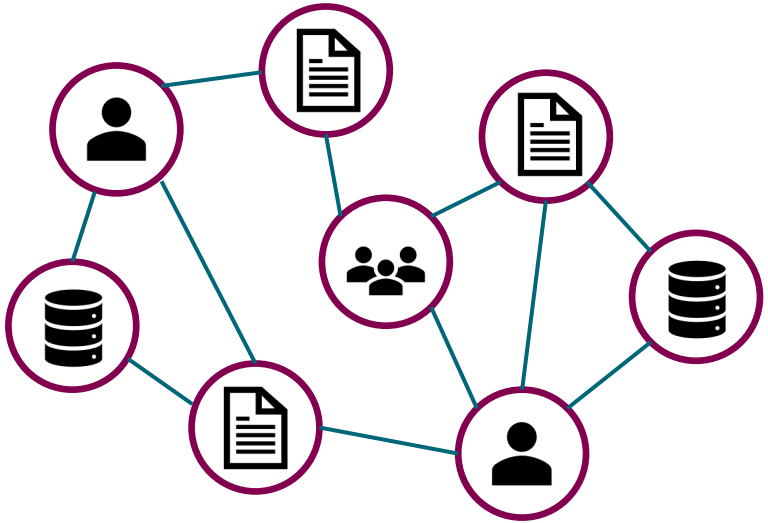
OPEN SCIENCE AND RESEARCH



By cea + from The Netherlands - Metadata is a love note to the future, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=46624252

Tutkimuksen metatietojen kaksi käyttökontekstia

Tutkimustieto / CRIS



Tutkimuksen hallinta, toistettavuus

Datan FAIR-periaatteet

Findable

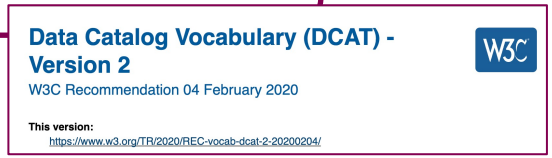
F1. (Meta)data are assigned a globally unique and persistent identifier

Cool URI

URN

Handle

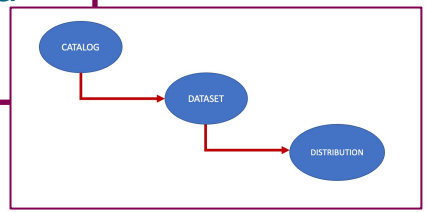
F3. Metadata clearly and explicitly include the identifier of the data they describe



F2. Data are described with rich metadata (defined by R1 below)

DataCite

F4. (Meta)data are registered or indexed in a searchable resource



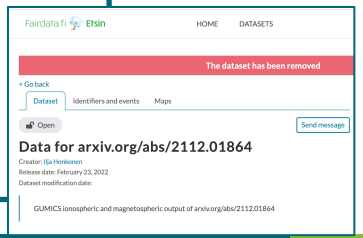
Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

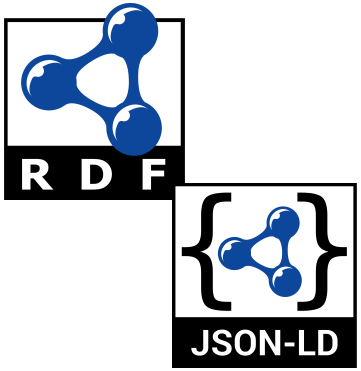
A2. Metadata are accessible, even when the data are no longer available



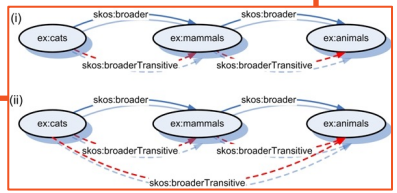
Interoperable

Semantic artefacts

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.



I2. (Meta)data use vocabularies that follow FAIR principles



I13. (Meta)data include qualified references to other (meta)data

Reusable

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

R1.2. (Meta)data are associated with detailed provenance.

CC-BY

R1.3. (Meta)data meet domain-relevant community standards.

