TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Metadatakatalogit; Yhteiskuntatieteet, esimerkkinä Tietoarkisto ja CESSDA Data Catalogue

Laadukas metadata - tutkimusaineistojen kuvailu ja semanttinen yhteentoimivuus

26.4.2022

Tuomas J. Alaterä, Tietoarkisto

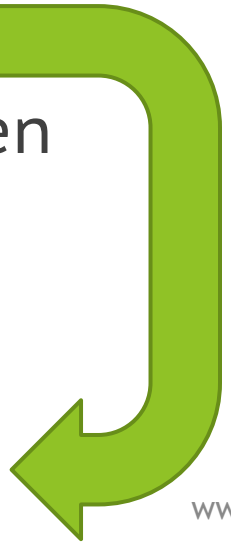
Metadatakatalogi yhteiskuntatieteissä

- Yhteiskuntatieteellisen datan saatavuus on usein käyttötarkoitusperusteella tai tietosuojasyistä rajattu
  - Metadata on lähes aina avointa ja/tai avoimesti lisensoitua
  - Metadata on yleensä aina(kin) DDI-formaatissa (study/variable level)
  - Metadata on yksityiskohtaista, usein muuttujatasoista
  - Metadataa voidaan uudelleenkäyttää ja käsitellä automaattisesti
- Datan löydettävyys ja hyödynnettävyys perustuu metadatan löydettävyyteen ja kattavuuteen metadatakatalogeista
  - Datan laadun, tarkoituksenmukaisuuden, kontekstin arviointi ja haut
  - Pitkä perinne: yhteinen metadaformaatti ja kansalliset data-arkistot

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Tietoarkisto ja CESSDA

- Tietoarkisto on erityisesti SSH-alojen tutkimuksen ja opetuksen palveluinfrastruktuuri, joka arkistoi ja välittää tutkimusaineistoja uudelleenkäytettäväksi
  - FAIR-periaatteet keskeisiä palvelun ja ”FAIRiyden” ylläpitämiseksi
  - Metadatan tuottaminen ja aineistojen kuratointi keskeisiä
  - Aineistoluettelo (Aila)
- CESSDA ERIC on eurooppalaisten yhteiskuntatieteellisten data-arkistojen muodostama tutkimusinfrastrutuuri
  - Suomen palveluntuottaja Tietoarkisto
  - Tutkimusdata käyttöön riippumatta tutkijan tai datan paikasta
  - Aggregoiva aineistoluettelo CESSDA Data Catalogue



Tutkimusaineiston kuvailu

- Hyvä tutkimusdatan kuvailu kuvaa tutkimusaineistoa ja sen muodostamista, ei tutkimuksen tuloksia
  - Miten, milloin ja missä tutkimus toteutettiin?
  - Mikä oli tutkimuksen perusjoukko, miten otanta muodostettiin?
  - Mikä/mitkä olivat datankeruun instrumentit?
  - Datatiedostojen ja rakenteen kuvailu
  - Muuttujien kuvailu
  - Miten aineistoa on käsitelty arkistointiprosessissa?
  - Miten tutkimusaineisto on saatavilla/lisensoitu?
  - Muu kontekstoiva tieto, paradata, itse kuvailun tiedot
- Data ja metadata matkaavat yhdessä, rakenteisessa muodossa

DDI (Data Documentation Initiative)

- DDI Codebook (2.x) datan pitkäaikaissaatavuuden metatiedoille (Tietoarkisto)
  - https://ddialliance.org/Specification/DDI-Codebook/2.5/
  - https://doi.org/10.25504/FAIRsharing.EZCpPd
- DDI Lifecycle (3.x) datan koko elinkaaren hallintaan
- DDI-C käsittää n. 300 elementtiä
  - Document Description
    - Itse kuvailun perustiedot, tunniste ja lisenssi
  - Study Description
    - Datan kuvailu, mm. tekijät, asiasanat, abstrakti, otanta, keruun kuvaus, havaintoyksiköt, perusjoukko, saatavuus, käyttöehdot
  - Data Files Description
    - Tiedostojen ja rakenteen kuvailu, mm. formaatti, muuttujien ja havaintojen määrä, koko
  - Variable Description
    - Muuttujien kuvailu, muuttujien arvot ja selitteet, kysymystekstit
  - Other Study-Related Material

