Bibliographic Data Science: from records to research

Oct 9, 2019, CERL Annual Seminar, Göttingen
Mikko Tolonen & Leo Lahti
University of Helsinki | University of Turku
Openly available bibliographic records: crossing the borders

- **Overcoming the nationally delineated perspective.** Not a new phenomena: this is how our history has been written since the early modern period. We have to change it.
- **We need a genuinely cross-European take on public discourse of the past** (metadata work enables this), **combined** when possible to opportunity for text mining.
- **But, we are nowhere near yet that this would be reality.** It has to be accepted.
To understand public communication covering the early modern Europe

**Movement of ideas**
- Metadata work based on several different library catalogues
- genres (poetry, pamphleteering); intellectual traditions (natural law tradition, ancient texts)
- text reuse: genres (historical works, quoting practices)

**Conceptual change**
- concepts are crucial, but not directly jumping into this for various reasons
- Theoretical underpinning (historians + linguists)
- Concepts as linguistic objects (linguists + historians + CS)

**Research data releases**
- ESTC; Fennica; Kunglica; CERL; ECCO text reuse (+ EEBO text reuse); Finnish Newspapers

**Tools for others**
- UIs, APIs, shiny apps etc.
Research potential of library catalogues has been debated for decades.

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 Peckham³—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.”
Bibliographic Data Science and the History of the Book (c. 1500–1800)

Leo Lahti\textsuperscript{a}, Jani Marjanen\textsuperscript{b}, Hege Roivainen\textsuperscript{b}, and Mikko Tolonen\textsuperscript{b}

\textsuperscript{a}Department of Mathematics and Statistics, University of Turku, Finland; \textsuperscript{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
From library catalogues to research & reports

Research potential

Open bibliographic data science ecosystem

Research cases

Original Articles

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

Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen

Pages 5-23 | Received 07 Jul 2018, Accepted 10 Oct 2018, Published online: 07 Jan 2019
Octavo and the Enlightenment & vernacularization in Europe
“I have observed that the Author of a Folio, in all Companies and Conversations, sets himself above the Author of a Quarto; the Author of a Quarto above the Author of an Octavo; and so on, by a gradual Descent and Subordination, to an Author in Twenty Fours. This Distinction is so well observed, that in an Assembly of the Learned, I have seen a Folio Writer place himself in an Elbow-Chair, when the Author of a Duo-decimo has, out of a just Deference to his superior Quality, seated himself upon a Squabb. In a word, Authors are usually ranged in Company after the same manner as their Works are upon a Shelf.”

– Joseph Addison, The Spectator (6 November, 1712)
Shakespeare was made big by small books!

Figure: Mikko Tolonen | Data: ESTC
Measuring printing activity: quantitative indicators

Page count estimation from MARC standards: “[4],vii-xii,[4],222p.,plate” → 240 pages
## Unit tests for quality control

<table>
<thead>
<tr>
<th>Original</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>6:o</td>
<td>6to</td>
</tr>
<tr>
<td>46 cm (2°)</td>
<td>2fo</td>
</tr>
<tr>
<td>29-40 cm. (4°; 2°)</td>
<td>NA</td>
</tr>
<tr>
<td>4°. '</td>
<td>4to</td>
</tr>
<tr>
<td>2° (2 half-sheets)</td>
<td>2fo</td>
</tr>
<tr>
<td>1/2°; 2°.</td>
<td>2fo</td>
</tr>
</tbody>
</table>
Quality control / Bias detection

ECCO/ESTC page count comparison (n = 183777)

Documents with missing page counts (original; n=18266)

Author life span (year)
Linked data science?

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

Data → Standardization → Validation → Integration → Analysis
(Open) bibliographic data science ecosystem

Bibliographic metadata

Full texts: books, newspapers...

Supporting data

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

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
Interactive tools or algorithmic workflows?