Kansainvälinen ja kansallinen tutkimusdata

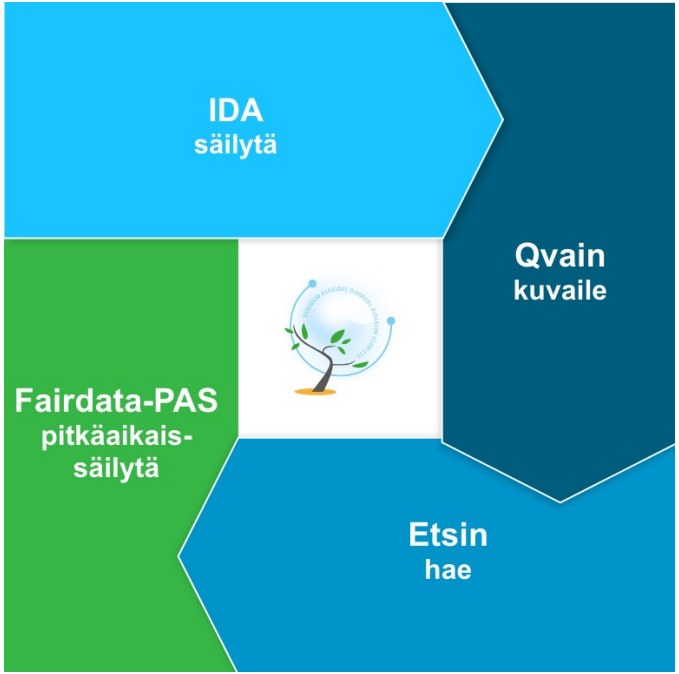
Kansalliset tutkimusinfrastruktuurit

- The Accelerator Laboratory of the University of Jyväskylä (JYFL-ACCLAB)
- ALD center Finland -research infrastructure for atomic layer deposition and etching
- Biobanking and Biomolecular Resources Research Infrastructure of Finland (BBMRI.fi)
- Biocenter Finland (BF)
- Bioeconomy Infrastructure (BIOECONOMY RI)
- CSC's Research Infrastructure Services
- Common Language Resources and Technology Infrastructure (FIN-CLARIAH)
- Earth-space research ecosystem (E2S)
- Euro-BioImaging: Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences (EuBI-Fi)
- European Infrastructure of Screening Platforms for Chemical Biology (EU-OS FI)
- European Life-Science Infrastructure for Biological Information (ELIXIR)
- European Plate Observing System (FIN-EPOS)
- European Social Survey (ESS)
- Finnish Biodiversity Information Facility (FinBIF)
- EuroFinnish Computing Competence Infrastructure (FCCI)
- The Finnish Infrastructure for Public Opinion (FIRIPO)
- Finnish Marine Research Infrastructure (FINMARI)
- Finnish National Infrastructure for Light-Based Technologies (FinnLight)
- Finnish Social Science Data Archive & CESSDA ERIC's Finnish Service Provider (FSD)
- Integrated Atmospheric and Earth System Science Research Infrastructure (INAR RI)
- Integrated Structural Biology Infrastructure (FinStruct & Instruct-ERIC Centre FI)
- Measuring Spatiotemporal Changes in Forest Ecosystem (Scan4estEcosystem)
- Otaniemi Micro- and Nanotechnology Research Infrastructure (OtaNano)
- Partnership for Advanced Computing in Europe (EuroHPC)
- Printed Intelligence Infrastructure (PII)
- RawMatTERS Finland Infrastructure (RAMI RI)
- Research Infrastructure for Future Wireless Communication Networks (FUWIRI)

SUOMEN AKATEMIA

Fairdata-palvelukokonaisuus

- Fairdata-kokonaisuus varmistaa tutkimuksen todennettavuuden, toistettavuuden, sekä turvaa tietoaineistojen pitkäaikaisen saatavuuden
- Sujuva siirtyminen eri palveluiden välillä
  - Kertakirjautuminen palvelukokonaisuuteen
  - Kuvailutietojen siirtyminen palveluiden välillä
- Kokonaisuuteen kuuluu
  - Säilytyspalvelu **IDA**
  - Kuvailupalvelu **Qvain**
  - Hakupalvelu **Etsin**
  - Pitkäaikaissäilytyspalvelu **Fairdata-PAS**

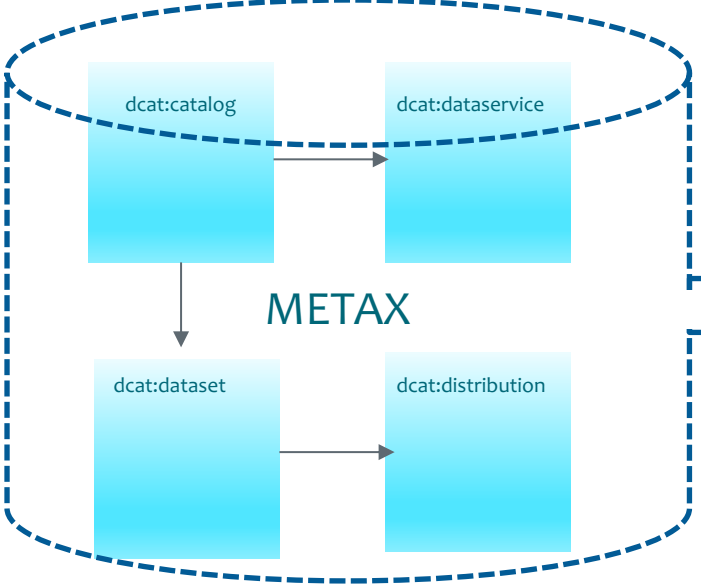


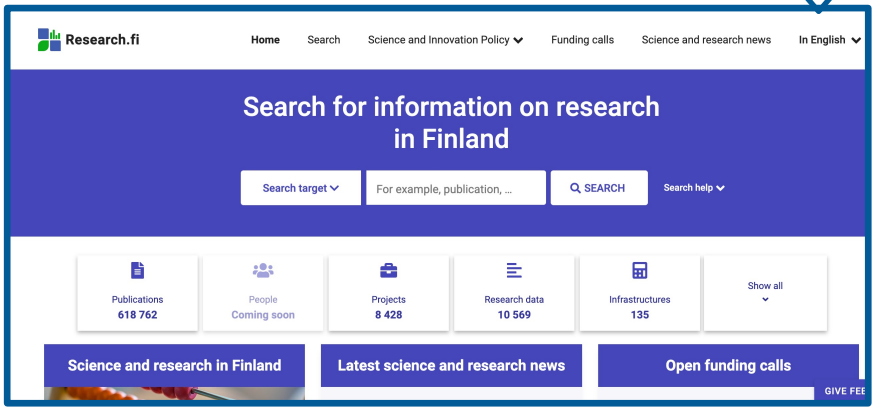
Metatiedot fairdata-palveluissa

- Metatiedoissa ei joko tai vaan sekä että, tärkeintä on tarkkuus ja kontekstin merkitseminen
- Metax hyödyntää mm DCAT, DataCite elementtejä sekä CreDIT-taksonomiaa
- Metax tukee linkitettyä dataa ja FAIR-periaatteita hyvin
  - Runsaasti mahdollisuuksia hyödyntää pysyviä tunnisteita
  - Tunnisteiden välille pystyy luomaan tyypiteltyjä relaatioita
  - Useita eri referenssimetatieto aineistoja integroituna

Metax ja tutkimustieto

- Metax pyrkii keräämään mahdollisimman paljon tunnisteita ja niiden välisiä relaatioita
- Tunnisteet ja referenssimetadatat kuten koodistot ja kuvailussa käytettävät käsitteet suunnitellaan yhdessä tutkimustietovarannon kanssa
- Näin varmistetaan koneluettavan metatiedon uudelleenkäytettävyyttä
- Tunnisteiden käyttö ja hallinta keskeistä





ESFRI & EOSC & SRIA

EUROPEAN OPEN SCIENCE CLOUD

SRIA Structure

- Section 1 - New ways of science
- Section 2 - EOSC in the making
- Section 3 - Strategic objectives of the European Open Science Cloud
- Section 4 - Guiding principles
- Section 5 - Implementation challenges
- Section 6 - Boundary conditions
- Section 7 - Expected impacts
- Section 8 - Roadmap
- Section 9 - Conclusions

