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

Routledge
Taylor & Francis Group

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

Helsinki Computational History Group

https://comhis.github.io/

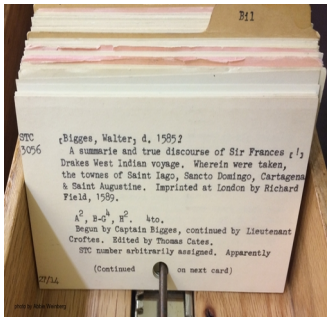
Computer scientists researching open workflows, algorithms and interfaces for humanities text and metadata

Linguists exploring the relationship between words and concepts

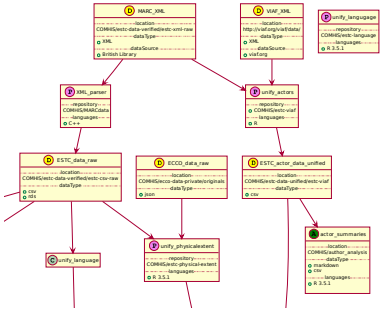
Historians interested in conceptual and actual historical processes

From library catalogues to research reports

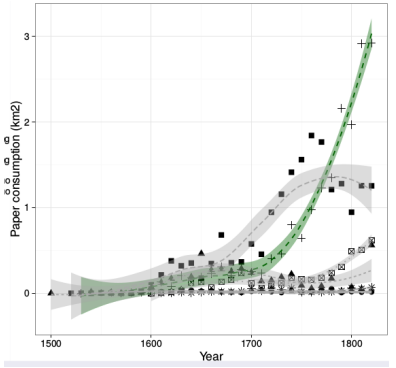
Research potential



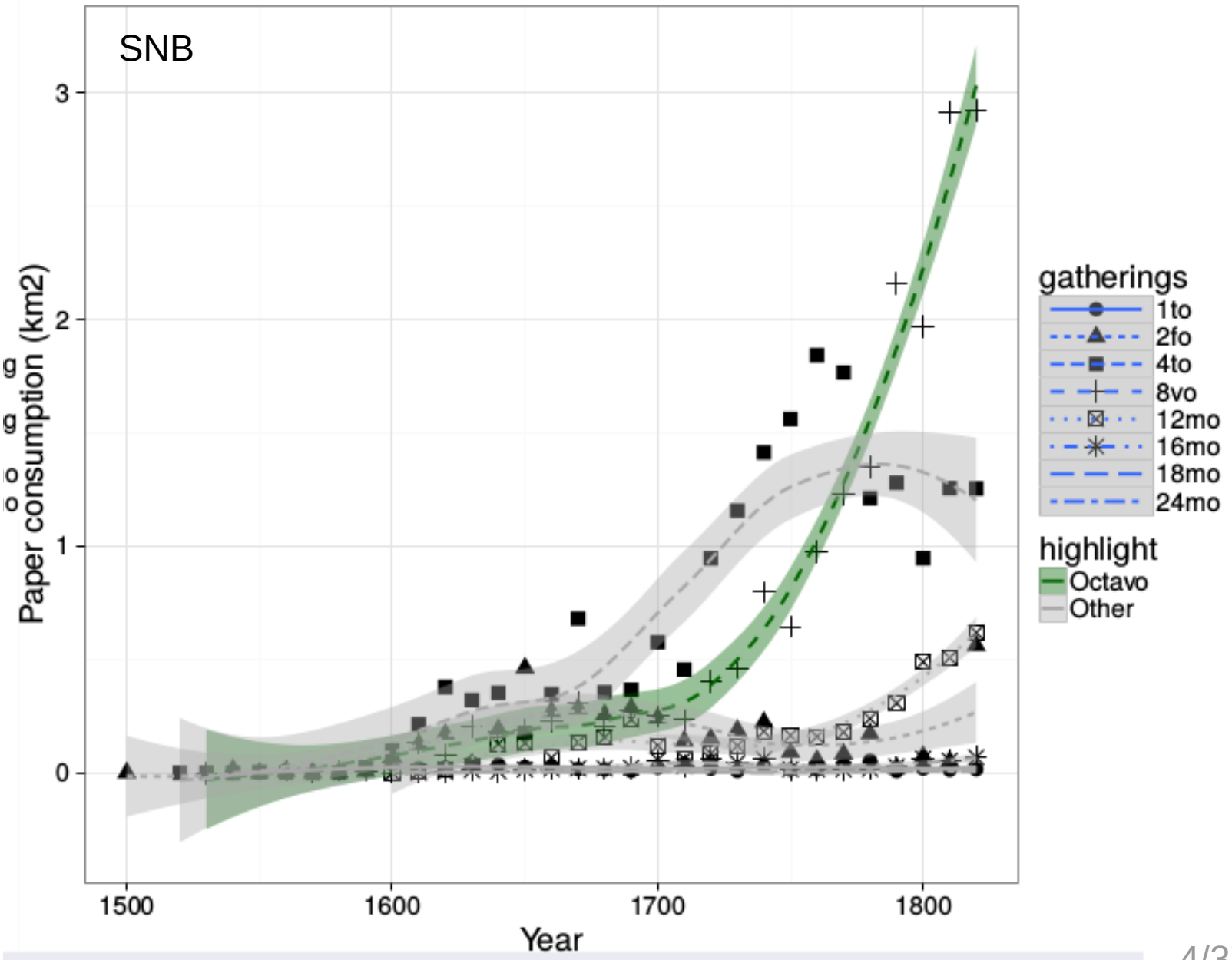
Open data science ecosystem



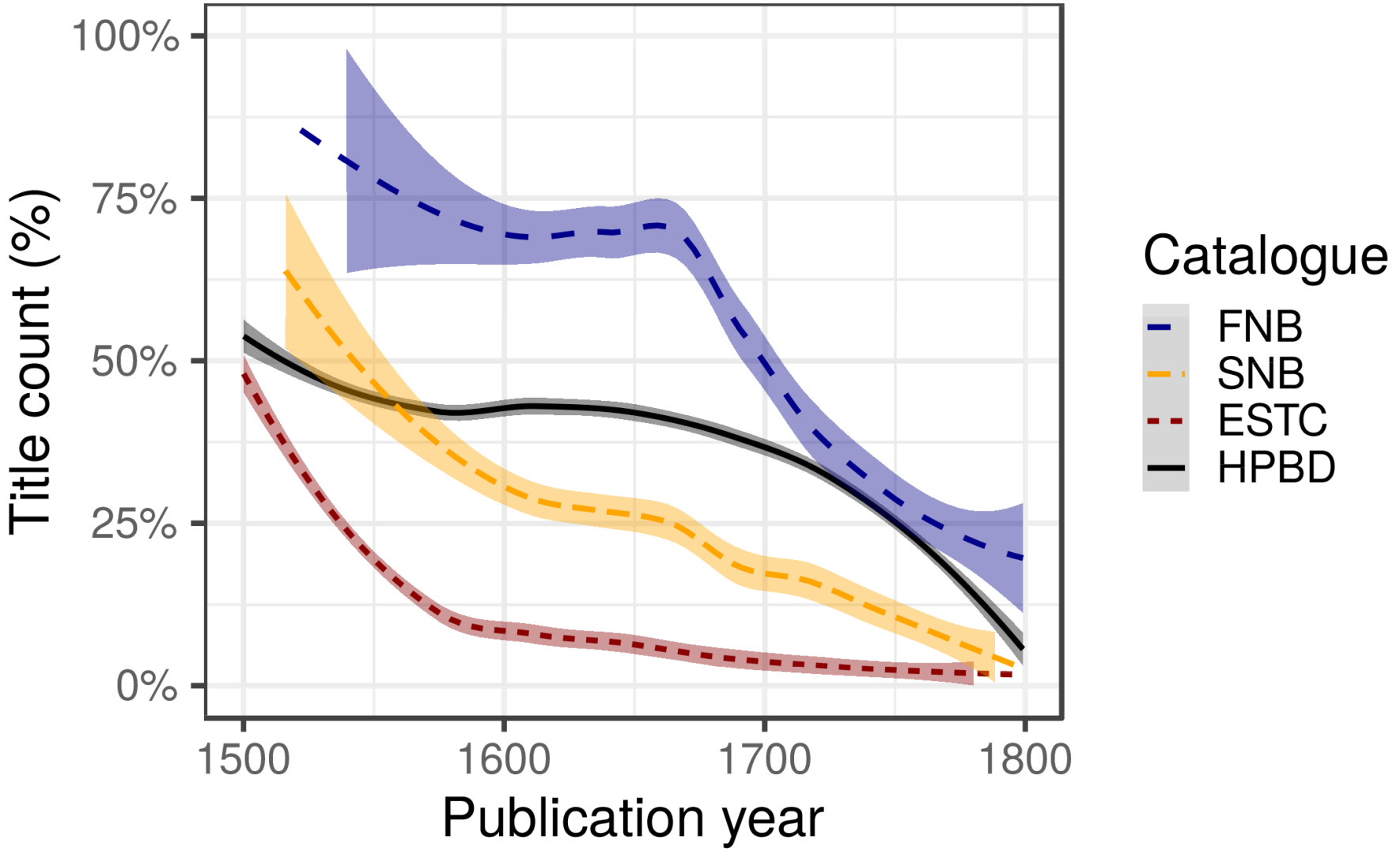
Research cases



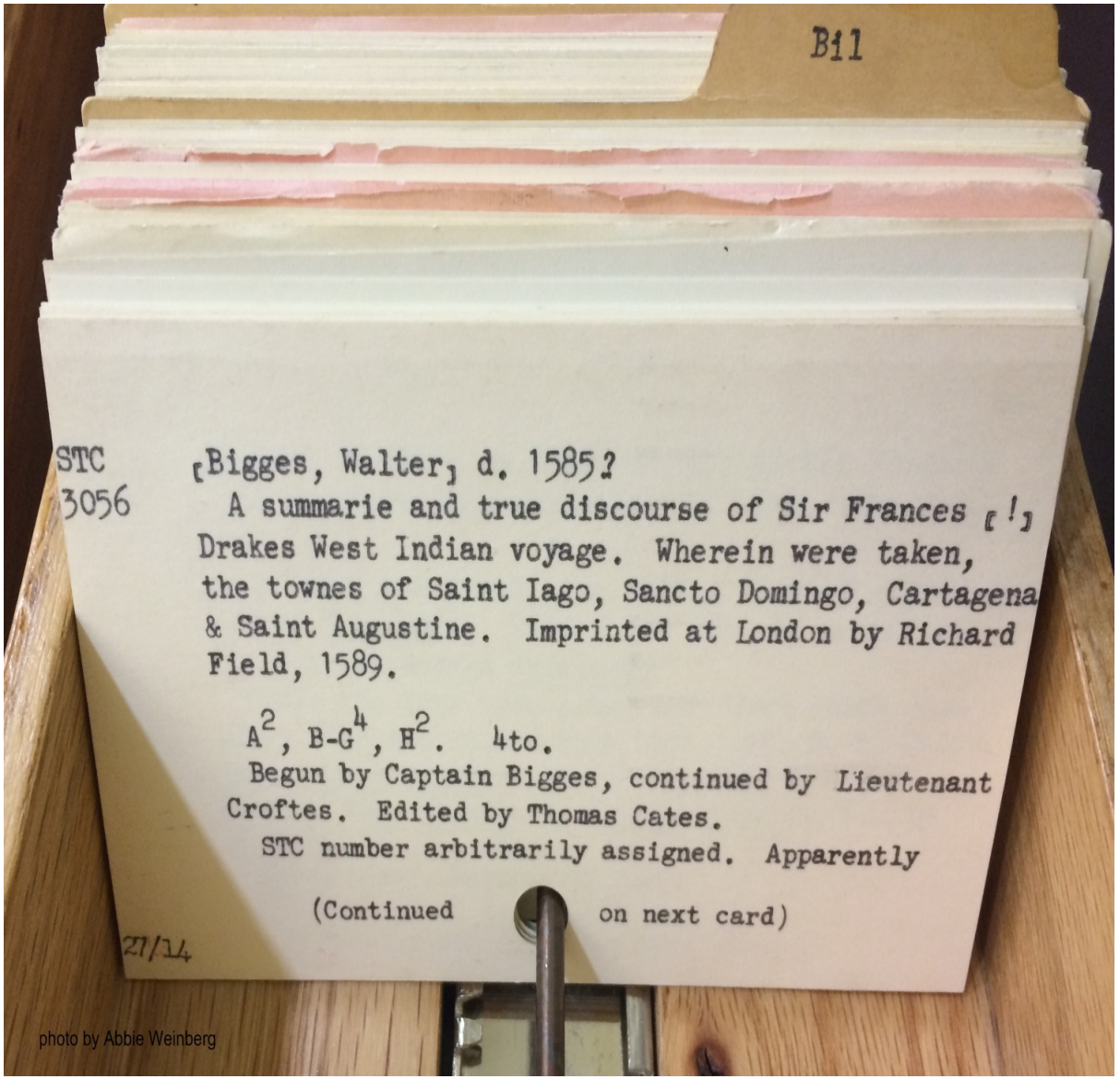
The rise of Octavo: paper consumption



Title count share for books in Latin (primary language)



