Suomalaiset ja virolaiset runoaineistot FILTER-hankkeessa

Kati Kallio (SKS, HY) & Eetu Mäkelä (HY)

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HUMAN SCIENCES — COMPUTING INTERACTION

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HUMAN SCIENCES — COMPUTING INTERACTION

FILTER: Formulaic intertextuality, thematic networks, and poetic variation across regional cultures of Finnic oral poetry, 2020–2024

I. K. Inha 1894, SKS; Honkola, Terhi, et al. 2019. https://doi.org/10.3176/lu.2019.3.01

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Kalevalamittainen runous (runolaulu, regilaul)

- kansallisesti keskeisenä pidetty ilmiö Suomessa, Virossa ja Venäjän Karjalassa
- tieteellisesti monelta kannalta kiinnostava
- tallennettu runsaasti 1800–1900-luvulla
  - ensimmäiset tallenteet 1500–1600-luvulta
- 1800-luvulla kansanrunokokoelmia, Kalevala, laulukirjoja, kirjallista runoutta
- 1900-luvun alussa Suomen Kansan Vanhat Runot (SKVR) ja Vana Kannel
- kalevalamittaisen runouden pääaineistot digitoitu 1990-luvun lopulla
  - samankaltainen rakenne
  - kaksi erilaista runotyyppihakemistoa
- kolme kansallista korpusta ja tutkimusperinnettä

Korpukset ja niiden puute

- suomalainen SKVR-korpus 89 000 tekstiä (suomi, karjala, inkeroinen, vatja)
- virolainen ERAB-korpus 100 000 (-> 150 000) tekstiä (viro, eteläviro, inkeroinen)
- suomalainen julkaisemattomien runojen korpus 60 000 tekstiä (ei vielä käytössä)

monesta relevantista aineistosta ei ole korpusta

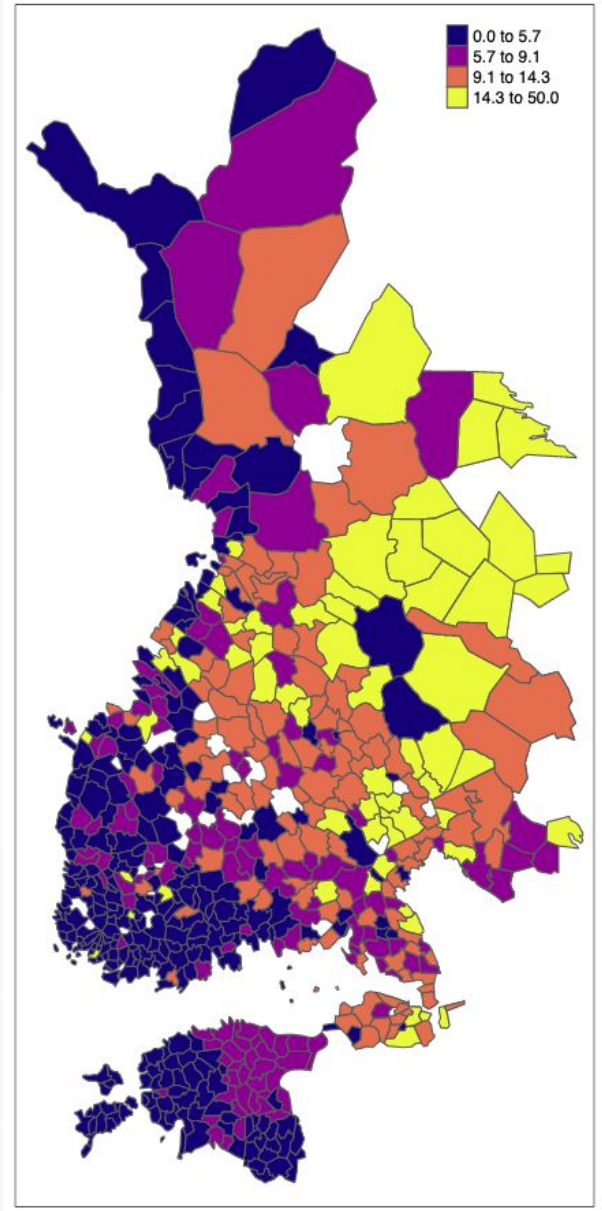
- etenkin venäjänkarjalaiset kokoelmat käsikirjoituksina, kirjoina, äänitteinä
- kirjalliset kalevalamittaiset runot ja pienet kansanrunojulkaisut (1690–)
  - yksittäisiä teoksia sähköisenä (Kotus, Kansalliskirjasto ym.)
- muut perinteen lajit: itkut, rekilaulut, karjalaiset joiut, uskomustarinat...
  - manuaalisia arkistokortistoja, kirjoja
- ei-kalevalamittaiset itämerensuomalaiset traditiot: vepsäläiset, liiviläiset
  - arkistokokoelmia, kirjoja

