Bibliographic Data Harmonization in Research

open ecosystems for scalable collaboration

HELDIG Summit

Helsinki

7 November 2019

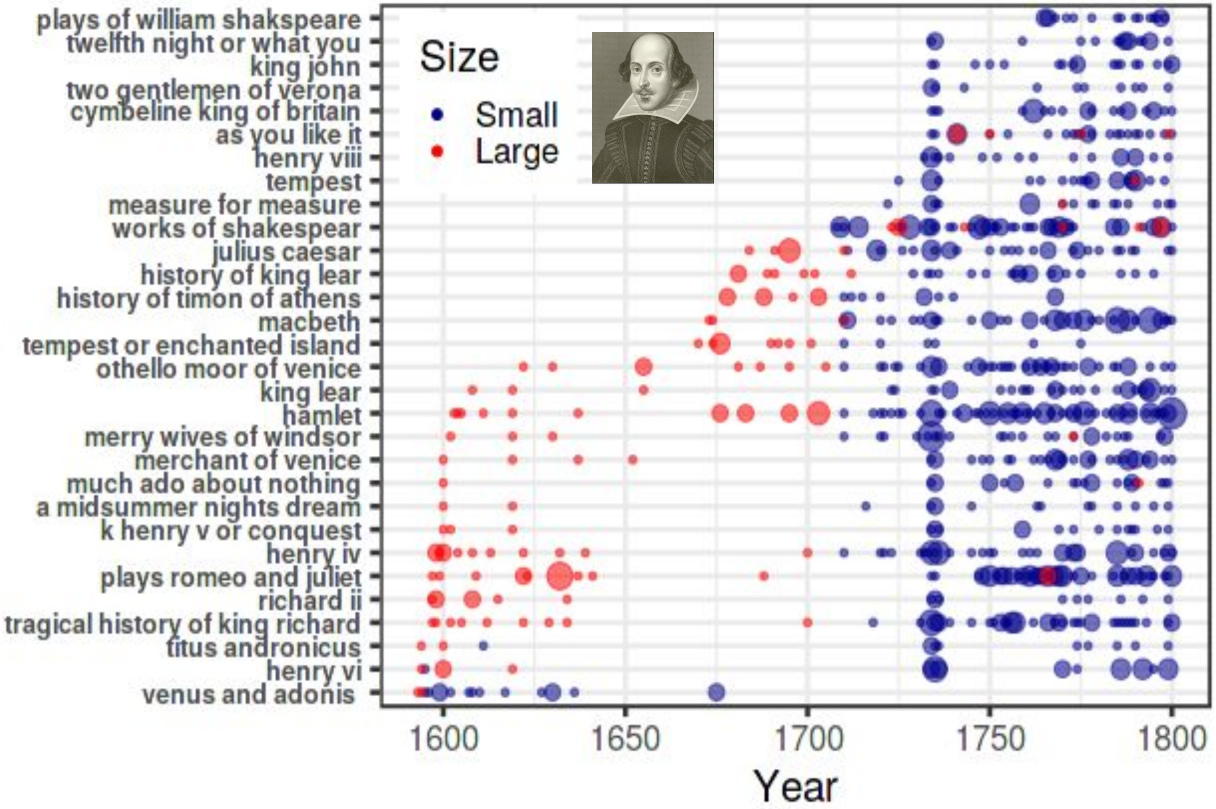
Leo Lahti (University of Turku)

leo.lahti@iki.fi | @openreslabs

Shakespeare was made big by small books!

Drastic shift from large (2fo/4to) to small (8vo/12mo) books observed around 1700’s.

… how reliable and representative this data set is?



Data: ESTC | Figure: DH2019 (best!) poster, Utrecht.

