Bibliographic Data Science:
Open Ecosystems for Scalable Collaboration

Better Science Through Better Data 2019; #scidata19

Wellcome Collection, London

6 November 2019

Leo Lahti (University of Turku, Finland)

leo.lahti@iki.fi | @openreslabs

Case study
publishing in the early modern era
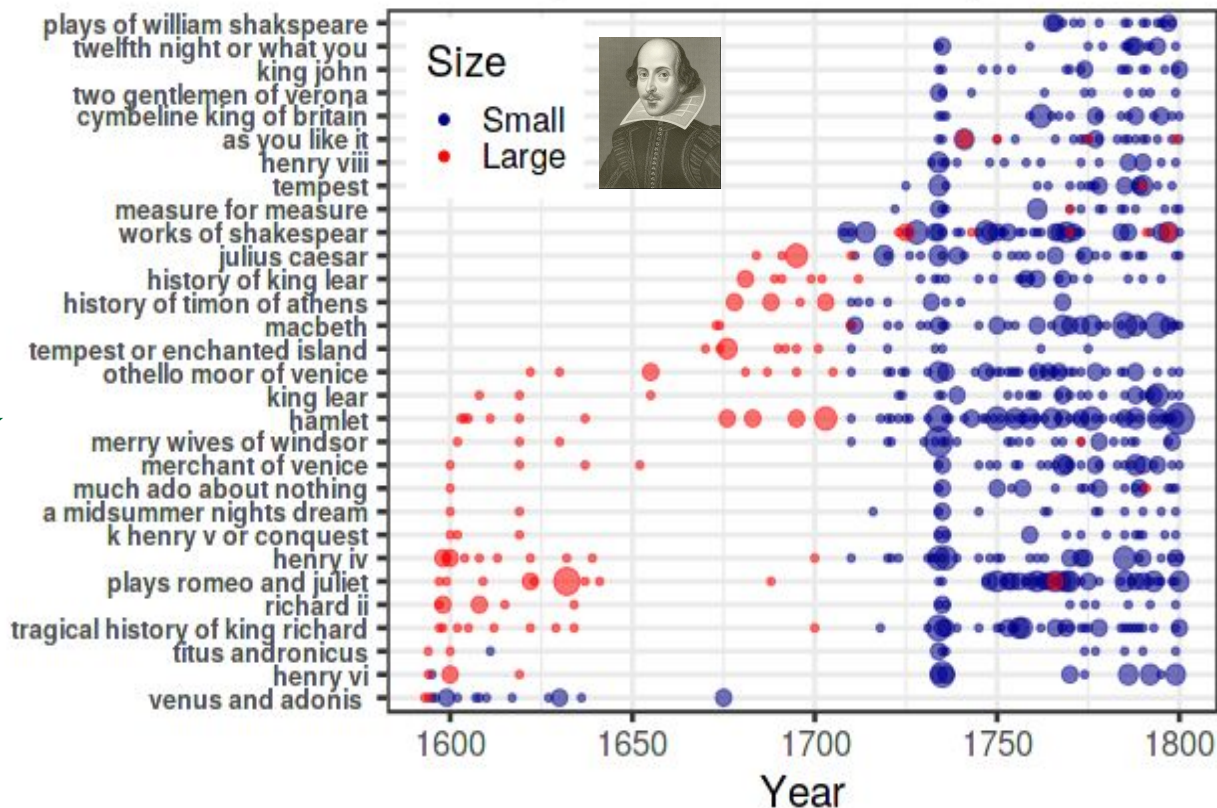
Data
English Short Title Catalogue

Methods
open data science & collaboration

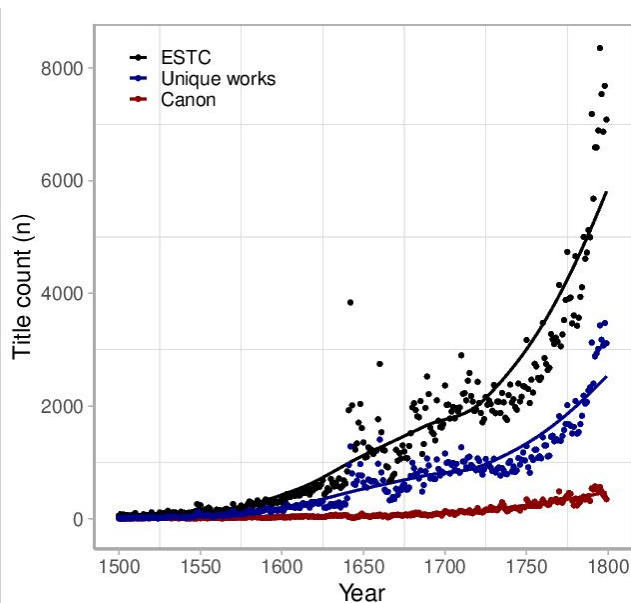# Shakespeare was made big by small books!

ESTC indicates drastic shift from large (2fo/4to) to small (8vo/12mo) book format around 1700's.