The (meta)data



ESTC System No.	006196908
ESTC Citation No.	S722
Author - personal	Bigges, Walter, -1586.
Title	A **summarie** and **true discourse** of **Sir Frances Drakes** VVest **Indian voyage**. VVherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint Augustine.
Variant title	**Summarie** and **true discourse** of **Sir Frances Drakes West Indian voyage**
	Sir Frances Drakes VVest **Indian voyage**
	Sir Frances Drakes West Indian voyage
Publisher/year	Imprinted at London : By Richard Field, dwelling in the Blacke-Friars by Ludgate, 1589.
Physical descr.	[4], 52 p. ; 4º.
General note	"Begun by Captaine Bigges ... the same being afterwardes finished (as I thinke) by his lieutenant Maister Croftes, or some other, I knowe not well who"--A2r.
	Editor's dedication signed: Thomas Cates.
	Running title reads: **Sir Frances Drakes** VVest **Indian voyage.**
	Signatures: A² B-G⁴ H².
	Another state (STC 3056.5) has three additional lines in the title and a line of errata on the last page.
	Often bound with maps, which were evidently sold separately. Those with letterpress English captions are separately listed as STC 3171.6, which see for information on states and combinations.
	Stationers' Register: Entered to W. Ponsonby 26 November 1588.
Uncontrolled note	Signatures from DFo.
Citation/references	STC (2nd ed.), 3056
	Luborsky & Ingram. Engl. illustrated books, 1536-1603, 3056
Surrogates	Microfilm. Ann Arbor, Mich. University Microfilms International, 1983. 1 microfilm reel ; 35 mm. (Early English books, 1475-1640; 1772:10).
Person as subject	Drake, Francis, **Sir**, 1540?-1596.
Subject	Explorers -- England -- Biography -- Early works to 1800.
	West Indies Expedition, 1585-1586 -- Early works to 1800.
Subject	America -- Discovery and exploration -- English -- Early works to 1800.
Added name	Croftes, Lieutenant.
	Gates, Thomas, **Sir**, -1621, ed.
Copies - Brit.Isles	British Library
	Glasgow University Library
	Oxford University Bodleian Library (includes The Vicar's Library, ST. Mary's Church, Marlborough)
Copies - N.America	Folger Shakespeare
	Henry E. Huntington Library and Art Gallery
	Massachusetts Historical Society
	New York Public Library
	United States, Library of Congress
	University of Virginia
Electronic location	Library of Congress Digital Collections ; { Source Library: nDLC : E129.D7 B5 1589 }
	http://gateway.proquest.com/openurl?ctx_ver=Z39.88-2003&res_id=xri:eebo&rft_val_fmt=&rft_id=xri:eebo:image:23483 ; { Reproduction of original in the Henry E. Huntington Library and Art Gallery. }

FMT	BK
LDR	cam a2200469 4500
001	006196908
003	Uk-ES
005	20130916220616.0
008	900830s1589 enk||| 00| ||eng c
009	S722
035	|a (CU-RivES)S722
040	|a CU-RivES |c CU-RivES |d CStRLIN |d Uk-ES |e dcrb
1001	|a Bigges, Walter, |d -1586.
24512	|a A summarie and true discourse of Sir Frances Drakes VVest Indian voyage. |b VVherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint Augustine.
2463	|a Summarie and true discourse of Sir Frances Drakes West Indian voyage
2463	|a Sir Frances Drakes VVest Indian voyage
2463	|a Sir Frances Drakes West Indian voyage
260	|a Imprinted at London : |b By Richard Field, dwelling in the Blacke-Friars by Ludgate, |c 1589.
300	|a [4], 52 p. ; |c 4^{o}.
500	|a "Begun by Captaine Bigges ... the same being afterwardes finished (as I thinke) by his lieutenant Maister Croftes, or some other, I knowe not well who"--A2r.
500	|a Editor's dedication signed: Thomas Cates.
500	|a Running title reads: Sir Frances Drakes VVest Indian voyage.
500	|a Signatures: A^{2} B-G^{4} H^{2}.
500	|a Another state (STC 3056.5) has three additional lines in the title and a line of errata on the last page.
500	|a Often bound with maps, which were evidently sold separately. Those with letterpress English captions are separately listed as STC 3171.6, which see for information on states and combinations.
500	|a Stationers' Register: Entered to W. Ponsonby 26 November 1588.
509	|a Signatures from DFo.
5104	|a STC (2nd ed.), |c 3056
5104	|a Luborsky & Ingram. Engl. illustrated books, 1536-1603, |c 3056
533	|a Microfilm. |b Ann Arbor, Mich. |c University Microfilms International, |d 1983. 1 microfilm reel ; 35 mm. |f (Early English books, 1475-1640; 1772:10).
60010	|a Drake, Francis, |c Sir, |d 1540?-1596.
648 7	|a 1473-1640 |2 local
650 0	|a Explorers |z England |v Biography |v Early works to 1800.
650 0	|a West Indies Expedition, 1585-1586 |v Early works to 1800.
651 0	|a America |x Discovery and exploration |x English |v Early works to 1800.
7001	|a Croftes, |c Lieutenant.
7001	|a Gates, Thomas, |c Sir, |d -1621, |e ed.
752	|a Great Britain |b England |d London.
852	|a bL |b British Library |e London, England, U.K. |j [Shelfmark not available] |x C> |q imp., e [CM] |r 1116038
852	|a nDFo |b Folger Shakespeare |e Washington, District of Columbia |j STC 3056 |x V> |q HH27/14. Stained, t.p. torn and mended, affecting imprint. Bernard Quaritch pencilled note. |p bookplate of Gloddaeth library; small label: Edwd. Parry ...; signature of Persivall Golding; Baron Mostyn - Harmsworth copy |r 1116040
852	|a bGu |b Glasgow University Library |e Glasgow, Scotland |j [Shelfmark not available] |x C> |q [CM] |r 1116037
852	|a nCSmH |b Henry E. Huntington Library and Art Gallery |e San Marino, California |j 18731 |x P> |q Binding signed by F. Bedford. +Port. and 4 plates with Latin and French engraved captions |p bookplates of Marshall Clifford Lefferts and E.D. Church |r 1116036
852	|a nMHi |b Massachusetts Historical Society |e Boston, Massachusetts |j [Shelfmark not available] |x C> |q [CM] |r 1116041
852	|a nNN |b New York Public Library |e New York, New York |j [Shelfmark not available] |x C> |q b [CM] |r 1116042
852	|a bO |b Oxford University Bodleian Library (includes The Vicar's Library, ST. Mary's Church, Marlborough) |e Oxford, England |j Antiq.e.E.1589.5 |x V> |r 1116039

English Short Title Catalogue (ESTC)

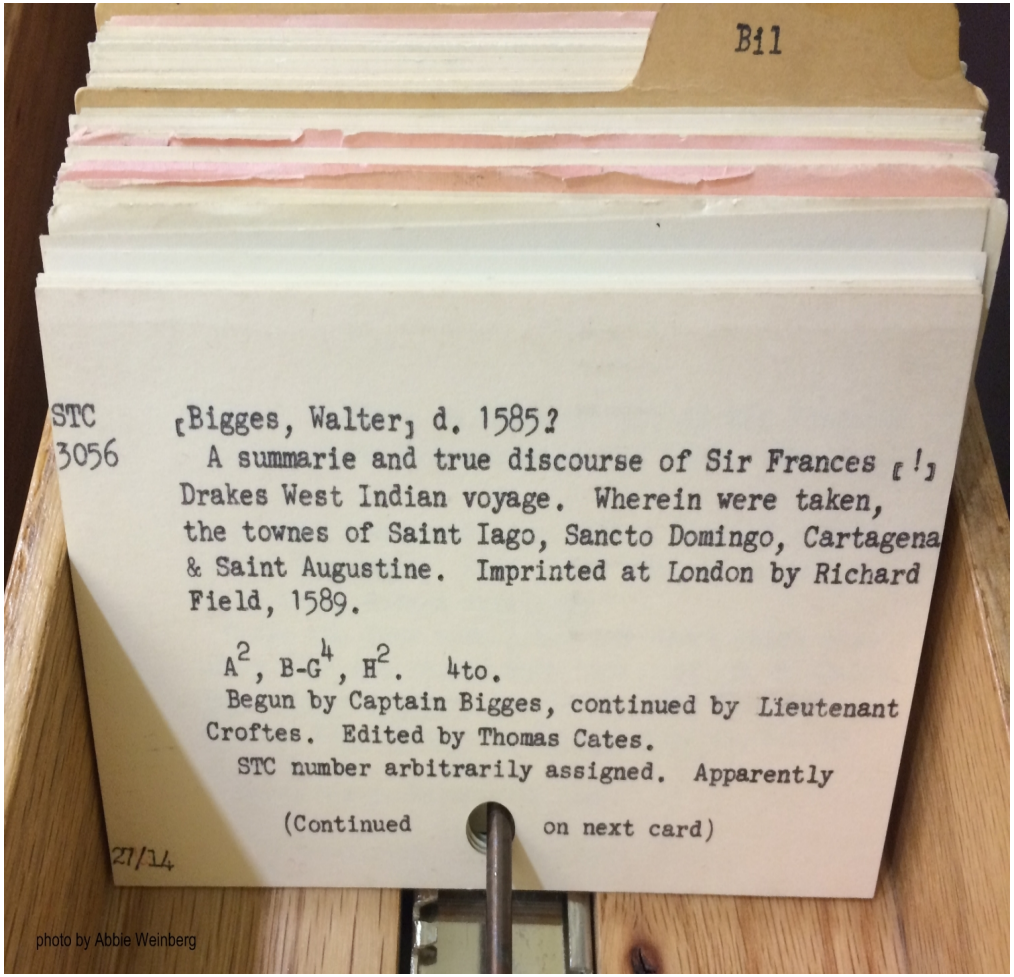
- English literature
- ~0.5M entries 1470-1800

Online version:

- Entire catalogue *browsable* at http://estc.bl.ac.uk

Offline version:

- Not open or publicly available
- One (non-standard) XML file
- 1,450,034 lines
- Data entries discrete & non-harmonized



LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

A quantitative study of history in the English short-title catalogue (ESTC)

Leo Lahti, Niko Ilomäki, Mikko Tolonen

LIBER Quarterly 25(2), 2015

Scaling

Combining catalogues, harmonizing formats; language differences..

Automation

Optimizing analysis algorithms (speed, accuracy, generalizability.)

Extensions

Support full-text analyses

Library catalogues

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%

FNB (Fennica)

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**

ESTC British Library: N > 500,000

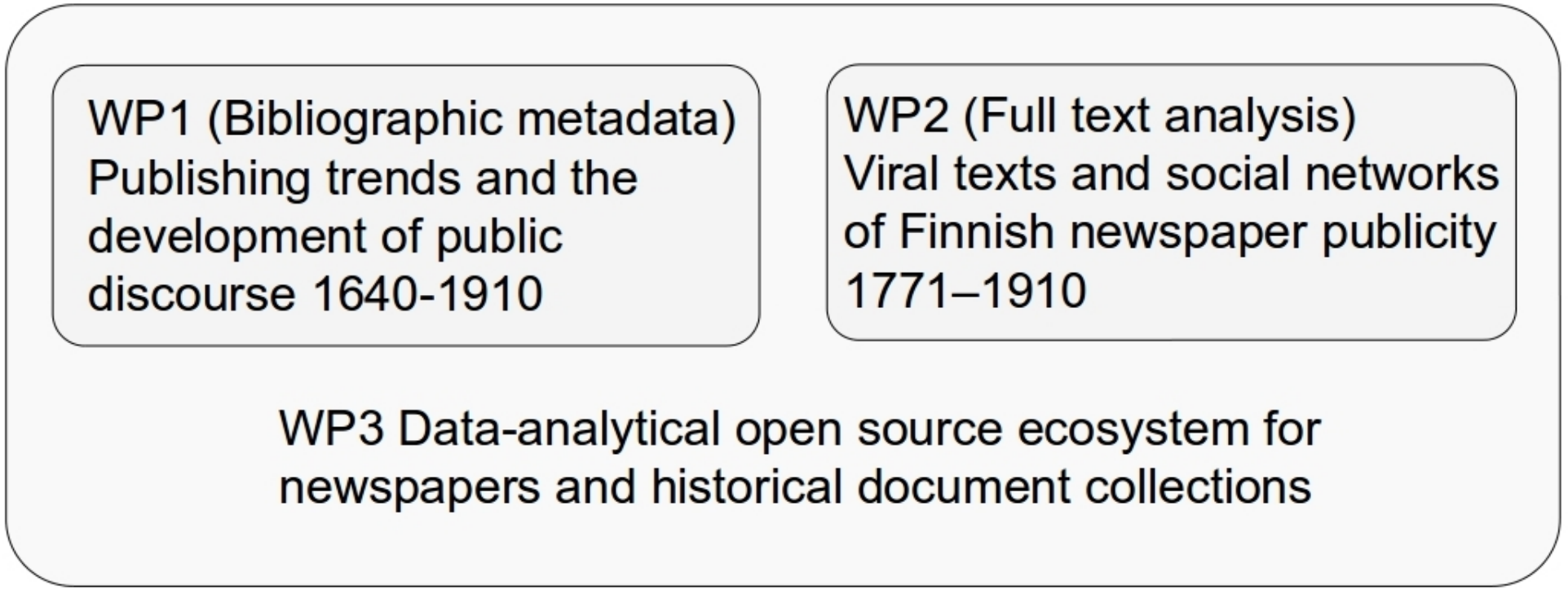
SNB (Kungliga): Swedish National bibliography N >18 M entries

