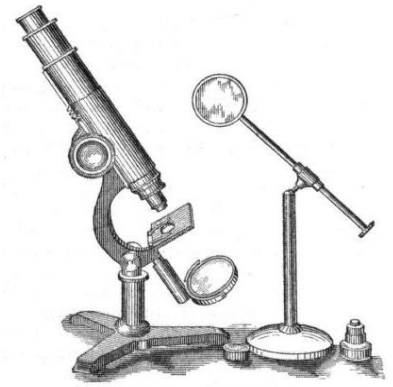
Open Data Science



Leo Lahti
University of Turku, Finland

@antagomir

Turun yliopisto
University of Turku

Transparent reporting and communication were part of academic culture since the early days

Source: Wikimedia Commons / Public domain

Alchemy & algorithms: perspectives on the philosophy and history of open science

Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenoja, Mikko Tolonen

Beyond Open Access - The Changing Culture of Producing and Disseminating Scientific Knowledge

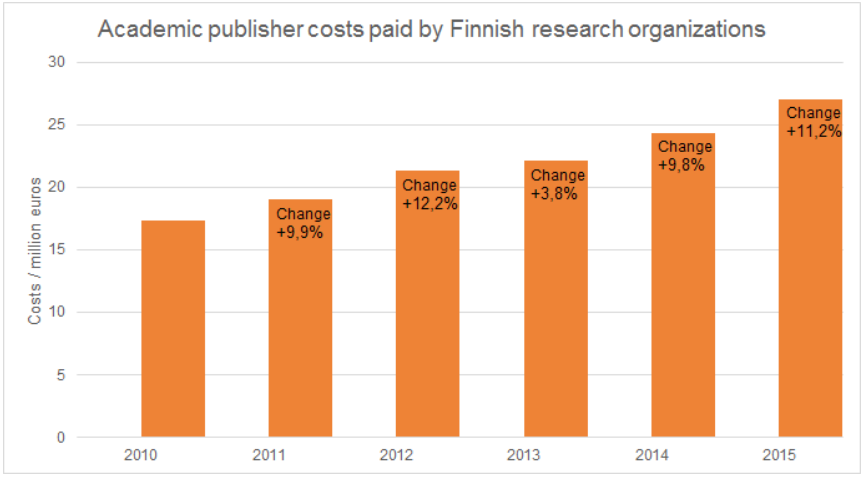
Heidi Laine Leo Lahti Anne Lehto

Finland takes leading role in the openness of academic journal pricing

June 13, 2016

Journal subscription prices 2010-2017 for ~70 Finnish institutions released as open data from MoE. Only after Freedom of Information request and court decision..

Rapid increase in subscription journal costs (roughly 10% in Finland, every year).



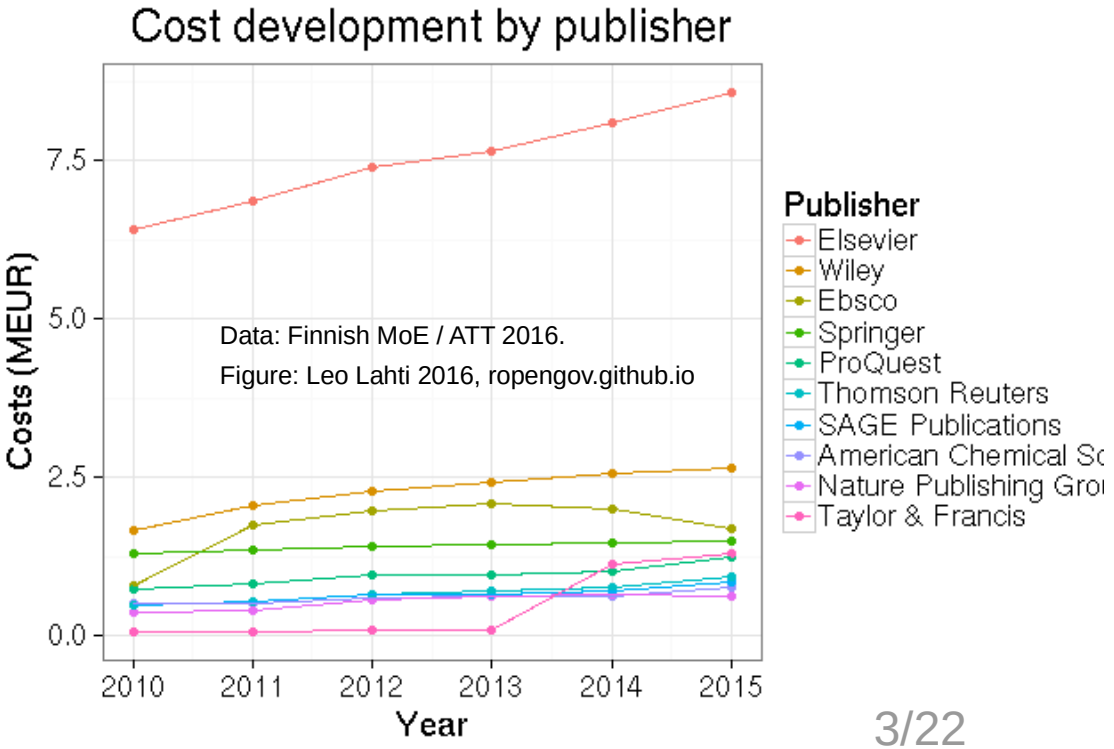


Figure & Data release: Ministry of Education and Culture of Finland / Open Science and Research Initiative 2014–2017

Computational workflows have now increasingly central role in research: challenge & opportunity

1) data gathering and storage
2) access, documentation
3) harmonization & enrichment
4) quality control
5) custom analysis tools and workflows
6) reporting & dissemination

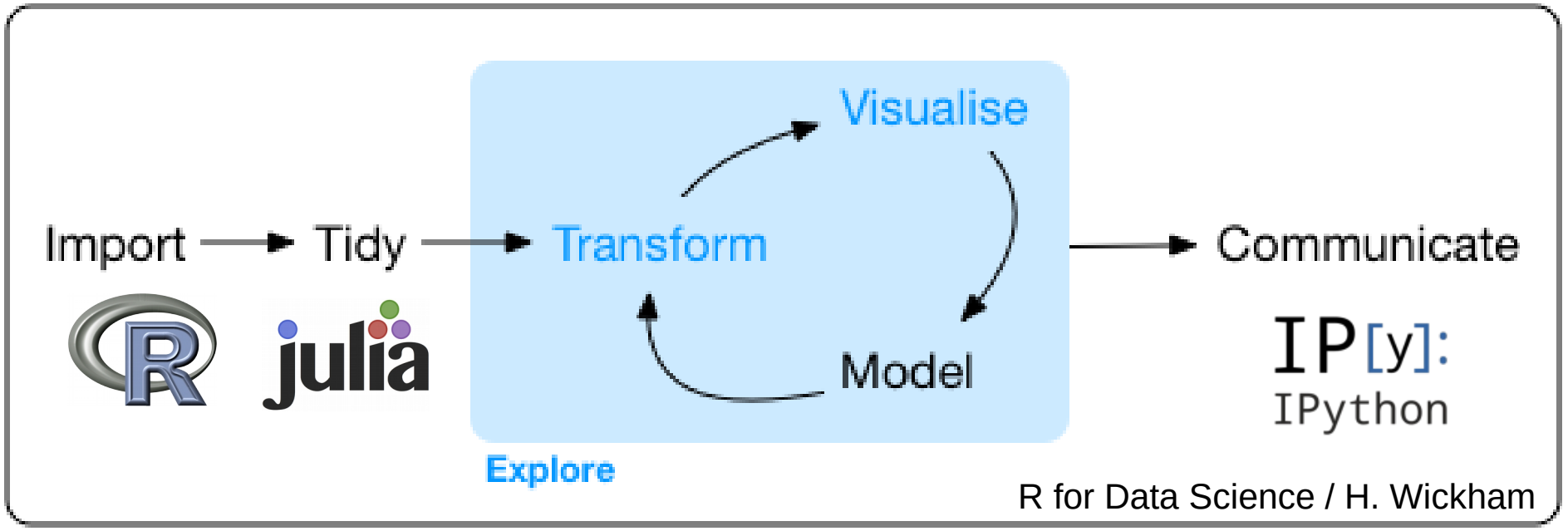
Science 13 April 2012:
Vol. 336 no. 6078 pp. 159-160
DOI: 10.1126/science.1218263