Top works

Critical aspects

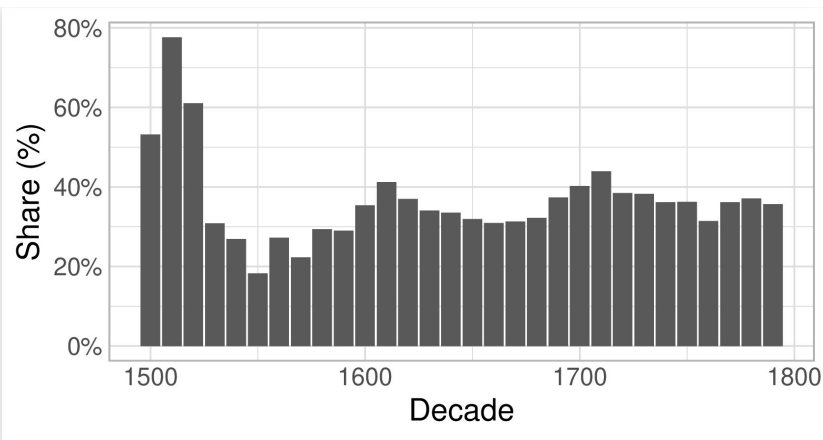**How many editions** were printed for <u>unique works</u> (documents, works, canon 1500-1800)

Who published literature? reverse *author centric approach* & revise book history!

Publications by the largest publishers (top-1% percentile)

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

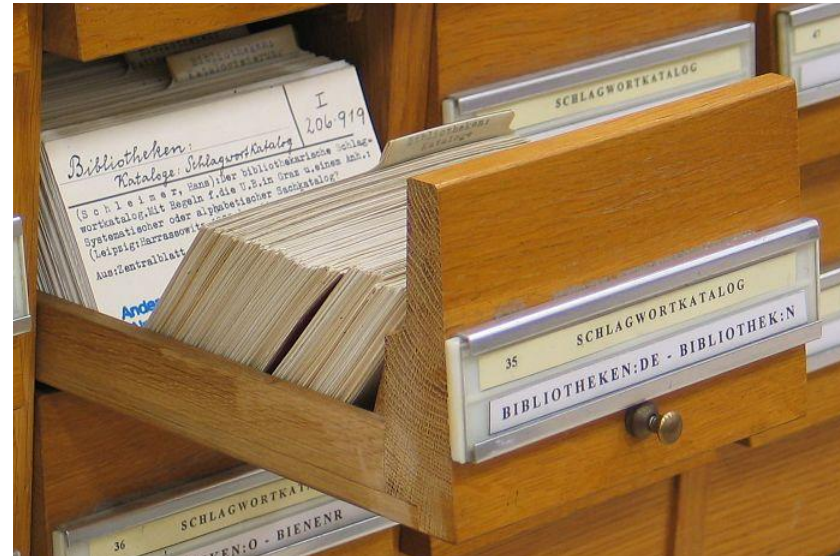**Authors:** Leo Lahti, Niko Ilomäki, Mikko Tolonen ✉

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

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

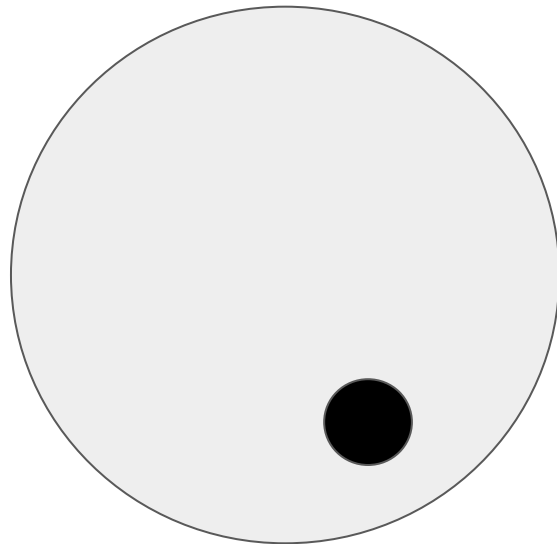shakespeare, william, 1564-1616., (adaptations)

shakespeare, william, 1564-1616, (adaptations)

shakespeare, william, 1564-1616., (adaptions)
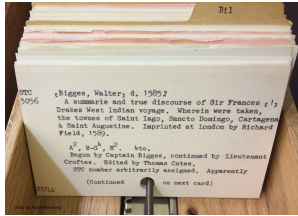
shakespeare, william, 1564-1616., (selections)

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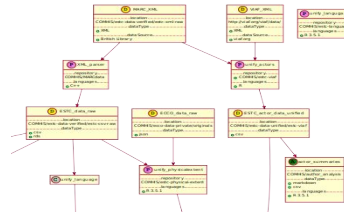- 558,243 original
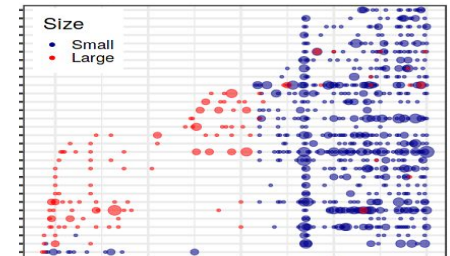- 92,044 (16%) harmonized

From library catalogues to research reports?

