TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Kontrolloidut sanastot tutkimusaineistojen kuvailussa

Erityisasiantuntija Taina Jääskeläinen, Tietoarkisto

Miksi sanastoja kannattaa käyttää?

- Koneluettavuus (machine actionability)
- Eri organisaatioiden metadatan tekninen ja semanttinen yhteensopivuus, kielten yli
- Rajausmahdollisuudet hakuliittymissä
- Metadatan ymmärrettävyys, keskinäinen vertailtavuus ja johdonmukaisuus
- Datan etsijöille sanastotermit antavat nopean kuvan aineiston pääpiirteistä
- Helpottaa ja nopeuttaa aineistojen kuvailua

Miksi sanastoja kannattaa käyttää?

- Metadataa menee myös oman organisaation ulkopuolella
- Yhteisluettelot, varsinkin jos monta kieltä ja maata, kolme vaihtoehtoa:
  - Standardoitu metadata sanastoineen
  - Metadata ’puhdistetaan’ yhteisluettelon päässä, usein joko algoritmeilla tai pudottamalla pois ei-standardisoidut tiedot
    - Vaatii resursseja, algoritmien tulokset eivät täydellisiä
  - Sekalainen metadata, käyttäjä voi olla hukassa
- Tieteenalan/metadatastandardin sanastot

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

ISO-standardit

- Päivämäärätiedot
  - ISO-standardien mukaiset päivämäärät sallivat rajaukset esim. keruuajan mukaan tai aineistojen lajittelun julkaisupäivämäärän mukaan (uusimmat ensin).
- Maatieto
  - ISO 3166 kaksinumeroinen maakoodi ’FI’, ’SE’ jne.
  - Sallii yhteisluetteloissa rajausmahdollisuuden keruumaan mukaan
  - Metadatan maanimissä variaatioita, esim. The U.K., The UK, United Kingdom, GB, Great Britain. Vaikka Suomen nimessä ei, yhteisluettelojen filtterit eivät toimi ilman ISO-koodia tai algoritmeja, variaation vuoksi metadatassa olevaan maannimeen ei voi luottaa.
- Kielikoodit
  - ISO 639-1 metadatakuvailuille. Jos monikielinen luettelo, metadatan kielikoodit sallivat käyttäjän valita haluamansa kuvailukielen.
  - Itse datan kieli voi olla eri ja voi sisältää historiallisia tai paikallisia kieliä. Eri ISO-kielikoodistandardien sijaan voi IANA – Language Subtag Registry olla hyödyllisempi. Se sisältää kaikki eri ISO-standardit.

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

DDI-sanastot

- Yhteiskuntatieteellisten data-arkistojen eurooppalainen infrastruktuuri CESSDA käyttää DDI-metadatastandardia sekä sen sanastoja.
- Lyhyt esittely joistakin yleisemmin käytetyistä DDI-sanastoista. Monet näistä ovat saatavina usealla eri kielellä.
- Sopivia myös mm. terveystieteisiin.
- Tutkimusmenetelmien luokittelua. Termeillä yleensä määritelmät.
- Tutkijat pystyvät näiden avulla myös nopeasti kuvailemaan tutkimusaineistonsa luonteen.
- Linkit lisätty sanaston nimeen otsikkoon.

Keruumenetelmä (ModeOfCollection)

- Kuvailtavan aineiston ensisijainen keruumenetelmä (noin 50 termiä)
  - Jos eri menetelmiä käytössä, kuvaillaan kukin omalla termillään.
  - Jos tarkka menetelmä ei tiedossa, voi käyttää laajempaa termiä. Laajempia termejä käytetään usein filttereissä.
- Haastattelu (esim. kasvokkainen, tietokoneavusteinen, puhelin, verkkohaastattelu)
- Itsetäytettävä lomake (esim. paperi tai verkko)
- Oma kirjoitus tai päiväkirja
- Havainnointi
- Kontrolloitu koe, Taltiointi, Automatisoitu tiedonpoiminta
- Sisällönkoodaus (kvalista kvanti), Kooste- tai synteesiaineisto, Aggregointi, Simulointi jne.

