Kirjallisuutta ja digiä!

Loppuseminaari · Turun kaupunginkirjastossa 12.6.2026

Leo Lahti, Julia Matveeva

Digitaaliset menetelmät kotimaisen kirjallisuushistorian uudistajana

Konsortiohankkeen pyrkimyksenä on laajentaa merkittävästi vallitsevaa käsitystä Suomen kirjallisuushistoriasta uusien digitaalisten aineistojen ja menetelmien avulla. Kansalliskirjaston, Turun yliopiston ja Itä-Suomen yliopiston yhteistyönä kartoitamme suomen-ja ruotsinkielisen kaunokirjallisuuden taustatietoja 1800-luvulta yhtenäiseen muotoon, joka mahdollistaa aineiston laajamittaisen tilastollisen analysoinnin. Laadimme aineiston tutkimuskäyttöä varten erityisesti suunnitellun avoimen datatieteen työvirran.

Datalähtöinen kokonaiskuva kotimaisen julkaisutoiminnan historiasta 1800-luvulla.

Digitaaliset menetelmät laajentavat käsitystä kotimaisesta kirjallisuushistoriasta.

Jalostimme kaunokirjallisuuden teosluettelot muotoon, joka mahdollistaa laajamittaiset tilastolliset analyysit ja rikastaa lähiluentaa.

Digitaalisten kokoelmien ja datatieteen avulla laajennetaan käsitystä 1800-luvun suomalaisesta kirjallisuudesta — kartoitetaan suomen- ja ruotsinkielinen kaunokirjallisuus, jonka aiempi tutkimus sivuutti, eli se mitä Margaret Cohen kutsuu nimellä *”suuri lukematon”.*

Uudistetaan 1800-luvun kotimainen kirjallisuushistoria -
luetaan järjestelmänä, ei suurten kirjailijoiden kaanonina

Kartoitetaan julkaistu kaunokirjallisuus:
lajityypit, kielet, kirjoittajuus ja salanimet

Rakennetaan kirjallisuushistorian tueksi
avoin bibliografisen datatieteen viitekehys

Digitaaliset menetelmät Suomen kirjallisuushistorian uudistajina

Suomen Akatemia, 2022–2026 · sites.utu.fi/digital-history-literature-finland

WP1 · Kirjallisuushistoria
PI Kati Launis (UEF), Aino Mäkikalli, Viola Parente-Čapková & Veli-Matti Pynttäri

WP2 · Datatieteet
PI Leo Lahti (TY), Julia Matveeva, Jeba Akewak, Ville Laitinen

WP3 · Digitaaliset resurssit
PI Osma Suominen (Kansalliskirjasto)

cc

SUOMEN AKATEMIA
FINLANDS AKADEMI · ACADEMY OF FINLAND

1640 KANSALLIS KIRJASTO

UNIVERSITY OF EASTERN FINLAND

UNIVERSITY OF TURKU

CATALOGING & CLASSIFICATION QUARTERLY
https://doi.org/10.1080/01639374.2018.1543747

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Bibliographic Data Science and the History of the Book (c. 1500–1800)

Leo Lahti[a], Jani Marjanen[b], Hege Roivainen[b], and Mikko Tolonen[b]

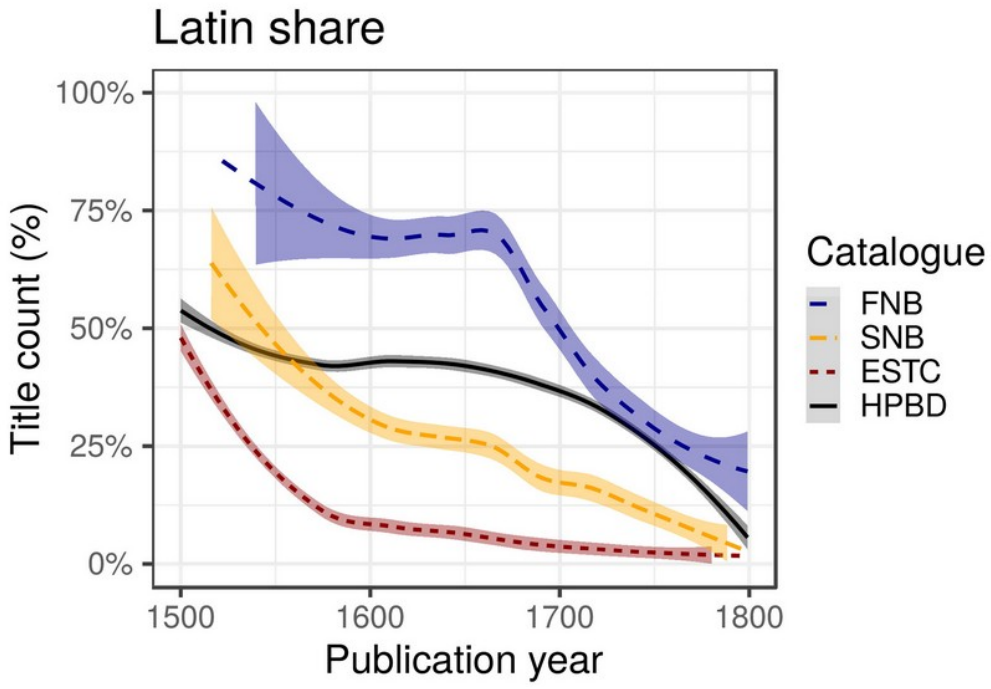
[a]Department of Mathematics and Statistics, University of Turku, Finland; [b]Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Finland

ABSTRACT

National bibliographies have been identified as a crucial resource for historical research on the publishing landscape, but using them requires addressing challenges of data quality, completeness, and interpretation. We call this approach *bibliographic data science*. In this article, we briefly assess the development of book formats and the vernacularization process in early modern Europe. The work undertaken paves the way for more extensive integration of library catalogs to map the history of the book.

ARTICLE HISTORY
Received July 2018
Revised September 2018
Accepted October 2018

KEYWORDS
National bibliography; data ecosystem; publishing history; digital humanities; open science



Subject catalogue of the University Library of Graz.
Source: Wikimedia Commons.

Laskennallisten ihmistieteiden asema vahvistuu ja muokkaa Suomen tiedekenttää

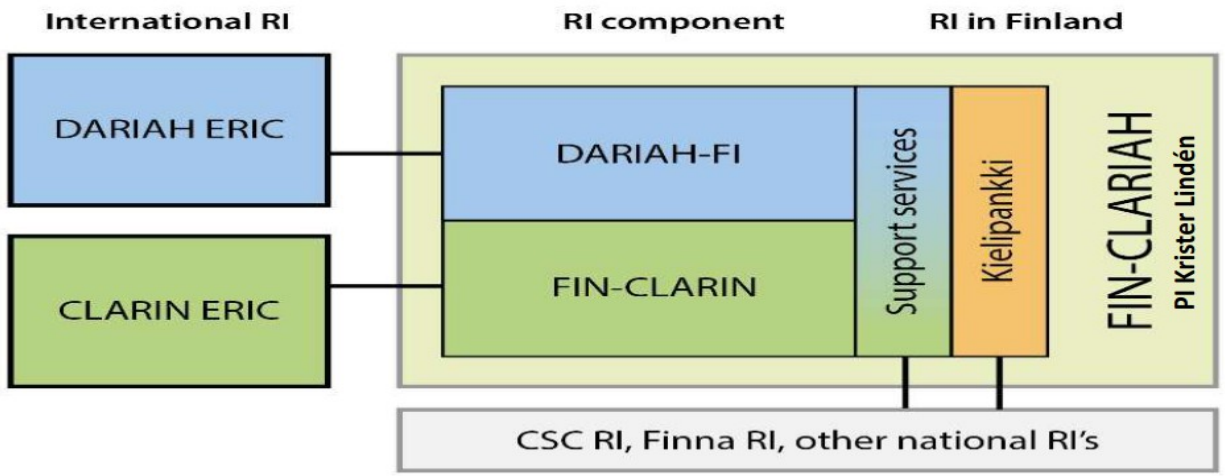
20.12.2021 UUTINEN

Useiden suomalaisten yliopistojen yhteistyönä kehitettävä digitaalinen infrastruktuuri tukee modernin datatieteen vahvistumista ihmistieteiden keskeisenä menetelmänä humanistisilla ja yhteiskuntatieteellisillä aloilla. Hankkeessa karttuvilla digitaalisilla tutkimusaineistoilla ja -menetelmillä voidaan ymmärtää kulttuurin ja yhteiskunnan tilaa ja muutoksia historiasta nykypäivään.