HPBD Heritage of the printed book, Göttingen: N > 6M entries

COMHIS consortium (Academy of Finland 2016-2019)

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

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



The role of computational / data science workflows

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

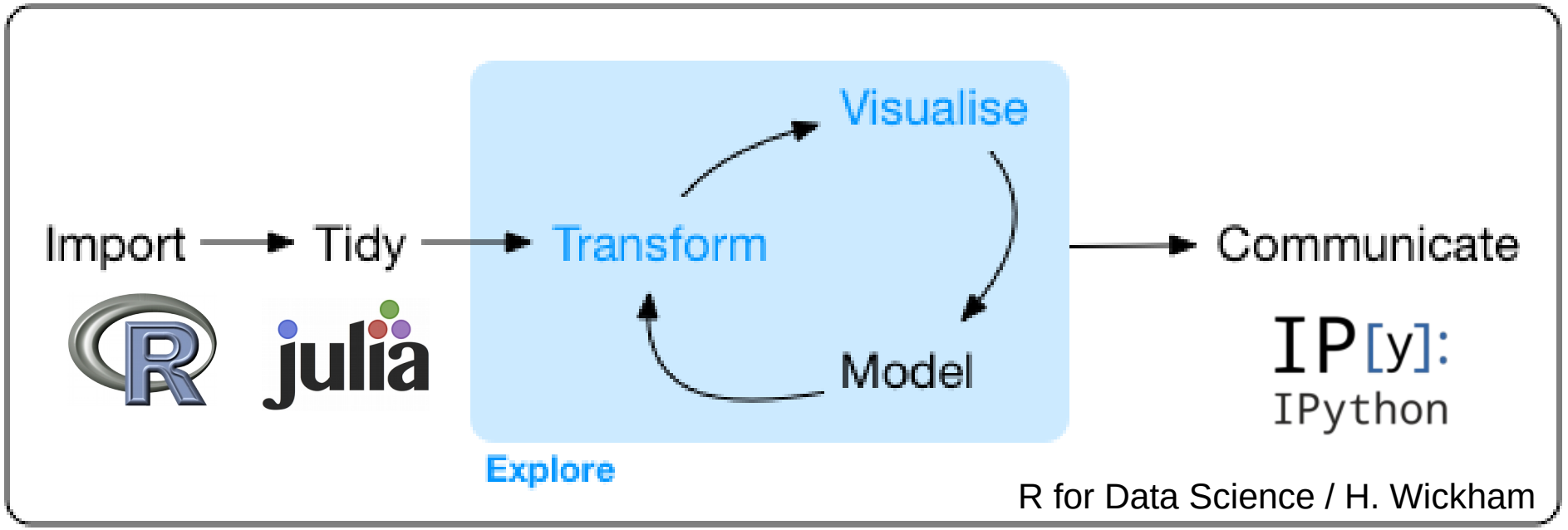
Science 13 April 2012:
Vol. 336 no. 6078 pp. 159-160
DOI: 10.1126/science.1218263

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,*]



R for Data Science / H. Wickham

Elements of bibliographic data science

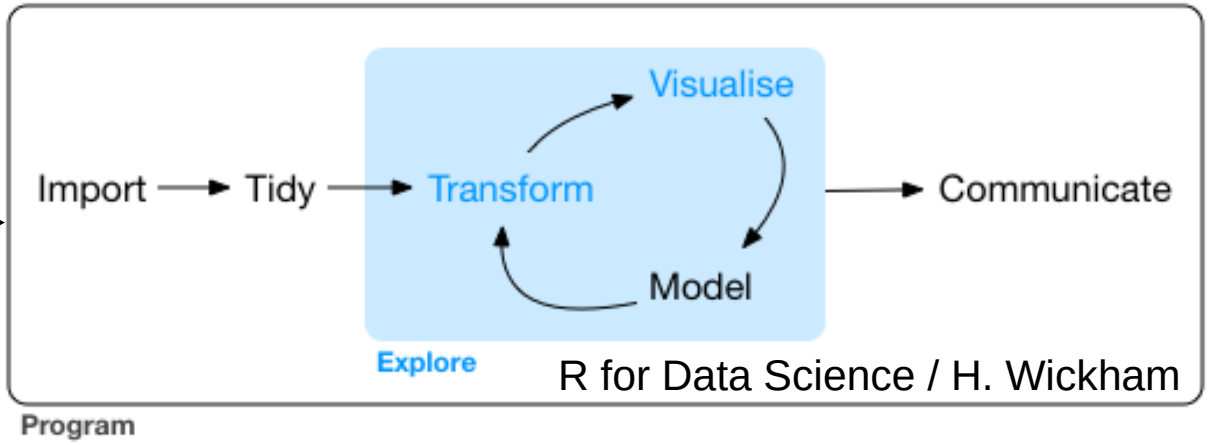
Quantitative frame for qualitative research
Beyond counting titles (volume, time, geography…)

Bibliographic metadata

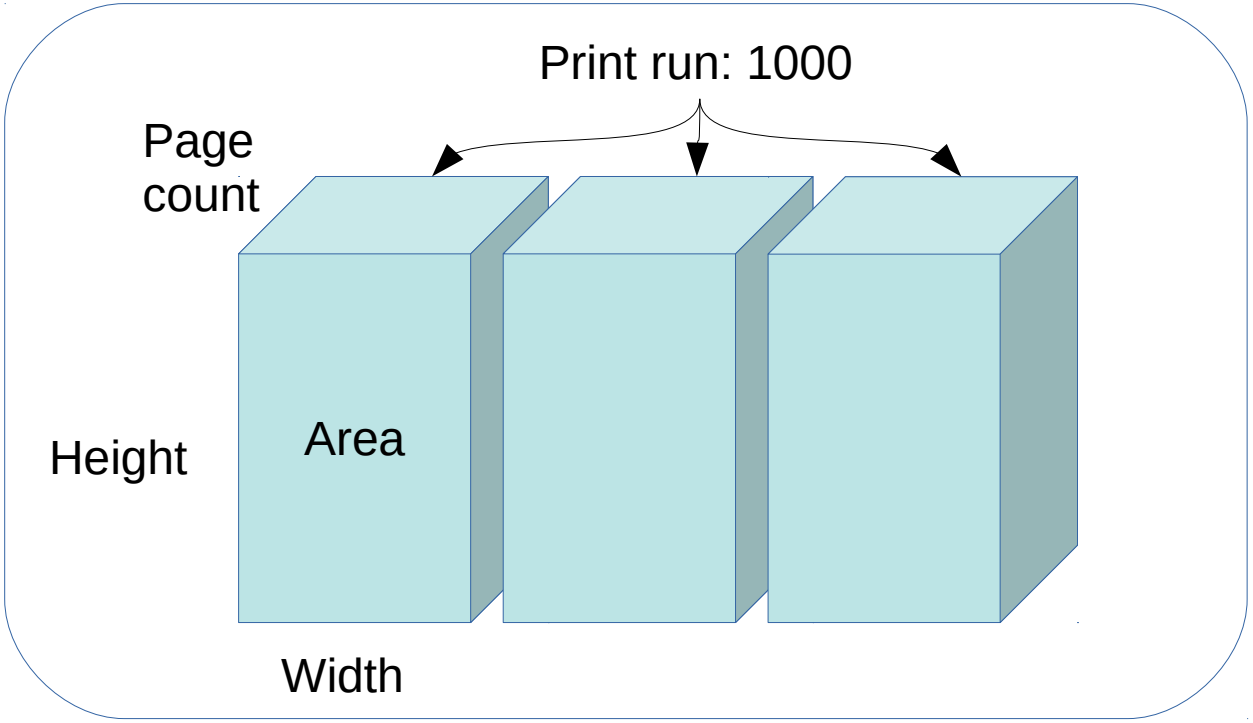
Full texts: books, newspapers...

Supporting data

Open data science ecosystem



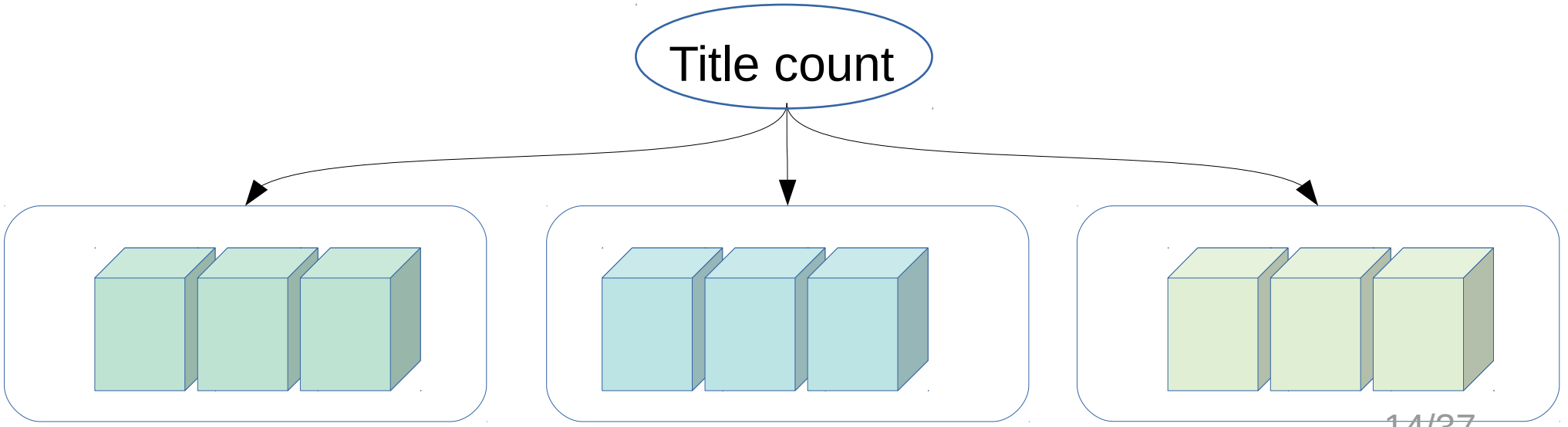
Measuring printing activity: quantitative indicators