Research potential

Open bibliographic data science ecosystem
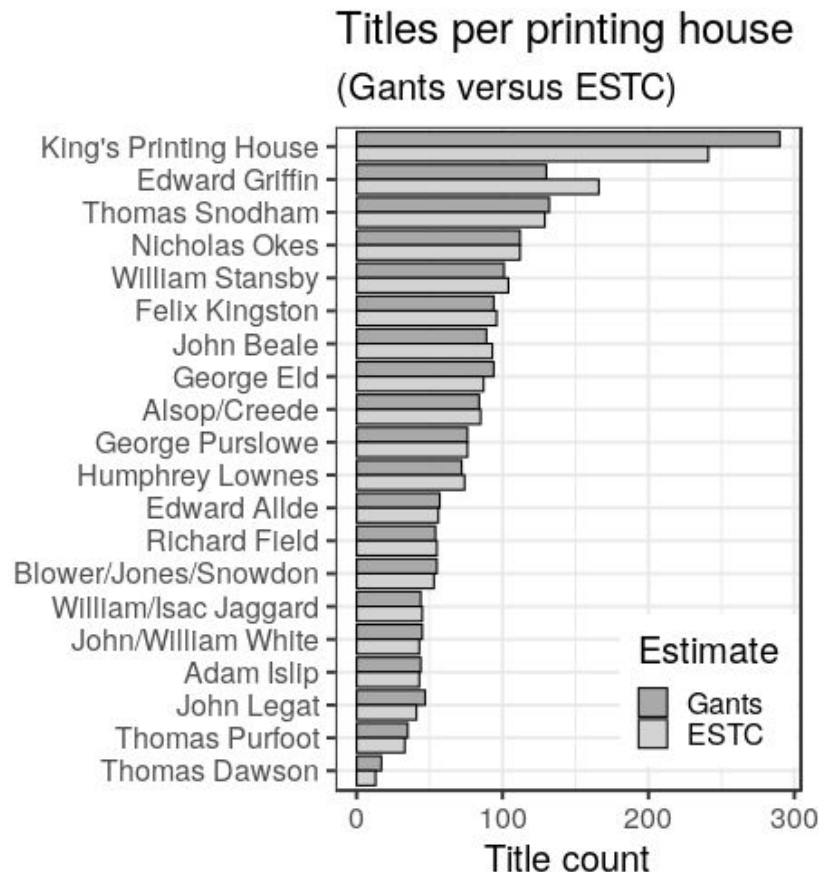

Research cases

Boost quality control & scalability by automation

Counting editions by various publishers in London 1637-1662
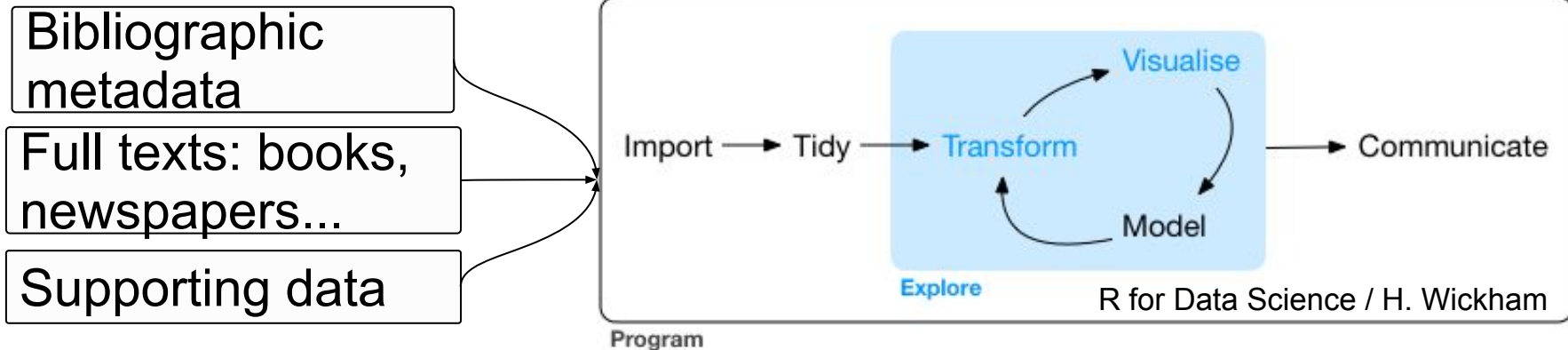
- Manual curation (David Gants)
- Automated analysis (Iiro Tiihonen)

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

Titles per printing house
(Gants versus ESTC)

(Open) bibliographic data science ecosystem

Bibliographic metadata

Full texts: books, newspapers...

Supporting data

R for Data Science / H. Wickham

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

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

Source: Wikimedia Commons / Public domain

Routledge
Taylor & Francis Group

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

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

[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.

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

Thanks!

Comhi/
Helsinki Computational History Group

Turun yliopisto
University of Turku

SUOMEN AKATEMIA
FINLANDS AKADEMI · ACADEMY OF FINLAND

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI