



# Onko data riittävän **FAIR?**

Datan elinkaari 19.4.2018, Kansalliskirjasto  
Esa-Pekka Keskitalo, [esa-pekka@helsinki.fi](mailto:esa-pekka@helsinki.fi)

# Toteutamme FAIR-periaatteita

<https://www.force11.org/group/fairgroup/fairprinciples>

Löydän

Saan

Liitän



FAIR = järkevät käytännöt 3+2+1

- ▶ Metatieto
- ▶ Tunnisteet
- ▶ Linkitys

- ▶ Lisenssit
- ▶ Tunnistaminen

• Standardit

▶ **TO BE FINDABLE:**

- ▶ F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- ▶ F2. data are described with rich metadata.
- ▶ F3. (meta)data are registered or indexed in a searchable resource.
- ▶ F4. metadata specify the data identifier.

▶ **TO BE ACCESSIBLE:**

- ▶ A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
  - ▶ A1.1 the protocol is open, free, and universally implementable.
  - ▶ A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- ▶ A2 metadata are accessible, even when the data are no longer available.

▶ **TO BE INTEROPERABLE:**

- ▶ I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- ▶ I2. (meta)data use vocabularies that follow FAIR principles.
- ▶ I3. (meta)data include qualified references to other (meta)data.

▶ **TO BE RE-USABLE:**

- ▶ R1. meta(data) have 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 their provenance.
  - ▶ R1.3. (meta)data meet domain-relevant community standards.

## FAIR GUIDING PRINCIPLES

1. To be **Findable** any Data Object should be uniquely and persistently identifiable.
  - 1.1. The same Data Object should be re-findable at any point in time, thus Data Objects should be **persistent**, with emphasis on their metadata. **Persistence is an organizational property.**
  - 1.2. A Data Object should minimally contain basic machine actionable metadata that distinguishes it from other Data Objects.
  - 1.3. Identifiers for any concept used in Data Objects should therefore be **Unique** and **Persistent**.
2. Data is **Accessible** in that it can be always obtained by machines and humans
  - 2.1 Upon appropriate authorization
  - 2.2 Through a well-defined protocol
  - 2.3 Thus, machines and humans alike will be able **to judge the actual accessibility** of each Data Object.
3. Data Objects can be **Interoperable** only if:
  - 3.1. (Meta) data is **machine-actionable**
  - 3.2. (Meta) data formats utilize shared vocabularies and/or ontologies
  - 3.3 (Meta) data within the Data Object should thus be both **x§§**
4. For Data Objects to be **Re-usable** additional criteria are:
  - 4.1 Data Objects should be compliant with
  - 4.2 (Meta) data should be sufficiently **well-described and rich** that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources
  - 4.3 Published Data Objects **should refer to their sources** with rich enough metadata and provenance to enable proper citation.

[q4\_1] Radio- ja televisiotoimittajien tulisi teitellä kaikkia yli 40-vuotiaita haasteltaviaan