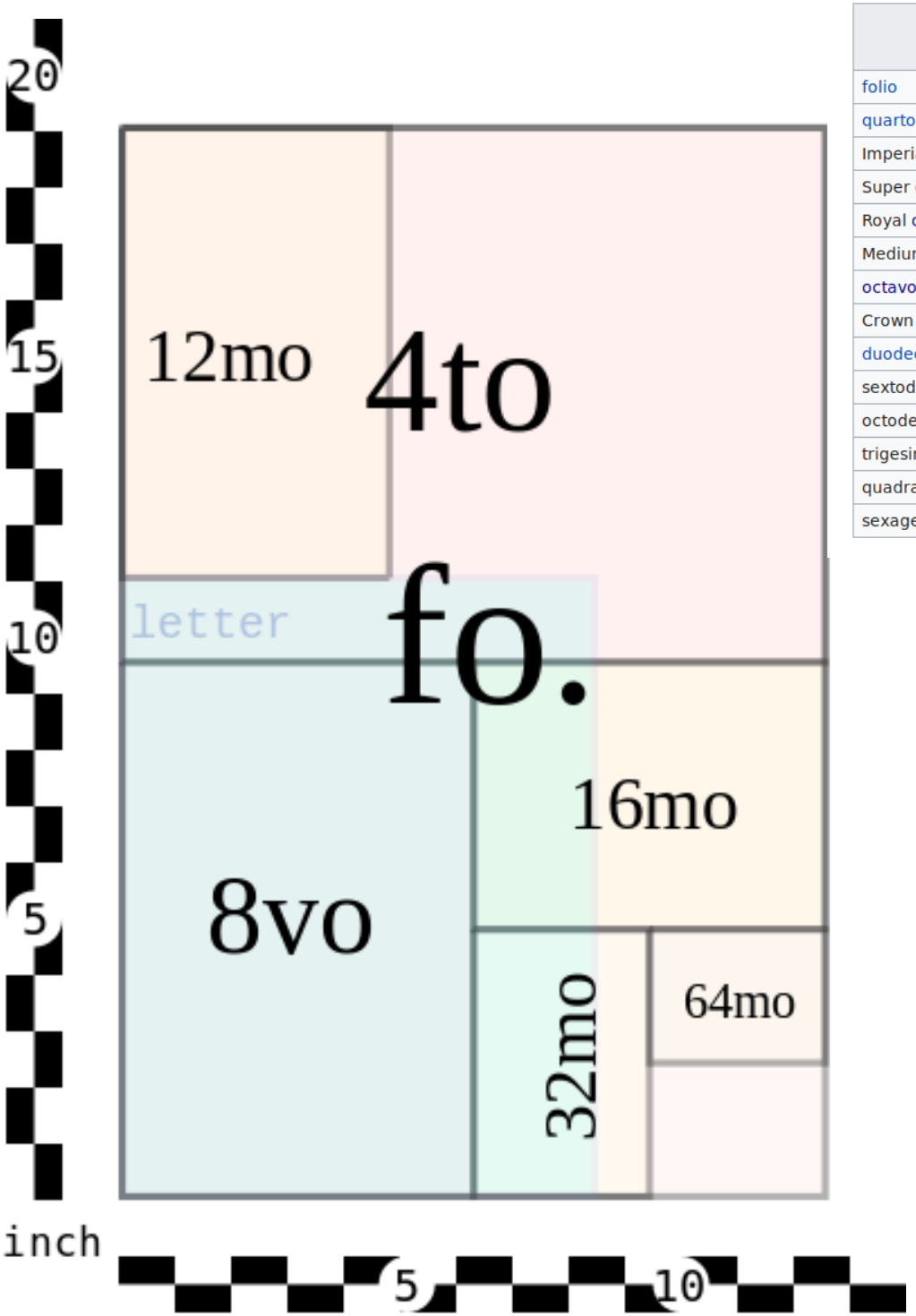


Title count

Print area

Paper consumption



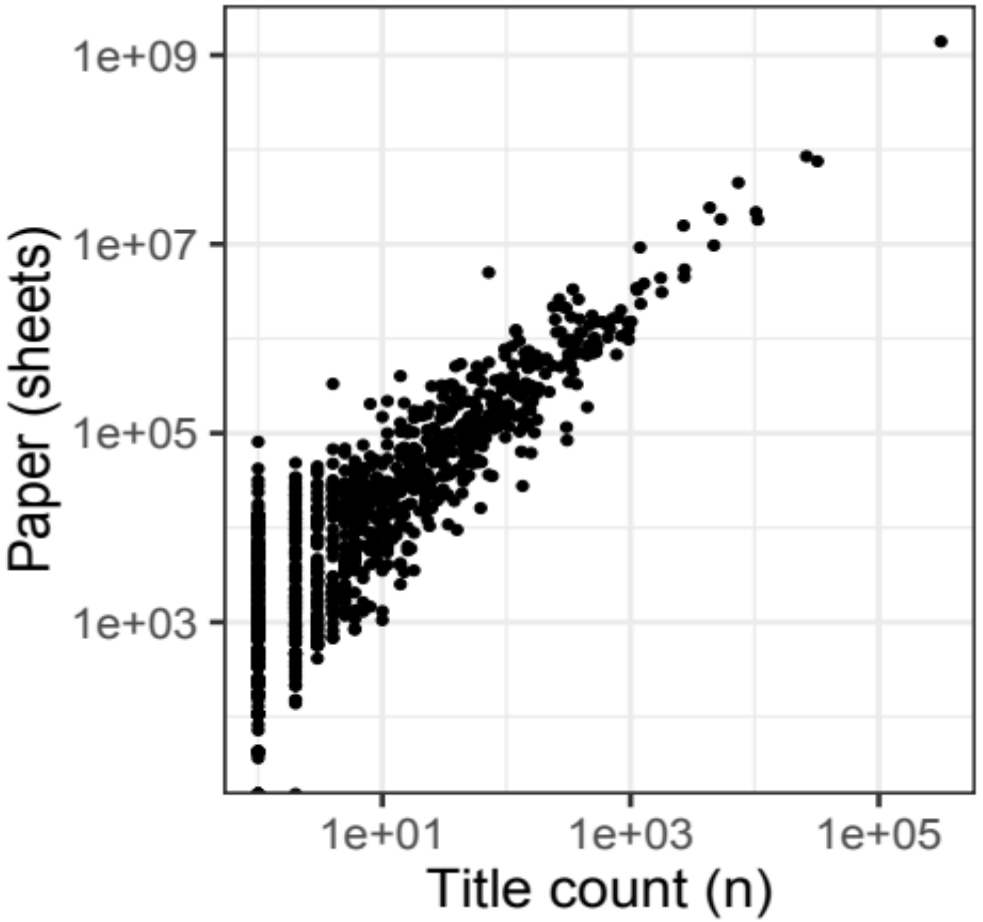


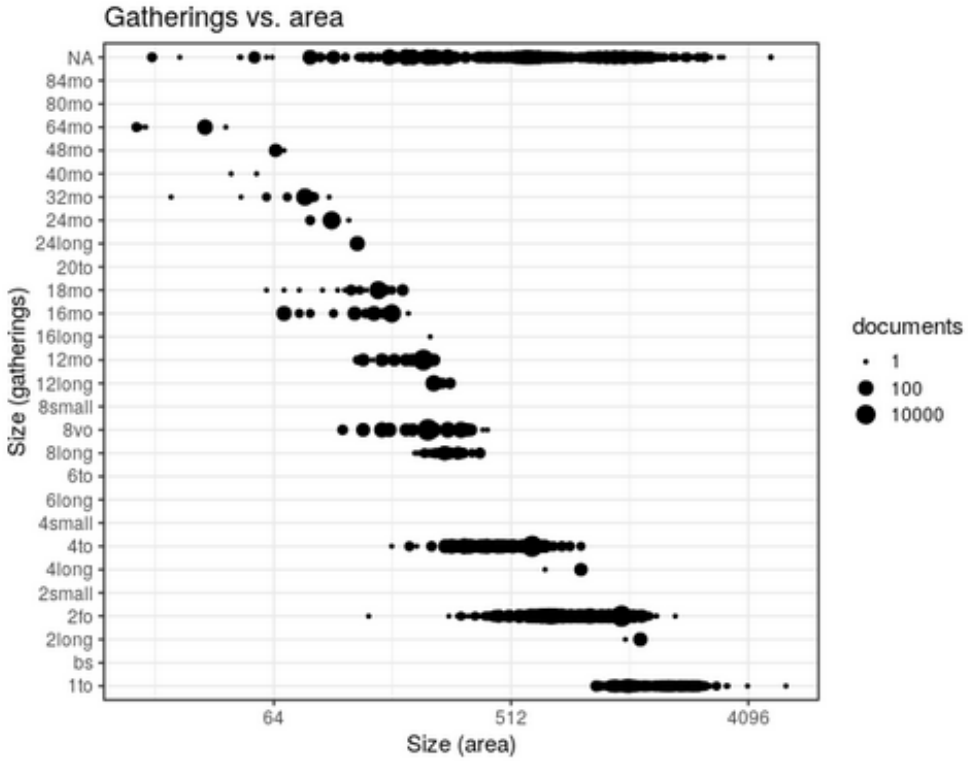
Name	Abbreviations	Leaves	Pages	Approximate cover size (width × height)	
				inches	cm
folio	2º *or* fo	2	4	12 × 19	30.5 × 48
quarto	4º *or* 4to	4	8	9½ × 12	24 × 30.5
Imperial octavo	8º *or* 8vo	8	16	8¼ × 11½	21 × 29
Super octavo	8º *or* 8vo	8	16	7 × 11	18 × 28
Royal octavo	8º *or* 8vo	8	16	6½ × 10	16.5 × 25
Medium octavo	8º *or* 8vo	8	16	6½ × 9¼	16.5 × 23.5
octavo	8º *or* 8vo	8	16	6 × 9	15 × 23
Crown octavo	8º *or* 8vo	8	16	5⅜ × 8	13.5 × 20
duodecimo *or* twelvemo	12º *or* 12mo	12	24	5 × 7⅜	12.5 × 19
sextodecimo *or* sixteenmo	16º *or* 16mo	16	32	4 × 6¾	10 × 17
octodecimo *or* eighteenmo	18º *or* 18mo	18	36	4 × 6½	10 × 16.5
trigesimo-secundo *or* thirty-twomo	32º *or* 32mo	32	64	3½ × 5½	9 × 14
quadragesimo-octavo *or* forty-eightmo	48º *or* 48mo	48	96	2½ × 4	6.5 × 10
sexagesimo-quarto *or* sixty-fourmo	64º *or* 64mo	64	128	2 × 3	5 × 7.5

Document size matters

Printing activity: quantitative indicators

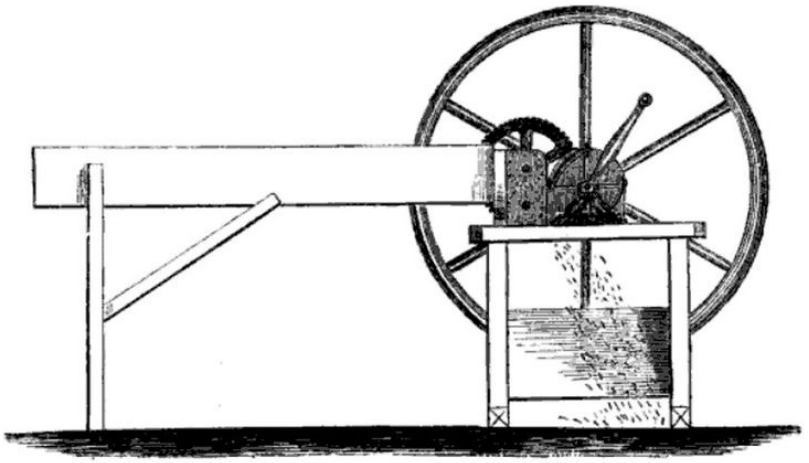
- Title count (number of unique titles)
- Print area (width x height x title count)
- Paper (width x height x page count x title count x print run)





Data harmonization:
estimating page counts from MARC cataloguing standards

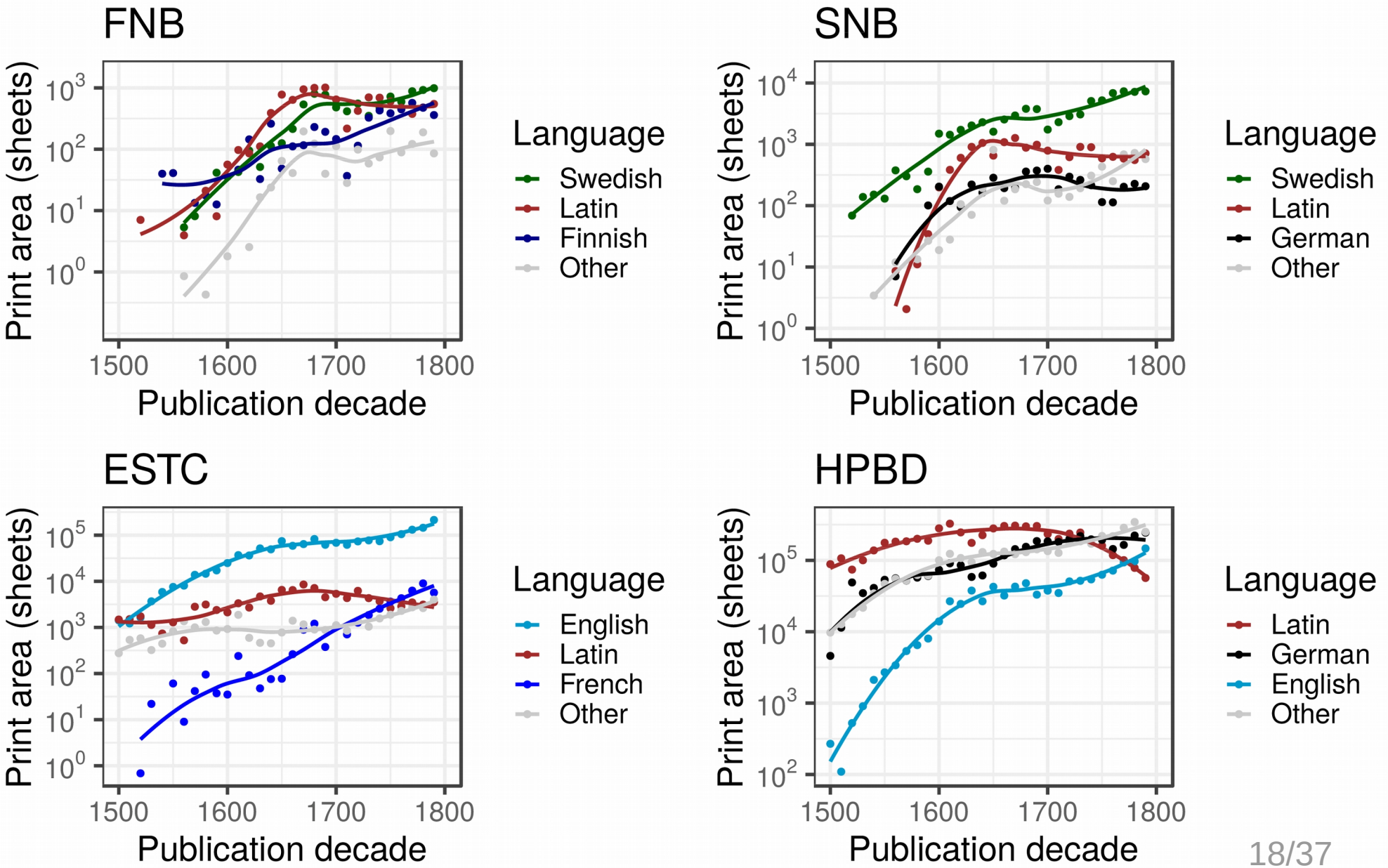
“[4],vii-xii,[4],222p.,plate” → 240 pages



```
polish_physical_extent("iii-xxiv, 118, [2] p.")$pagecount
```

```
## [1] 142
```

