Big data ownership in data-intensive research: life sciences & humanities

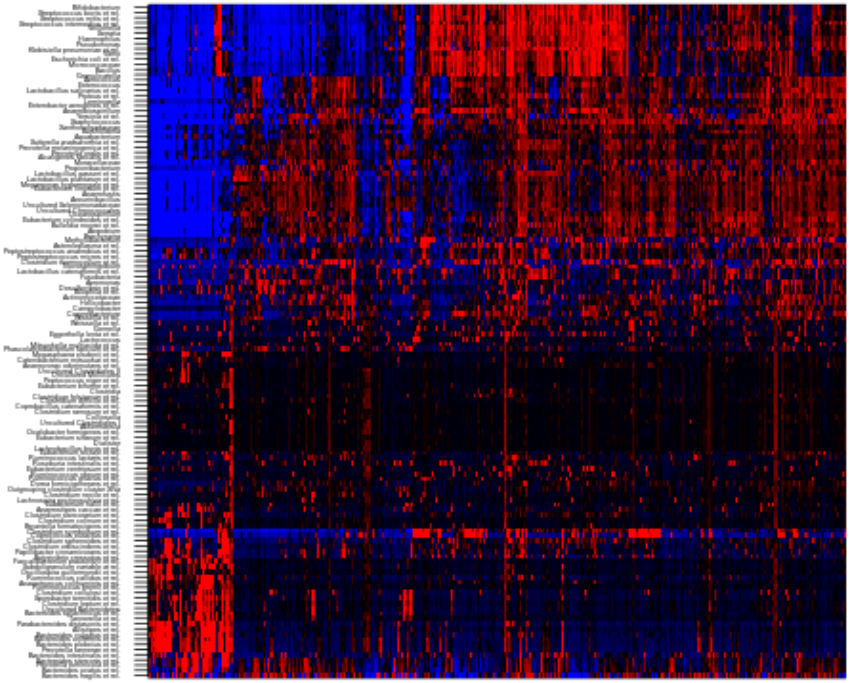
Leo Lahti, Adj. Prof. / D.Sc.(Tech.)
Dpt Mathematics and Statistics
University of Turku, Finland
leo.lahti@iki.fi | @antagomir

Turun yliopisto
University of Turku

Open genomic data repositories and open source revolutionized molecular biology: (e.g. human genome sequencing project)

Population cohort studies in biomedicine

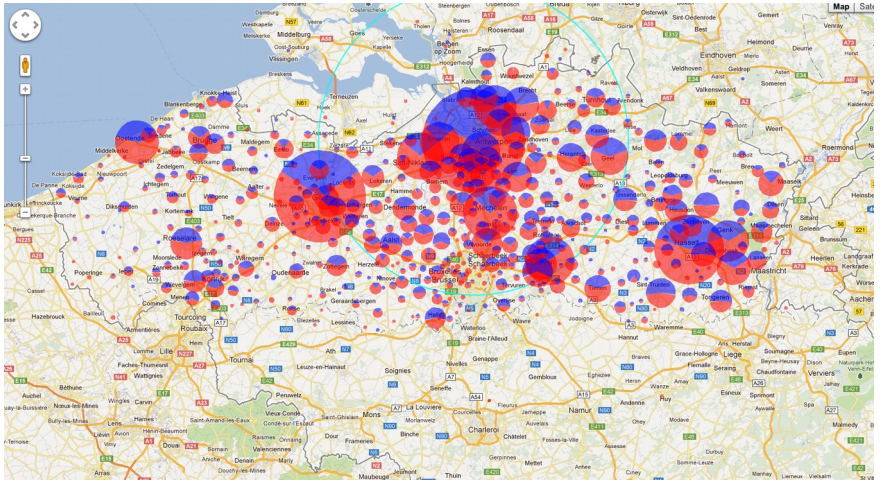
HITChip Atlas / Wageningen
N 10,000+
20+ nationalities
Highly standardized
Phylogenetic microarrays



Blue: Low abundance
Red: High abundance

Lahti et al. Nat. Comm. 2014

Flemish Gut Flora Raes Lab, VIB/KU Leuven
Cross-sectional & longitudinal
Comprehensive metadata
Focused geographical coverage