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

Authors: Leo Lahti, Niko Ilomäki, Mikko Tolonen

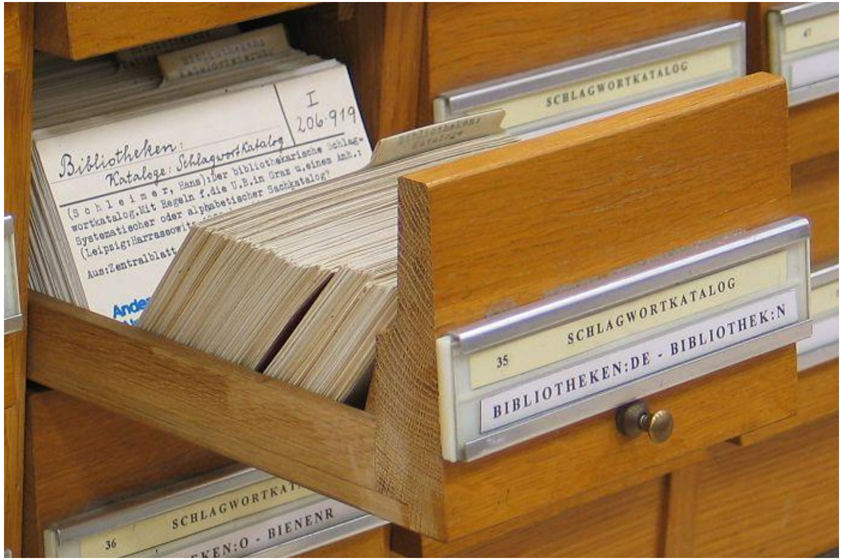
One (non-standard) XML file

~480 000 entries (1470-1800)

Designed for information retrieval rather than quantitative analysis

Not openly available

Browsable online: http://estc.bl.ac.uk



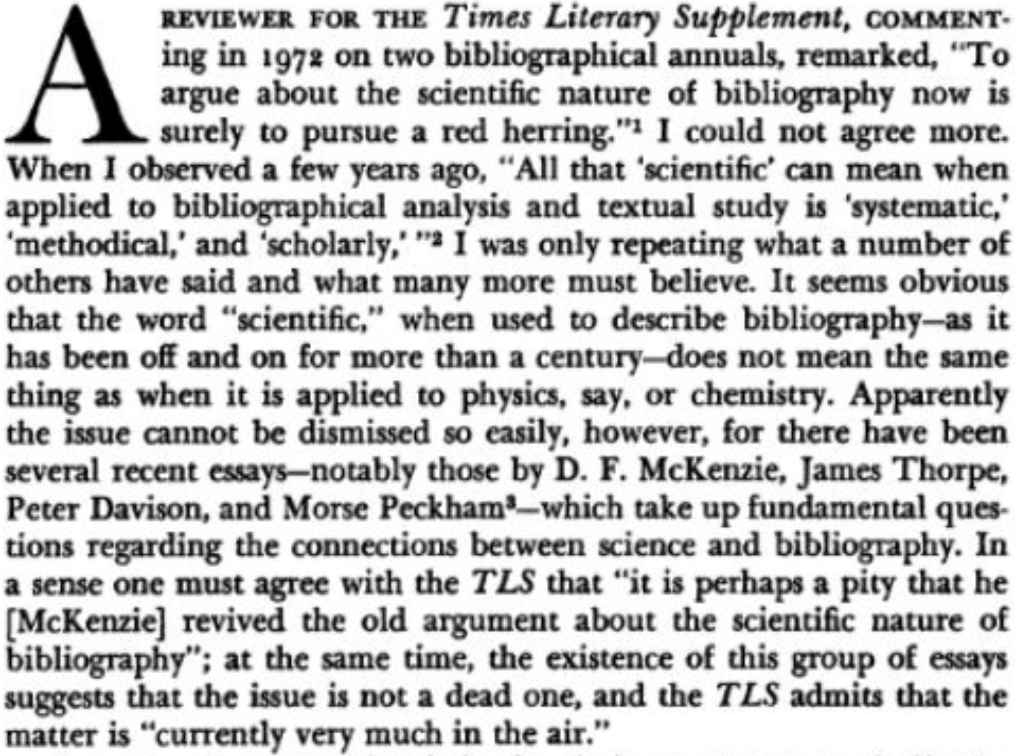
Subject catalogue of the University Library of Graz.
Source: Wikimedia Commons.

Research potential of library catalogues has been debated for decades

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

Published by: Bibliographical Society of the University of Virginia

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."