Print area for top languages



ESTC Actor Fields (100, 110, 700, 710)

- Extracted 557,847 actors from 397,061 documents (for which there were named actors)
- Cleaned up and standardized unicode.
- Created individual actor records per document.
- Assigned roles when known.
- Harmonized by string matching (when appropriate) and with Virtual International Authority File (VIAF)
  - Problems: VIAF often has duplicate records; single records are clearly for multiple individuals, IDs change.
- **558,243 actor records harmonized into 92.044 unique actors.**

Data harmonization: Mark Hill, Ville Vaara, Ali Ijaz

Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material

Expert curated

Automatically constructed

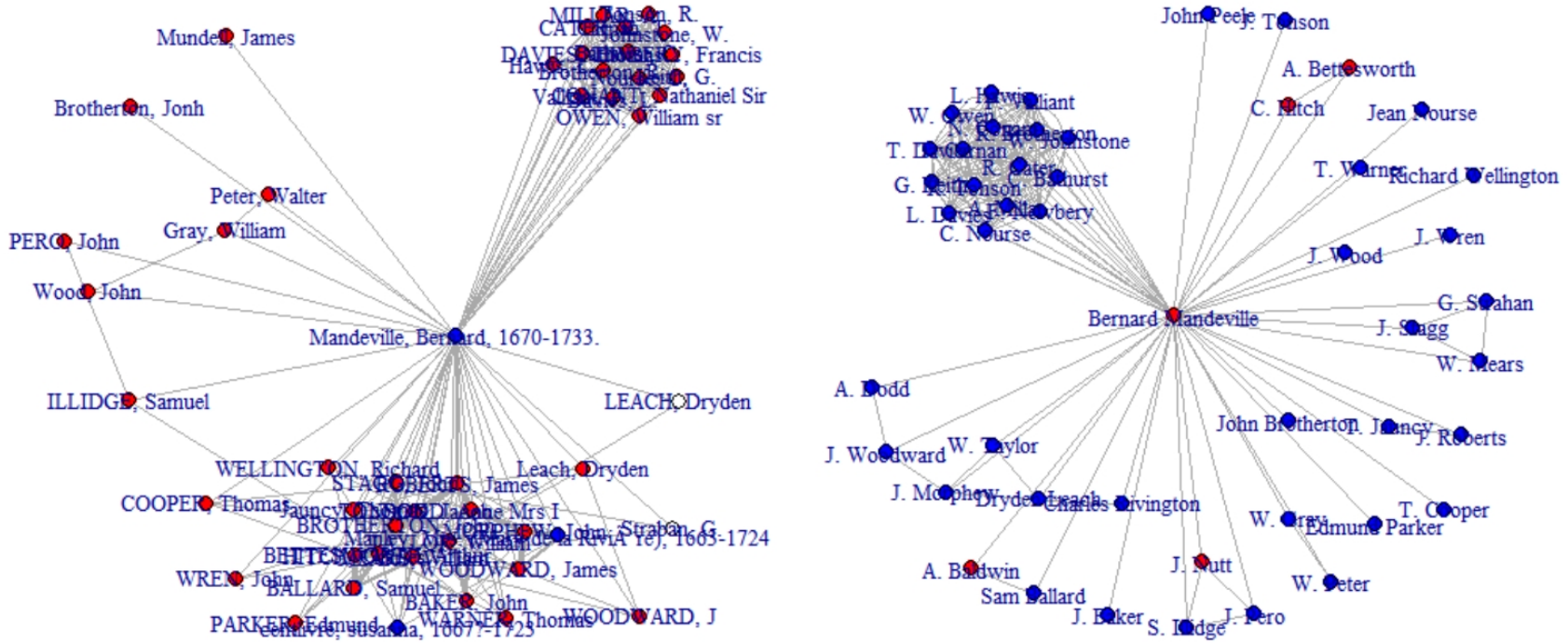


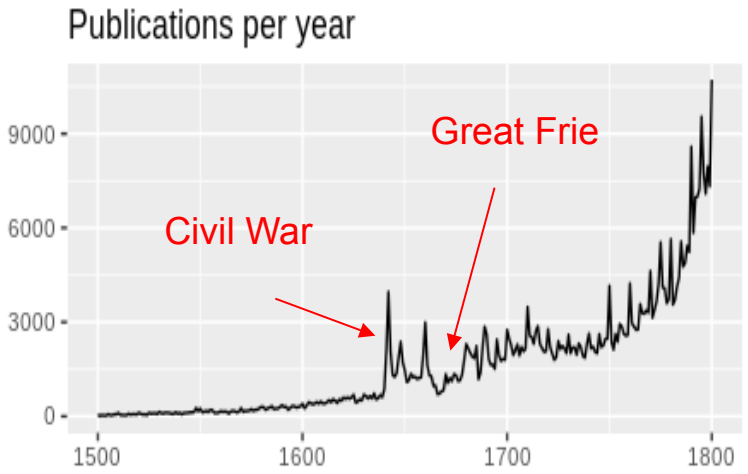
Fig. 4. Bernard Mandeville's (1670-1733) professional network. Left: constructed by a historian of Mandeville.[15] Right: constructed with parsed ESTC metadata.

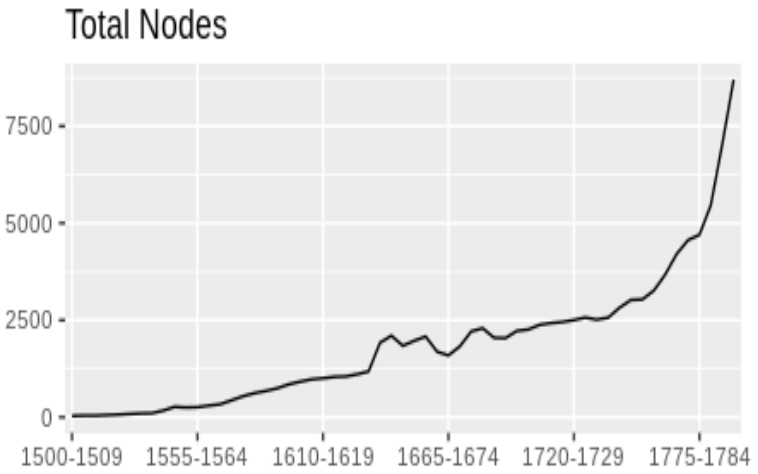
Mark Hill et al. DHN 2019.

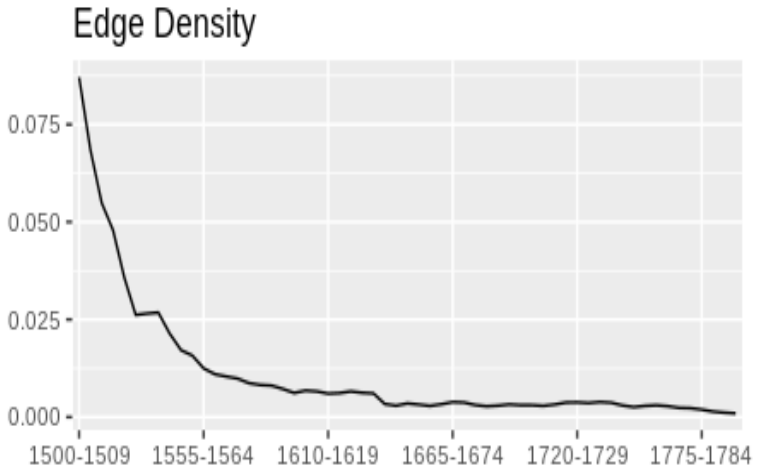
Imprint Field (260)

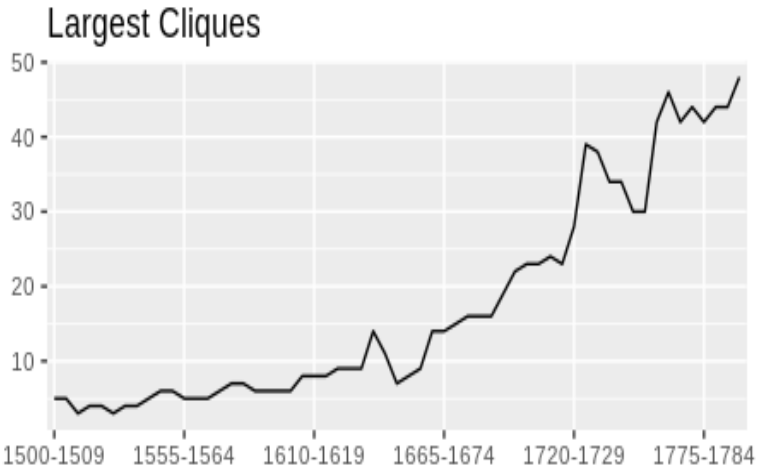
- Named entities from ESTC field 260 (*imprint/publisher statement*)
  - *printed for Bernard Lintott at the Cross-Keys, between the two Temple-Gates, in Fleet-Street. The Double Gallant: Or, the Sick Lady's Cure. A Comedy. Written by Mr. Cibber*
  - *printed by E: Coates. 1655. Sould by Thomas Heath in Covent garden, and Henry Herringman at the Ancker on the lowest side of the New-Exchange*.
- Assigned roles (publisher, printer, bookseller) and addresses
- Corrected and enriched names
- Using town, address, matching initials and name, name combinations, years of activity, etc, harmonized and expand on existing named entities.
- Verified against BBTI (British Book Trade Index)
- Similar issues to VIAF - in particular duplicated entries
- **332,410 named actors in tag 260 unified as 35,252 unique actors..**

ESTC network changes: 1500-1799









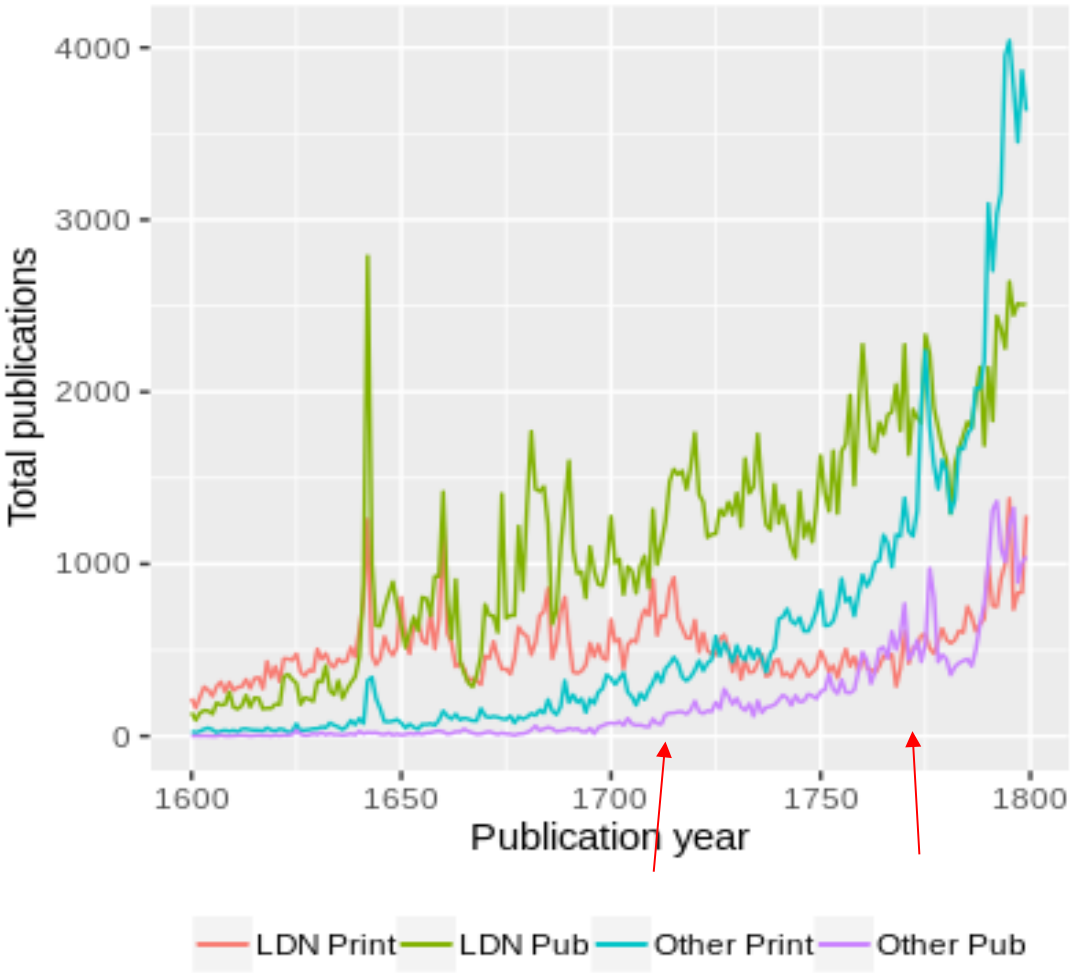
As publications increase, so too do actors (nodes).