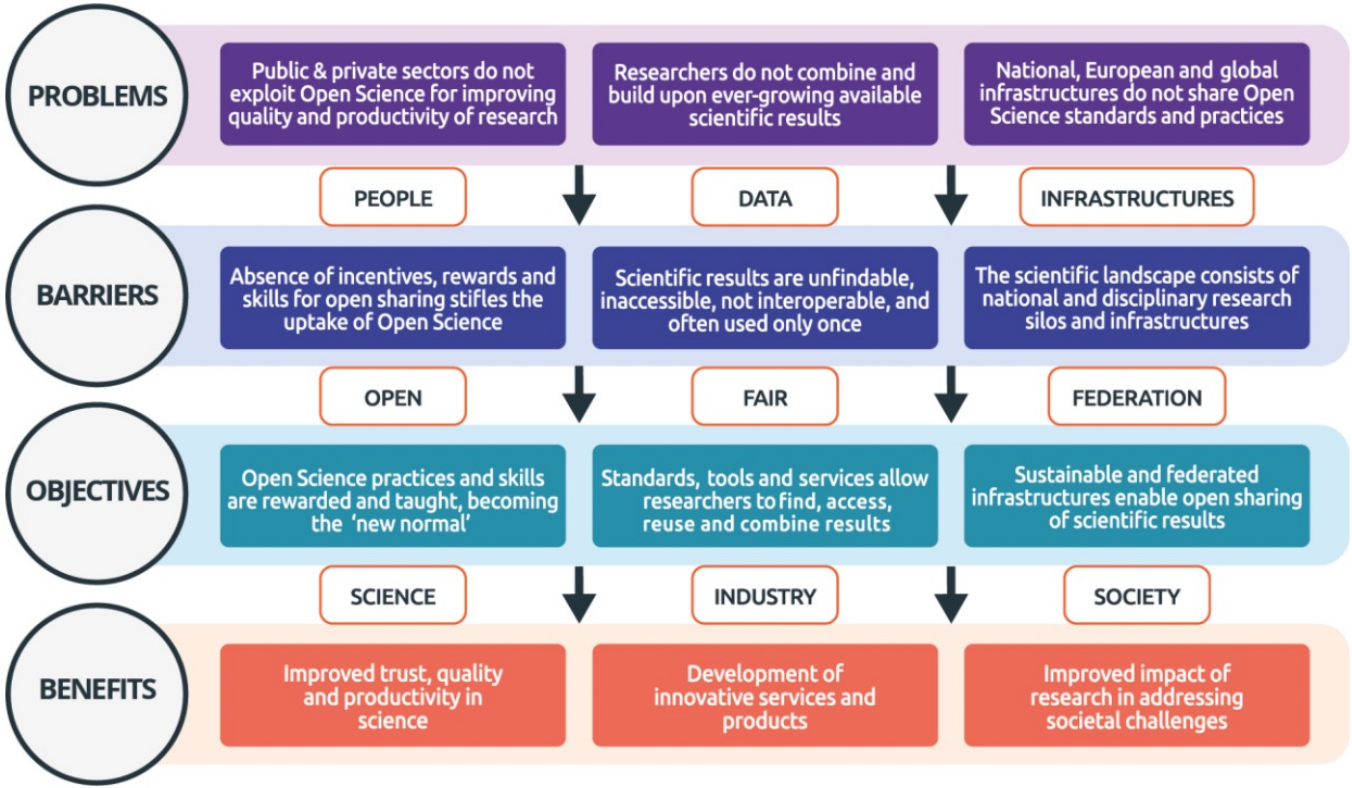


Figure 0.1: European Open Science Cloud Objectives Tree

https://www.eosc.eu/sria

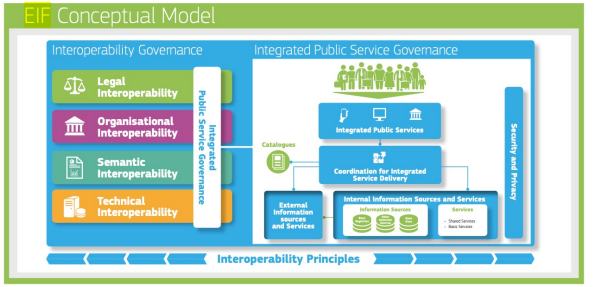
Ohjaavat periaatteet

- Sidosryhmien osallisuus
- Avoimuus: Niin avoin kuin mahdollista, niin suljettu kuin välttämätöntä
- FAIR-periaatteet: tieteestä läpinäkyvää ja toistettavaa
- Infrastruktuurien federointi
- Avoimen tieteen palvelut
- Politiikkasuositukset tueksi

Implementoinnin haasteista uusia hankkeita 2022

- Tunnisteet
- Metatiedot ja ontologiat
- FAIR-metriikka ja sertifiointi
- Käyttäjä- ja pääsynhallinta
- Ympäristöt käyttäjille
- Palvelutarjoajien ympäristöt
- Yhteentoimivuus eri tasoilla (EOSC Interoperability Framework)