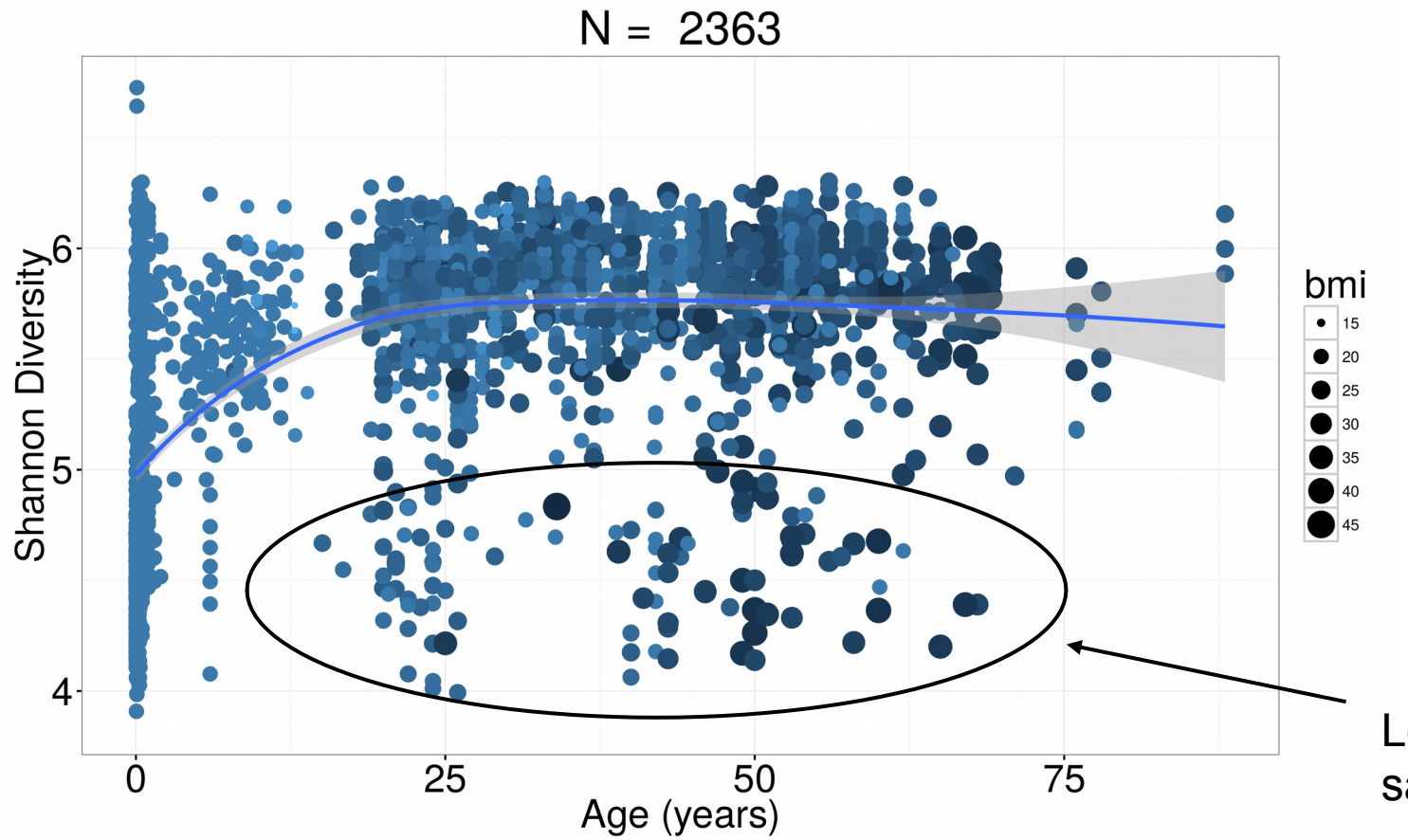


Falony et al. Science 352, 2016

Software: microbiome.github.io

HMP, MetaHIT, LLDeep, TwinsUK, American Gut..

Gut microbiome ecosystem diversity and ageing: healthy & normal obese subjects (HITChip Atlas)



Review

Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

Reuse of public genome-wide gene expression data

Johan Rung[1] & Alvis Brazma[1] About the authors

top

Our understanding of gene expression has changed dramatically over the past decade, largely catalysed by tec experiments — microarrays and large amounts of genome-wide archives. Added-value database to make them accessible to eve the gene expression data that a making use of these data. Reus many obstacles in data prepara results. We will discuss these c believe can improve the utility of such data.

Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas

Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang & Adam A Margolin

Affiliations | Contributions | Corresponding authors

Nature Genetics **45**, 1121–1126 (2013) | doi:10.1038/ng.2761
Published online 26 September 2013

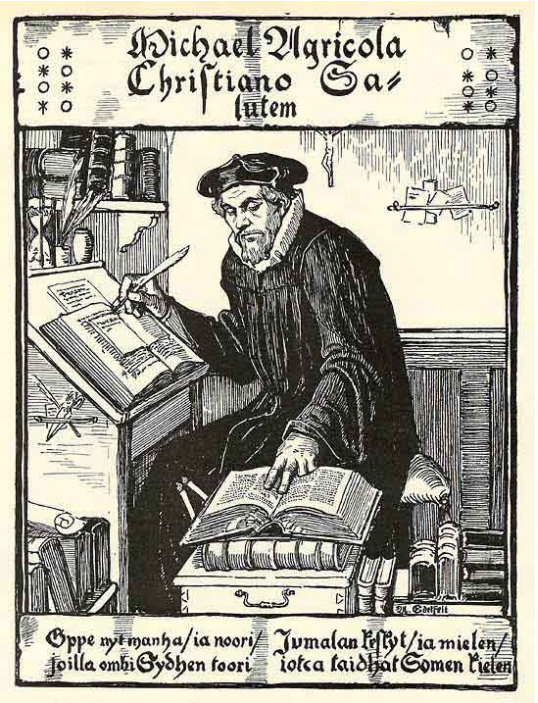
Automated multidimensional phenotypic profiling using large public microarray repositories

Min Xu[a,1], Wenyuan Li[a,1], Gareth M. James[b], Michael R. Mehan[a], and Xianghong Jasmine Zhou[a,2]

[a]Molecular and Computational Biology, Department of Biological Sciences, and [b]Marshall School of Business, University of Southern California, Los Angeles, CA 90089

Big data in the humanities (and social sciences)

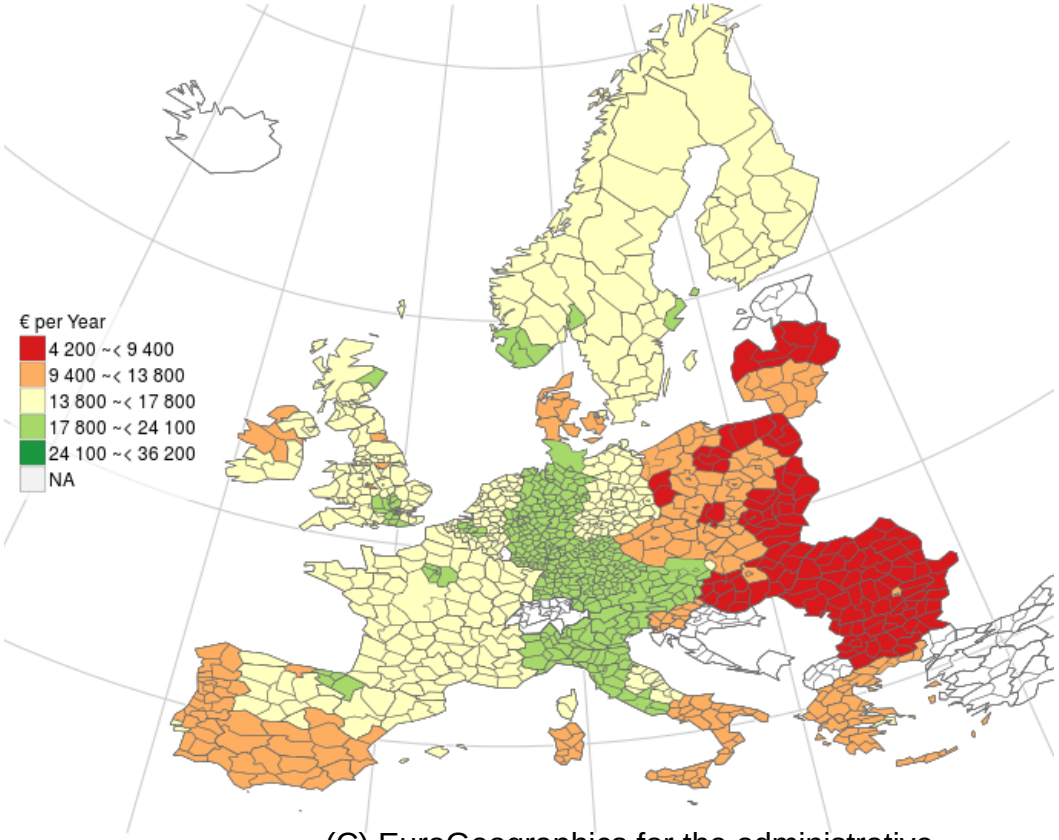
- Structured metadata catalogues (libraries, museums..)
- Full texts (books)
- Text streams (online discussion, social media..)
- Images
- Audiovisual material
- Descriptions of three-dimensional objects