Funding awarded to major national project in the humanities

The Academy of Finland has granted EUR 4.6 million in funding for the FIN-CLARIAH infrastructure project, which will make available to researchers extensive collections of texts and multimodal data, as well as tools for the analysis and enrichment of the data. The funding enables humanities research that can take advantage of neural network technology, high-powered computing and extensive datasets.

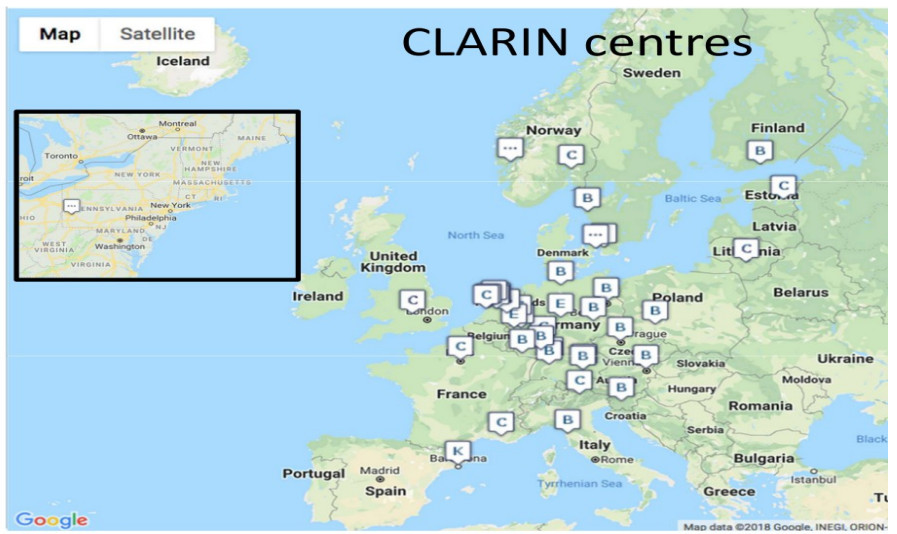
CLARIAH Finland – a Research Infrastructure for Social Sciences and Humanities (2020)



CLARIN

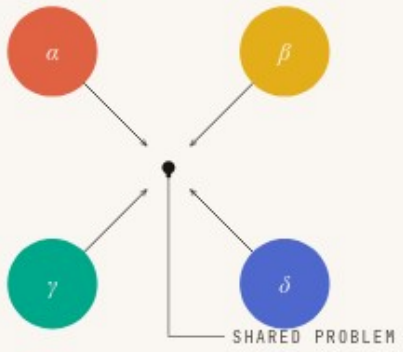
Member countries:

The Netherlands
Austria
Bulgaria
Czech Republic
Denmark
DLU
Estonia
Finland
Germany
Greece
Hungary
Italy
Latvia
Lithuania
Norway
Poland
Portugal
Slovenia
Sweden
France
UK
USA / CMU



Monitieteinen
(Multidisciplinary)

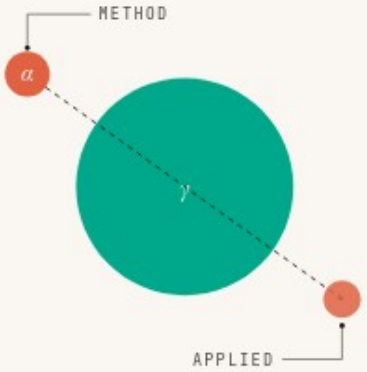
“lisäävä” (additive)



Tieteenalat työskentelevät rinnakkain saman ongelman parissa, kukin itsenäisesti omia menetelmiään tai viitekehyksiään muuttamatta. Tulokset kertyvät yhteen mutta pysyvät erillisinä.

Poikkitieteellinen
(Cross-disciplinary)

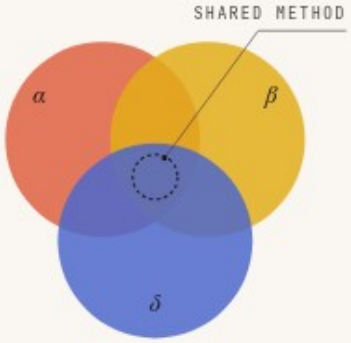
“risteävä” (crossing)



Yhden tieteenalan työkaluja tai käsitteitä lainataan ja sovelletaan toisella alalla, mutta aloja ei sulauteta yhteen.

Tieteidenvälinen
(Interdisciplinary)

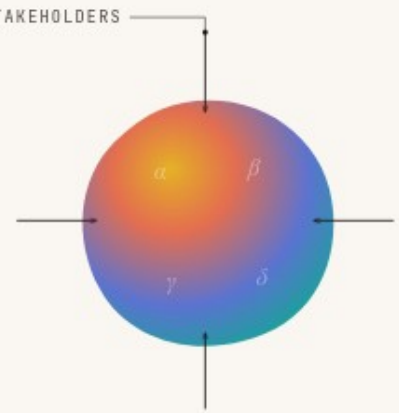
“vuorovaikutteinen” (interactive)



Tieteenalat yhdistävät menetelmiä ja näkökulmia ja suunnittelevat yhdessä lähestymistapoja, joita kumpikaan ei voisi kehittää yksin. Yhteistyö muokkaa ymmärrystä.

Tieteidenylinen
(Transdisciplinary)

“kokonaisvaltainen” (holistic)



Tieteenalojen rajat häviävät kokonaan, ja mukaan voidaan ottaa myös akateemisen maailman ulkopuolisia sidosryhmiä. Tavoitteena uudet viitekehykset, jotka ylittävät yksittäisen alan näkökulman.

cc

See e.g. Choi & Pak (2006)

Ohjelma

klo 10.15 – 15.00

10.15

Alkusanat

Leo Lahti

10.30–12.00

Osahankkeiden tulosten esittely

Datatiede kirjallisuushistorian tutkimusvälineenä Leo Lahti

Kansallisbibliografia Fennica uudenlaiseen hyötykäyttöön Osma Suominen

Uusi katse kirjallisuushistoriaan Kati Launis

12.00

Kommentti- ja tutkijapuheenvuoro

Hannu Salmi · professori

12.30–13.30

Lounastauko

omakustanteinen *

13.30

Vihreä sali ja saksalaiset puutarhurit

Digitaalisten arkistojen anti fiktiivisten puutarhojen luomisessa · kirjailija Sirpa Kähkönen

---

14.30

Kahvit

15.00

Toden ja epätoden rihmastot

Bibliomania 1800-luvun sanomalehdistössä · Hannu Salmi

** Lounas omakustanteinen*

WP 2: Datatiede

Turun yliopisto, Tietotekniikan laitos

Hankkeen Digitaaliset menetelmät kotimaisen kirjallisuushistorian uudistajana loppuseminaari

12.06.2026

TURUN YLIOPISTO

Digitaaliset menetelmät Suomen kirjallisuushistorian uudistajina

Suomen Akatemia, 2022–2026 · sites.utu.fi/digital-history-literature-finland

WP1 · Kirjallisuushistoria
PI Kati Launis (UEF), Aino Mäkikalli, Viola Parente-Čapková & Veli-Matti Pynttäri

WP2 · Datatieteet
PI Leo Lahti (TY), Julia Matveeva, Jeba Akewak, Ville Laitinen

WP3 · Digitaaliset resurssit
PI Osma Suominen (Kansalliskirjasto)

CC BY

UNIVERSITY OF TURKU

Suomen Kansallisbibliografia Fennica

kansalliskirjasto.fi/en/services/fennica-finnish-national-bibliography

Suomen kansallisbibliografia, joka sisältää tietoa kirjoista vuodesta 1488, lehdistä vuodesta 1711, sarjajulkaisuista, kartoista, audiovisuaalisista ja elektronisista aineistoista.