POLICY FORUM

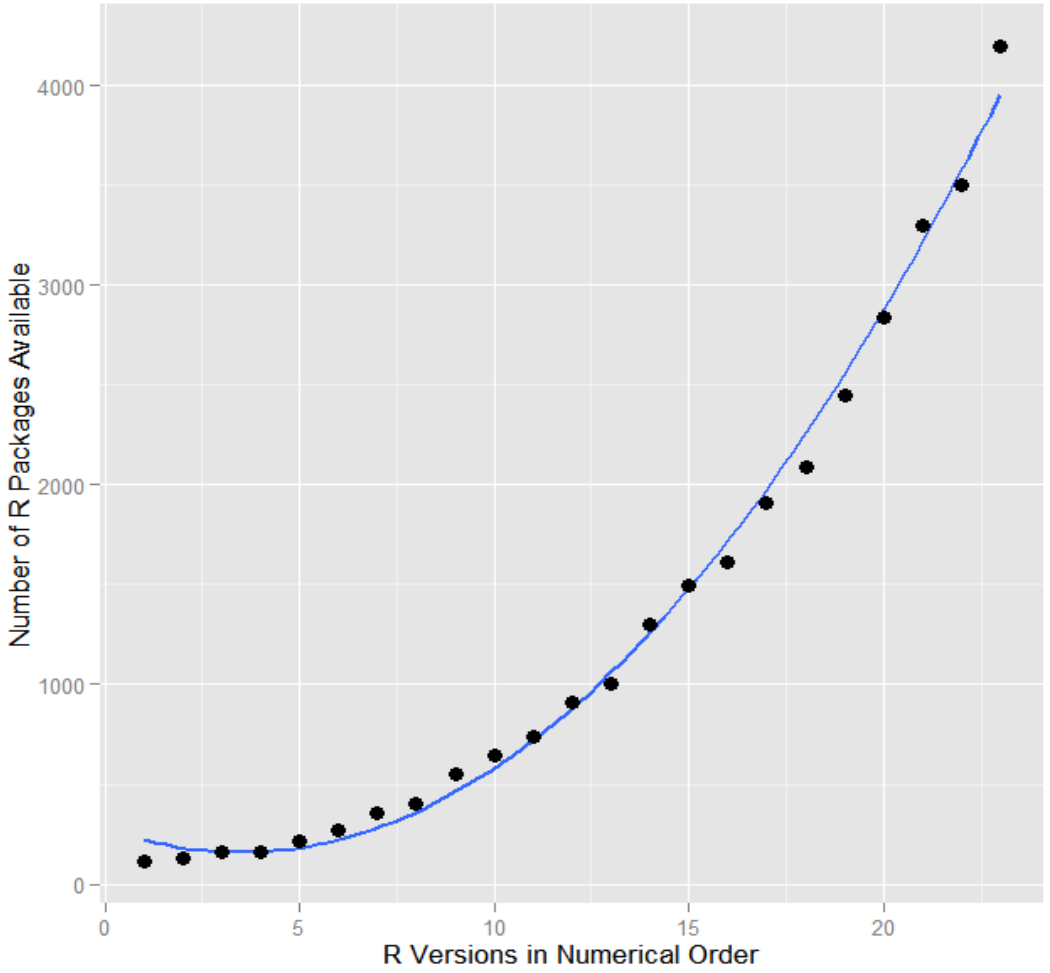
RESEARCH PRIORITIES

Shining Light into Black Boxes

A. Morin[1], J. Urban[2], P. D. Adams[3], I. Foster[4], A. Sali[5], D. Baker[6], P. Sliz[1,*]



Number of open analysis tools has grown exponentially



rOpenGov

ROpenSci

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

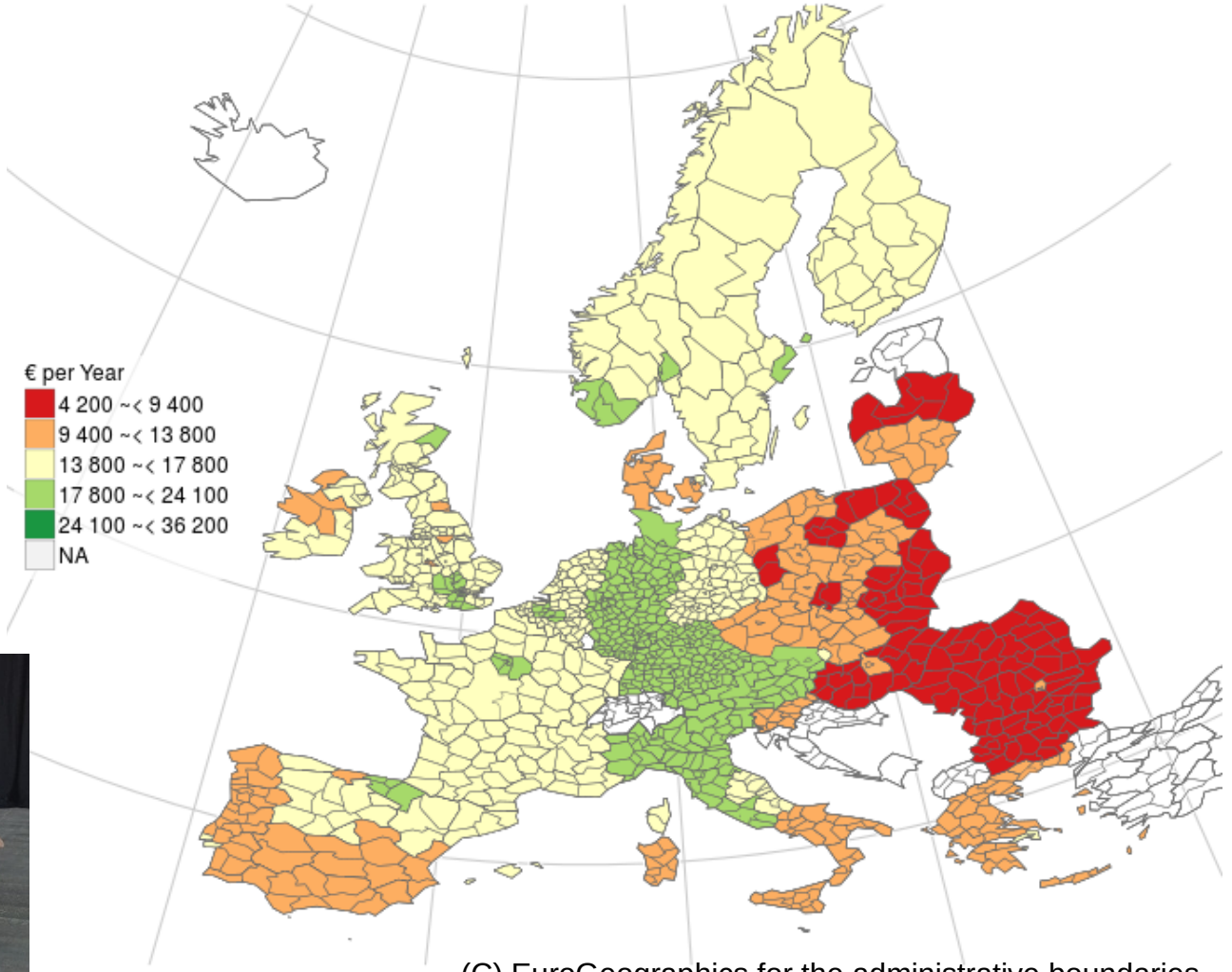
Retrieval and Analysis of Eurostat Open Data with the eurostat Package

by Leo Lahti, Janne Huovari, Markus Kainu, and Przemysław Biecek

rOpenGov

International network for open government data analytics

- 20+ R packages
- 100k+ downloads
- open collaboration



(C) EuroGeographics for the administrative boundaries

Eurostat open data: average household expenditure in 2011

The *eurostat* package

R tools to access open data from Eurostat database

Search and download

Data in the Eurostat database is stored in tables. Each table has an identifier, a short table_code, and a description (e.g. *tsdtr420* - People killed in road accidents).
Key eurostat functions allow to find the table_code, download the eurostat table and polish labels in the table.

Find the table code

The **search_eurostat(pattern, ...)** function scans the directory of Eurostat tables and returns codes and descriptions of tables that match pattern.

```
library("eurostat")
query <- search_eurostat("road", type = "table")
query[1:3,1:2]
##                                 title     code
## 1           Goods transport by road ttr00005
## 2   People killed in road accidents tsdtr420
## 3 Enterprises with broadband access tin00090
```

Download the table

The **get_eurostat(id, time_format = "date", filters = "none", type = "code", cache = TRUE, ...)** function downloads the requested table from the *Eurostat bulk download facility* or from *The Eurostat Web Services JSON API* (if **filters** are defined). Downloaded data is cached (if **cache=TRUE**). Additional arguments define how to read the time column (**time_format**) and if table dimensions shall be kept as codes or converted to labels (**type**).

```
dat <- get_eurostat(id="tsdtr420", time_format="num")
head(dat)
##     unit    sex    geo  time values
## 1     NR      T     AT  1999   1079
## 2     NR      T     BE  1999   1397
## 3     NR      T     CZ  1999   1455
## 4     NR      T     DK  1999    514
## 5     NR      T     EL  1999   2116
## 6     NR      T     ES  1999   5738
```

Add labels

The **label_eurostat(x, lang = "en", ...)** gets definitions for Eurostat codes and replace them with labels in given language ("en", "fr" or "de").

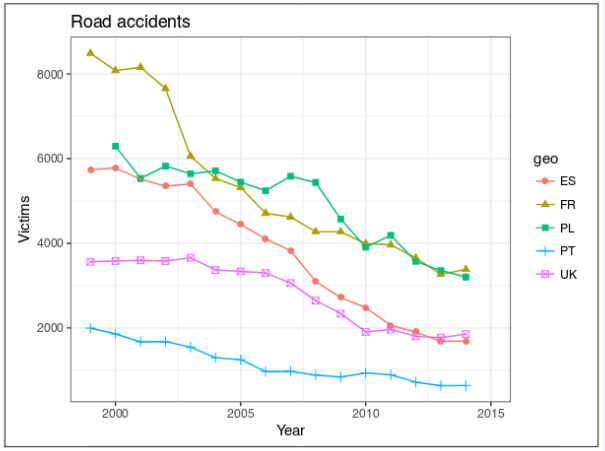
```
dat <- label_eurostat(dat)
head(dat)
##     unit    sex            geo  time values
## 1 Number  Total        Austria  1999   1079
## 2 Number  Total        Belgium  1999   1397
## 3 Number  Total Czech Republic  1999   1455
## 4 Number  Total        Denmark  1999    514
## 5 Number  Total         Greece  1999   2116
## 6 Number  Total          Spain  1999   5738
```

eurostat and plots

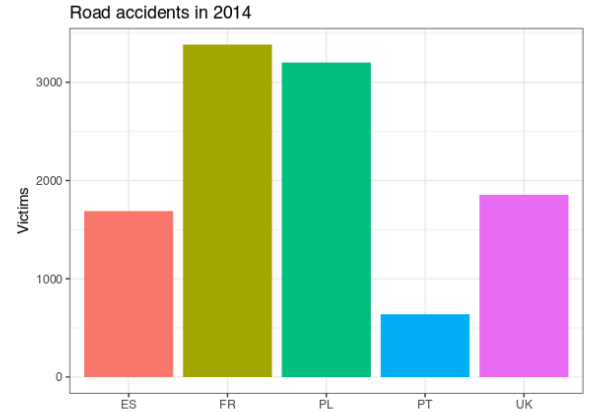
The **get_eurostat()** function returns tibbles in the long format. Packages **dplyr** and **tidyr** are well suited to transform these objects. The **ggplot2** package is well suited to plot these objects.

```
t1 <- get_eurostat("tsdtr420", filters =
            list(geo = c("UK", "FR", "PL", "ES", "PT")))

library("ggplot2")
ggplot(t1, aes(x = time, y = values, color = geo,
               group = geo, shape = geo)) +
  geom_point(size = 2) +
  geom_line() + theme_bw() +
  labs(title="Road accidents", x = "Year", y = "Victims")
```




```
library("dplyr")
t2 <- t1 %>% filter(time == "2014-01-01")
ggplot(t2, aes(geo, values, fill=geo)) +
  geom_bar(stat = "identity") + theme_bw() +
  theme(legend.position = "none")+
  labs(title="Road accidents in 2014", x="", y="Victims")
```



Documentation & cheat sheets

Online tutorials & blog posts

Issue tracker

Automated unit tests

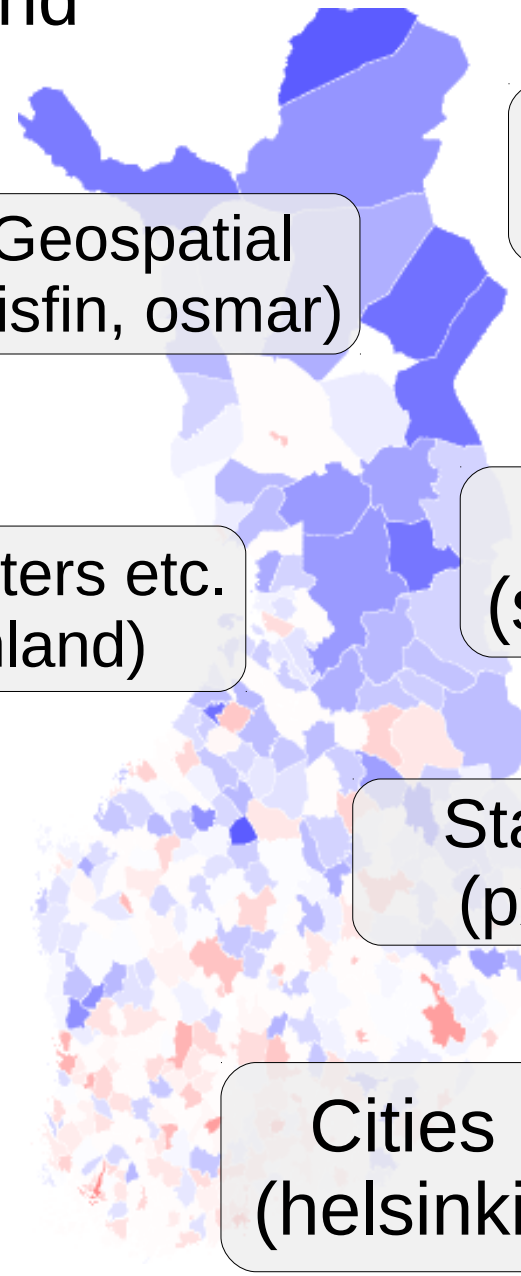
Project homepage

This onepager presents the **eurostat** package
Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek 2014-2017 package version 2.2.43 URL: *https://github.com/rOpenGov/eurostat*

From specific packages to package ecosystems

Open Street Map
Helsinki (osmar)

Algorithms for open data in Finland



Weather (fmi)

Geospatial (gisfin, osmar)

Health (sotkanet)

Registers etc. (finland)

Statistics (pxweb)

Cities (helsinki)

pxweb for PX-Web/PC-Axis data from stats authorities in: Denmark, Finland, Greenland, Iceland, Latvia, Norway, Sweden..