Retrieval and Analysis of Eurostat Open Data with the eurostat Package

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

Eurostat open data portal: average household expenditure in 2011



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:

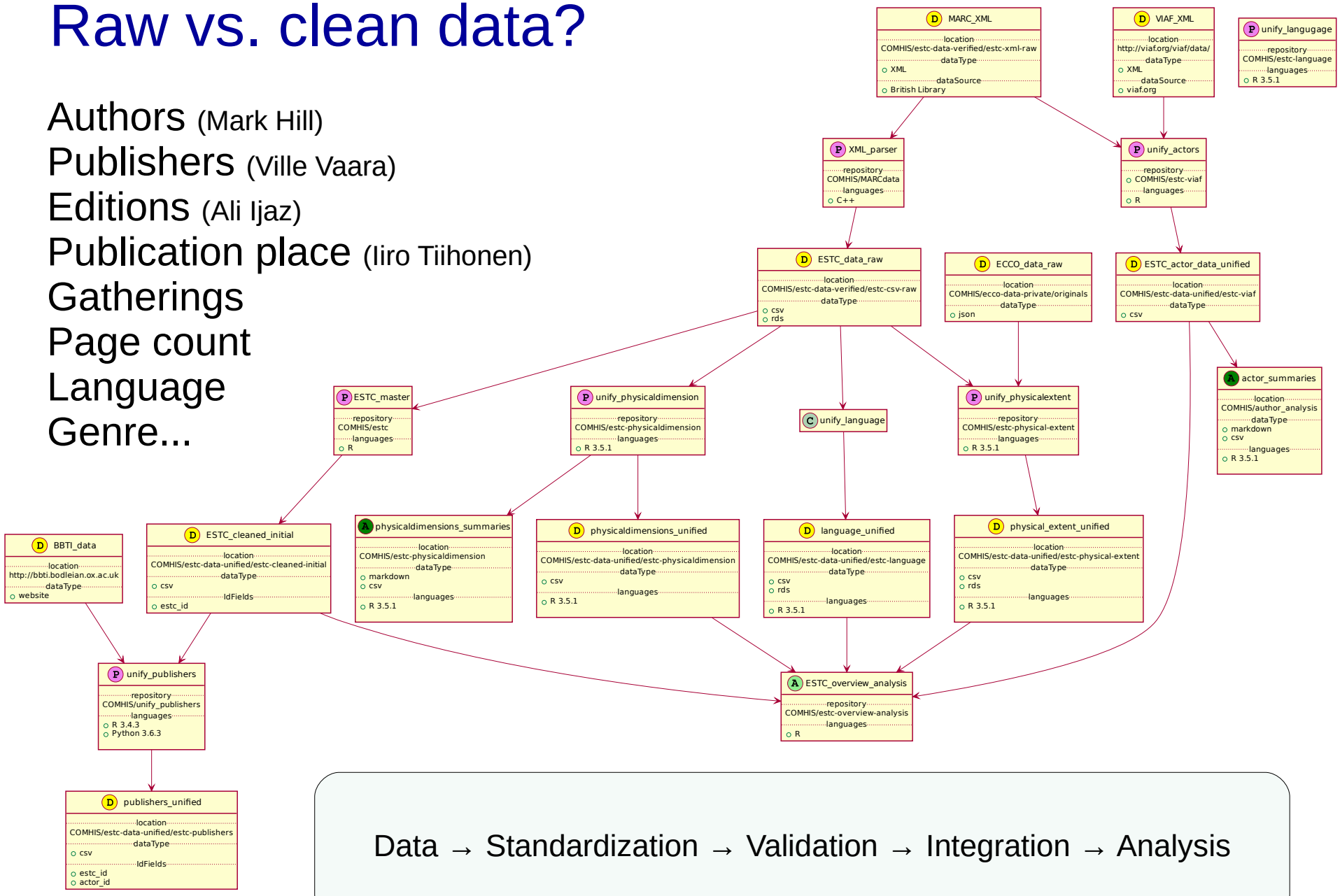
- Fennica: a generic overview
- Fennica: a specific overview (Fennica specific preprocessing notes)
- Presentation slide templates (PDF) and code
- A Quantitative Approach to Book Printing in Sweden and Finland, 1640–1828 Source code for the figures
- Knowledge production in Finland 1470-1828: Digital 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