Paljon aineistoja tuhoutui Turun palossa 1827.

14.12.2017

Suomen kansallisbibliografia julkaistu avoimena datana

Suomen kansallisbibliografia Fennica on julkaistu avoimena datana. Kansalliskirjasto toivoo avatuille tietovarannoilleen uusia käyttäjiä ja uusia käyttötapoja.

kansalliskirjasto.fi/en/services/fennica-finnish-national-bibliography

CC BY

UNIVERSITY OF TURKU

Metadatan luonteesta ja tutkimisesta

Metadata eli bibliografinen kuvailutieto kertoo

- Tekijästä ja nimimerkeistä
- **Teoksesta:** koko, muoto, sivumäärä, kieli
- **Julkaisemisesta:** kustantaja
- **Otsikoista:** otsikoiden pituus, alaotsikot, lajimääreet

Bibliografiat luovat kirjallisuushistoriaa, eivät ainoastaan toimi tiedonhaun välineinä.

Ne sisältävät paljon informaatiota määrälliseen analyysiin.

Kirjastoluetteloiden tutkimusmahdollisuudet hyvin tunnistettuja

Studies in Bibliography
Vol. 27 (1974), pp. 55-89 (35 pages)

Published by: Bibliographical Society of the University of Virginia

https://www.jstor.org/stable/40371588

Bibliography and Science

by

G. THOMAS TANSELLE

A REVIEWER FOR THE *Times Literary Supplement*, COMMENTing in 1972 on two bibliographical annuals, remarked, "To argue about the scientific nature of bibliography now is surely to pursue a red herring."[1] I could not agree more. When I observed a few years ago, "All that 'scientific' can mean when applied to bibliographical analysis and textual study is 'systematic,' 'methodical,' and 'scholarly,' "[2] I was only repeating what a number of others have said and what many more must believe. It seems obvious that the word "scientific," when used to describe bibliography—as it has been off and on for more than a century—does not mean the same thing as when it is applied to physics, say, or chemistry. Apparently the issue cannot be dismissed so easily, however, for there have been several recent essays—notably those by D. F. McKenzie, James Thorpe, Peter Davison, and Morse Peckham[3]—which take up fundamental questions regarding the connections between science and bibliography. In a sense one must agree with the *TLS* that "it is perhaps a pity that he [McKenzie] revived the old argument about the scientific nature of bibliography"; at the same time, the existence of this group of essays suggests that the issue is not a dead one, and the *TLS* admits that the matter is "currently very much in the air."

Kansallisbibliografian jalostus laajamittaiseen tilastolliseen analyysiin

Sisältö

- Fennica sisältää yli 1,3 miljoonaa bibliografista tietuetta vuodesta 1488 nykypäivään

Haasteet

- Esitysmuoto (MARC21) on suunniteltu luettelointiin, ei tilastolliseen analyysiin
- Metadata on usein epäjohdonmukaista, puutteellista, hierarkkista ja vaikeasti skaalattavaa

Tavoite

- Luodaan toistettava työnkulku, joka muuntaa tiedot analyysivalmiiksi aineistoksi.
- Tuetaan ja toteutetaan datalähtöistä kirjallisuustutkimusta

FENNICA-tietokanta - Suomen kansallisbibliografia on suomalaisen julkaisutuotannon tietovaranto, jota tuotetaan Kansalliskirjastossa.

Fennica on nyt julkaistu avoimena datana CC0-lisenssillä, joka antaa mahdollisuuden sen vapaaseen hyödyntämiseen esimerkiksi sovelluksissa ja datan visualisoinneissa.

Suomen kansallisbibliografia sisältää seuraavia kuvailutietoja:

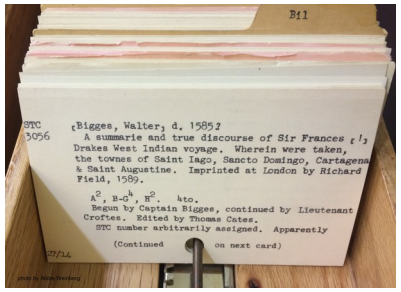
- suomalaiset kirjat vuodesta 1488 (N > 1,2 milj.)
- sarja-aineistot, mm. lehdet vuodesta 1771 (N > 73 000)
- kartat 1540-luvulta lähtien (N ~ 60 000)
- audiovisuaaliset aineistot
- digitoituja vanhoja aineistoja
- pienpainatekokoelman aiheenmukaiset ryhmäluettelot
- verkkosivujen teemakeräyksistä tehtyjä ryhmäkuvailuja vuodesta 2008
- kustantajien ennakkotiedot tulevista julkaisuista
- valikoidut e-kirjat n. vuodesta 2008 alkaen.

The database is founded on the materials provided pursuant to provisions in the Act on Collecting and Preserving Cultural Materials (Laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä1433/2007) and it complies with international recommendations on national bibliographies.

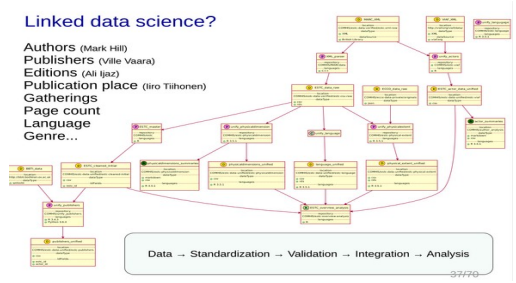
The database grows annually by approximately 40,000 monograph and 700 continuing publication records. The database comprises over 1,200,000 books and other monograph records (printed or electronic materials), approximately 73,000 continuing publications (journals or serial publications) and approximately 60,000 maps.

Kirjastoluetteloista tutkimusraportteihin

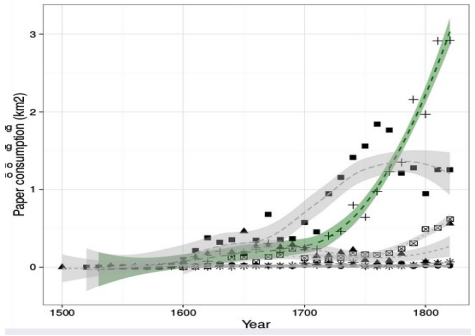
Tutkimusmahdollisuudet



Datatiede



Tutkimuskysymykset



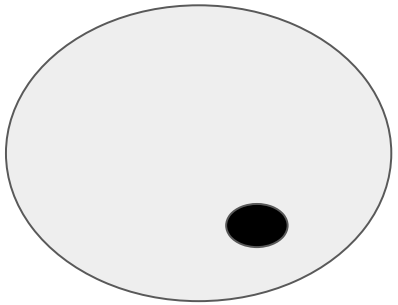
Työvaiheita

- *Datan jalostus* laajamittaiseen tilastolliseen analyysiin
- *Otoksen valinta kansallisbibliografiasta*: automatisoitu ja manuaalinen kuratointi
- *Tilastollinen analyysi* (esim. taustatrendien tunnistus)

Datan “jalostus” ja “rikastus”?
Alkuperäinen data ei aina sovellu suoraan tutkimuskäyttöön

Frenckell-julkaisijat Fennicassa
[J. C. Frenckell]
typis Frenckellianis
Frenckell
[J. C. Frenckells änka]
tryckt hos Johan Christopher Frenckell
impressit Joh. Christoph. Frenckell
[Frenckell]
typis Frenckelliorum
in officina Frenckelliana
tryckt i Frenckellska boktryckeriet
wastuudest prändätty Frenckellin kirja-prändisä
.. 484 spellings ..