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Tutkimuksen aikaulottuvuus (TimeMethod)

- Sanastossa aikaulottuvuus käytännölliseltä kannalta.
- Voiko aineistoa käyttää pitkittäistutkimukseen eli toistuvatko ainakin osin samat kysymykset/moduulit/mittaukset yms?
- Ei välttämättä samat vastaajat mutta sama populaatio tai kohortti.
- Pitkittäisaineisto (kohortti/tapahtuma-aineisto, trendiaineisto, paneeliaineisto)
- Aikasarja-aineisto (jatkuva, diskreetti)
- Poikkileikkausaineisto, Poikkileikkausaineisto: täydennys/seurantakeruu

Otantamenetelmä (SamplingProcedure)

- Sallii filtterit, joissa erotellaan todennäköisyysotannat muista otoksista. Tarjoaa menetelmäluokituksen, tarkemmat tiedot otannasta tai otoksesta vapaatekstinä.
- Kokonaisaineisto
- Todennäköisyysotannan eri menetelmät: satunnaisotanta, systemaattinen/ositettu/suhteutettu ositettu/suhteuttamaton ositettu/ryväsotanta/ositettu ryväsotanta/monivaiheinen todennäköisyysotanta.
- Ei-todennäköisyysotantamenetelmät: itsestään muotoutunut näyte/harkinnanvarainen poiminta/kiintiöpoiminta/osallistuja-avusteinen poiminta
- Hyvät määritelmät, avuksi myös tutkijoille kuvaamaan omassa tutkimuksessaan käytettyä otantaa

Havainto/aineistoyksikkötyyppi (AnalysisUnit)

- Havaintoyksikkö, jota tutkimuksessa havannoidaan tai josta aineistoa kerätään. Huom: voi olla eri kuin tutkimuksen analyysiyksikkö.
- Samassa tutkimuksessa voi olla useampia esim. sekä Henkilö että Perhe tai Henkilö ja Organisaatio.
- Henkilö / Organisaatio / Perhe / Kotitalous
- Tapahtuma/prosessi/toiminta
- Maantieteellinen yksikkö / Poliittis-hallinnollinen alue / Aikajakso / Ryhmä / Eloton objekti/esine
- Mediatyyppi / Mediatyyppi: ääni / Mediatyyppi: still-kuva / Mediatyyppi: teksti / Mediatyyppi: liikkuva kuva

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Aineiston formaatti (GeneralDataFormat)

- Tutkimusaineiston fyysinen formaatti eli esim. mitä formaattia aineistotiedostot ovat, mitä jatkokäyttäjä saa aineiston ladatessaan.
- Numeerinen / Teksti
- Kuva / Audio / Video
- Paikkatieto / Ohjelmisto / Vuorovaikutteinen / 3D
- DDIn sanastotyöryhmä ottaa mielellään vastaan ehdotuksia, jos joku tuntuu puuttuvan.

Aineiston alkuperä (DataSourceType)

- Aineiston alkuperän tyyppiluokittelu. Käytetään jos tutkimusaineisto perustuu jo aiemmin olemassa olevaan aineistoon.
- Rekisterit/asiakirjat ja niiden eri alatyypit: esim. hallinnolliset, historialliset, juridiset, lääketieteelliset/kliiniset, taloudelliset, oppimis-, soveltuvuus- ja kykytestit, henkilökohtainet dokumentit, äänestystulokset
- Viestintä: esim. julkinen tai henkilöiden välinen
- Tutkimusaineisto (olemassa oleva): julkaistu/julkaisematon
- Väestöryhmä, Maatieteellinen alue
- Fyysiset objektit (esim. aineistot jossa kuvaillaan taideteosten tai kivinäytteiden fyysisiä ominaisuuksia)
- Biologiset näytteet

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Keruuväline (TypeOfInstrument)

- Tutkimusaineiston muodostamisessa käytetyt keruuvälineet, -ohjeistukset tai –suunnitelmat. Karkea tyyppiluokitus.
- Lomake ja sen eri tyypit (strukturoitu, puolistrukturoitu, strukturoimaton)
- Haastatteluteemat tai –runko
- Aineistokeruuohjeistus
- Osallistujatehtävä
- Tekniset keruuvälineet objektiivisten havaintojen keräämiseen
- Ohjelmointikoodi

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

ContributorRole

- Ei suomennosta tällä hetkellä, Tietoarkisto saattaa tehdä.
- Hieman laajempi sanasto kuin DataCiten ’contributorType’, jota esim. Open Aire käyttää laajennettuna muutamilla CReditT:n (Contributor Roles Taxonomy) käsitteillä. Qvain.
- Määritelmät sisältävät mäppäykset DataCiten käsitteisiin milloin mahdollista.
- Ovatko DataCiten tai CRediT –sanastot suomennettu?