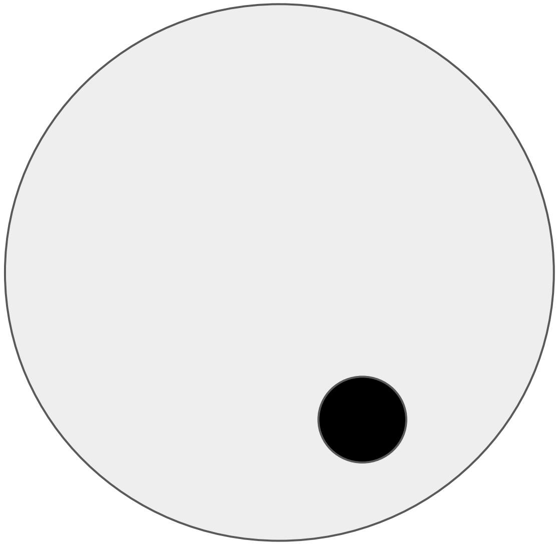
Original data not ready for analysis

Variants of *Shakespeare* in ESTC

ghost of shakespeare
kenrick, william shakespeare
shakespeare, john
shakespeare room (birmingham, england)
shakespeare, thomas, active 1598
shakespeare, william
shakespeare, william, 1564-1616
shakespeare, william, 1564-1616., (adaptations)
shakespeare, william, 1564-1616, (adaptations)
shakespeare, william, 1564-1616., (adaptions)
shakespeare, william, 1564-1616., (selections)

Actors:

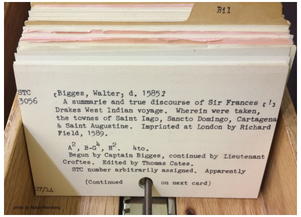
- 558,243 original
- 92,044 (16%) harmonized



Actor harmonization: Mark Hill, Ville Vaara

From library catalogues to research reports?

Research potential



B11

STC 3056

[Bigges, Walter, d. 1585?
A summarie and true discourse of Sir Frances [!] Drakes West Indian voyage. Wherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint Augustine. Imprinted at London by Richard Field, 1589.

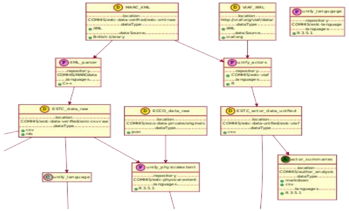
A^2, B-G^4, H^2. 4to.
Begun by Captain Bigges, continued by Lieutenant Croftes. Edited by Thomas Cates.
STC number arbitrarily assigned. Apparently

(Continued on next card)