Julkaisijat
- 13 345 alkuperäistä nimeä
- <10% yhdistettyä nimeä

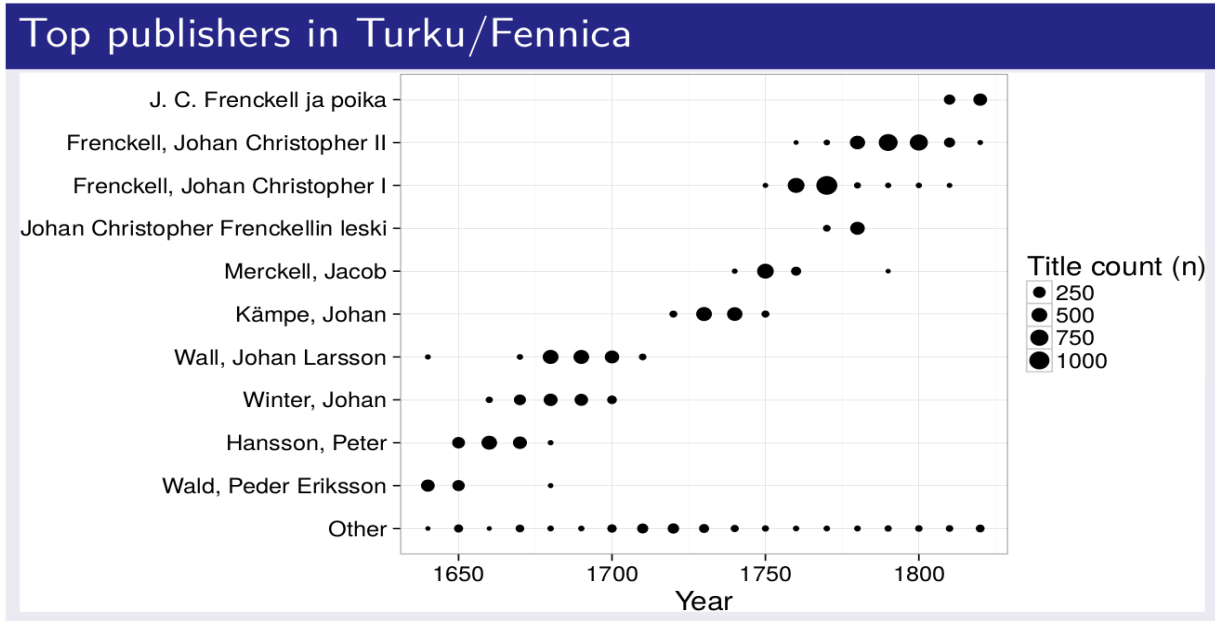


HISTORICAL METHODS
2018, VOL. 52, NO. 1, 57–78
https://doi.org/10.1080/01615440.2018.1526657
Routledge Taylor & Francis Group
OPEN ACCESS Check for updates

A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828

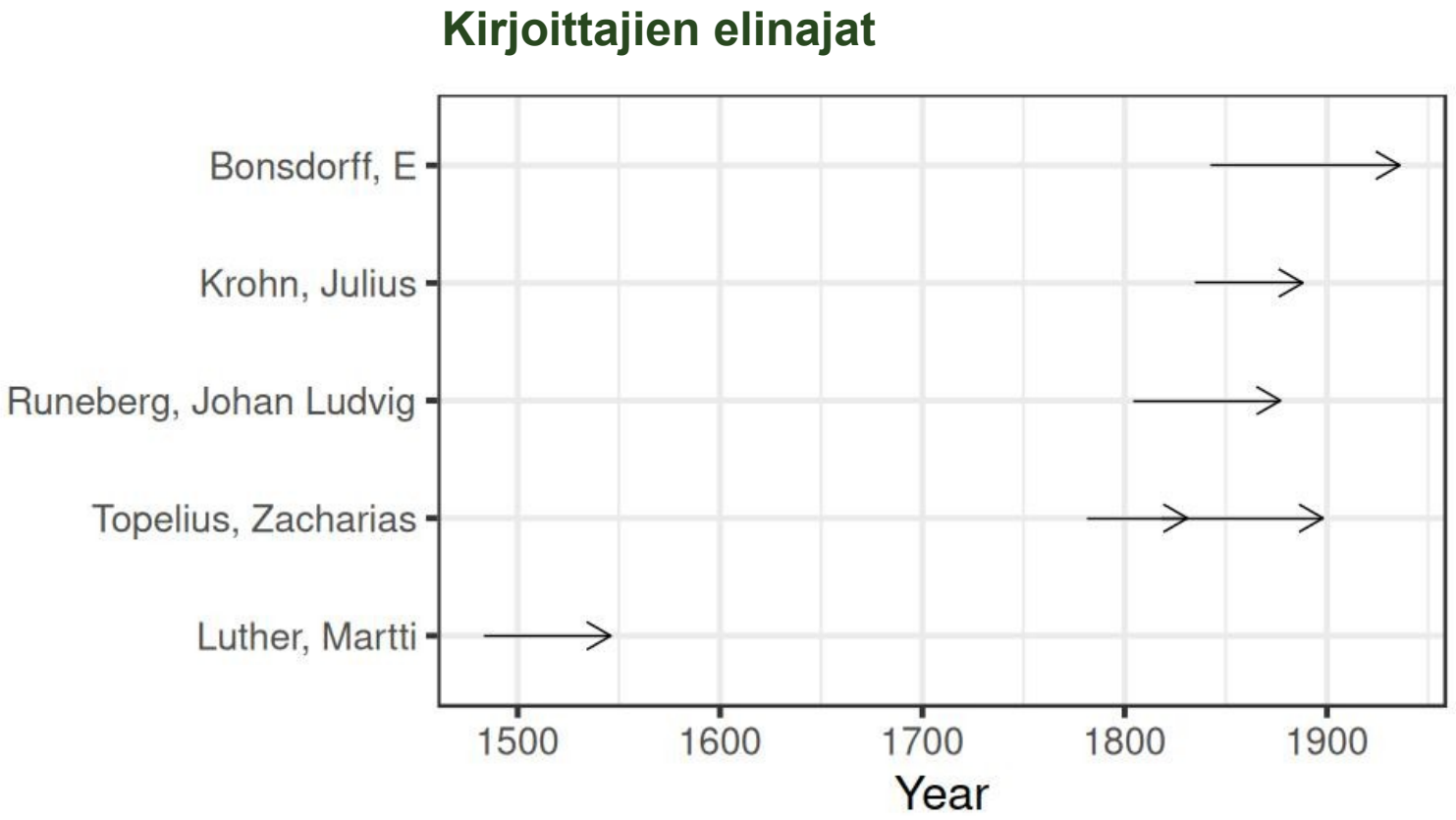
Mikko Tolonen[a], Leo Lahti[b], Hege Roivainen[a], and Jani Marjanen[a,*]

[a]Department of Digital Humanities, University of Helsinki, Finland; [b]Department of Mathematics and Statistics, University of Turku, Finland

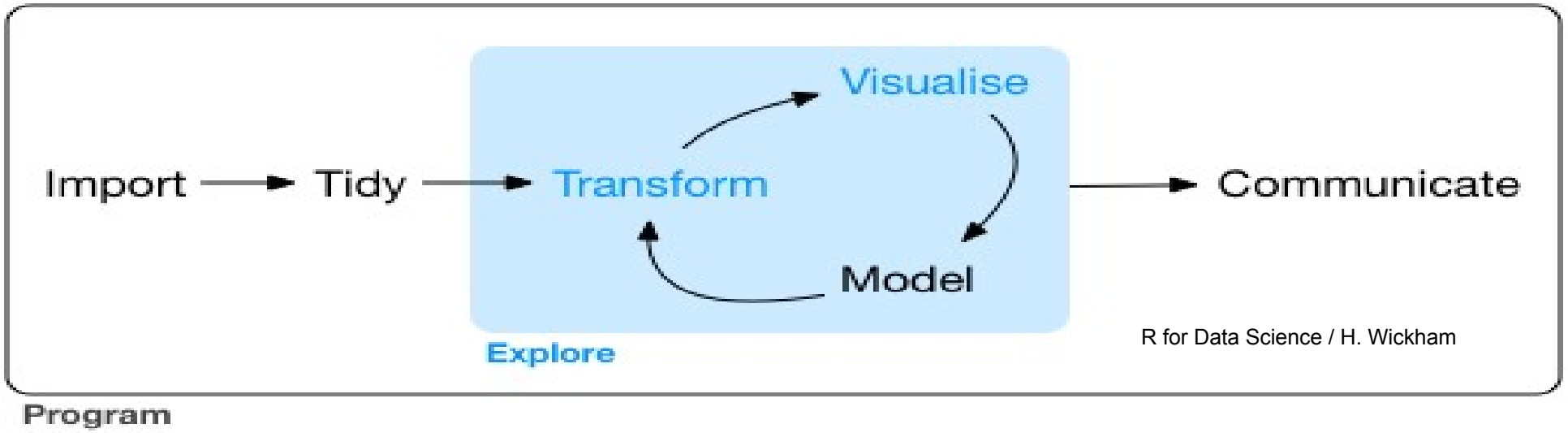


Sivumäärän arviointi?

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

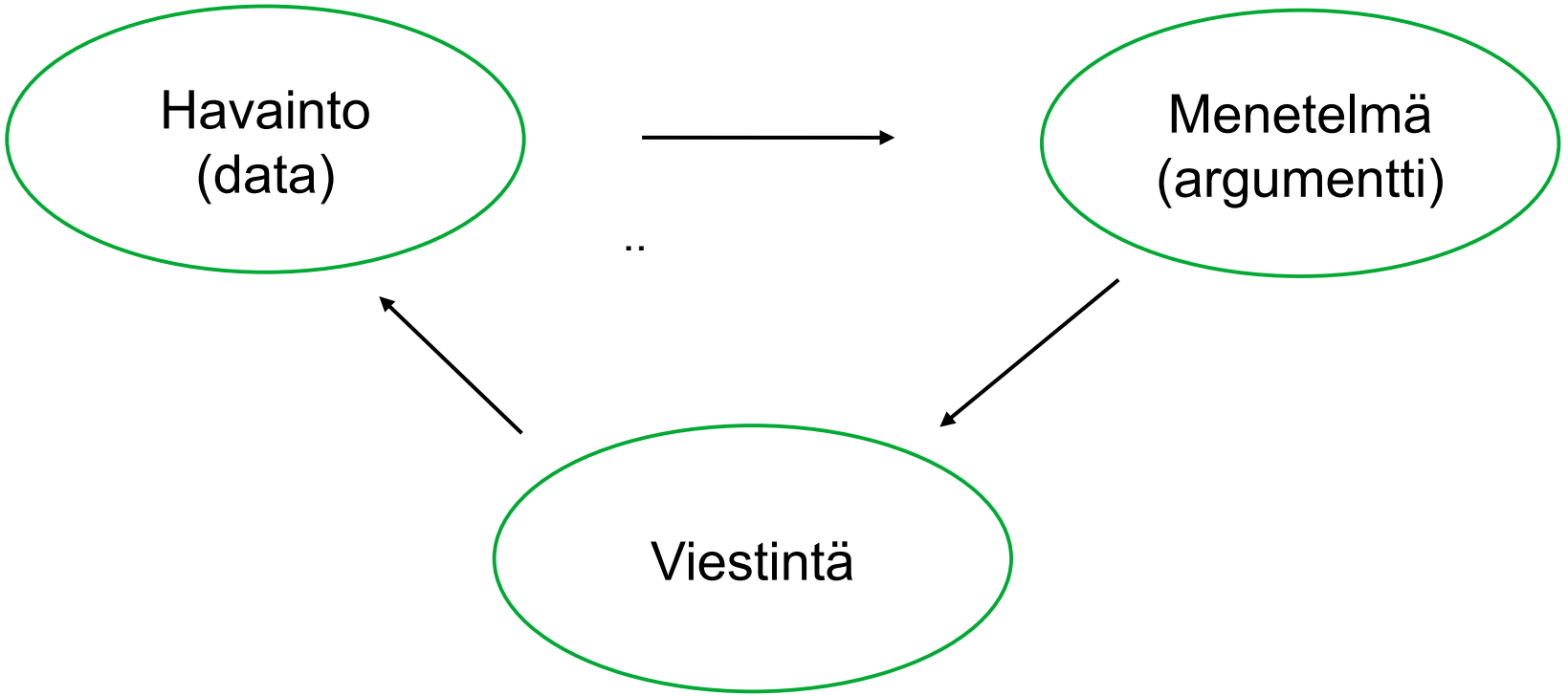


Datatieteen työvirta



R for Data Science / H. Wickham

Avoimen tieteen kansallinen julistus ja linjaukset



avointiede.fi

Toistettavat työvirrat CSC:n koneilla

https://fennica-fennica.2.rahtiapp.fi/

Julia Matveeva

Harmonized Finnish National Bibliography

Fennica metadata conversions: statistical monitoring and analysis

AUTHOR
Turku Data Science Group

PUBLISHED
February 13, 2025

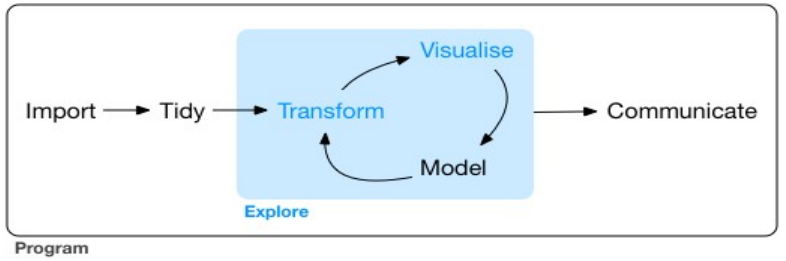
1 Preface chapter

It is imperative to underscore that this bookdown project is an evolving **work-in-progress**, and several additional fields will be incorporated into the dataset as they undergo the rigorous processes of data cleaning and harmonization.

Within the FIN-CLARIAH Metadata Harmonization and Analysis work package, we leverage the Finnish national bibliography (FNB) Fennica dataset to develop a harmonized dataset, serving research purposes and laying the groundwork for further infrastructure iterations. The project's outcomes will be instrumental in supporting the DHL-FI project funded by the Research Council of Finland. The FNB encompasses metadata for over a million documents, including books, newspapers, maps, etc., with records spanning from 1488 to the present. For more details about Fennica, visit The National Finnish Library website.