Yhteentoimivuuden tasot



EIF

Organisational interoperability

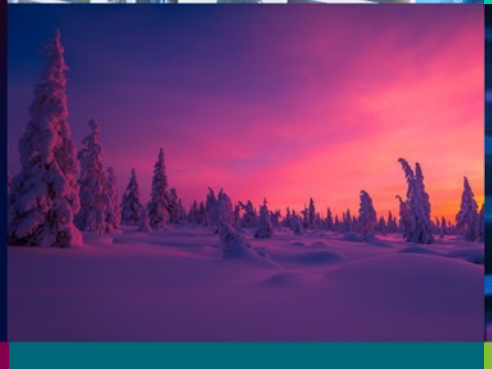
Semantic interoperability

Legal interoperability

Metadata

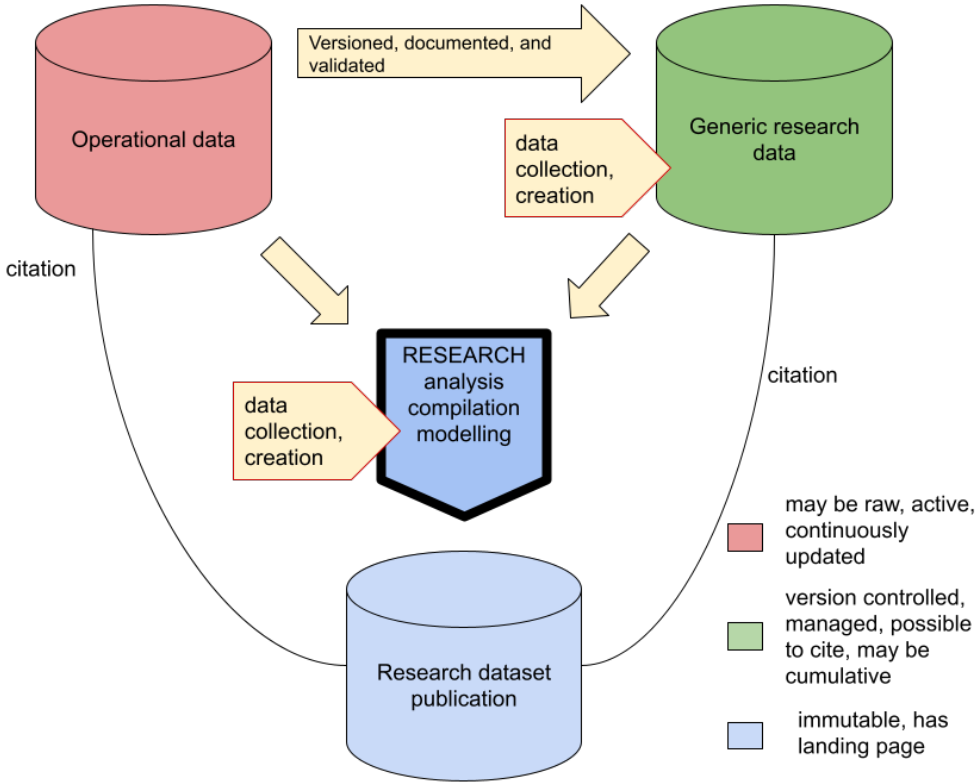
Technical interoperability

Tutkimuksen digitaaliset aineistot

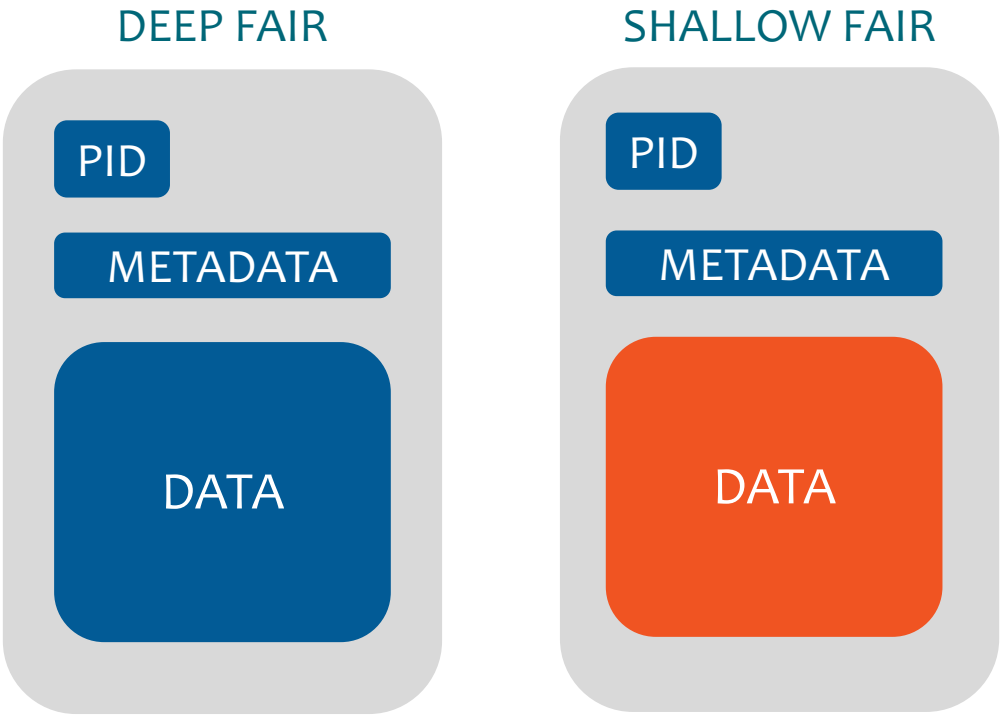


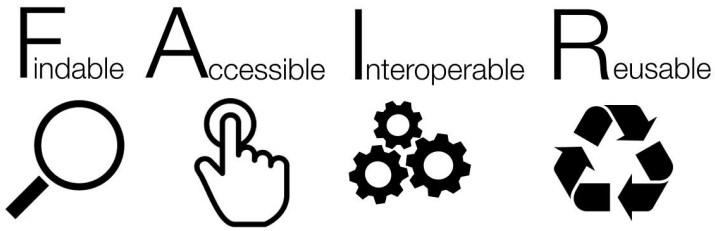
Aineiston elinkaari

Uudenlaisia aineistotyyppejä

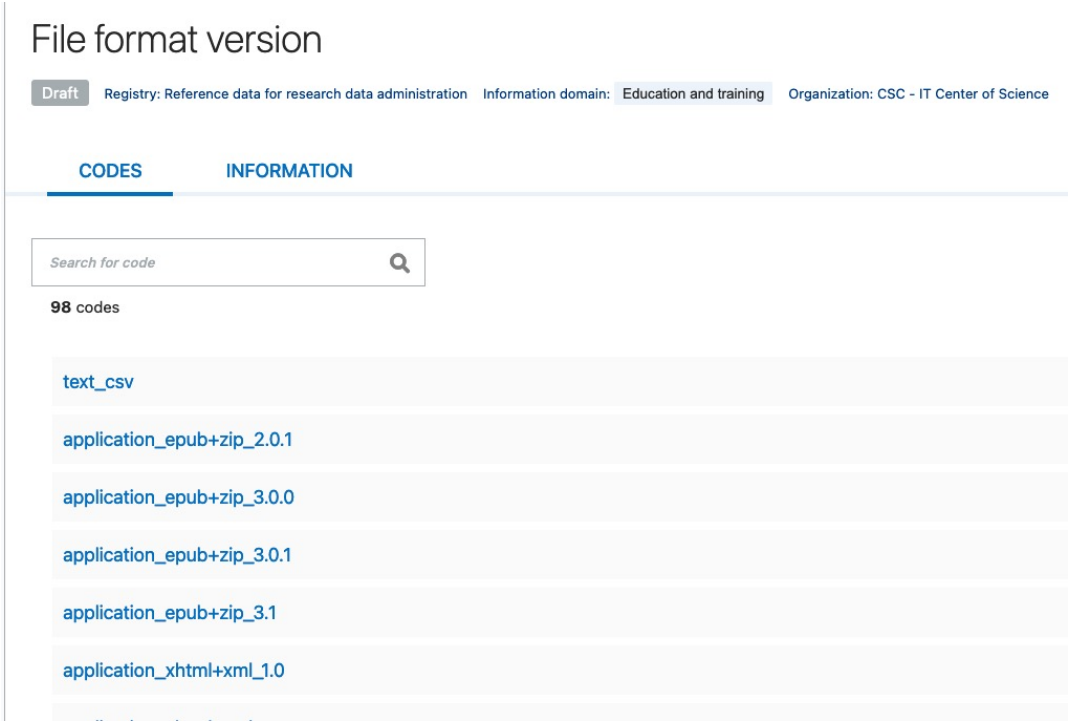


Koneluettavuus ja viittaaminen





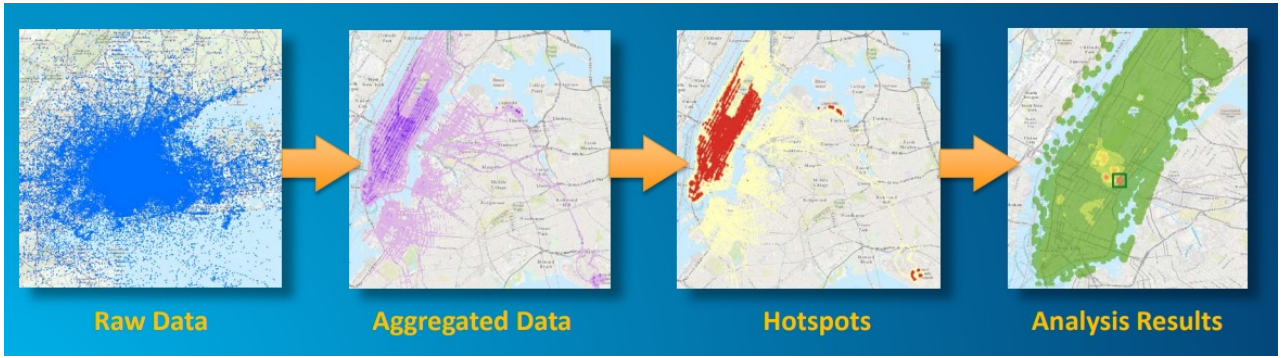
Suuri määrä tiedostoformaatteja



Kompleksinen data

Finnish Digital Agency	Addresses of buildings		14.8.2020	SHAPE	ETRS-TM35FIN	Download	Description
Finnish Digital Agency	Addresses of buildings		15.8.2016	SHAPE	ETRS-TM35FIN	Download	Description
Finnish Digital Agency	Addresses of buildings		15.8.2018	SHAPE	ETRS-TM35FIN	Download	Description
Finnish Food Authority	Field plots	1:5 000 - 1:250 000	2012	SHAPE	YKJ	Download	Description
Finnish Food Authority	Field plots	1:5 000 - 1:250 000	2016	SHAPE	ETRS-TM35FIN	Download	Description
Finnish Meteorological Institute	Daily global radiation, 10km	10 km x 10 km	1961-2019	TIFF	ETRS-TM35FIN	Download	Description
Finnish Meteorological Institute	Daily global radiation, 10km	10 km x 10 km	1961-2019	NetCDF	ETRS-TM35FIN	Download	Description
Finnish Meteorological Institute	Daily maximum temperature predictions	10 km x 10 km	1981-2100	NetCDF	WGS84/ETRS-TM35FIN	Download	Description
Finnish Meteorological Institute	Daily maximum temperature, 10km	10 km x 10 km	1961-2019	NetCDF	ETRS-TM35FIN	Download	Description