[q4\_2] Myyjä voi käyttää moi-tervehdystä kaikenikäisten asiakkaiden kanssa

[q4\_3] Hyvää päivänjatkoa on sopiva toivotus asiakaspalvelutilanteissa

## **Tapaus 1: Kielenkäyttäjien mielipiteitä suomen kielen tilasta 2013 – Tietoarkisto**

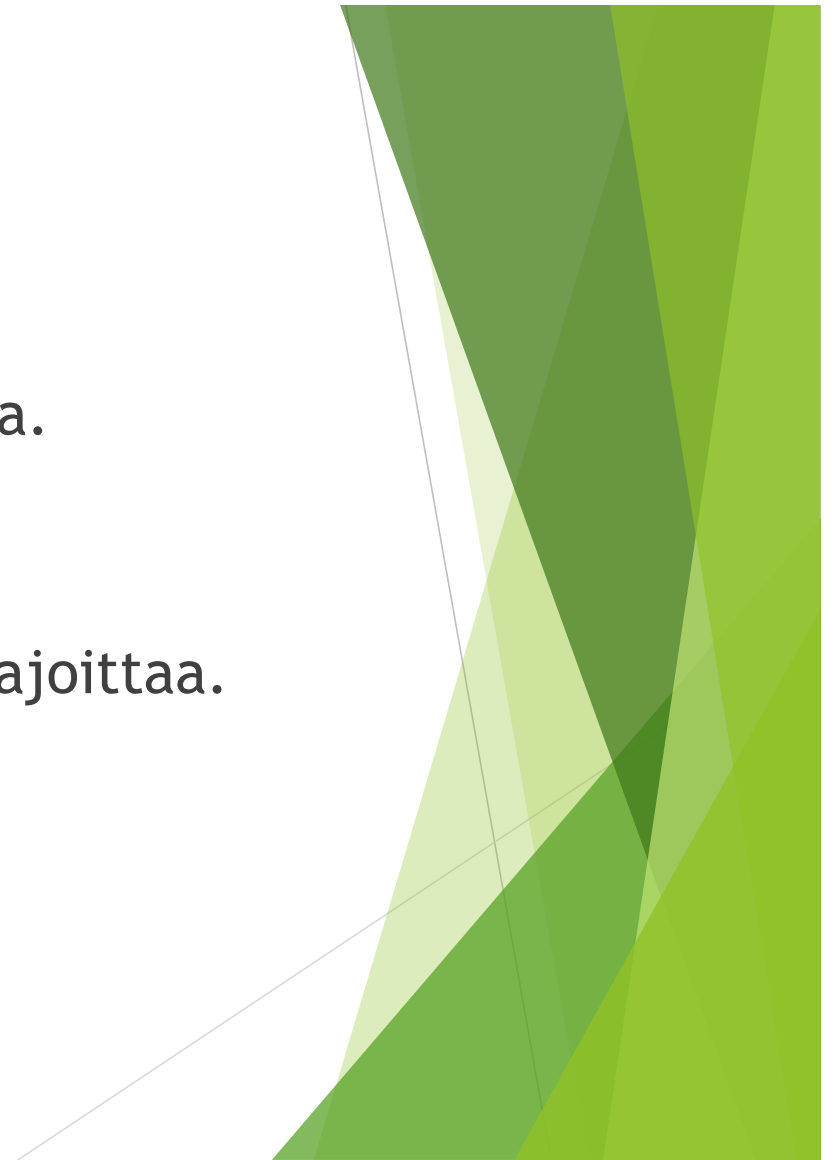
<http://urn.fi/urn:nbn:fi:fsd:T-FSD2979>

# Löydettävyys

- ▶ + Aineistolla on pysyvä tunniste
  - ▶ ? Pitäisikö olla oma tunniste metatiedolle?
- ▶ + Aineiston metatieto on rikasta; sekä ihmis- että koneluettavissa.
- ▶ + Metatieto löytyy helposti (Aila, Etsin, Google...)
- ▶ + Tunniste käy ilmi metatiedosta
  - ▶ Käytetty laiskasti viittauksissa, vaikka malliviittauksessa. Arkiston sisäinen tunnus löytyy useammin (koska osana otsikkoa?)

## Saatavuus

- ▶ + Aineiston saa käsiin tunnisteiden avulla.
- ▶ + Protokolla on avoin, maksuton ja yleiskäyttöinen
- ▶ + Aineistojen saatavuutta voi Ailassa rajoittaa. Lupaa voi hakea.
- ▶ + Aineistoista jää hautakivi.





## Yhteentoimivuus

- ▶ + Data ja metatieto ovat käyttökelpoisessa, yleisessä, formaalissa muodossa.
- ▶ Käytetyt sanastot noudattavat FAIR-periaatteita
  - ▶ + OKM:n tieteenalaluokitus (Fintossa)
  - ▶ - CESSDA (lista FSD:n sivuilla)
  - ▶ - Tietoarkiston oma tieteenalaluokitus (lista)
- ▶ Metatiedossa on viittauksia muihin lähteisiin
  - ▶ ? Koneluettavuus

# Uudelleenkäytettävyys

- ▶ + Metatiedossa on linkittäviä, tarkkoja attribuutteja.
  - ▶ Ks. <https://opendata.uta.fi/en/apis/study-descriptions-by-finnish-social-science-data-archive/>
- ▶ + Metatieto on CC-lisensoitu.
- ▶ ? Datan käyttöehdot ovat selkeät mutta epästandardit. Mutta mikä edes olisi se standardi.
- ▶ ? Datassa on versiotieto mutta ei juuri muuta provenienssia.

```
3020010 *v& veCer& !xva roZdstva :: ::
3020100 *byst& Ze v$ d$ni t& . Jzide povel@nie ot& kesar@ av&gusta .
3020110 napisati v$sO uselenOjO .
3020200 se napisanie pr&voe byst& . vladOStju sUriejO i kUriniju .
3020300 J id@axO v$si k$Z&do napsati sE v& svoi grad& ::
3020400 v$zide Ze iosif& ot& galilejE . Jz grada nazaret&ska .
3020410 v$ ijudeO v& grad& !dv& . JZe naricaat& sE vitleem& .
3020420 zane b@aSe ot& domu i ot$C$stvi@ !davdva .
3020500 napisati sE s& marie^jO obrOCenOjO (emu ZenojO .
3020510 sOStejO neprazd&no^jO .
3020600 byst& Ze egda byste tu . Jspl&niSE sE d$nie roditi ei .
3020700 J rodi !sn& svoi pr&v@nec& . J povit& I . J polozi i v$ @slex& .
3020710 zane ne b@ ima m@sta v& obit@li ::
```