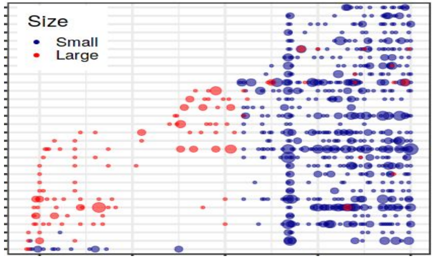
27/14

photo by Abbie Weinberg

Open bibliographic data science ecosystem



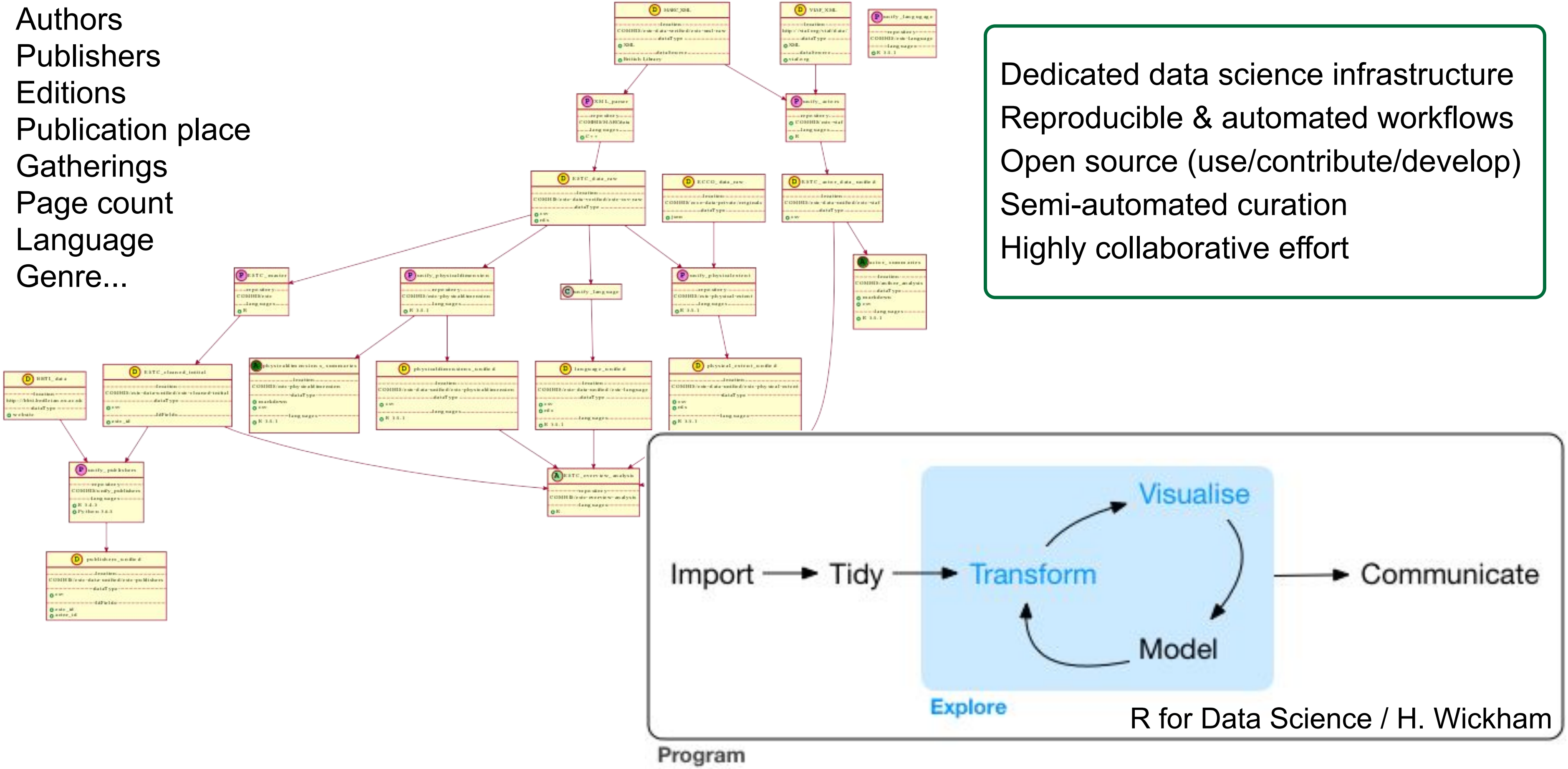
Research cases



Open data science ecosystem?

Authors
Publishers
Editions
Publication place
Gatherings
Page count
Language
Genre...

Dedicated data science infrastructure
Reproducible & automated workflows
Open source (use/contribute/develop)
Semi-automated curation
Highly collaborative effort