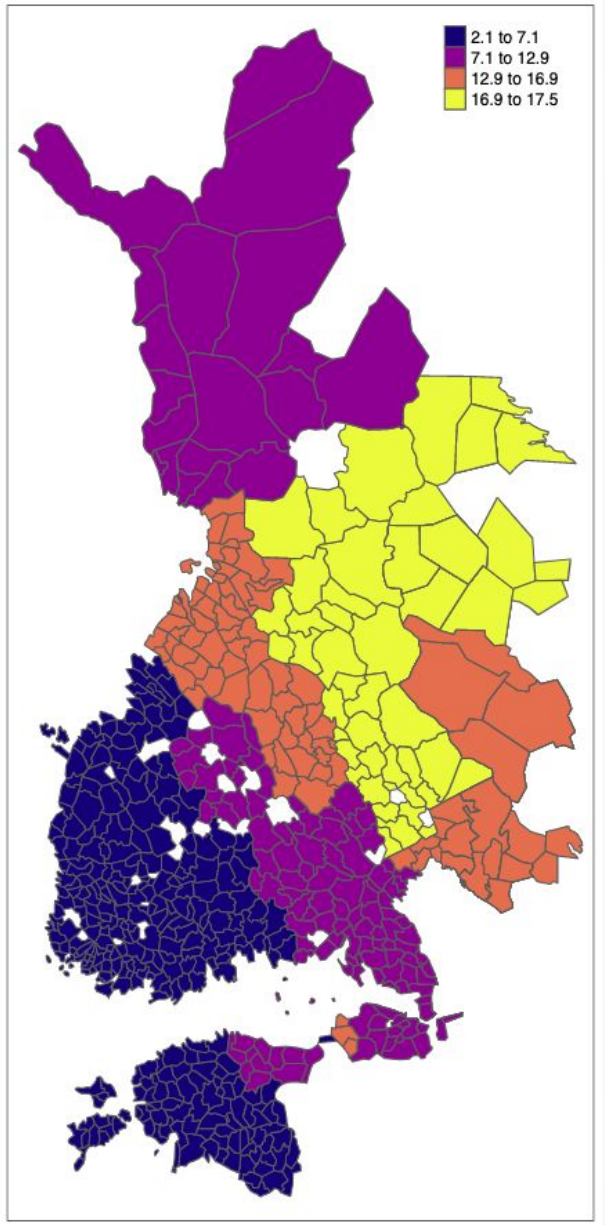
- poikkeuksellinen keskeisaineisto!
- alkuperäisaineistot ja digitoidut korpukset instituutioiden vastuulla
- FILTERin työvaiheet ja menetelmät dokumentoituja ja avoimesti käytössä
- aineiston korjaukset ym. takaisin arkistoille
- dataopas muille käyttäjille
- osa aineistoista vain manuaalisia

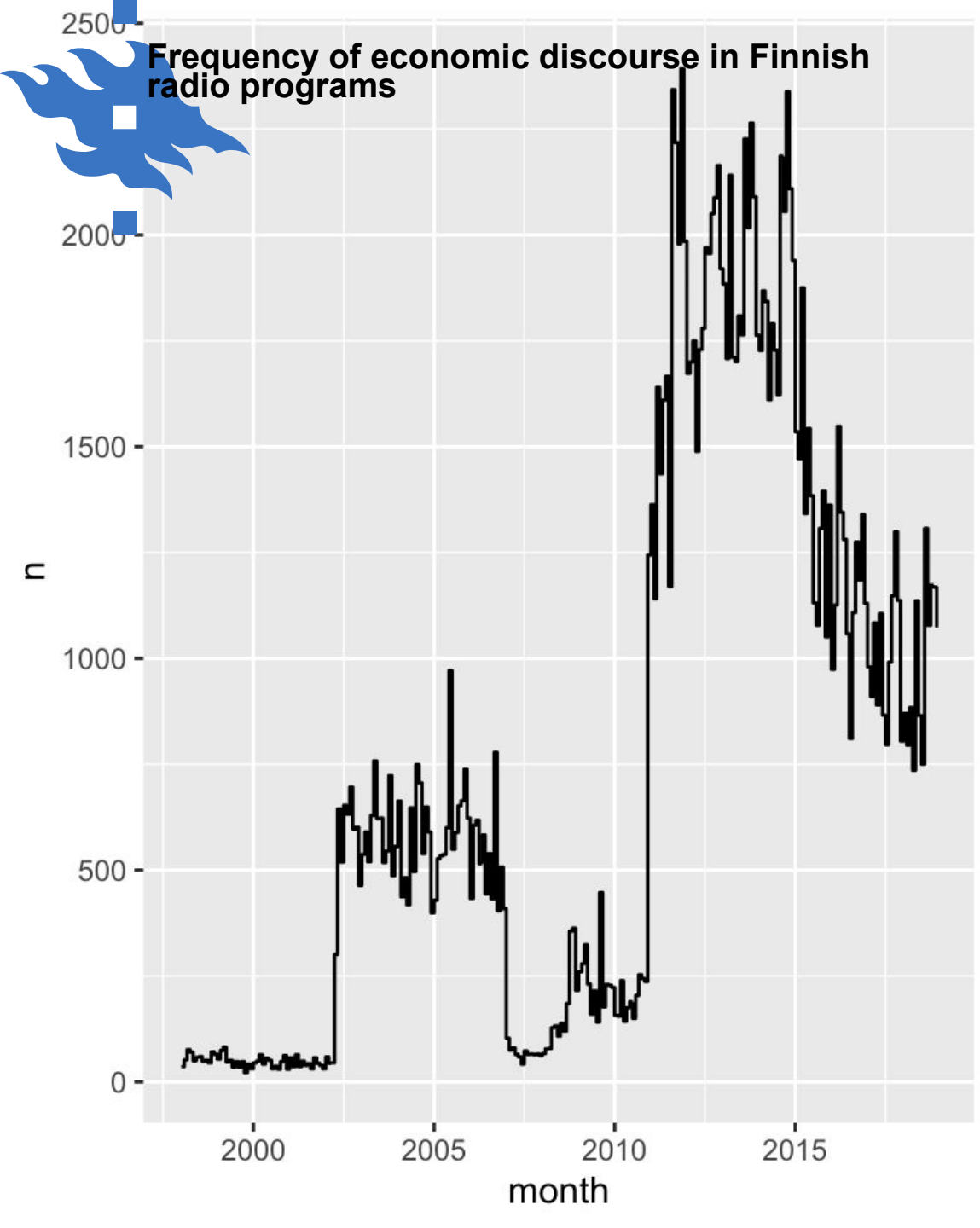
- hankkeessa eri osa-aineistot tuntevia tutkijoita
- keskeiskokoelmia hallitsevat instituutiot mukana (SKS & Eesti Rahvaluule Arhiiv)
- aineistojen käyttöön saaminen silti paikoin hidasta