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Metadatan muotoutuminen Tietoarkistossa

- Datankäsittelijät: datan prosessointi ja mahdollinen anonymisointi → kuvailu käyttäen DDI-standardia ja kontrolloituja sanastoja yhteentoimivuuden takaamiseksi
  - YSO (https://finto.fi/yso/fi/)
  - DDI Controlled Vocabulary (https://vocabularies.cessda.eu/)
    - https://doi.org/10.25504/FAIRsharing.y4RpVy
  - CESSDA Topic Classification – datan aihepiiriluokitus (https://vocabularies.cessda.eu/vocabulary/TopicClassification)
  - Tieteenalaluokitus (https://finto.fi/okm-tieteenala/fi/)
  - ELSST (European Language Social Science Thesaurus) 14 kieltä (https://thesauri.cessda.eu/elsst/en/)
    - https://fairsharing.org/581 (Awaiting DOI)
  - Käyttöesimerkit Aineistonhallinnan käsikirjassa https://www.fsd.tuni.fi/fi/palvelut/aineistonhallinta/sanastot/

CESSDA Data Catalogue

- Kontrolloituihin sanastoihin perustuva metatieto:
  - Mahdollistaa tarkat haut ja filtteröinnit suuresta aineistomassasta
  - Lisää ymmärrettävyyttä, vertailtavuutta ja yksikäsitteisyyttä
  - Mahdollistaa kielirajat ylittävät monikieliset haut (termit / käsitteet)
  - Liittyvien datasettin tarjoaminen
  - Ohjaa dataa säilyttävän arkiston kattavampaan kuvaukseen (PID)
- Koneluettavuus, tekninen ja semanttinen yhteentoimivuus

cessda
DC Data Catalogue

wages

English

Filter summary | Reset filters | Clear search

9 studies found in English from a total of 40735

User Guide | About

Topic

× Health care services and policies (9)

Collection years

1900 — 2022

2010 - 2022 Go

Country

Publisher

Results per page 30

Sort by Relevance

1

ARD-DeutschlandTrend 2020

ARD-Landesrundfunkanstalten; Infratest dimap Gesellschaft für Trend- und Wahlforschung, Berlin

Since 1997 the ARD-DeutschlandTREND is being conducted on behalf of the ARD (Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland - First German Public Broadcasting Association) as well as various print media by Infratest dimap. The monthly telephone survey with approx. 1,000 respondents (for party preferences approx. 1,500 respondents) per wave is based on representative samples and measures attitudes of the voting-age population in the Federal Repu...

Read more

Study description available in: DE EN Access data

ARD-DeutschlandTrend 2020

ARD-Landesrundfunkanstalten; Infratest dimap Gesellschaft für Trend- und Wahlforschung, Berlin

Since 1997 the ARD-DeutschlandTREND is being conducted on behalf of the ARD (Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland - First German Public Broadcasting Association) a

Send feedback

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

”FAIR” metadata Tietoarkistossa

From a FAIR perspective, metadata are more important than (your) data, because metadata would always be openly available and they link research data and publications [..]

While data documentation is meant to be [..] understood by humans, metadata [..] are primarily meant to be processed by machines.

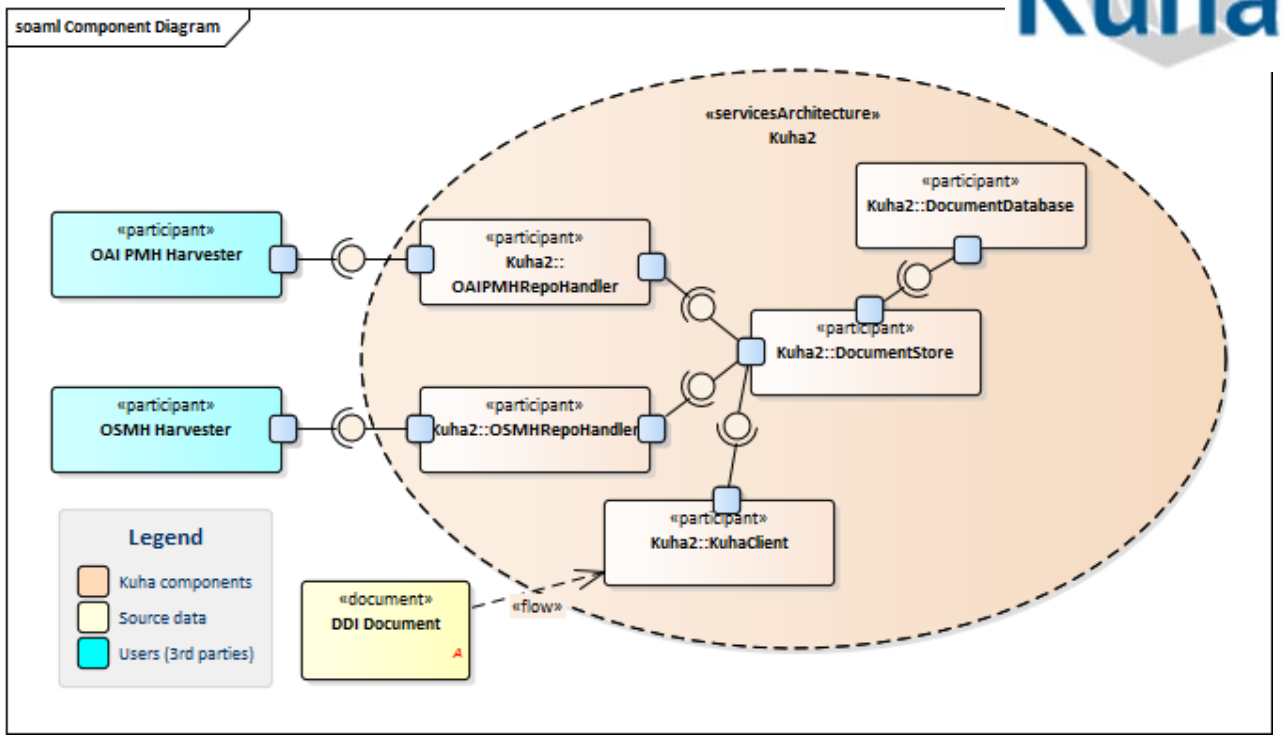
www.howtofair.dk

- Datan perustiedot JSON-LD -muodossa, kattavat XML-muodossa
  - schema.org–datatyypit, koneluettevat lisenssit, tunnisteet …
  - Jatkossa: signposting

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Haravointi CDC:hen – ja muualle

- CDC, Etsin, tiedejatutkimus.fi, Finna, Data Citation Index
  - Erilaisia haravointitoteutuksia, erilaisia formaatteja
  - Kuha2 OAI-PMH
    - DDI, DC, EAD
  - DDI XML mahdollistaa muiden metadataformaattien tuottamisen ”master-metadatasta
    - Vähentää virhemahdollisuuksia
    - Helpottaa päivitystä ja versiointia
    - Mahdollistaa automatisointia (ja rikastamista)
  - Koneluettava metadata on silti vain yhtä hyvää, oikein tai yhteentoimivaa kuin alkuperäinen



Metadatan kääntäminen

- Vaatii ymmärrystä aineistonkeruusta, käsittelystä ja analyysistä
- Kontrolloidut sanastot pohjana
- Perusta sille, että data Suomesta tai suomalaisista on käytettävissä kansainvälisesti – ei vain julkaisut

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Dublin Core <header>

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.title" content="FSD0000 Suomen FAIR Barometri 2022" />
<meta name="DC.creator" content="Jaska Jokunen" />
<meta name="DC.creator" content="Fairsingin yliopisto" />
<meta name="DC.publisher" content="Finnish Social Science Data Archive" />
<meta name="DC.type" content="Dataset" />
<meta name="DC.format" content="application/zip" />
<meta name="DC.language" content="fi" scheme="DCTERMS.RFC4646" />
<meta name="DC.date" content="2022-01-01" scheme="DCTERMS.W3CDTF" />
<meta name="DC.identifier" content="http://urn.fi/urn:nbn:fi:fsd:T-FSDOOOO" scheme="DCTERMS.URI" />
<meta name="DC.rights" content="info:eu-repo/semantics/restrictedAccess" scheme="DCTERMS.URI" />
<meta name="DCTERMS.license" content="https://creativecommons.org/licenses/by/4.0/"
scheme="DCTERMS.URI" />
<meta name="DCTERMS.publisher" content="Finnish Social Science Data Archive">
<meta name="DCTERMS.modified" content="2022-04-26">
...
```

Typed links / Signposting <header>

https://signposting.org/FAIR/

<link rel="cite-as" href=" http://urn.fi/urn:nbn:fi:fsd:T-FSDOOOO" />

<link rel="type" href="https://schema.org/CreativeWork" />
<link rel="type" href="https://schema.org/Dataset" />
<link rel="type" href="https://ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/" />

<link rel="author" href="https://orcid.org/0000-0000-0000-0000" />
<link rel="author" href="https://ror.org/040af2s02" />

<link rel="license" href="https://creativecommons.org/licenses/by/4.0/" />

<link rel="item" type="application/zip" href="https://services.fsd.tuni.fi/catalogue/FSD0000?tab=download" />
<link rel="item" type="application/pdf" href="https://services.fsd.tuni.fi/catalogue/FSD0000/PIP/cbF0000e.pdf" />

<link rel="describedby" type="application/ld+json" href="/fsd0000.json" />
<link rel="describedby" type="application/xml" href="fsd0000.xml" />

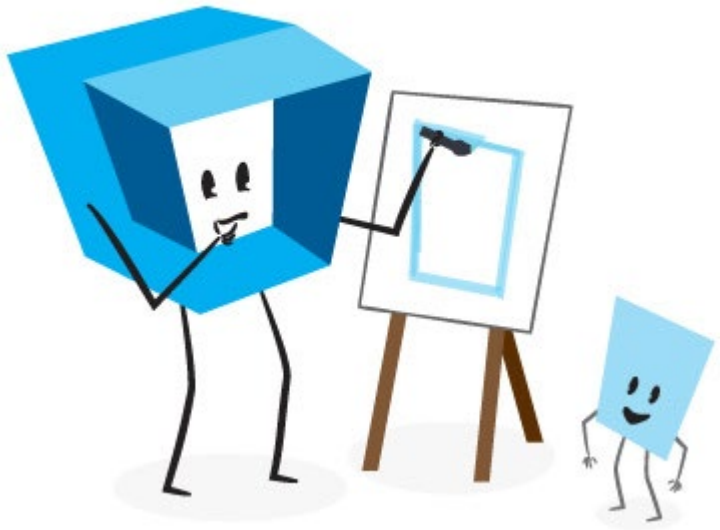
Upotettu / linkitetty JSON-LD

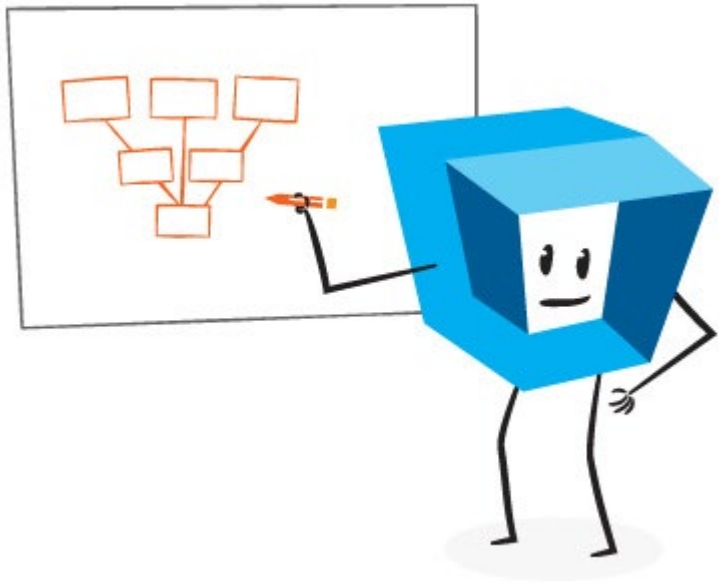

```
{
  "@context": {
     "@vocab": "https://schema.org/"
     },

  "@type": "Dataset",

  "@id": " http://urn.fi/urn:nbn:fi:fsd:T-FSDOOOO ",

  "name": [
     {
        "@value": " FSD0000 Finnish FAIR Barometer 2022 ",
        "@language": "en"
     },
     {
        "@value": " FSD0000 Suomen FAIR-barometri 2022 ",
        "@language": "fi"
     }
  ],
     ...
```

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Kysymyksiä tai keskustelua

https://www.fsd.tuni.fi/





Tuomas J. Alaterä
tuomas.alatera at tuni.fi
ORDIC 0000-0002-3448-344