## Tapaus 2: Corpus Cyrillo-Methodianum Helsingiense - Helsingin yliopisto

<http://www.helsinki.fi/slaavilaiset/ccmh/index.html>

# Löydettävyys

- ▶ — Aineistolla ei ole tunnisteita
- ▶ — Aineiston metatieto on niukkaa ja proosamuotoista.
- ▶ ? Metatieto löytyy Googlella, muttei esim. Finnalla tai Etsimellä
- ▶ — Tunnisteet eivät käy ilmi metatiedoista, koska niitä ei ole

## Saatavuus

- ▶ **-** Aineiston saa käsiin URLilla
- ▶ **+** Protokolla on avoin, maksuton ja yleiskäyttöinen
- ▶ **?** Saatavuutta ei voi säädellä, mutta ei ole tarviskaan
- ▶ **-** Aineistoista jää hautakivi.



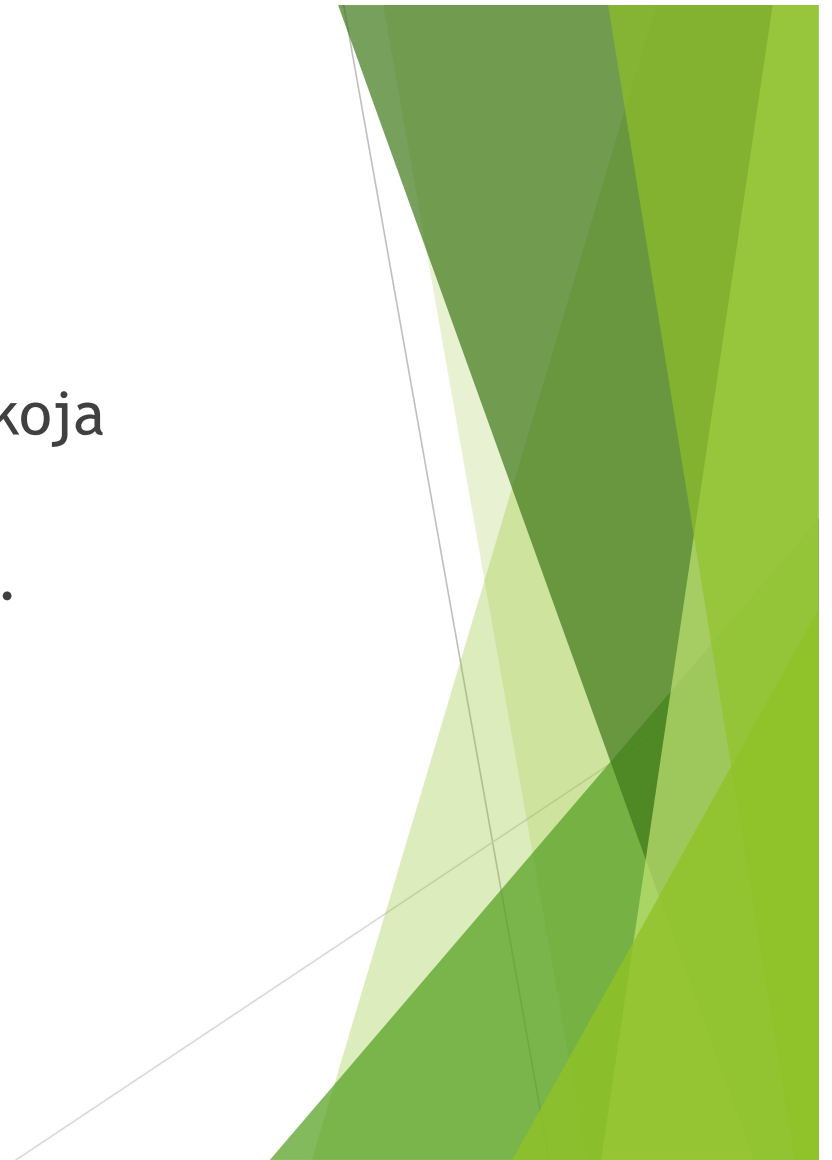
# Yhteentoimivuus

- ▶ — Metatieto on proosaa
- ▶ ? Data on toisaalta hyvin eksoottista, mutta toisaalta käyttökelpoisessa, yleisessä, formaalissa ja dokumentoidussa muodossa.
- ▶ — Ei ole käytetty sanastoja
- ▶ — Metatiedossa on epäformaaleja viittauksia muihin lähteisiin.



## Uudelleenkäytettävyys

- ▶ — Metatiedossa ei ole linkittäviä, tarkkoja attribuutteja.
- ▶ — Käyttöoikeuksista ei mainita mitään.
- ▶ — Ei provenienssia.



# Yhteenveto

## Kielenkäyttö

- ▶ ? Sanastojen FAIR-mukaisuus
- ▶ ? Viittaukset tekstipötköinä
- ▶ ? Aineiston proveniensi

## Corpus Cyrillo-Methodianum

- ▶ Työtä vielä tehtävänä...