Analysis & Figure: Janne Sinkkonen

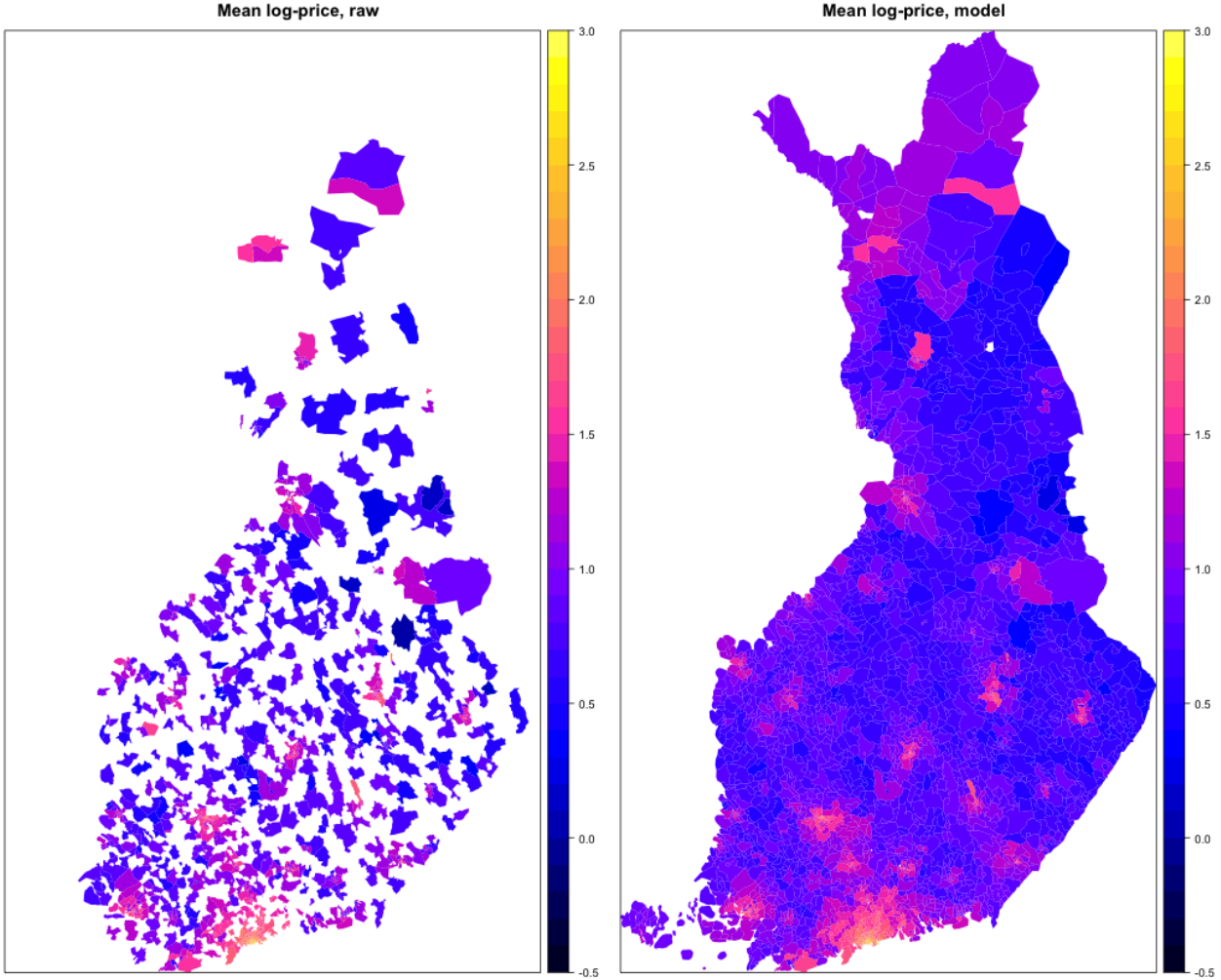
BY JANNE SINKKONEN — JUNE 11, 2015

A hierarchical model of Finnish apartment prices

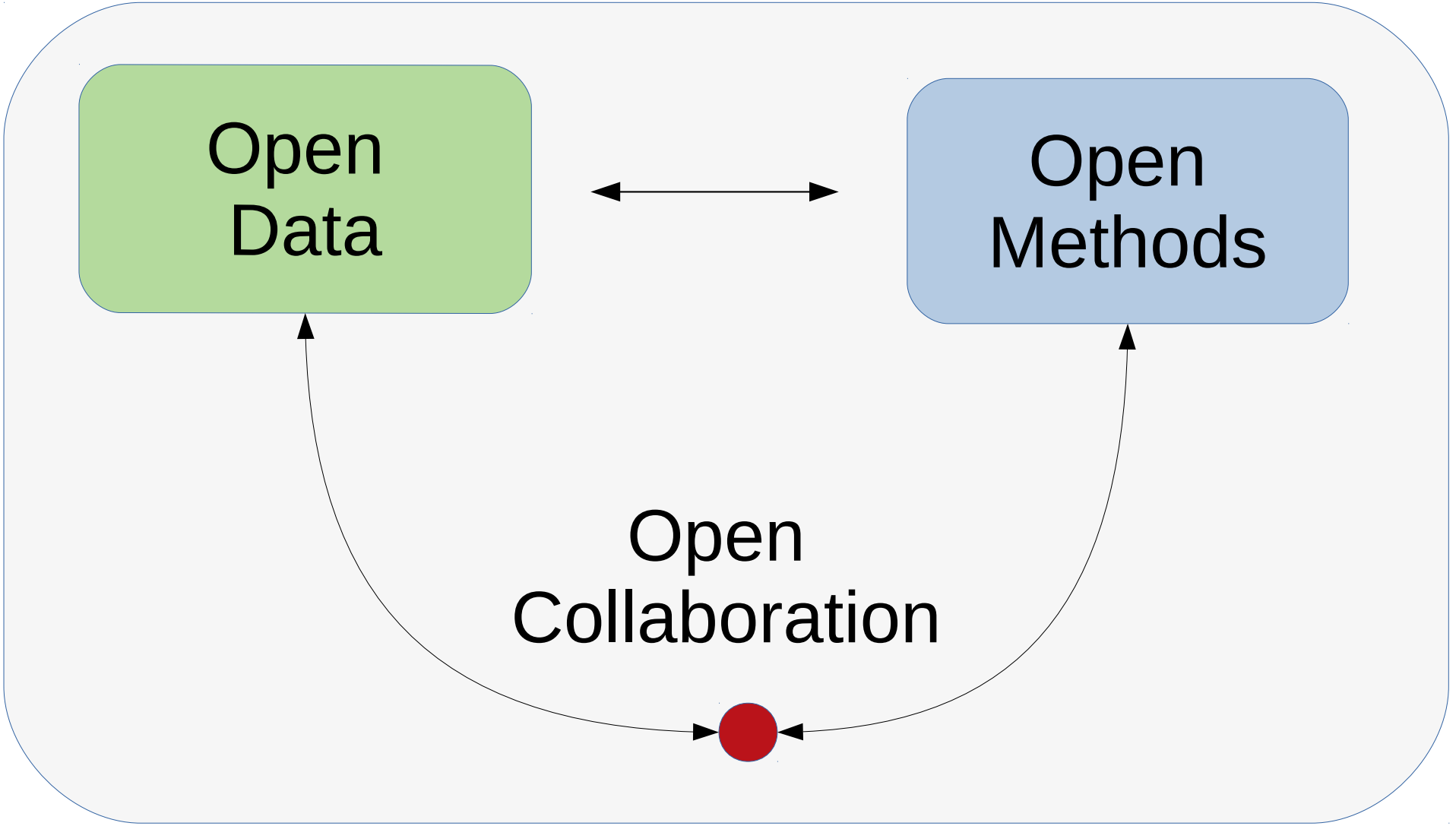
Basing on open data from Statistics Finland, we at Reaktor modelled Finnish apartment prices and their trends on zip-code level, in the years 2005–2014. Estimates from the model are available as an interactive visualization.