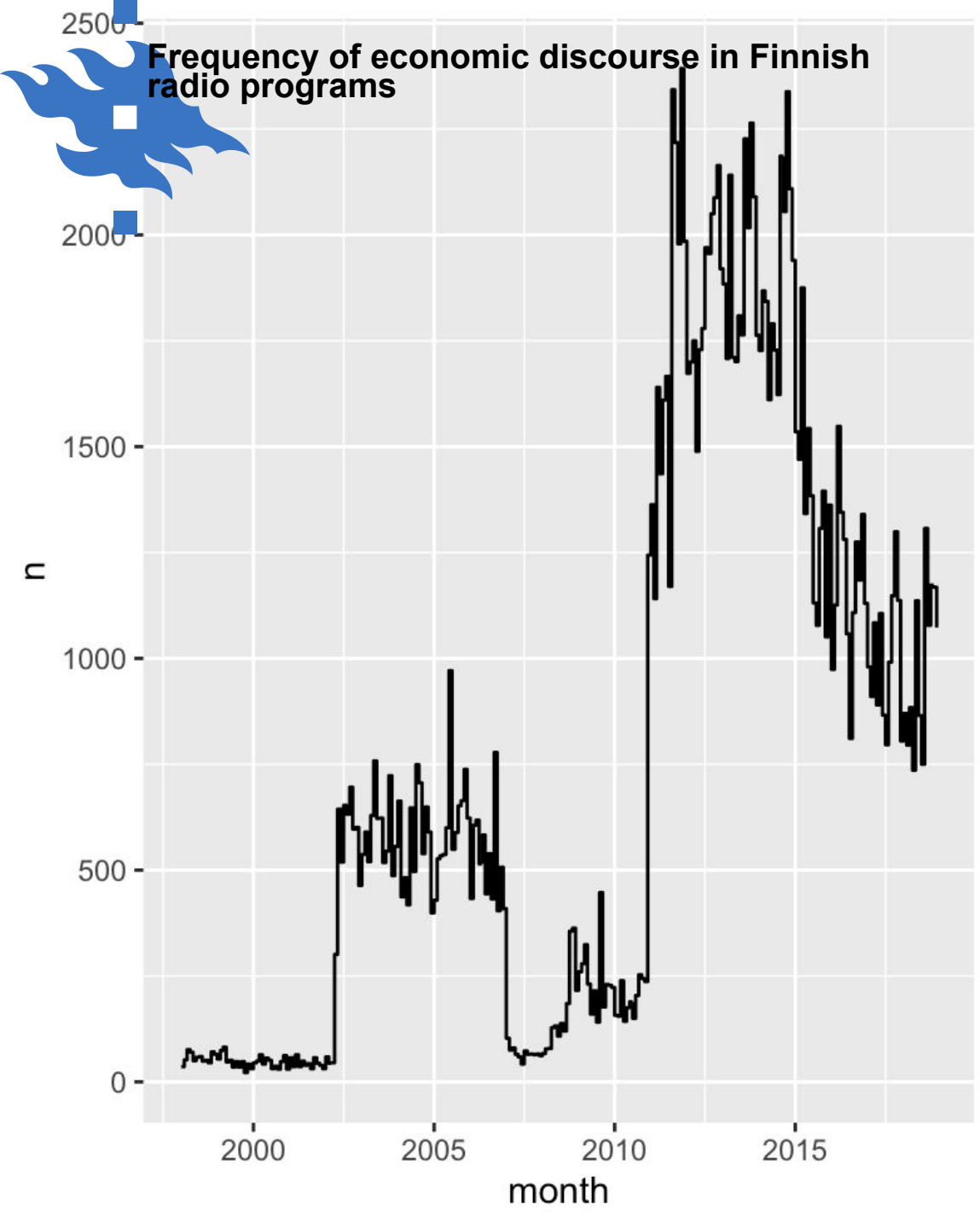
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