https://paituli.csc.fi



https://www.qualitylifemag.com/case-studies-big-data-analytics-gis/

Uudenlaisia tekijyyksiä: CRediT Taxonomy

Fairdata.fi

http://credit.niso.org/

Uudenlaisia rooleja: DataCite Contributor Types

Fairdata.fi

https://schema.datacite.org/

Yhteentoimivuus ja toistettavuus

FAIR-ohjeiden *Interoperable*-kohta sisältää kolme eri periaatetta, jotka keskittyvät semanttiseen yhteentoimivuuteen.

I1 Aineistot ja metatiedot ovat sisällöltään määrämuotoisia, monikäyttöisiä, saatavilla olevia ja jaettua kieltä käyttäviä

I2. Aineistoissa ja metatiedoissa käytetään sanastoja, jotka noudattavat FAIR-periaatteita

I3. Aineistoissa ja metatiedoissa on tyypiteltyjä viittauksia muihin resursseihin

https://wiki.eduuni.fi/x/TZTfDg

Semanttisen yhteentoimivuuden edistämisestä

Datan yhteentoimivuutta voidaan edistää ennen kaikkea hyvällä aineistonhallinnan suunnittelulla ja hyödyntämällä yhteisiä standardeja, käytäntöjä ja ratkaisuja.

Datanhallinnan asiantuntijan ja kouluttajan tehtävä on tutkijoiden tukeminen oikeiden formaattien, palveluiden ja kontrolloitujen sanastojen löytämisessä.

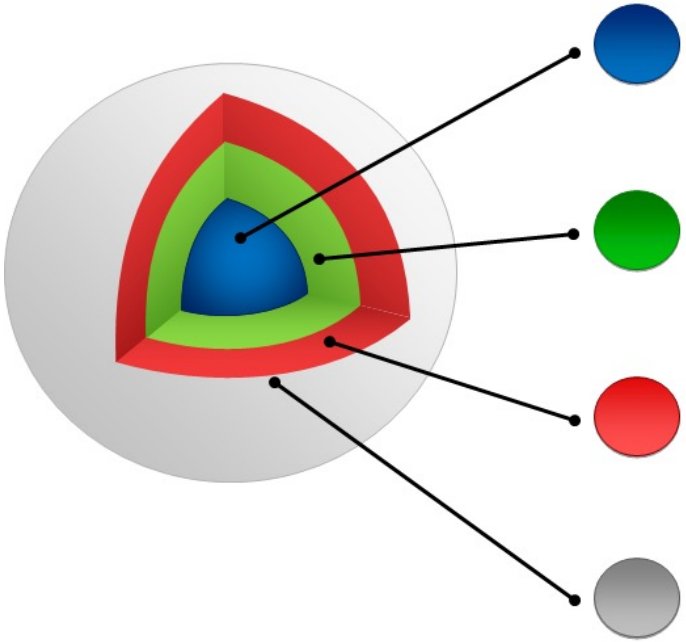
Data-asiantuntijan tulee ymmärtää metadatan, tunnisteiden ja elinkaarenhallinnan merkitys tutkimuksen toistettavuudelle sekä dataan viittaamisen käytännöt, ottaen huomioon viittaamisen.

https://wiki.eduuni.fi/x/TZTfDg

Tutkimuksen erilaiset tuotokset

- Metatietojen pitäisi sisältää linkkejä muihin objekteihin
- Tavoitteena mahdollisimman koneluettava, toiminnallinen dokumentaatio, esim. https://www.commonwl.org/
- Koodin dokumentoinnin RDA-suositus
  - Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A-L, Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., et al. (2022). **FAIR Principles for Research Software version 1.0. (FAIR4RS Principles v1.0).** Research Data Alliance. DOI: https://doi.org/10.15497/RDA00068
- Muita hyviä ohjeita esim. toiminalliset julkaisut: https://the-turing-way.netlify.app/welcome.html
- Kansallinen luonnos tutkimusmenetelmien ja -infrastruktuurien avoimen saatavuuden osalinjauksesta on avattu kommentoinnille 12.4.–24.5.2022.
  - https://avointiede.fi/fi/ajankohtaista/tutkimusmenetelmien-avoimuuden-osalinjausluonnos-kommentoitavana

The FAIR Digital Object



DATA

The core bits

At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve . Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