Density drop, clique growth: emerging network order representing specialization in trade.

Mark Hill et al. DHN 2019.

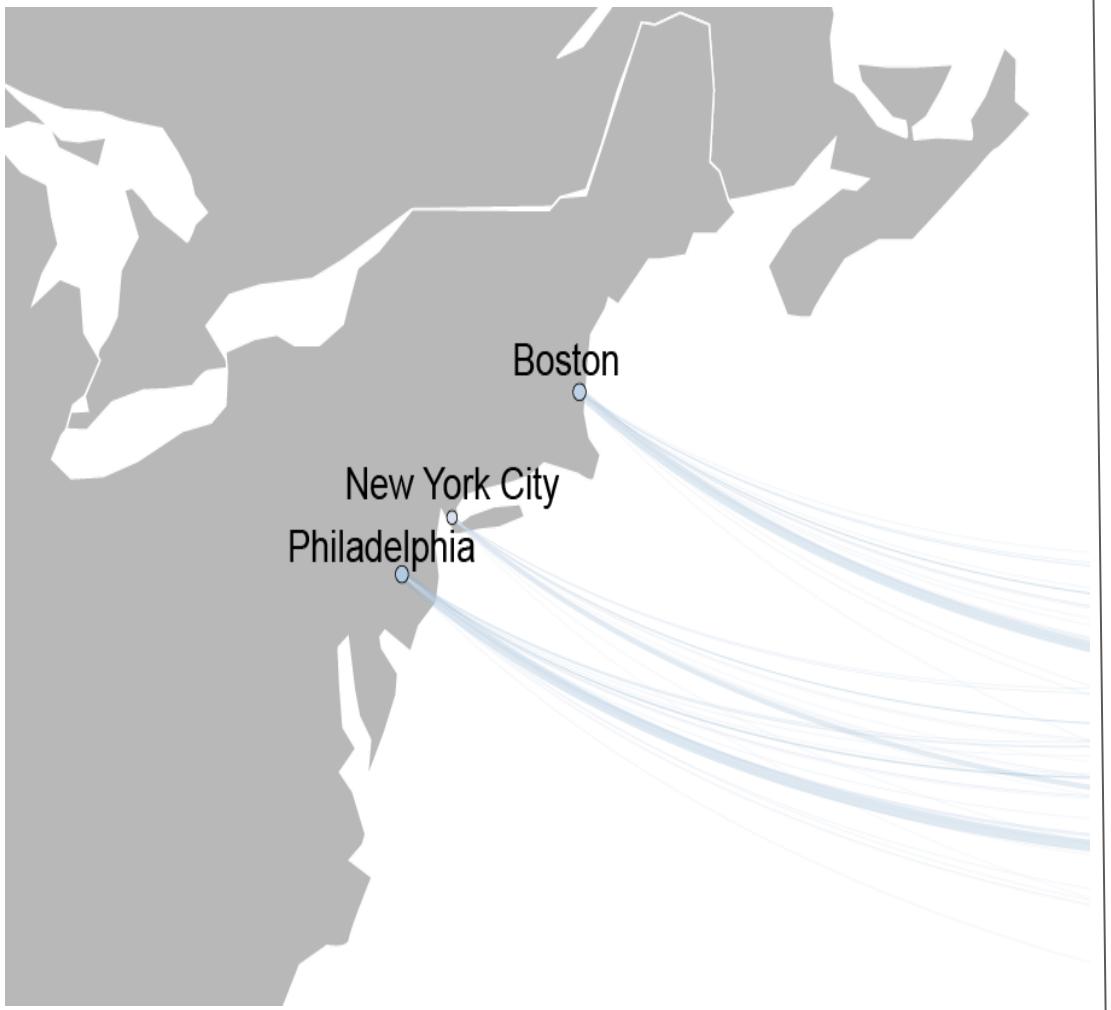
Increase in printers

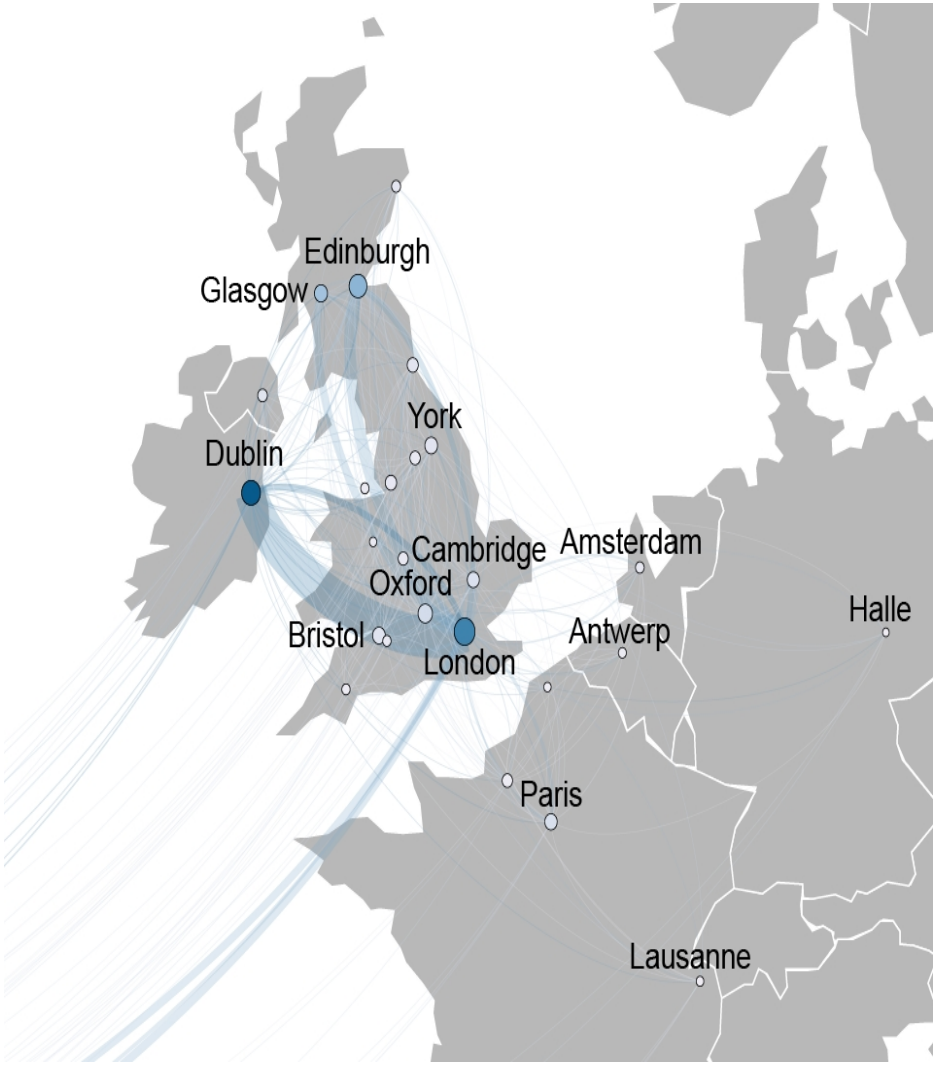
1. Growth of non-London based booktrade actors
2. 1710 Copyright Act ending perpetual copyright, and booms in reprints in 1740s and 1770s (red arrows)
3. Piracy? (still working out how to measure)



Mark Hill et al. 2019.

The Movement of Editions





Mark Hill et al. 2019.

PLOS | MEDICINE

Browse | For Authors | About Us | S

OPEN ACCESS

ESSAY

898,944 VIEWS | 1,119 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

How to Make More Published Research True

John P. A. Ioannidis

Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

RESEARCH PRIORITIES

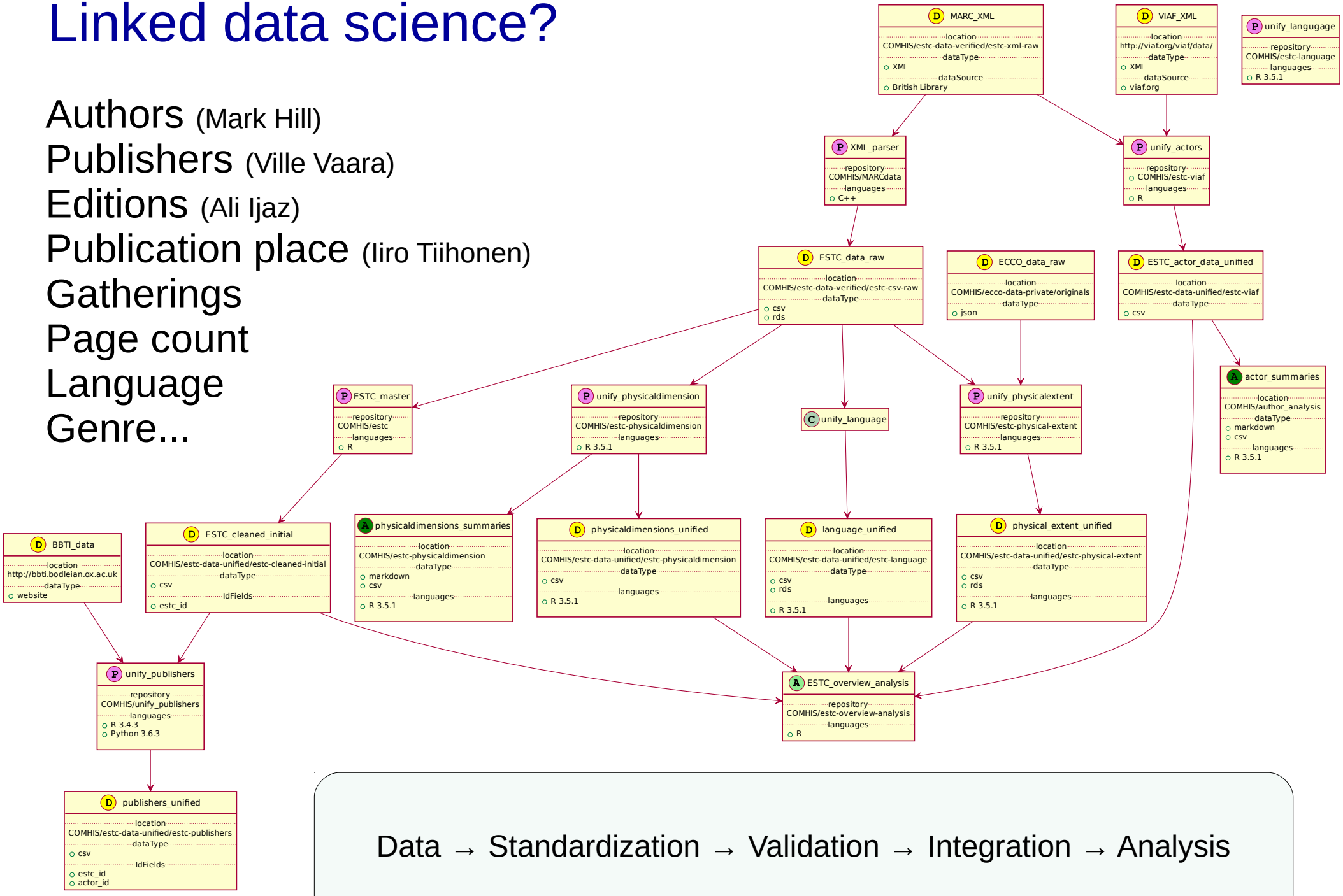
Shining Light into Black Boxes

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

25/37

Linked data science?