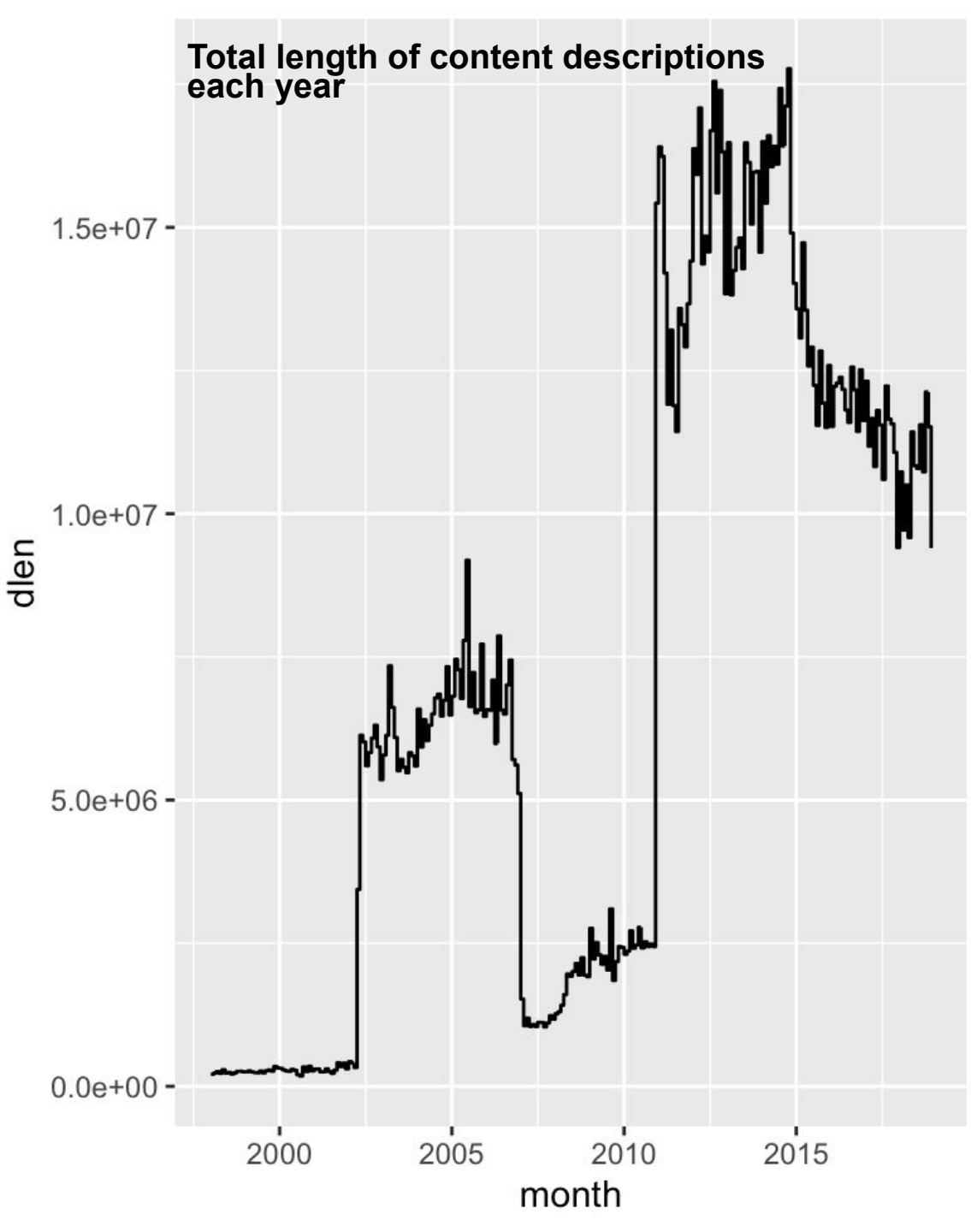
HUMAN SCIENCES — COMPUTING INTERACTION

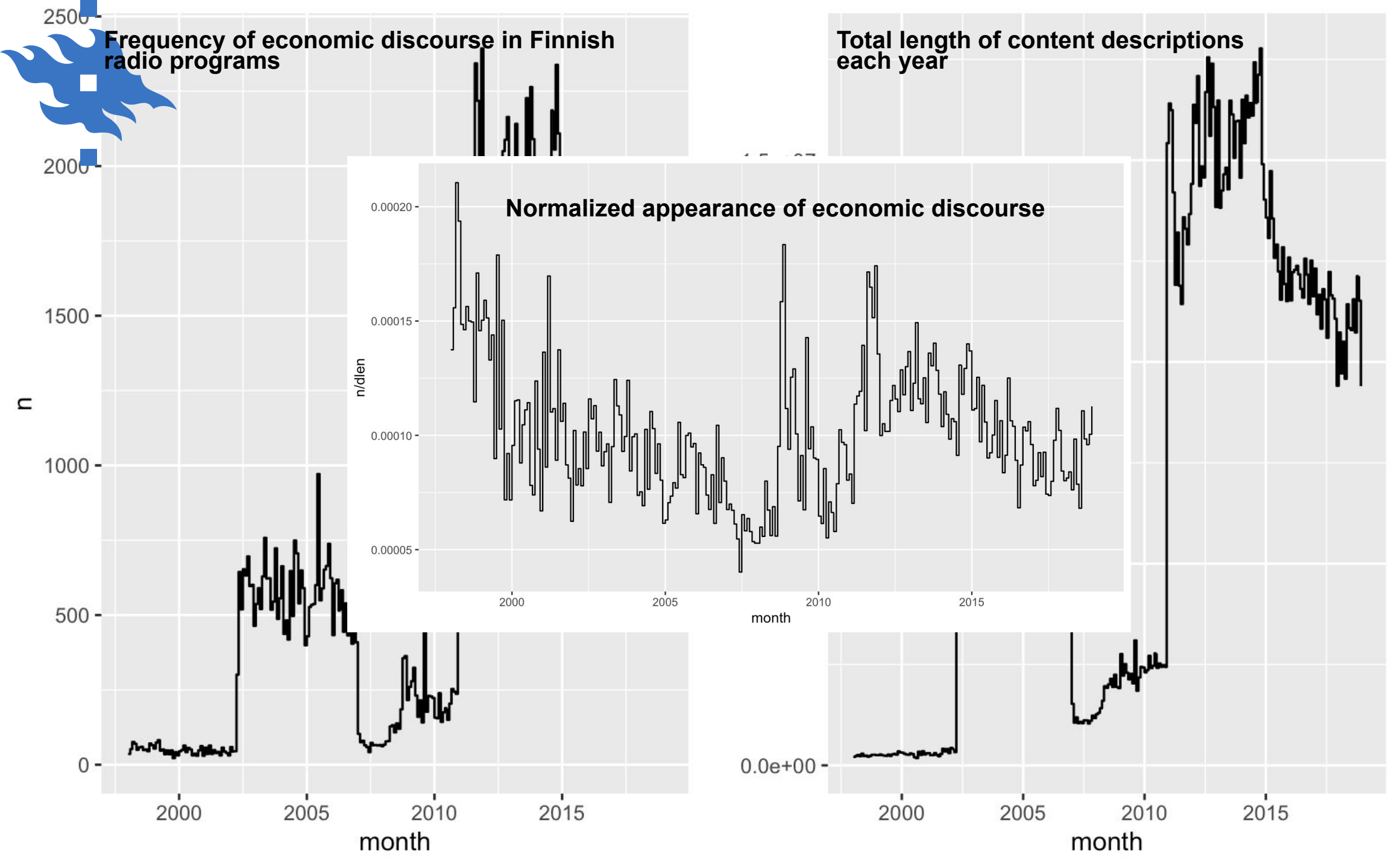

0.0 to 5.7
5.7 to 9.1
9.1 to 14.3
14.3 to 50.0


2.1 to 7.1
7.1 to 12.9
12.9 to 16.9
16.9 to 17.5

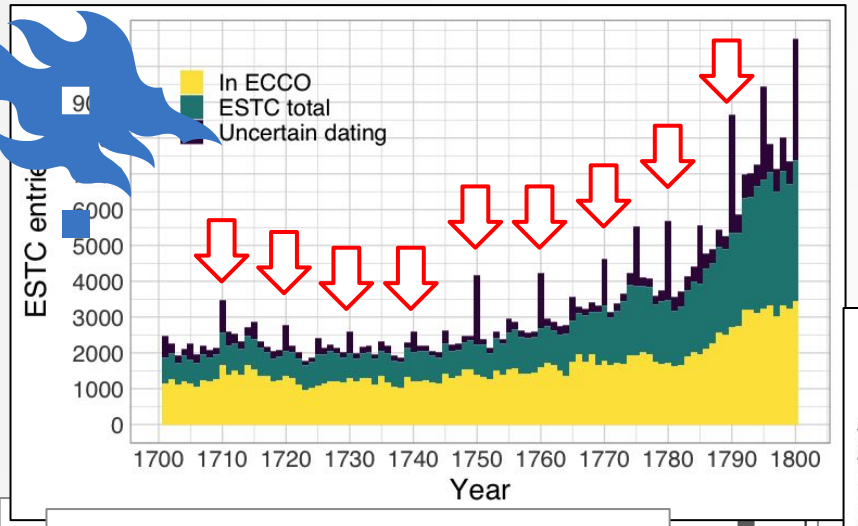

Frequency of economic discourse in Finnish radio programs
2500
2000
1500
1000
500
0
n
2000
2005
2010
2015
month

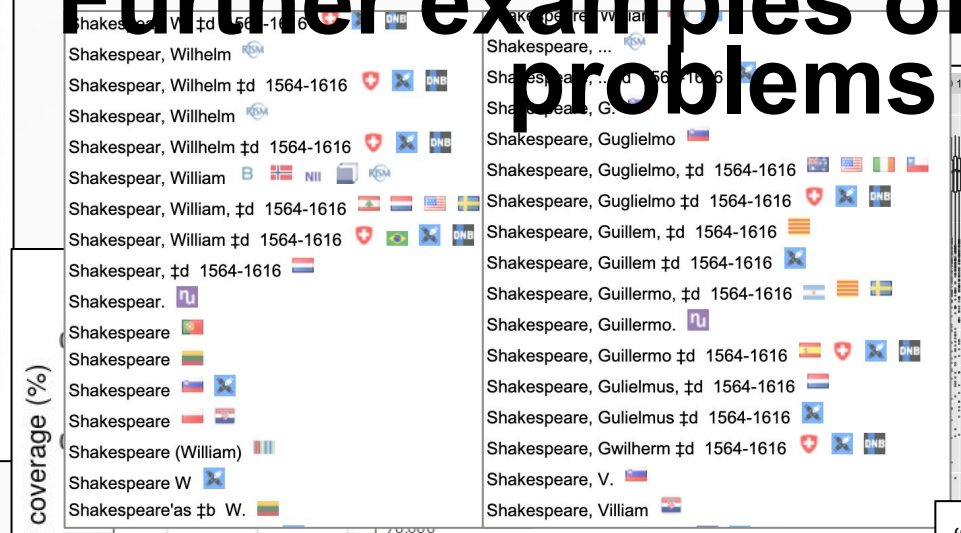

Frequency of economic discourse in Finnish radio programs
2500
2000
1500
1000
500
0
n
2000
2005
2010
2015
month

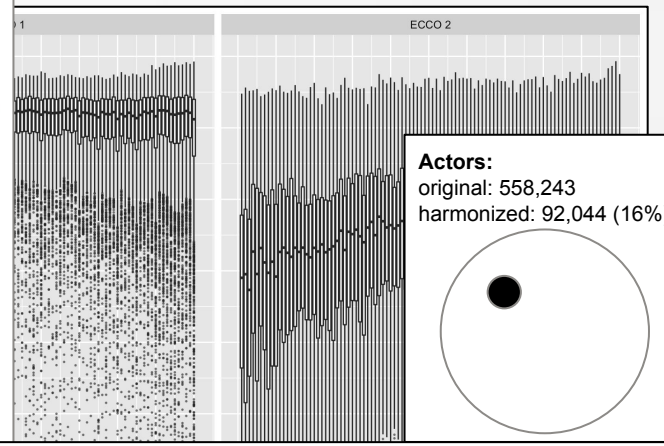

Total length of content descriptions each year
1.5e+07
1.0e+07
5.0e+06
0.0e+00
dlen
2000
2005
2010
2015
month


Frequency of economic discourse in Finnish radio programs
Total length of content descriptions each year
Normalized appearance of economic discourse
n
n/dlen
month
2500
2000
1500
1000
500
0
0.00020
0.00015
0.00010
0.00005
0.0e+00
2000
2005
2010
2015

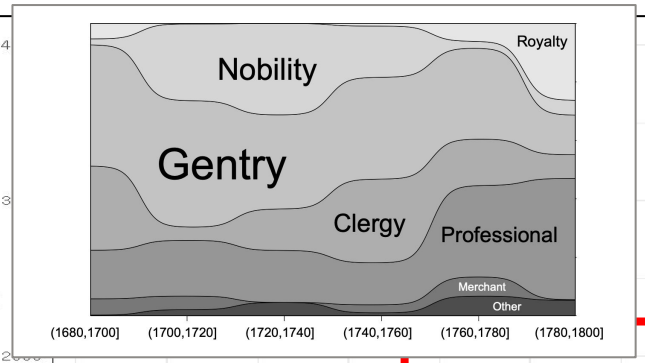
Further examples of bias and problems

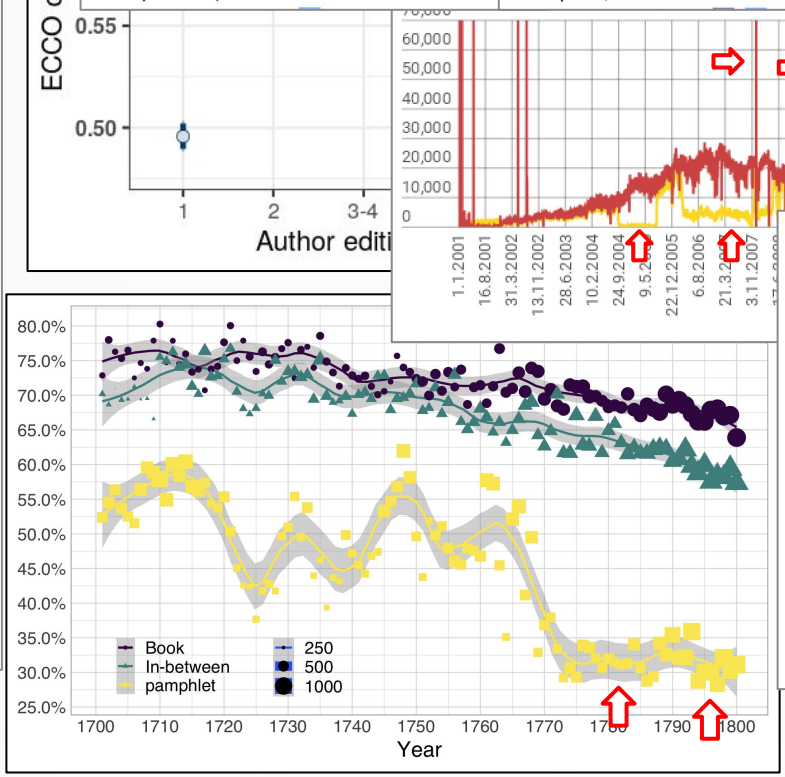


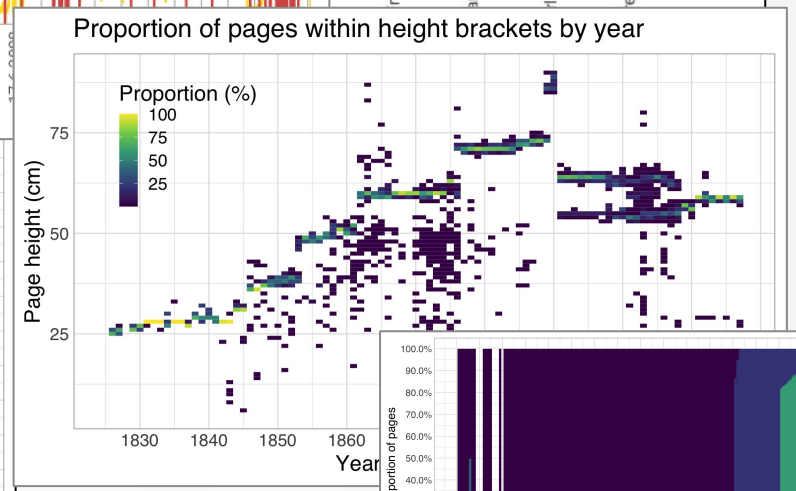


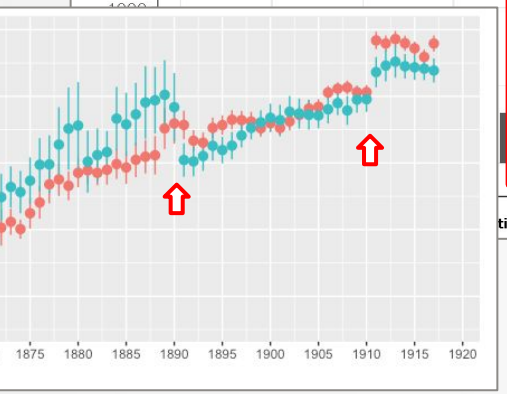


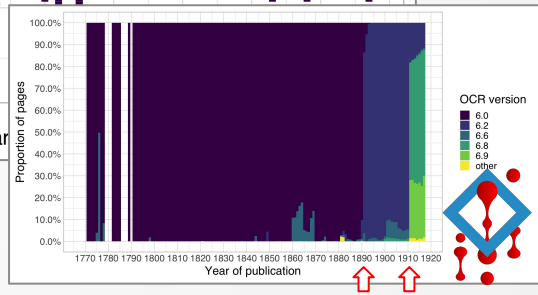
“[4],vii-xii,[4],222p.,plate” → 240 pages









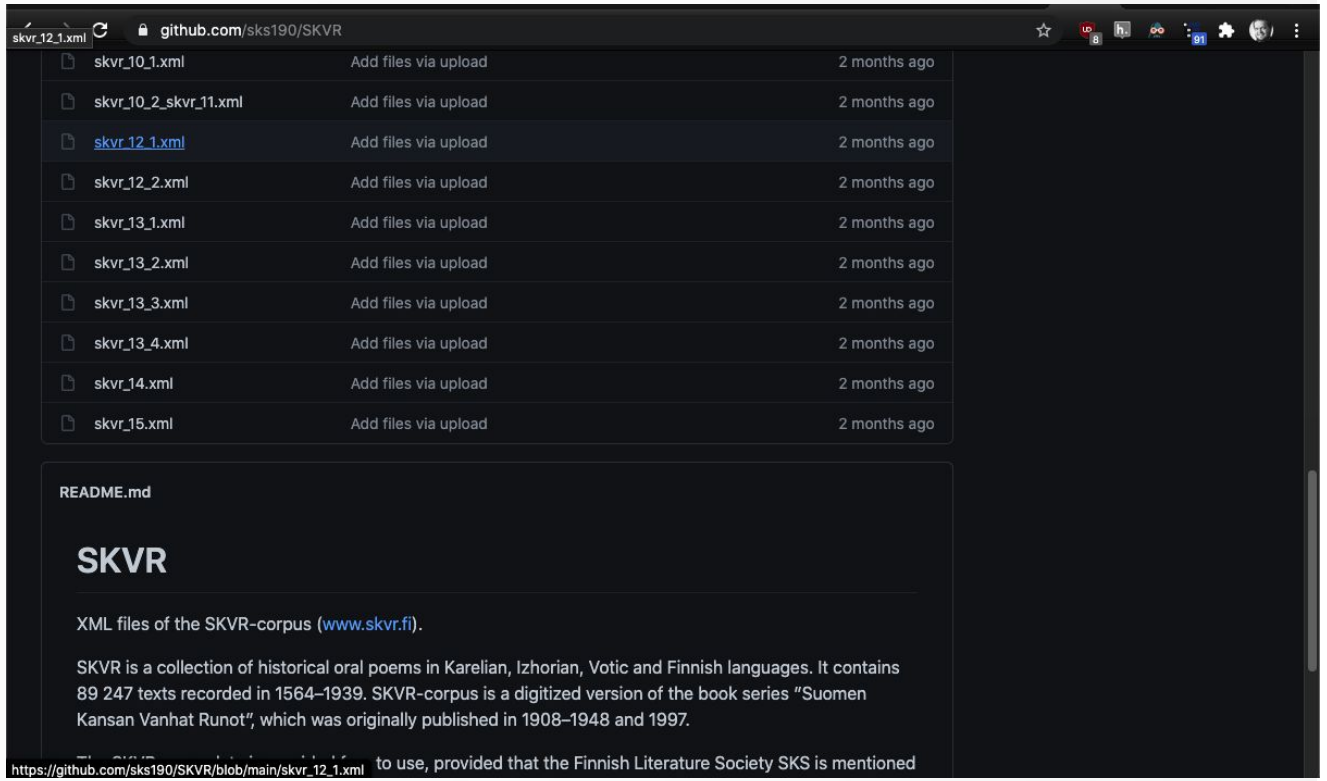


Hankaluuksia aineistonhankinnassa

- SQL-dumppi vs puutteellinen XML-rajapinta + ajantaisaisuus
- Julkaisemattomat aineistot

FAIR

1. Findable = Google/artikkelit
2. Accessible = GitHub/Zenodo/IDA
3. Interoperable = XML/CSV
4. Reusable = XML/CSV + kattava dokumentaatio



Turhaa:

1. Toiminnalliset aineistorepositoriot (Qvain/Etsin)
2. Aineistoa kuvaileva rakenteinen metadata
3. Semanttinen yhteentoimivuus

Missä reproducible?!

Aineistot käyttöön – kulttuuriperintölaitosten näkökulma

1. Tarjoa saataville aineiston alkuperäinen, rikkain versio, vaikka se olisi sotkuinenkin. (Sen sijaan että teet aineistolle “käyttöä helpottavia muunnoksia”, tarjoa sen sijaan alkup. aineisto ja koodi jolla tuon muunnoksen voi tehdä itse)
2. Mitä enemmän dokumentoit (tekstinä!) aineistoa, erityisesti sitä mitä aineisto **ei** sisällä, sen parempi. **Mutta:** parempi julkaista epäselvää sotkua heti kuin odottaa että saadaan kuvattua täydellisesti
3. Jos aineistosi muuttuu ja päivittyy usein, kiinnitä huomiota siihen että uudet versiot tulevat myös jakoon. **Mutta:** parempi julkaista kerran kuin odottaa että saadaan aineisto päivittymään jouhevasti