Original data: mean price per postal code area

Probabilistic model (Stan): smooth price estimates



Elements of open data science



L Lahti. Advances in Intelligent Data Analysis XVII. Springer, 2018.

Academy of Finland COMHIS consortium 2016-2019

University of Helsinki, University of Turku, National Library of Finland

Quantify early-modern knowledge production with massive bibliographic collections, full texts, and open data analytical infrastructure

Catalogue	Earliest Record	Records 1500-1800 (N)	Language available	Publication place available	Page count available	Gatherings available
FNB	1488	16365	100.0%	93.9%	99.9%	98.3%
SNB	1457	46764	100.0%	95.0%	99.9%	84.8%
ESTC	1473	479790	100.0%	99.4%	99.9%	97.0%
HPBD	1446	2095628	100.00%	86.7%	99.5%	45.3%

Fennica (FNB)

Finnish National bibliography

- >900,000 books and monographies (printed and electronic) since 1488
- >70,000 continuous publications (journals or series) since 1771
- Series, maps, audiovisual, and electronic material
- **Open data**

Other catalogues

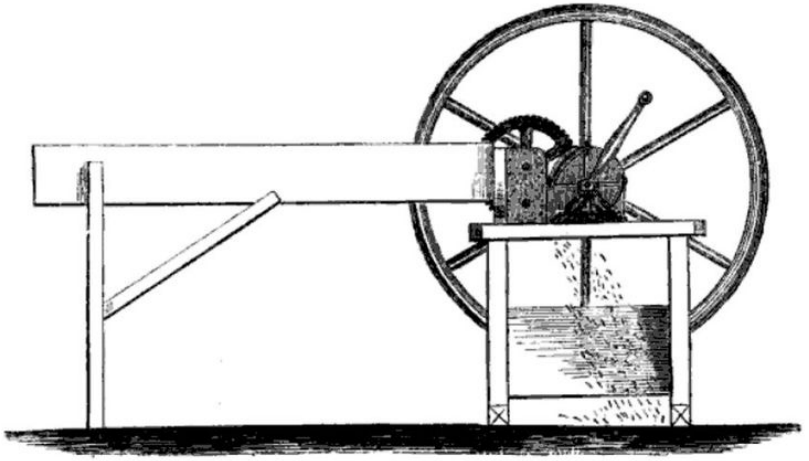
Swedish National Bibliography (>18M entries)

British Library (>0.5M)

Heritage of the Printed Book (Göttingen; 6M entries)

Data harmonization: estimating page counts from MARC cataloguing standards

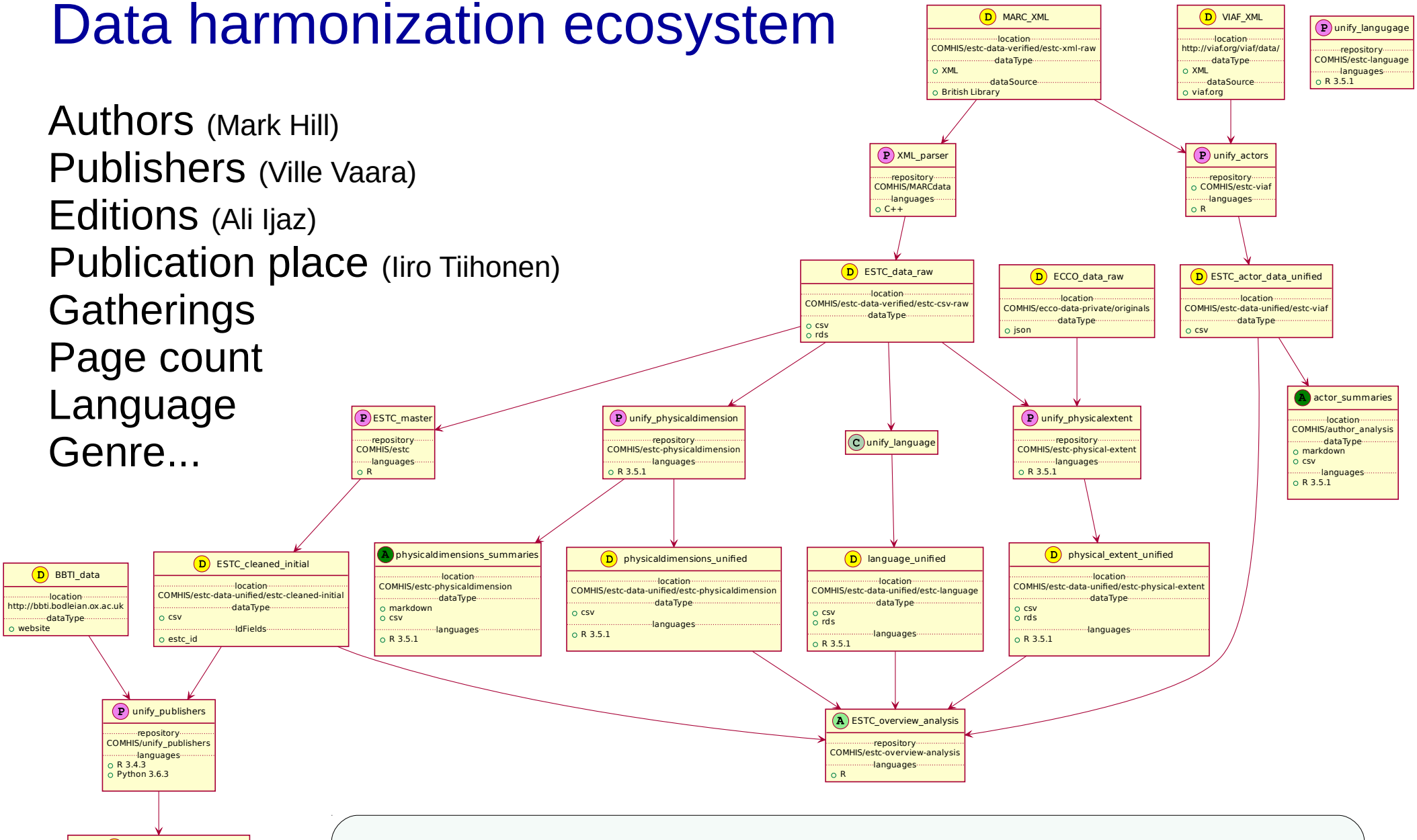
“[4],vii-xii,[4],222p.,plate”



→ 240 pages

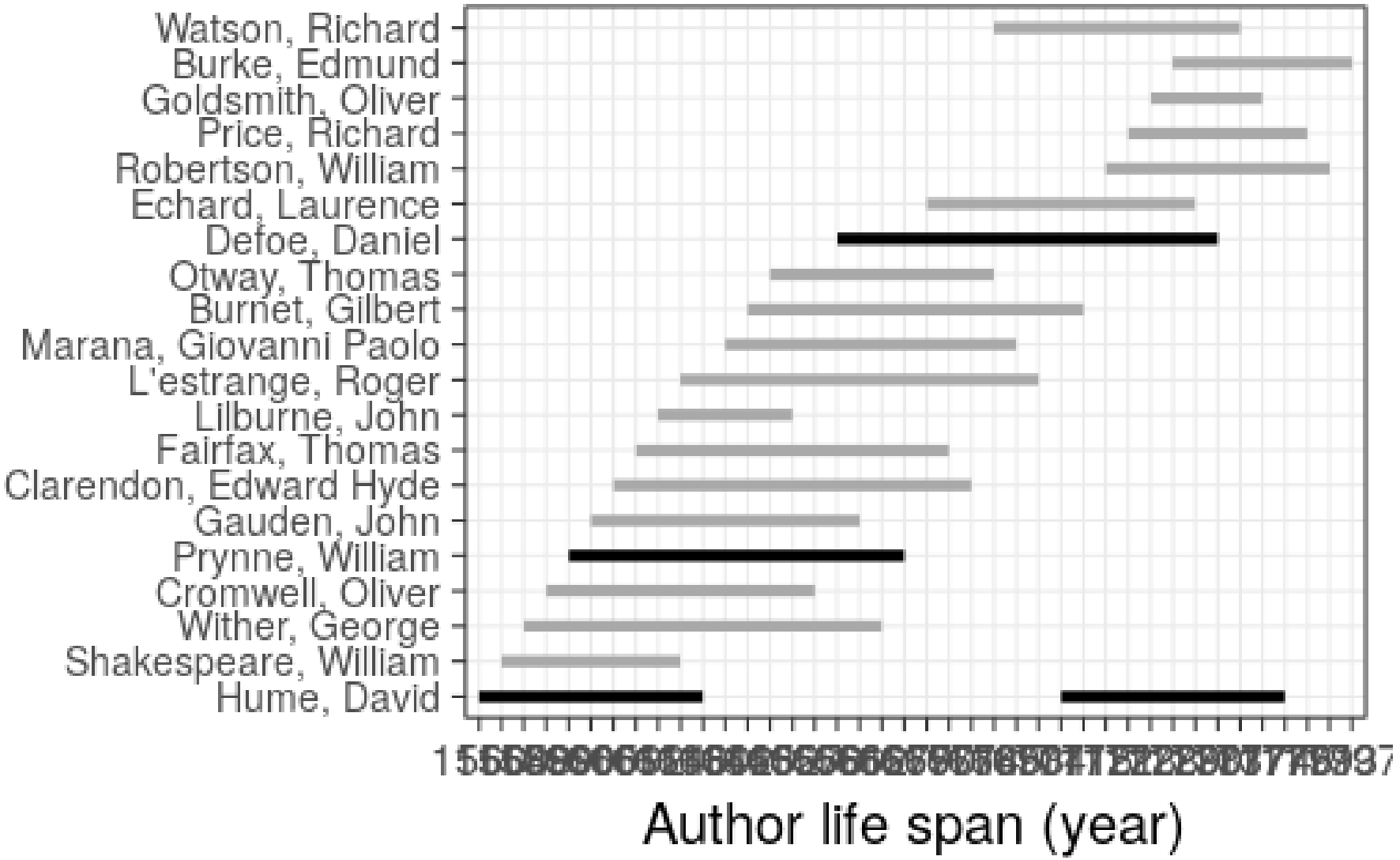
Data harmonization ecosystem

Authors (Mark Hill)
Publishers (Ville Vaara)
Editions (Ali Ijaz)
Publication place (Iiro Tiihonen)
Gatherings
Page count
Language
Genre...

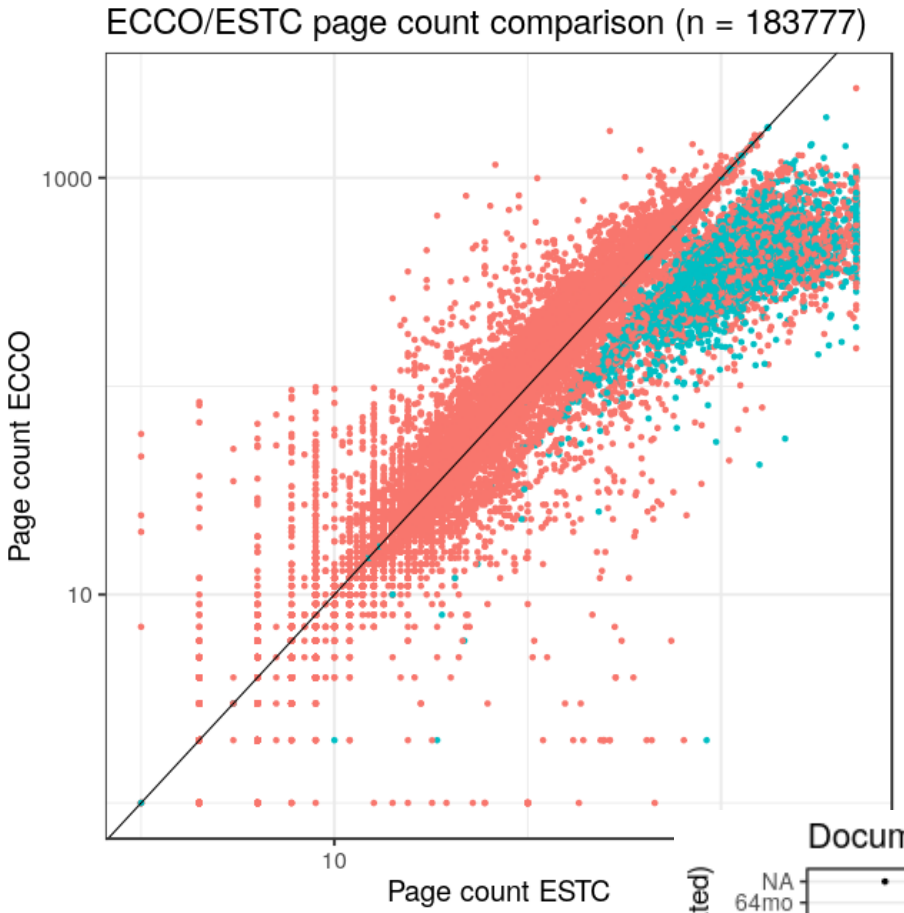


Data → Standardization → Validation → Integration → Analysis

Automated quality control: author life span



Validation: page count (ESTC ~ ECCO)



Good match:
estimated and
curated page counts

pagecount.estimated
FALSE
TRUE

Information for certain book formats and time periods is missing more often → bias?

Fennica: analysis of the Finnish national bibliography

This repository contains automated analysis of the Finnish national bibliography, Fennica. Fennica includes bibliographic metadata for over 70,000 documents between 1488-1955, representing the publishing activity in Finland during that period. This is analyzed in parallel with Kungliga, a related collection of bibliographic metadata from the Swedish National library.

The research project is funded by Academy of Finland 2016-2019.