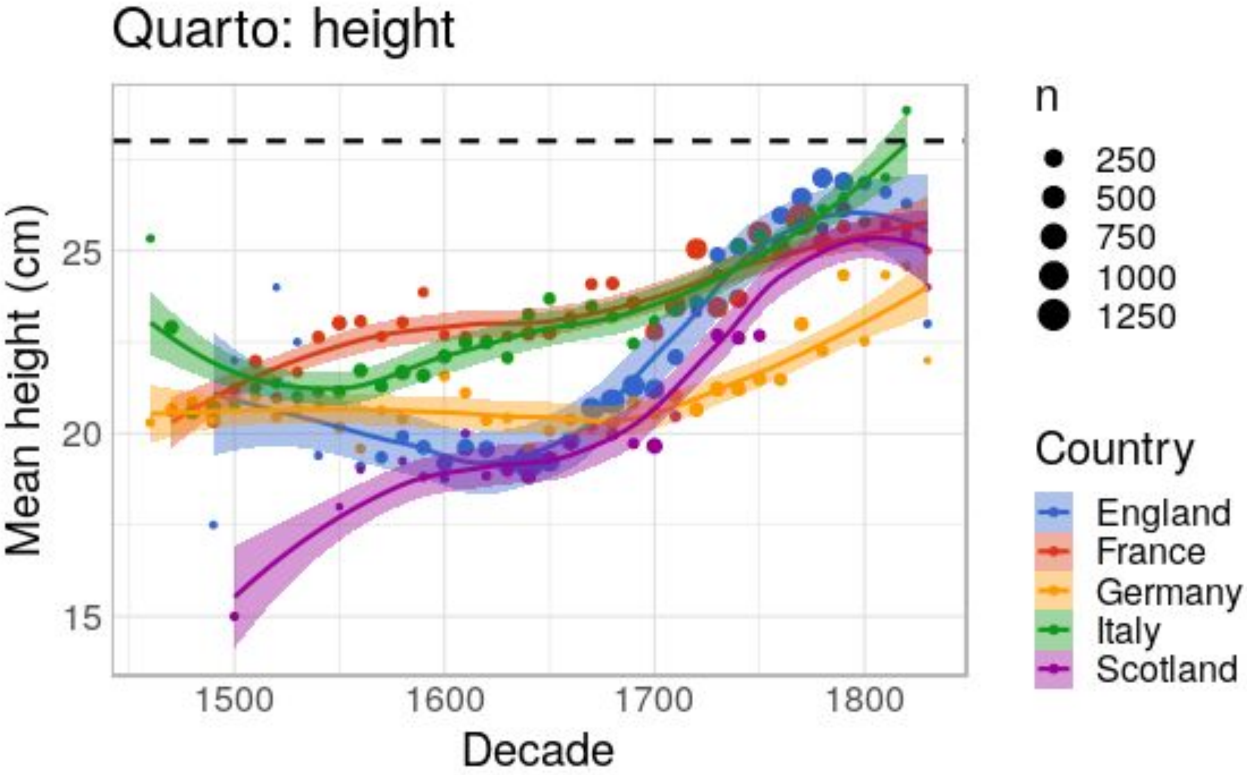


R for Data Science / H. Wickham

“Standard” doc sizes vary across time and space

Data availability (HPB):

- Gatherings: 22.5%
- Height: 11.6%
- Width: 1.1%



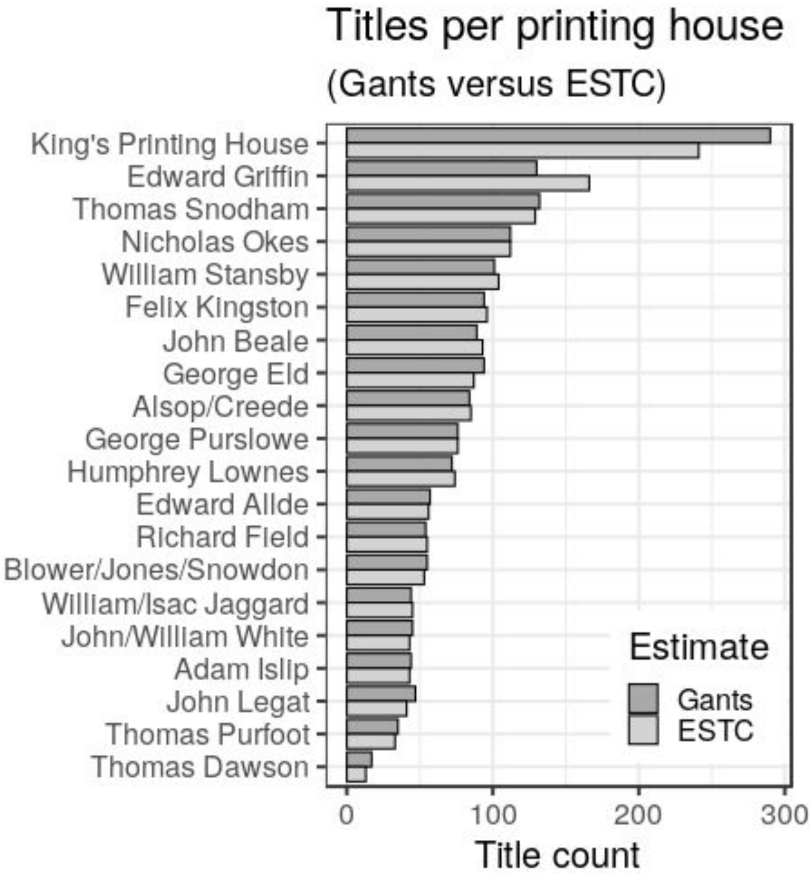
Boost curation & scalability by automation