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



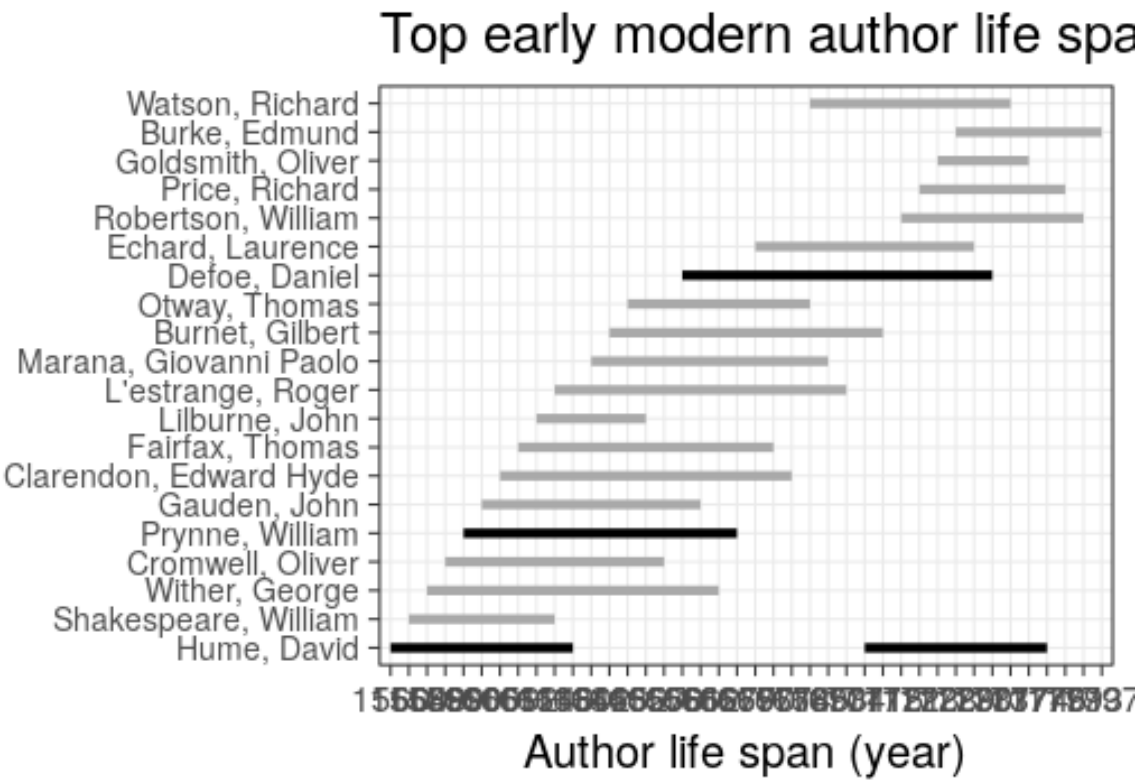
Data → Standardization → Validation → Integration → Analysis

Automated summaries for the unified data

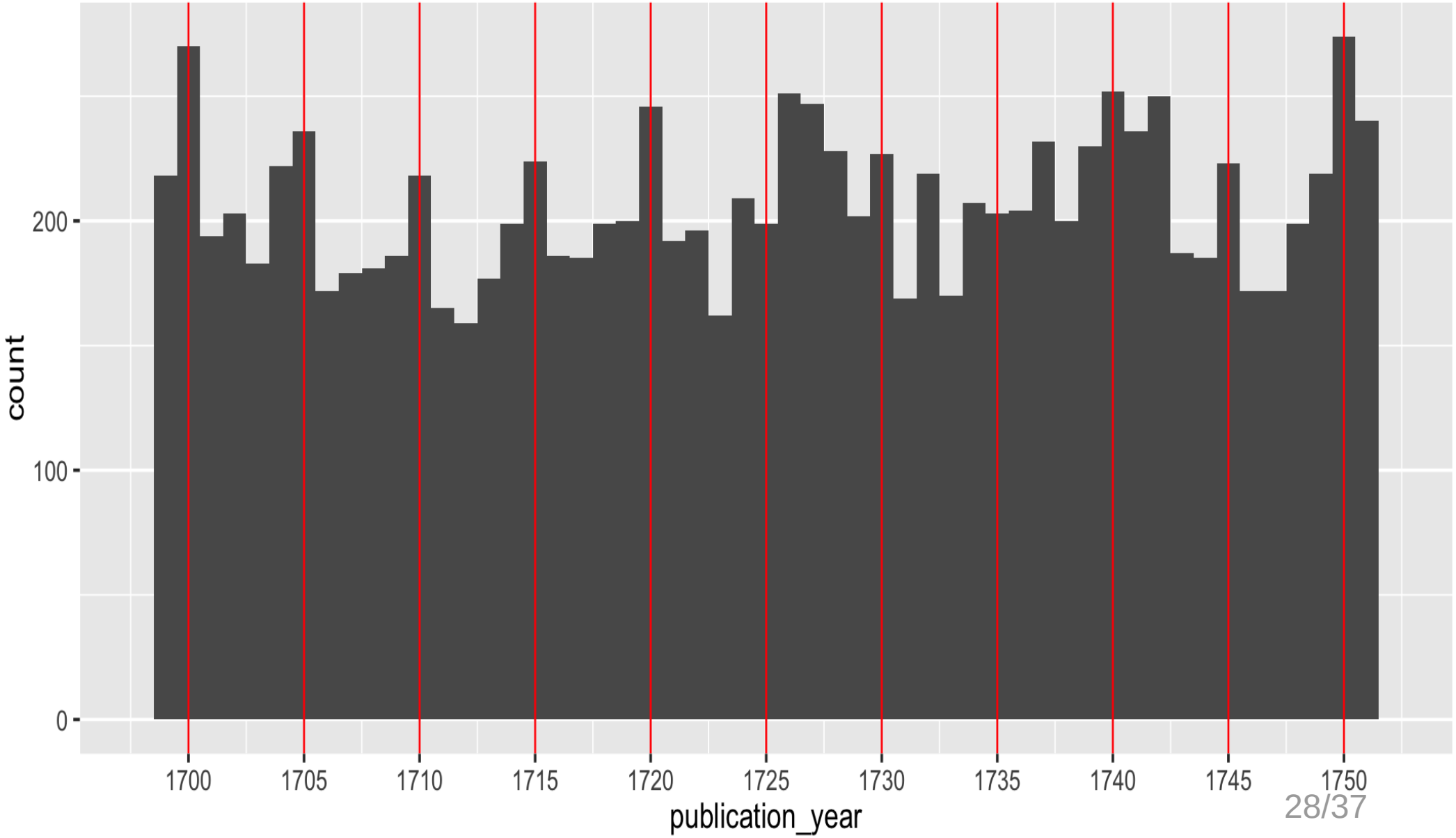
The data spanning years 1488-1955 has been included and contains 70451 documents on the data collection, see the source code for details.

Specific fields

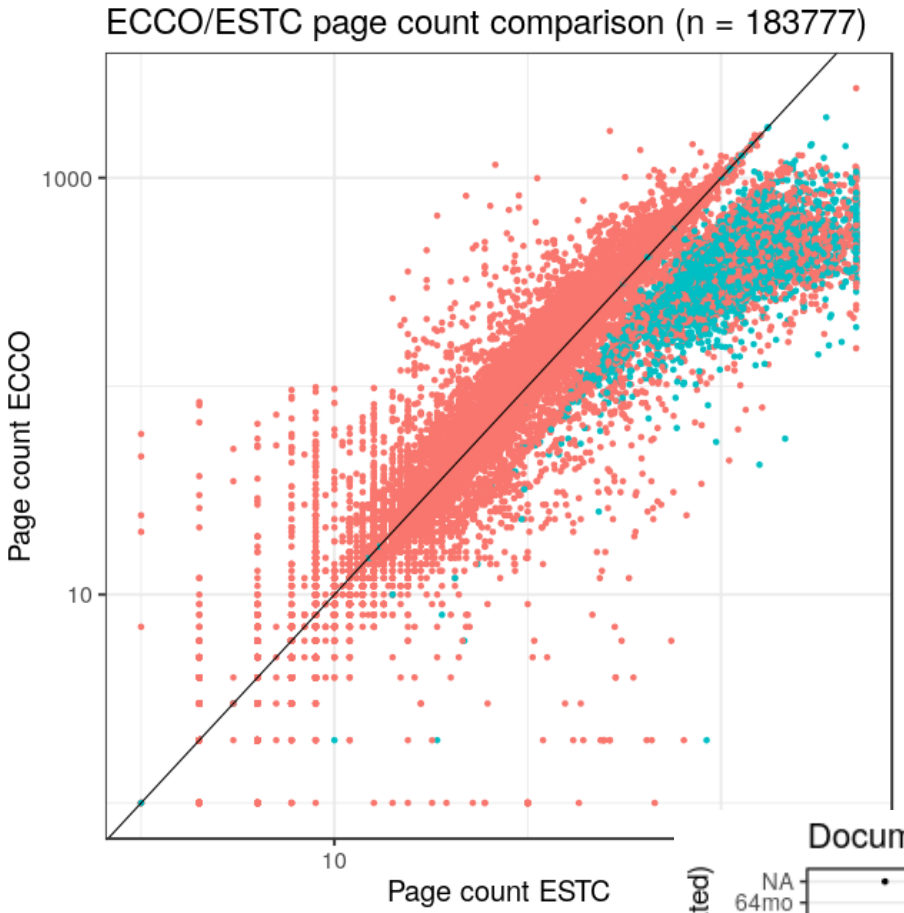
- Author info
- Gender info
- Publisher info
- Publication geography
- Publication year info
- Titles
- Page counts
- Physical dimension
- Document and subject topics
- Languages



Example of questions particular to data in DH collaboration:
The 5-year theory with respect to ESTC catalogue



Validation: page count (ECCO vs. ESTC)



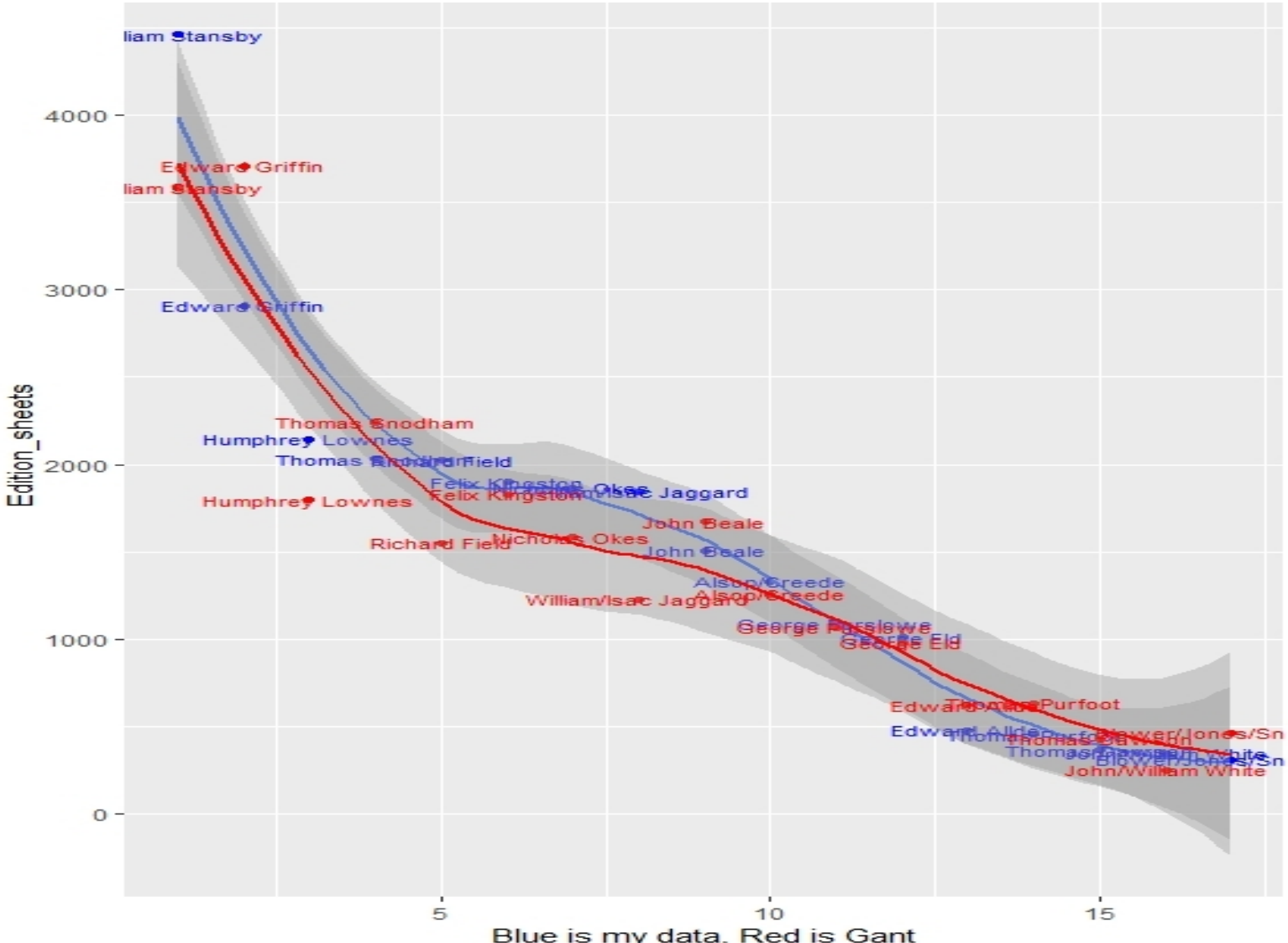
Clean up messy entries

```
polish_physical_extent("iii-xxiv, 118, [2] p.")$
```

```
## [1] 142
```

Scaling up by automated data harmonization: edition sheets: automated vs. curated?