Reproducible analysis

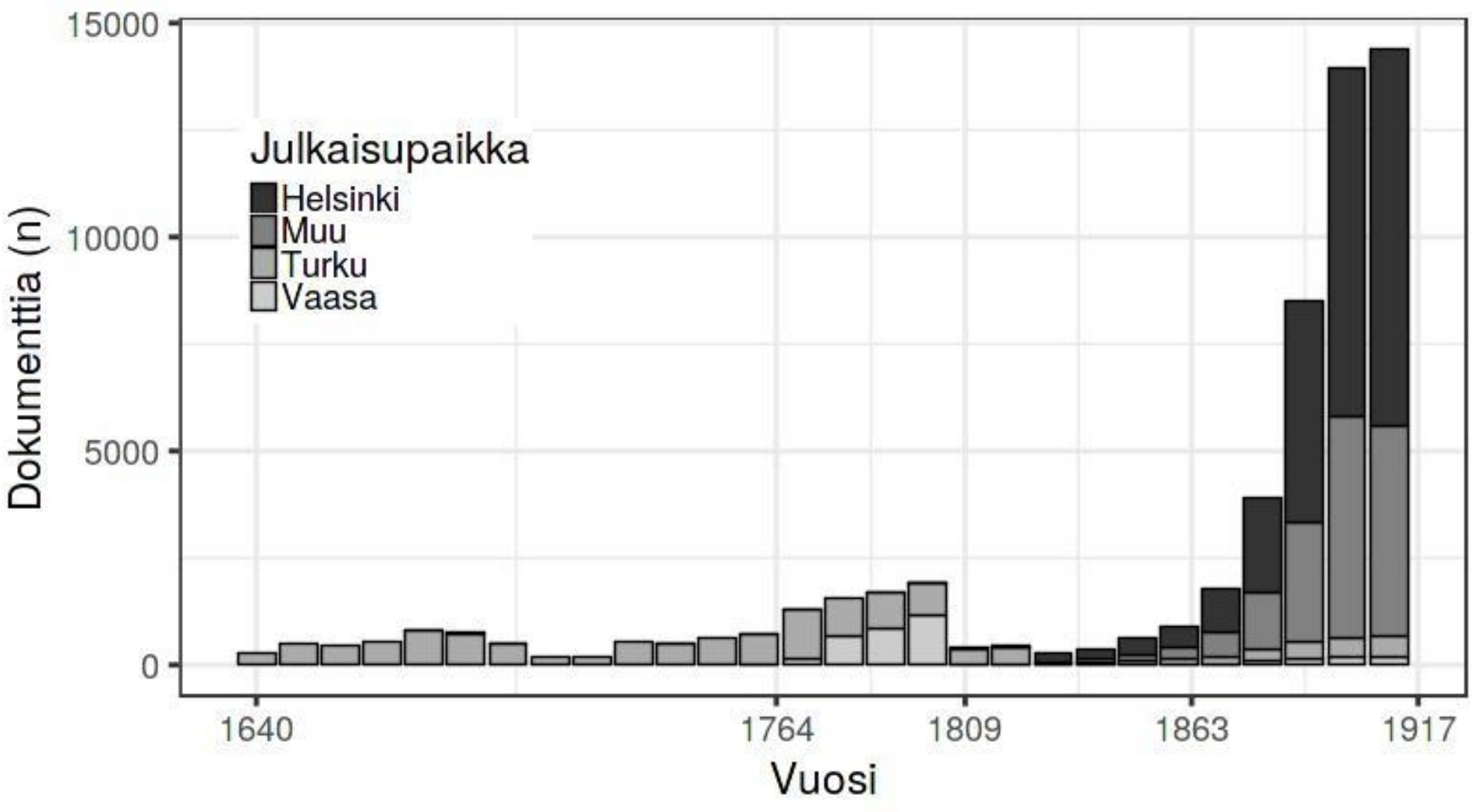
The data is summarized in the following automatically generated files:

- Fennica: a generic overview
- Fennica: a specific overview (Fennica specific preprocessing notes)
- Presentation slide templates (PDF) and code
- A Quantitative Approach to Book Printing in Sweden and Finland, 1640–1828 Source code for the figures
- Knowledge production in Finland 1470-1828: Digital Humanities 2016 conference presentation slides (PDF) and code
- Analyses on specific publication places and other topics (see the .md files)
- Figures and analyses for CCQ2019

The analyses cover several steps including XML parsing, data harmonization, removing unrecognized entries, enriching and organizing the data, carrying out statistical summaries, analysis, visualization and automated document generation. The analyses and full source code are provided in this repository and can be freely reused under the BSD 2 clause (FreeBSD) open source licence. The analyses are based on the R and rely on the custom bibliographica package for bibliographic data analysis, as well as many other R packages. The original raw data is available only on a separate agreement, so we are here publishing only the statistical summaries and our own analysis code.

github.com/COMHIS/fennica

Publishing in Finland 1640-1917 (title count)

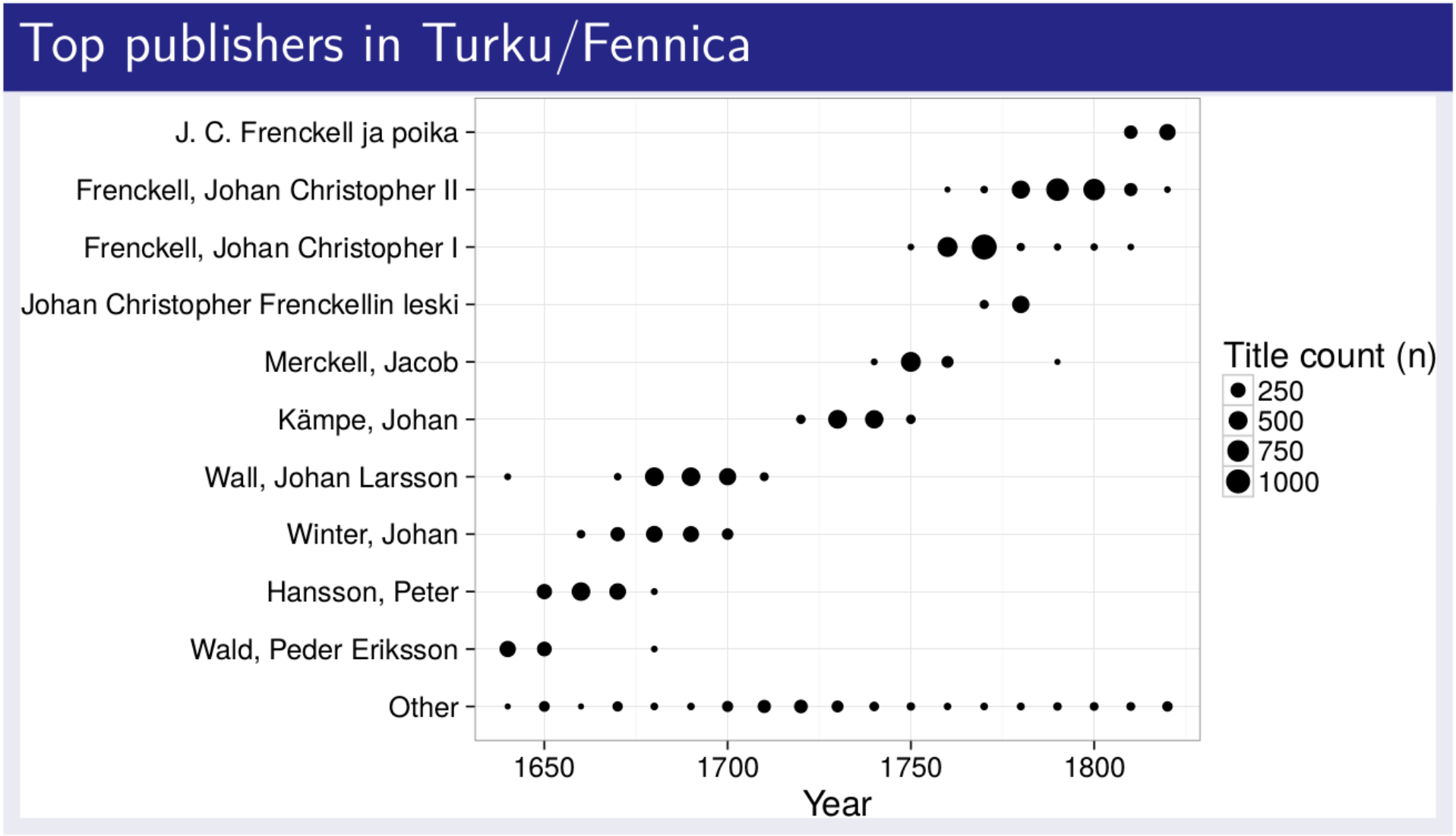


HISTORICAL METHODS
https://doi.org/10.1080/01615440.2018.1526657

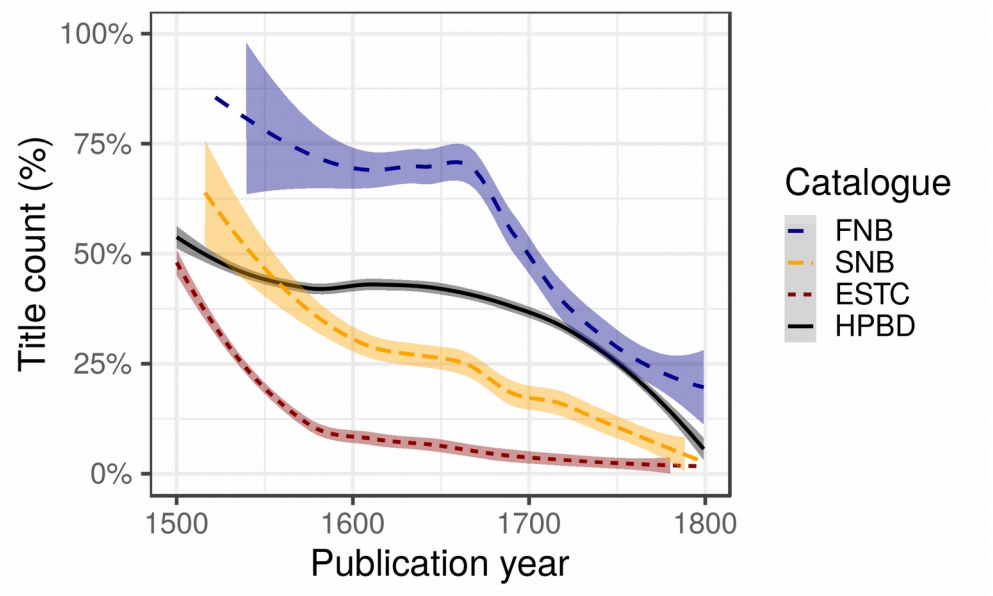
Routledge
Taylor & Francis Group

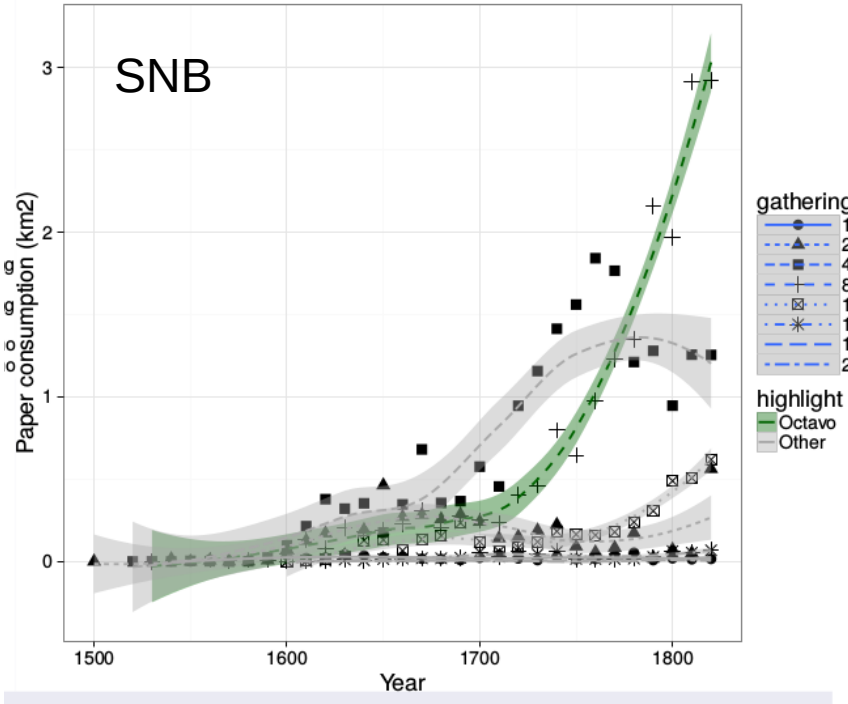
A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828

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



Decline in Latin, and the rise of Octavo





HISTORICAL METHODS
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,*]

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

Cataloguing & Classification Quarterly, 2019.

Leo Lahti, Jani Marjanen, Hege Roivainen, Mikko Tolonen

Helsinki Computational History Group

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 paper, 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 catalogues to map the history of the book.

REFEREE-ARTIKKELIT

AATEHISTORIA JA DIGITAALISTEN AINEISTOJEN MAHDOLLISUUDET

11.8.2015 MIKKO TOLONEN JA LEO LAHTI

Potential of data science in SSH research?

- New methods, classical questions
- Entirely new scales of quantitative analysis
- Transparent conclusions
- Quality through collaboration

Pitfalls of data science in SSH research?

- Data quality overlooked
- Expertise lacking
- Tools drive research
- Unrealistic expectations

Thank You!

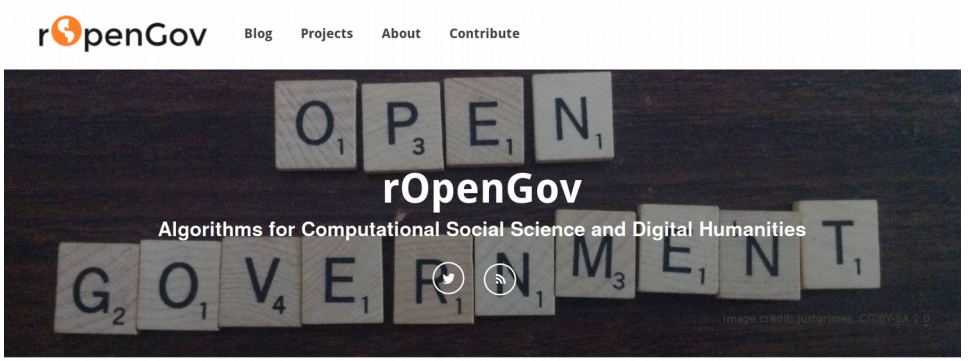
Publications, code, contact:
openresearchlabs.github.io

COMHIS
Mikko Tolonen
Jani Marjanen
Mark Hill
Ville Vaara
Hege Roivainen

Louhos / rOpenGov
Juuso Parkkinen
Joona Lehtomäki
Markus Kainu

Turku
Hannu Salmi

and many others !



Turun yliopisto
University of Turku

rOpenGov

OPEN KNOWLEDGE FINLAND

SUOMEN AKATEMIA
FINLANDS AKADEMI • ACADEMY OF FINLAND