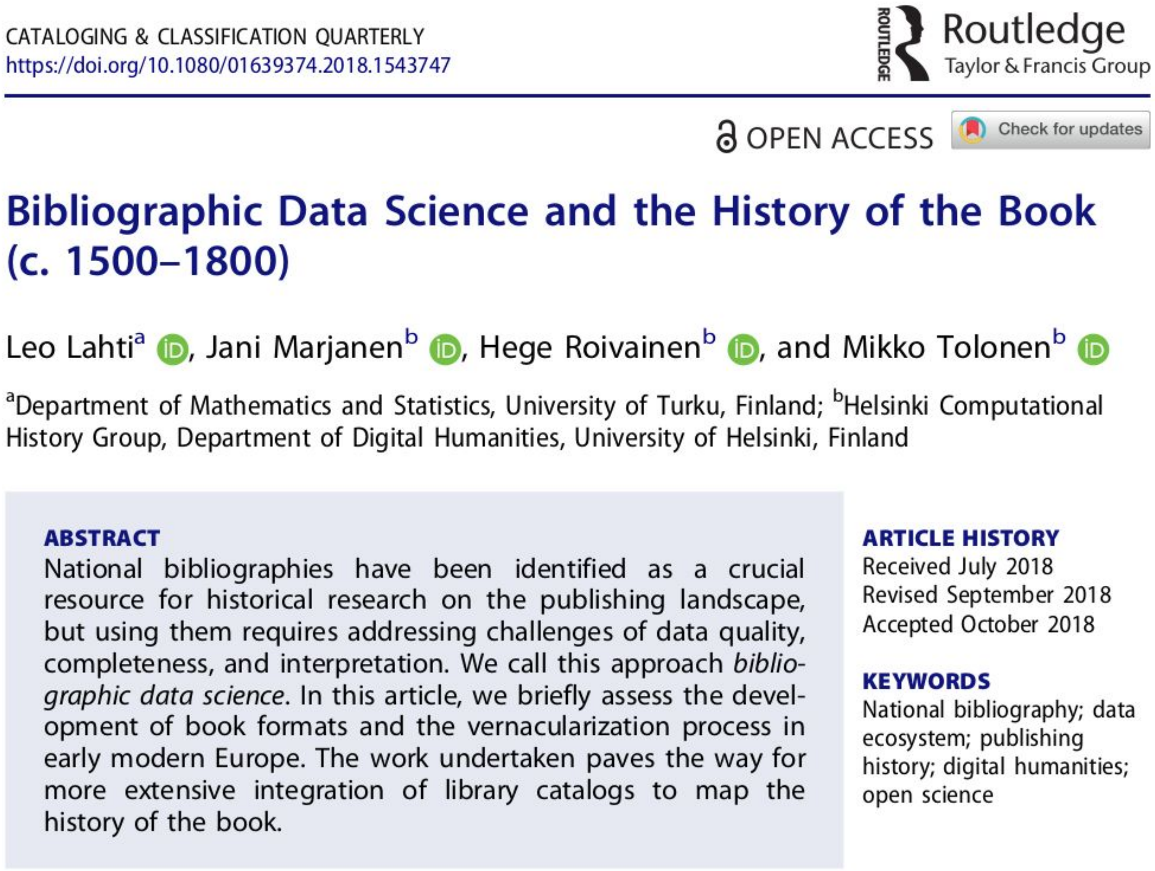
Counting editions by publishers in London 1637-1662

- Manual curation (David Gants)
- Automated analysis (Iiro Tiihonen)

Good correspondence supports our automated approach.

Manually curated data from: David Gants. A Quantitative Analysis of the London Book Trade. *Studies in Bibliography* 55:185-213, 2002





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

Thanks!

Comhis
Helsinki Computational History Group

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

SUOMEN AKATEMIA
FINLANDS AKADEMI • ACADEMY OF FINLAND

Turun yliopisto
University of Turku

Material for the slides contributed by: Mikko Tolonen, Leo Lahti, Jani Marjanen, Mark Hill, Ali Ijaz, Ville Vaara, Hege Roivainen, Iiro Tiihonen

Helsinki Computational History Group:
https://www.helsinki.fi/en/researchgroups/computational-history