Fig: Iiro Tiihonen. Printers and Publishers in London From 1637 to 1662: a Quantitative Approach



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

31/37

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

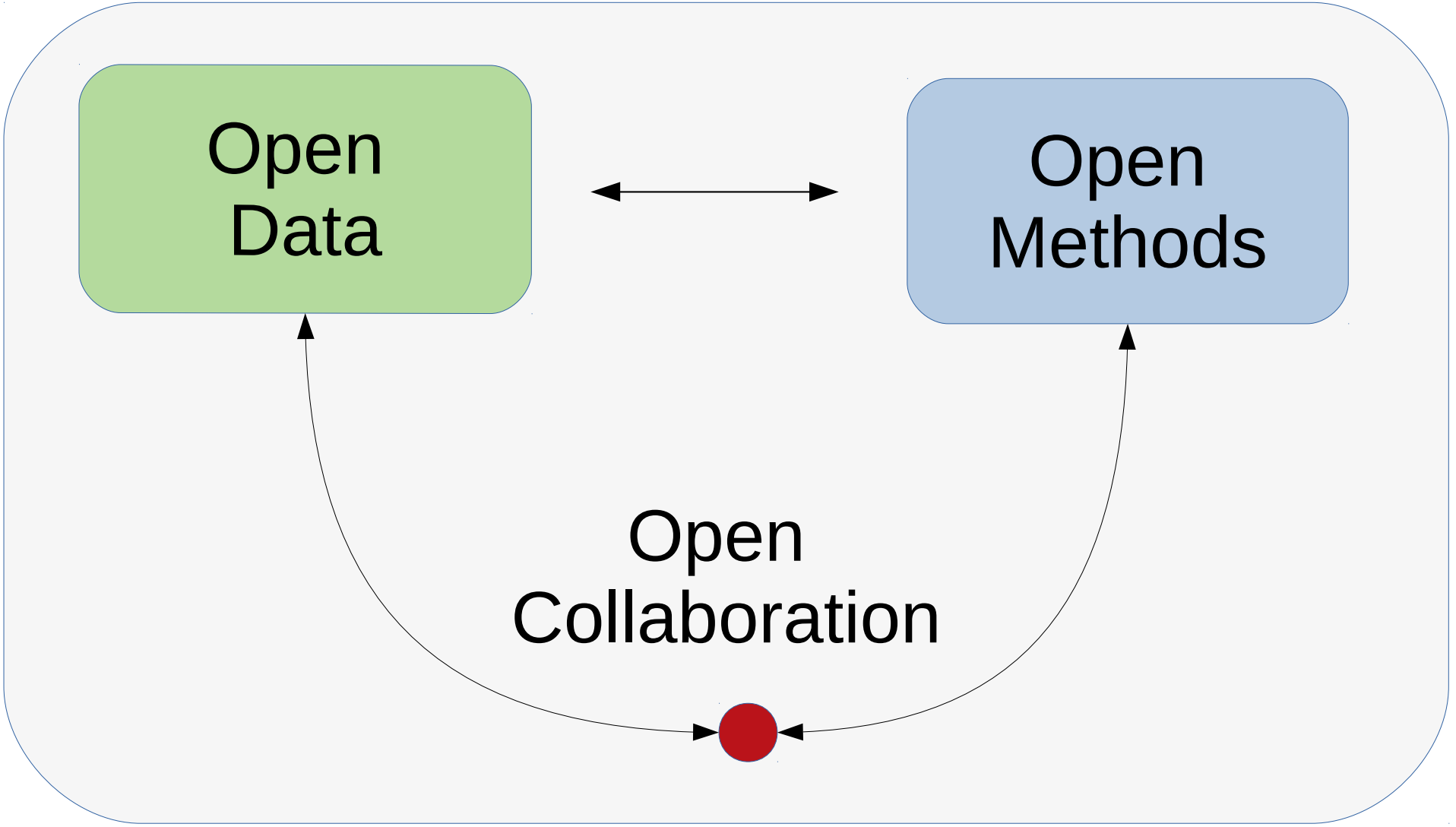
The data is summarized in the following automatically generated files:

- 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

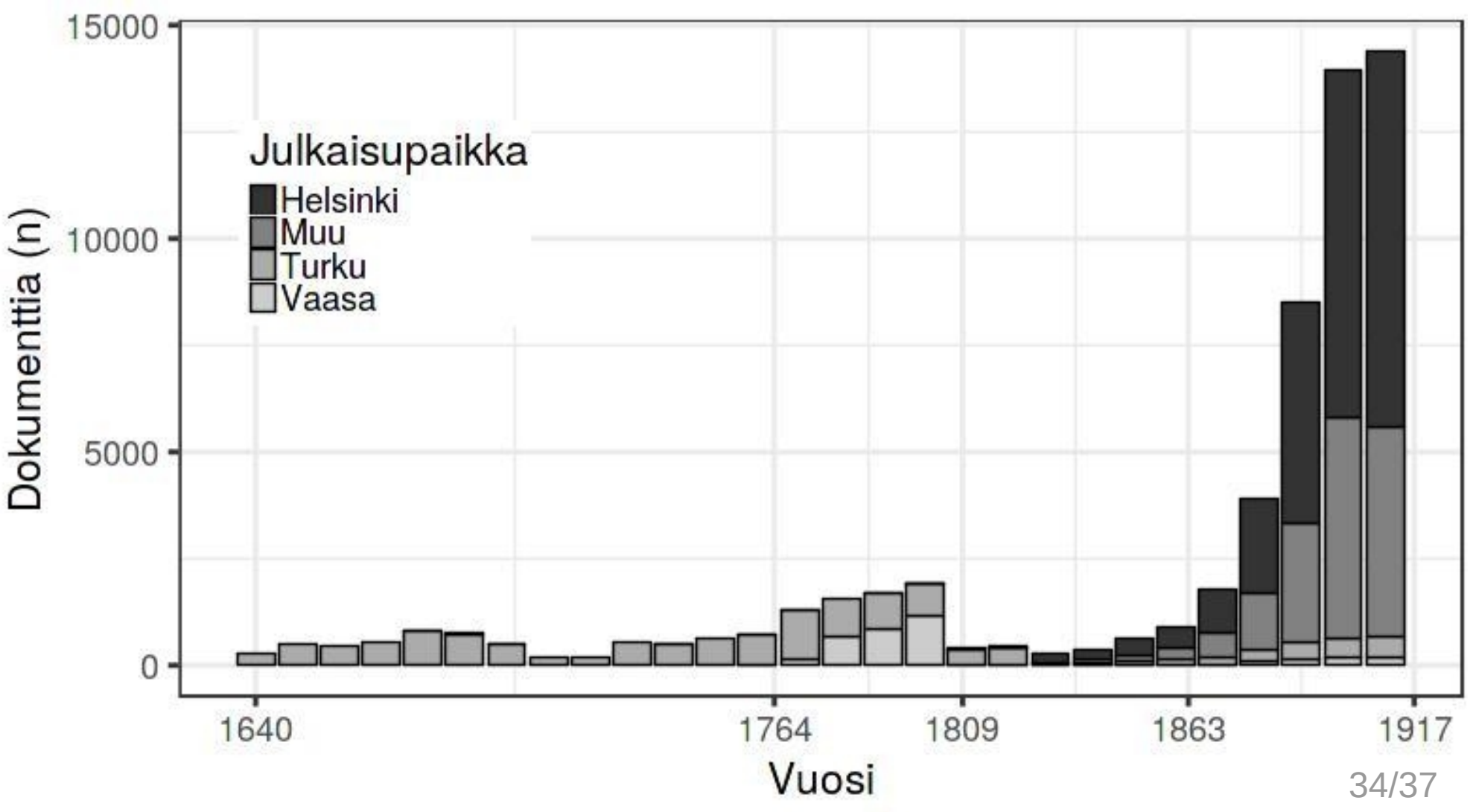
The analyses cover several steps including XML parsing, data harmonization, removing unrecognized entries, enriching and organizing the data, carrying out statistical summaries, analysis, visualization and automated document generation. The analyses and full source code are provided in this repository 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

Elements of open data science



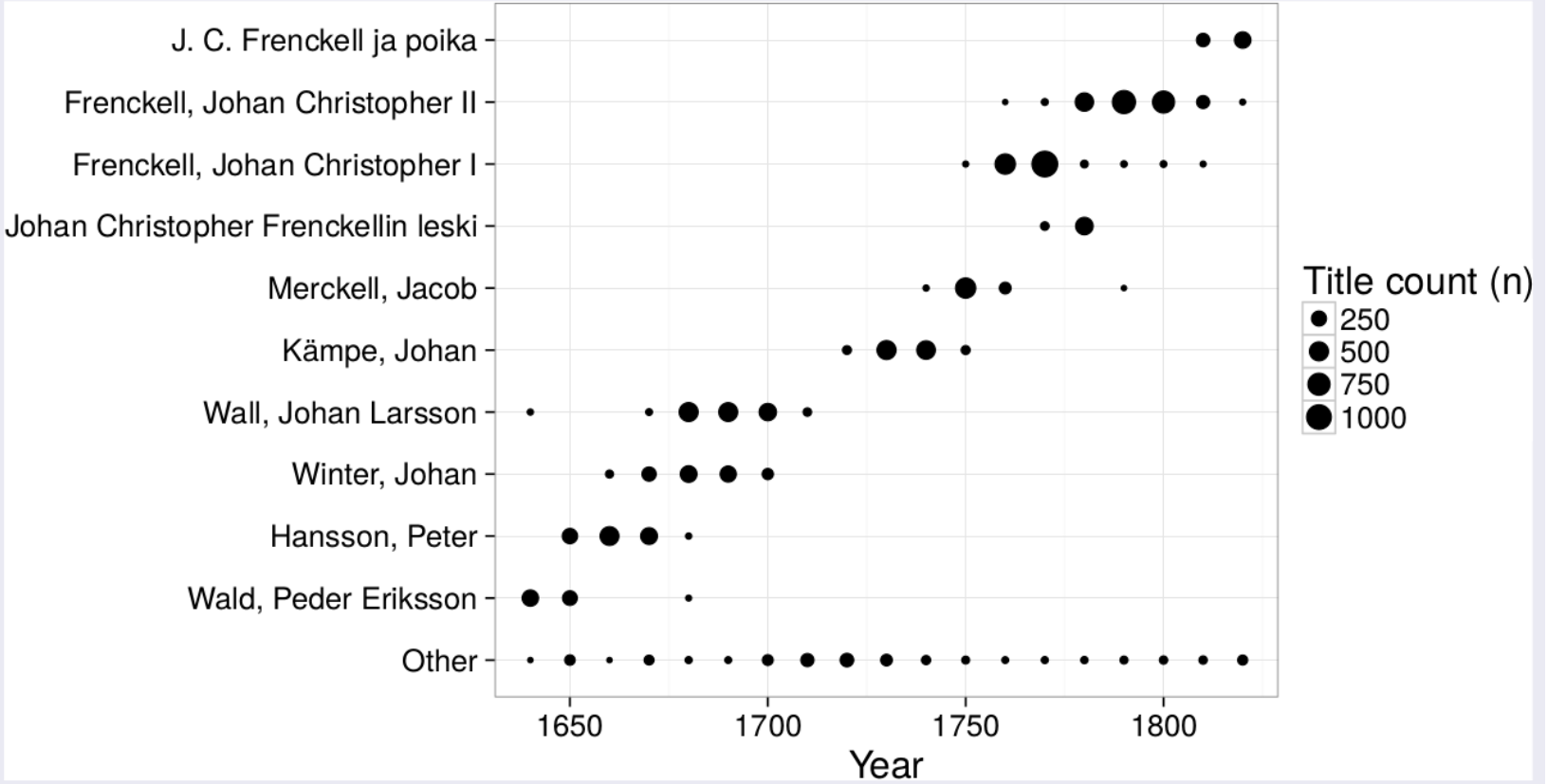
Publishing in Finland 1640-1917 (title count)



Publisher landscape in Turku 1640-1820

Author & publisher networks
Spatio-temporal dynamics
History, philosophy, religion, science..

Top publishers in Turku/Fennica



Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material

Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640-1828

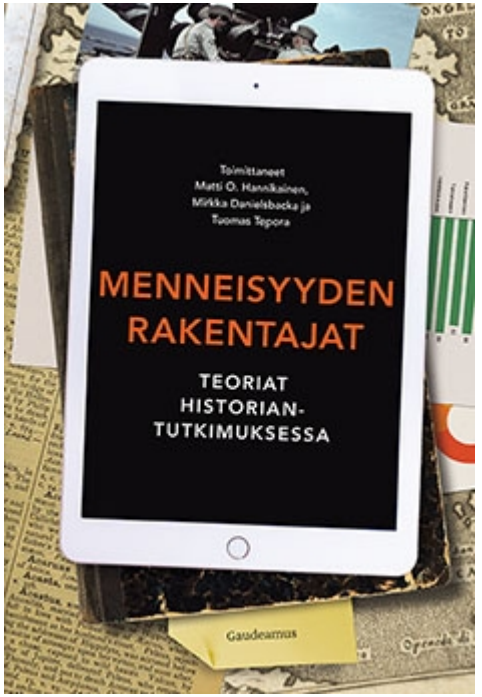
A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828

Mikko Tolonen, Leo Lahti, Hege Roivainen & Jani Marjanen

REFEREE-ARTIKKELIT

AATEHISTORIA JA DIGITAALISTEN AINEISTOJEN MAHDOLLISUUDET

11.8.2015 MIKKO TOLONEN JA LEO LAHTI



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

A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800

Authors: Leo Lahti, Niko Ilomäki, Mikko Tolonen

Helsinki Computational History Group

https://www.helsinki.fi/en/researchgroups/computational-history

Mikko Tolonen
Jani Marjanen
Mark Hill
Ali Ijaz
Ville Vaara
Hege Roivainen
Eetu Mäkelä
Tanja Säily
Antti Kanner
Simon Hengchen
Iiro Tiihonen

Metadata as research object

- Underestimated
- Objective
- Scalable

SUOMEN AKATEMIA
FINLANDS AKADEMI • ACADEMY OF FINLAND

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Turun yliopisto
University of Turku

16 40
KANSALLIS KIRJASTO