Currently, the bookdown project comprises a few distinct chapters. Notably, the harmonization process has been executed through the establishment of dual pipelines: **Complete FNB Pipeline** and **1809 to 1917 Period Pipeline** These chapters are dedicated to the specific metadata categories.

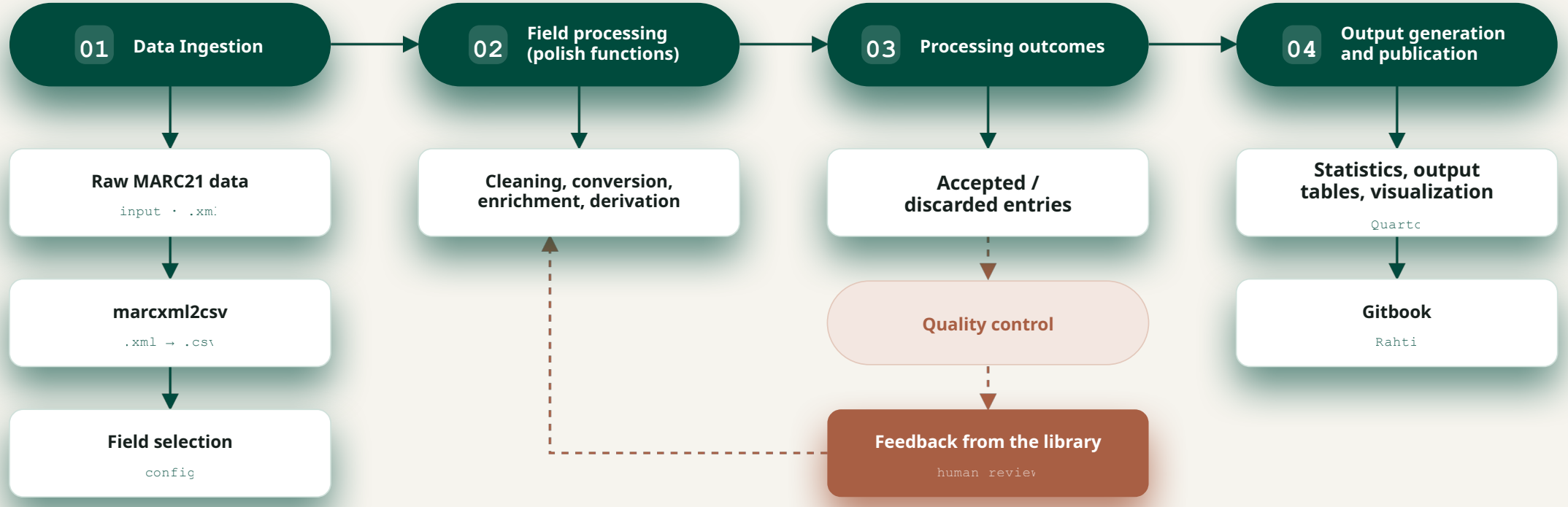
Grazin yliopiston kirjaston aiheluettelo – Wikimedia Commons.



Metadatan rikastaminen ja integrointi

Aineisto	Koko	Käyttö hankkeessa
Fennica – kansallisbibliografia	n. 60 000 tietuetta (1 milj. tietueesta)	Pääasiallinen bibliografinen analyysi
Melinda – yhteisluettelo	n. 60 000 tietuetta (16 milj. tietueesta)	Täydentää Fennicaa yksityiskohtaisemmilla tiedoilla muista bibliografisista tietokannoista
Kanto – kansallinen toimijanimitietokanta	n. 200 000 tietuetta	Tietoa kirjailijoista (myös salanimistä), muista henkilöistä, ryhmistä ja organisaatioista, joihin bibliografisessa metadatassa viitataan
Kansalliskirjaston digitoidut kaunokirjat	n. 1 000	Mesoanalyysi: lähiluvun ja määrällisen analyysin yhdistäminen Kansalliskirjaston Annif-palvelua hyödyntäen

Työnkulku & “human-in-the-loop”



Tiedonkulku — data flow Tarkistus & palaute — review loop Versionhallinta git

Automaattiset laadunvalvontaraportit

Avoin lähdekoodi

Palautesilmukka

Esimerkkejä rikastetuista kentistä

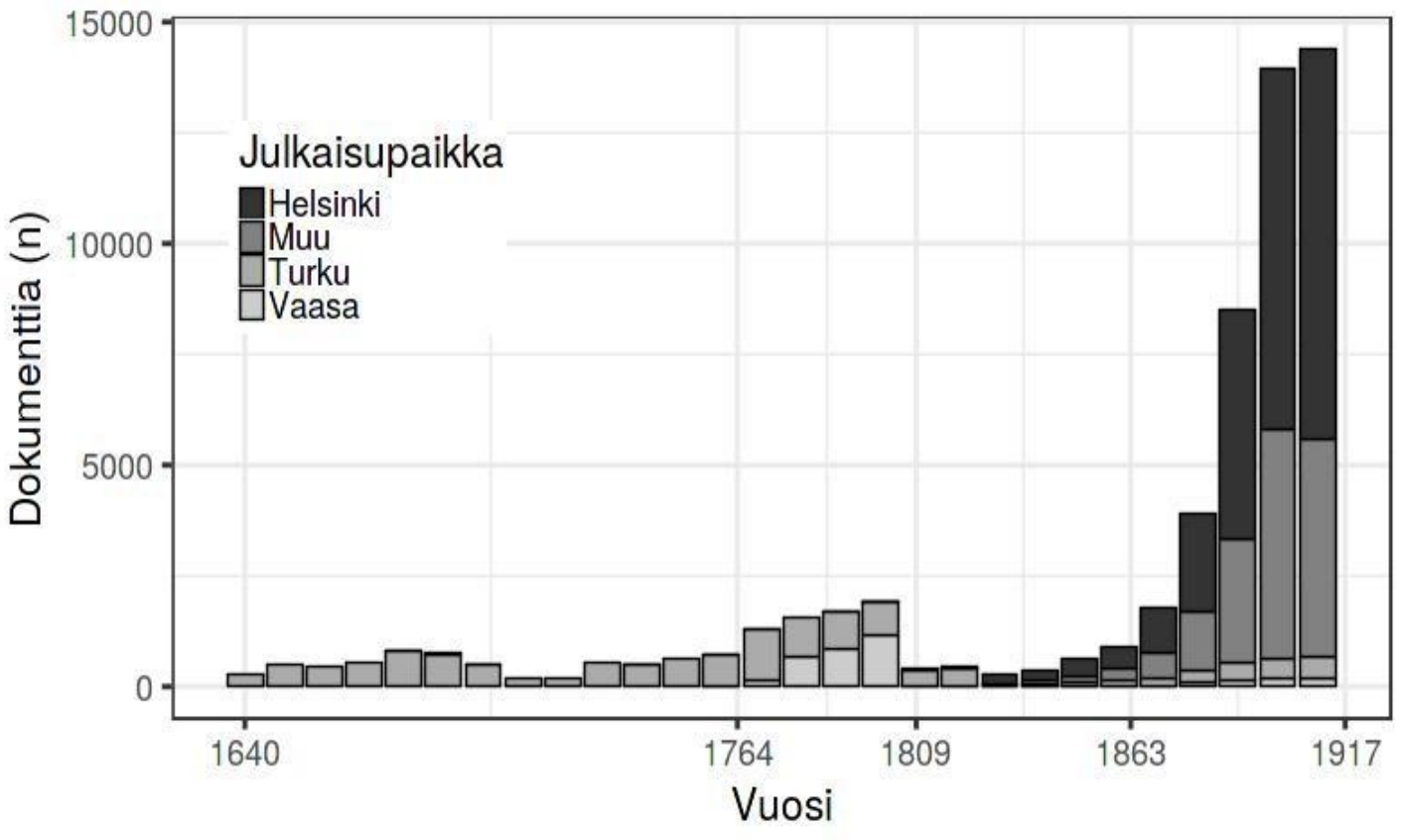
Tekijät	nimet, päivämäärät, ikä, sukupuoli
Nimekkeet	nimeke, alanimeke, sanamäärät
Kielet	kieli, pääkieli, alkukieli
Julkaisu	vuosi, vuosikymmen, paikka, maa, kustantaja
Lajityyppi/luokitus	MARC 008, MARC 655, UDK, hyllyluokka
Fyysinen muoto	sivut, niteet, osat, mitat
Tunnisteet	Melinda-ID, muut järjestelmätunnisteet

cc

Julkaisuaktiivisuus kaupungeittain

1809-1917 (N ~ 66,000 titles)