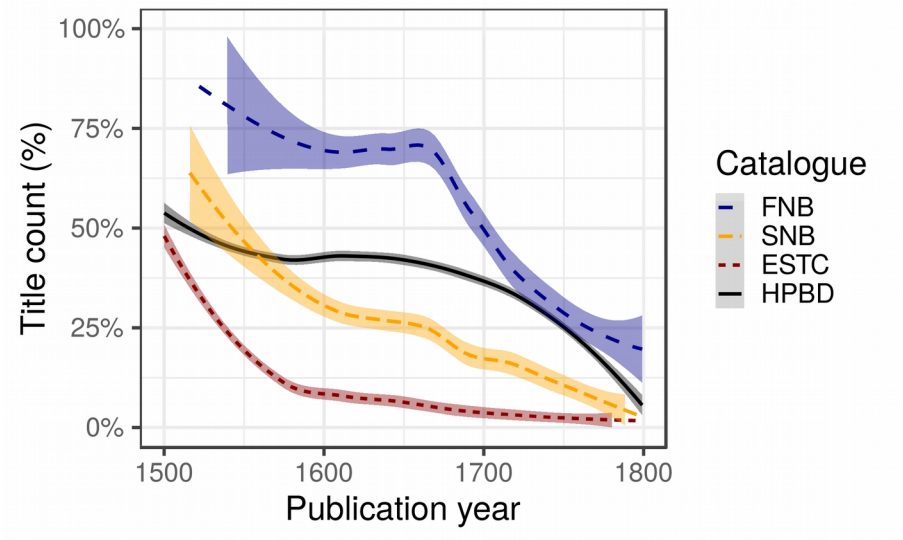
Raw vs. clean data?

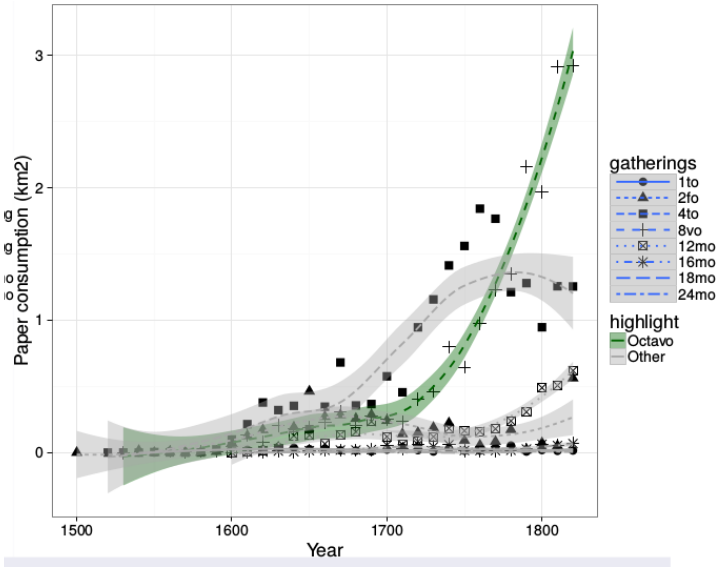
Authors (Mark Hill)
Publishers (Ville Vaara)
Editions (Ali Ijaz)
Publication place (Iiro Tiihonen)
Gatherings
Page count
Language
Genre...



Data → Standardization → Validation → Integration → Analysis

Decline in Latin, and the rise of Octavo





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

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,*]

To whom does the data belong?

- Life sciences: Research institutions, Healthcare system, Patients..

- Humanities: Memory organizations, companies, universities..

→ Licensing / Mydata ..?

Who gets the rights & burden of curation and storage?

- Open sharing vs. closed agreements?

- Raw vs. clean data?

Intellectual property rights?

- Patentability?

- Confidentiality & privacy

Challenge for today

Decentralizing data ownership?

- Distributed curation & maintenance?

- Open licensing?

- Mydata?

- Public commons?

- Pros & cons?

leo.lahti@iki.fi | openresearchlabs.github.io | @antagomir