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Asiasanastot Tietoarkistossa: YSO ja ELSST

- Sanastojen käytön etu:
  - Ei variaatioita kuten asumisoikeusasunto/Asumisoikeusasunnot, Biodiversiteetti/Biologinen monimuotoisuus
  - Sallii asiasanahaun tai sen mukaisen listauksen
- Asiasanat
  - Suomenkieliset kuvailut: YSO
  - Englanninkieliset: ELSST- asiasanasto, sisältää 16 kieltä
  - CESSDA käyttää ELSSTiä → englanninkielisissä kuvailuissa, noin 3300 käsitettä, sisältää lähinnä yhteiskunta- ja terveystieteiden käsitteitä.
    - Tutkijoille voi olla hyödyksi englanninkielisten asiasanojen löytämiseen hakemalla suomeksi ja katsomalla enkkukäsitteen.

TIETOARKISTO
FINNISH SOCIAL SCIENCE DATA ARCHIVE

CESSDAn ja Tietoarkiston käyttämät kv-sanastot

- Löytyvät CESSDA Vocabulary Service –palvelusta kieliversioineen paitsi asiasanastot. Sanastojen ylläpito ja kääntäminen työkalussa.
- Sanastoissa on ‘Muu’ –vaihtoehto mikäli käsite ei ole sanastossa.
- Termiehdotuksia tai kommentteja voi lähettää esim. ‘Send feedback’ nappulaa käyttäen (oikea alareuna) tai lähettämällä Tainalle meiliä.
- Export/download –valikossa voi valita latausmuodon (SKOS, PDF, HTML) ja halutun/halutut kielet.
- Rest API –linkki alareunassa.
- Sanastotyökaluun voi laittaa myös CESSDAn ja DDI:n ulkopuolisia sanastoja. Vaatimuksena on, että sanastolla on sitoutunut ylläpitäjä.

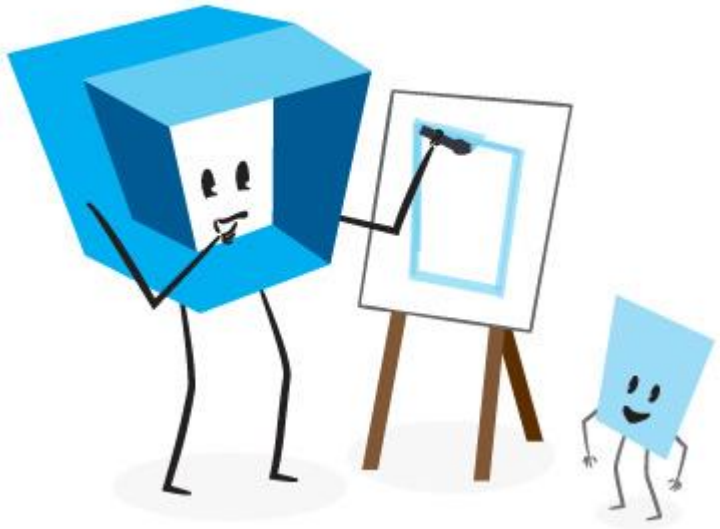
TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

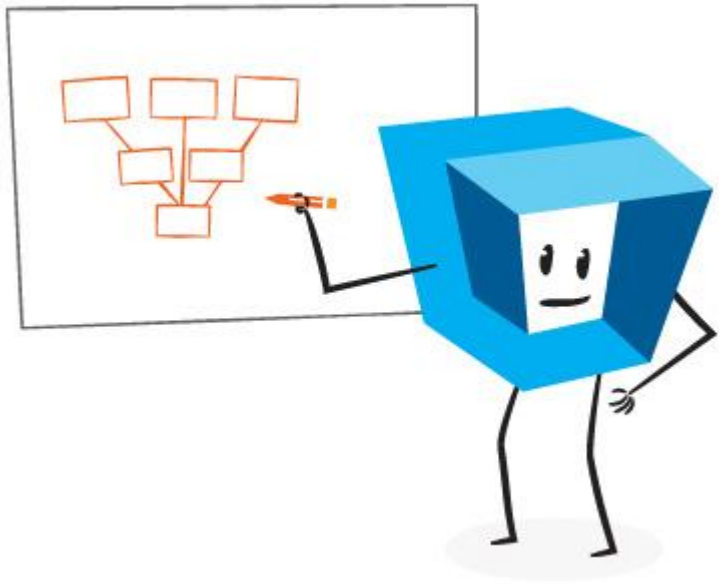
Linkki CESSDAn aineistoluetteloon

- Jos yliopistolla/organisaatiolla on tutkimusaineistosivuillaan linkkejä muihin luetteloihin, voisi lisätä tämän linkin:
  - Yhteiskuntatieteellisiä aineistoja Euroopasta: CESSDAn aineistoluettelo https://datacatalogue.cessda.eu/

TIETOARKISTO
FINNISH SOCIAL SCIENCE
DATA ARCHIVE

Jos kysyttävää, minulle voi lähettää sähköpostia





taina.jaaskelainen@tuni.fi