Kotimainen julkaiseminen 1640-1917 (nimikkeitten määrä)



Avoimen lähdekoodin “kirjastot”

Käyttö-, analyysi- ja visualisointityökalut (R-paketit)

1) **Fennica -** kansallisbibliografia
→ tilastollinen jalostus ja analyysi

2) **Finna** – kansallinen kulttuuriperinnön metadatakokoelma
→ laajentaminen muuhun metadataan (esim. musiikki / Viola; sisältää Fennican osajoukkona)

3) **Finto –** suomalainen asiasanasto- ja ontologiapalvelu
→ datan rikastaminen kirjailijoita, kustantajia ja paikkoja koskevilla yksityiskohtaisilla tiedoilla..

```r
library(geofi)
d1 <- get_municipalities(year = 2020)
d2 <- get_zipcodes(year = 2020)
d3 <- get_statistical_grid(resolution = 5)
d4 <- get_population_grid(resolution = 5)

library(ggplot2)
library(dplyr)
theme_set(
  theme_minimal(base_family = "Arial") +
  theme(legend.position= "none",
        axis.text = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank()
        )
)
p1 <- ggplot(d1, aes(fill = kunta)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "municipalities")
p2 <- ggplot(d1 %>% count(maakunta_code), aes(fill = maakunta_code)) + geom_sf(colour = alpha("white", 1/3)) + l
abs(subtitle = "Aggregated municipality data \nat region (maakunta) level \n(one of many!)")
p3 <- ggplot(d2, aes(fill = as.integer(posti_alue))) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "
zipcodes")
p4 <- ggplot(d3, aes(fill = nro)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "statistical grid")
p5 <- ggplot(d4, aes(fill = id_nro)) + geom_sf(colour = alpha("white", 1/3)) + labs(subtitle = "population gri
d")
p6 <- ggplot(municipality_central_localities, aes(color = as.integer(kuntatunnus))) + geom_sf() + labs(subtitle
= "Central municipality localities")

library(patchwork)
wrap_plots(list(p1,p2,p3,p4,p5,p6), ncol = 3) +
  patchwork::plot_annotation(title = "Spatial data in geofi-package")
```

finna

R-CMD-check (standard) passing | issues Unable to select next GitHub token from pool | pull requests 0 open

The `finna` package provides tools to access and analyze metadata from the Finna API, which aggregates content from Finnish archives, libraries, and museums.

finto 0.1.1 Reference Articles

finto

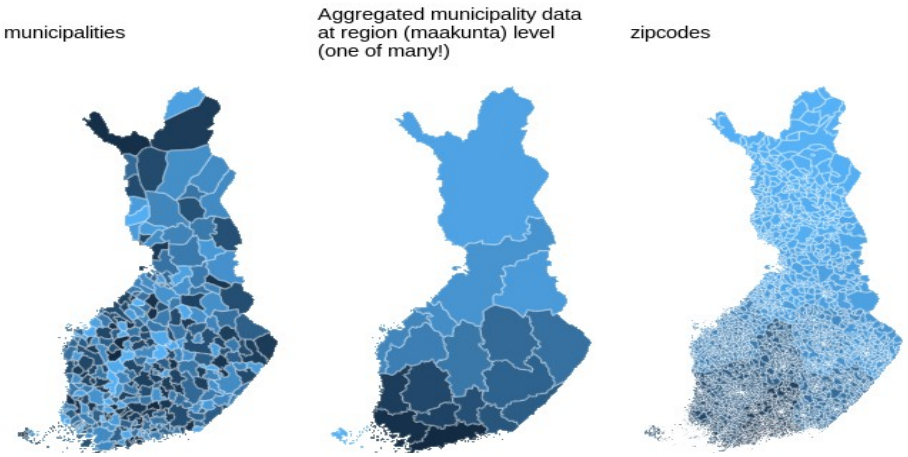
R-CMD-check (standard) passing | issues 0 open | pull requests 0 open

The `finto` package provides tools to access the serivice for interoperable thesauri, ontologies and classification schemes for different subject areas.

R-CMD-check passing | repo status Active | codecov 63% | chat on gitter | Watchers 30
CRAN 1.0.4 | downloads 3836 | downloads 398/month

geofi - Access Finnish Geospatial Data

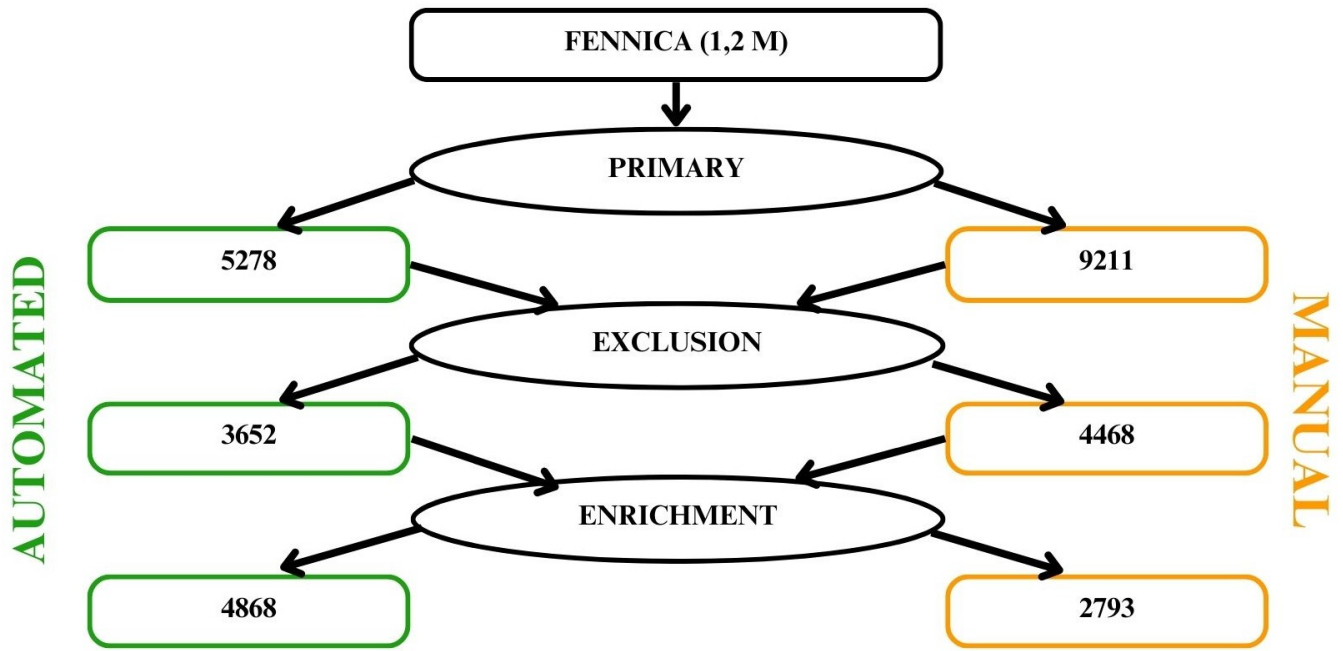
rOpenGov



Työvaiheita:

- *Datan jalostus* laajamittaiseen tilastolliseen analyysiin
- ***Otoksen valinta kansallisbibliografiasta*: automatisoitu ja manuaalinen kuratointi**
- *Tilastollinen analyysi* (esim. taustatrendien tunnistus)

→ *Datatieteen menetelmien kehitys*



Osajoukkojen valinta bibliografisessa tutkimuksessa: automaattisen ja manuaalisen kuratoinnin rajojen tarkastelu

Tehtävä: tunnista 1800-luvun suomalainen aikuisille kirjoitettu kaunokirjallisuus

	Manuaalinen työnkulku	Automaattinen työnkulku
Ensisijaiset kriteerit *aikakausi, kieli, signum*	Aikuisten kirjallisuus, julkaistu suomeksi ja ruotsiksi 1809–1917	
Poissulkemiskriteerit *esim. lastenkirjallisuus*	Käännösten ja lastenkirjallisuuden manuaalinen poisto	Sääntöpohjainen poissulkeminen metadatan avulla
Rikastamiskriteerit *esim. lajityypit, ensipainokset*	Reenpää-kokoelma ja asiantuntijalisäykset	UDK- ja lajityyppimetadatan rikastaminen
Ensipainokset	Asiantuntijavalidointi	Automaattinen heuristinen tunnistus
Tulos	**2 793 ensipainosta**	**4 868 ehdokasensipainosta**

CC BY

Havaintoja

- Automatisoitu otos toisti onnistuneesti 95 % asiantuntijoiden kuratoimasta korpuksesta.
- Metadatan rikastaminen paransi kattavuutta merkittävästi ja tunnisti **2 375 uutta ehdokasnimekettä**.
- Manuaalinen validointi on edelleen tarpeen tehtävissä, joita ei voida luotettavasti päätellä metadatasta, erityisesti ensipainosten tunnistamisessa.
- Metadatan epäjohdonmukaisuudet, puuttuvat arvot ja ristiriitaiset luokitukset vaikuttavat suoraan mukaanotto- ja poissulkemispäätöksiin.
- Automaattiset työnkulut voivat auttaa paljastamaan piileviä luettelointivirheitä ja metadatan laatuongelmia.
- Tarvittavan manuaalisen työn määrä riippuu pitkälti metadatan täydellisyydestä ja laadusta.
- Osajoukon valinta ei ole neutraali tekninen vaihe; se muokkaa korpusta ja vaikuttaa siten kirjallisuushistorialliseen tulkintaan.

Keskeinen havainto

- Tehokkain strategia on hybridityönkulku:

Ensin automaattinen valinta ja rikastaminen → sitten asiantuntijavalidointi.

- Automaatio tuo skaalautuvuutta, läpinäkyvyyttä ja toistettavuutta, kun taas asiantuntija-arviointi ratkaisee monitulkintaisuudet, joita bibliografinen metadata ei kata.

TURUN YLIOPISTO

Otoksen valinta bibliografisessa tutkimuksessa

automaattisen ja manuaalisen kuratoinnin vertailua

- Kirjallisuushistoriallinen tutkimus nojaa huolellisesti valittuihin otoksiin
- Perinteisesti teosten valinta on manuaalista asiantuntijatyötä

Manuaalinen kuratointi

- Aikaa vievä
- Vaikea toistaa
- Mahdollisesti subjektiivinen
- Asiantuntijoiden työmäärä rajallista
- Tarkkuutta vaativaa

Automaattinen kuratointi

- Skaalautuva
- Toistettavissa
- Läpinäkyvä
- Käsittelee suuria aineistoja
- Kriteerien johdonmukainen soveltaminen

Voidaanko asiantuntijatyö automatisoida, ja samalla parantaa läpinäkyvyyttä ja skaalautuvuutta?

Tapaustutkimus: kaunokirjallisuuden ensipainokset (1809–1917) kansallisbibliografiasta (Fennica)

CC BY

TURUN YLIOPISTO

Yhteenveto

Toteutuneita työvaiheita:

- *Datan jalostus* laajamittaiseen tilastolliseen analyysiin
- *Otoksen valinta kansallisbibliografiasta*: automatisoitu ja manuaalinen kuratointi
- *Tilastollinen analyysi* (esim. taustatrendien tunnistus)

→ *Datatieteen menetelmien kehitys*

Jalostetuilla aineistoilla voimme nyt tutkia aiempaa luotettavammin:

- lajityyppien kehitystä ajan myötä
- suomen- ja ruotsinkielisen kirjallisuuden eroja
- sukupuolen ja lajityypin suhteita
- julkaisemisen maantieteellistä vaihtelua
- julkaisutoiminnan dynamiikkaa

Huomioita:

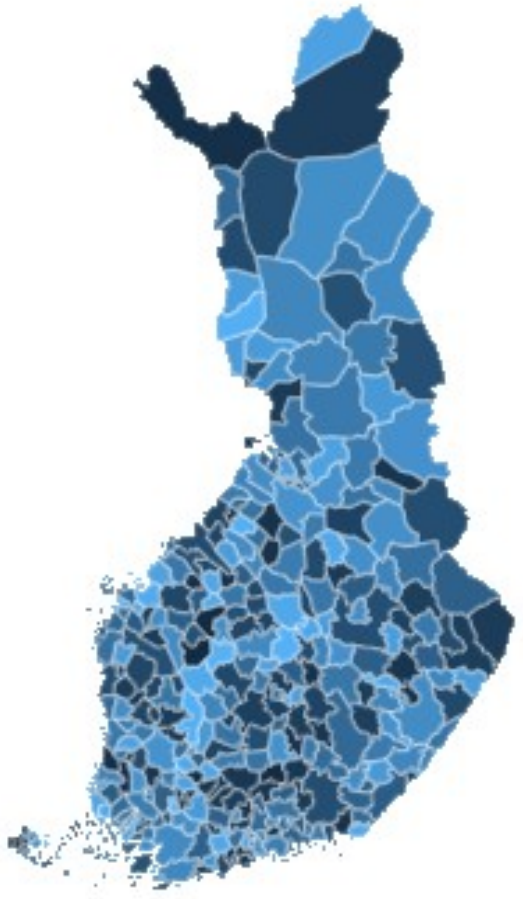
- Metadatan laatu tärkeä rajoittava tekijä
- Läpinäkyvyys, toistettavuus, ja skaalautuvuus paranevat
- Jatkuva kehitys ja uudelleenkäyttö
- Jalostus ja rikastaminen nostavat bibliografioiden tutkimusarvoa
- Tieteidenvälinen yhteistyö keskeistä

cc

TURUN YLIOPISTO

Kiitos!

- Artikkeleita mm. Data Science-lehdessä ja R Journalissa
- Australian tutkimusvierailut (Julia), konferensseissa, ja luennoilla
- Avoin datatieteen työvirta Fennicalle (Kansalliskirjaston & CSC:n kanssa)
- **Jatkossa: aineistojen ja analyysien syventäminen & laajentaminen**



Kati Launis (Itä-Suomen yliopisto)
Osma Suominen et co. (Kansalliskirjasto)
Veli-Matti Pynttäri (SKS), Viola Capkova (Turun yliopisto)
Julia Matveeva, Jeba Akewak, Pyry Kantanen, Ville Laitinen (Turun yliopisto)

CC BY

datascience.utu.fi

UNIVERSITY OF TURKU