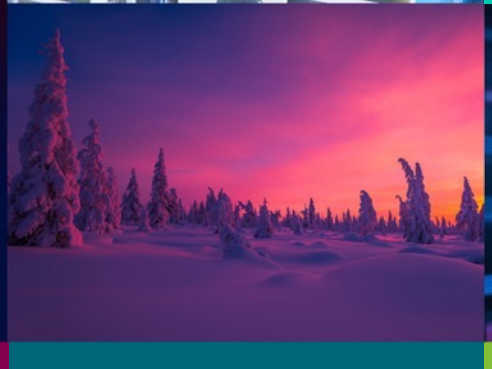
In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

Toistettavuuden varmistaminen

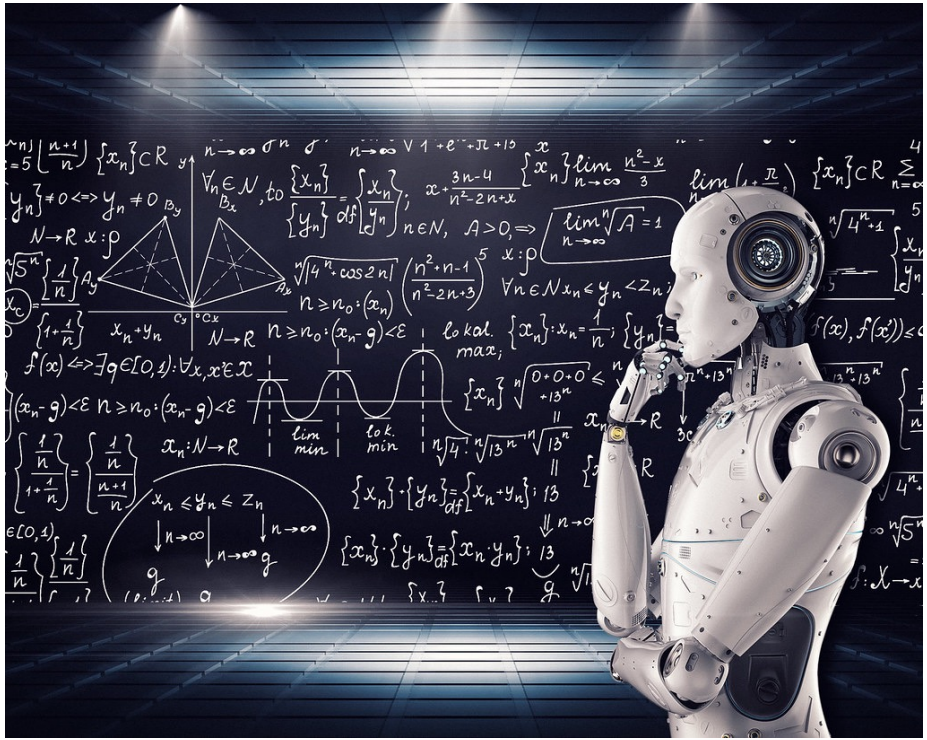
1. **Completeness:** The research compendium contains all of the objects needed to reproduce a predefined outcome.
2. **Organization:** It is easy to understand and keep track of the various objects in the research compendium and their relationship over time.
3. **Economy:** Fewer extraneous objects in the compendium mean fewer things that can break and require less maintenance over time.
4. **Transparency:** The research compendium provides full disclosure of the research process that produced the scientific claim.
5. **Documentation:** Information describing compendium objects is provided in enough detail to enable independent understanding and use of the compendium.
6. **Access:** It is clear who can use what, how, and under what conditions, with open access preferred.
7. **Provenance:** The origin of the components of the research compendium and how each has changed over time is evident.
8. **Metadata:** Information about the research compendium and its components is embedded in a standardized, machine-readable code.
9. **Automation**: As much as possible, the computational workflow is script- or workflow-based so that the workflow can be re-executed using minimal actions.
10. **Review:** A series of managed activities needed to ensure continued access to and functionality of the research compendium and its components for as long as necessary

https://www.rd-alliance.org/group/cure-fair-wg/outcomes/10-things-curating-reproducible-and-fair-research

Mitä seuraavaksi?



- Vahva usko koneisiin totuuden kaitsijoina
- Pysyvyyden ongelma, alkuperäisen ongelma
- Koneluettavuus ja tekoäly, mitä voi ratkaista ja miten voimme auttaa koneoppmista?
- Sanastojen ylläpito on aina käsityötä
- Manuaalinen, narratiivinen dokumentaatio ihmisille tärkeää kontekstia ja tulkintatietoa
- Kattavien kuvailuohjeiden puute monella alalla haaste yhteentoimivuudelle



"Artificial Intelligence & AI & Machine Learning" by mikemacmarketing is marked with CC BY 2.0.

Jessica PvE
@jpve

Tutkimusdatanhallinnan osaamiskeskus

Datatukiverkosto

Tutkimusdatanhallinnan tukipalvelut

Fairdata-verkosto

PID-palvelut

facebook.com/CSCfi

twitter.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi

asiakaspalvelu@csc.fi

Kuvat CSC:n arkisto, Adobe Stock ja Thinkstock