- **Algorithms**
  - Independent
  - Collaborative
  - Modular
  - Transferable

- **OpenRefine**
  - Quick
  - Easy
  - Intuitive
ESTC: attempt to write a new chapter in book history

- **Long tradition of analytical bibliography** studying material aspects of printing and composition and layout – we continue this work in a responsible, computational way. In a sense, trying to revive quantitative book history.
- **Publisher/printer information** will enable a **new perspective** on early modern intellectual history due to reversing the author centric approach and start studying **intellectual traditions through publishing networks** that had a greater impact on what was printed than has been realized.
- **Combining ESTC and ECCO** will open a new perspective to the use of both resources.
Case of data-driven approach to constructing and examining the English canon (ca. 1500-1800)

- Quantitatively constructed canon of works that were a) published most often, b) most frequently and c) for the longest period of time in Britain and North-America
- Making use of a processed version of the ESTC
- Keys to the analysis: 1) edition field information and 2) information extracted from imprints about publishers and printers
- Analyzing the canon in terms of time, people, places, and materiality.
  - Main interest: epistemological shifts during early modern era.
Total documents, works, and canon items in the ESTC per year (1500-1800)

The publishing timeline for the Full Canon. Canonical works have been sorted by the first publication year. The individual dots indicate the publishing year for the initial publication and all subsequent reprints.
Works that were most frequently printed at least during one decade between 1500 and 1800. The point size indicates the number of reprints for each work (rows) during the given decade (columns).
Posthumous publication of top authors between 1500 and 1800.

The point size indicates the number of publications for each author, including reprints (rows), per year (columns). The color indicates publication before and after death, respectively.
Works by female authors in the data-driven canon per decade.
The most popular subject-topics for the ten most printed works in each decade from 1500 to 1800.
Share of publications by the largest publishers (top-1% percentile)
Fraction of publications by place for the top publication places, excluding London, 1500–1800. Publication landscape becomes more diverse over time.
The share of canonical editions (of all editions).
Heritage of the Printed Book Database (HPB)

The HPB Database (previously called the Hand Press Book Database) is a steadily growing collection of files of catalogue records from major European and North American research libraries covering items of European printing of the hand-press period (c.1455-c.1830) integrated into one file. This makes it possible for information to be retrieved in one single search across all files. As the digitisation of collections in contributing libraries progresses, more and more catalogue records point to digital presentations of the early printed books.
HPBD documents in our data
- 6,004,893 initial docs
- 2,680,627 processed

Initially harmonized
Titles, editions
Authors / Genders
Publishers
Time & geography
Physical dimensions
Genre & language
Specific challenges (HPB)

**Scalability of data analysis:**

10x more data (ESTC ~0.5M -> HPBD ~6M)

**Harmonization within and across catalogues:**

47 catalogues & varying languages and notation conventions

**Interpretation of historical representativeness, reliability, and relevance**

Complex temporal & geographical publishing landscape
Variation in standard doc sizes across time and space

Data availability (HPB):
- Gatherings: 22.5%
- Height: 11.6%
- Width: 1.1%
HPB: current status & next steps

Done:
- Workflows implemented
- Reproducible summaries
- Team well prepared

Todo:
- Customize
- Speed up
- Research cases

https://github.com/COMHIS/cerl
Thank you!

“To talk about what one is doing can sometimes help one to proceed; but there are other times when it seems best to get on with the work and to define the work by doing it.” G. Thomas Tanselle in *Bibliography and Science*.

extras
Helsinki Computational History Group & integrated data-driven approach to history

- “Computational history” refers to a mixed methods approach to study large digitized historical sources.
- “Integrated” means that data science is combined to specialized subject knowledge; in the case of COMHIS, intellectual history and book history.

http://helsinki.fi/computational-history
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

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
Data - what is the English Short Title Catalogue (ESTC)?

- Bibliographic database
- Chronologically, its scope extends from the earliest printed work in British Isles (ca. 1473) through the last item printed in 1800
- Geographically:
  - British Isles
  - North America
  - British governed territories
  - Items printed in English, any part of the world
- Held by over 2000 institutions in North America, the United Kingdom, Europe, Australia and New Zealand
- 483,331 documents
Author gender distribution over time. Note that the name-gender mappings change over time and geography but this has not been taken into account here.
Publication places

- 31939 unique publication places; available for 2372974 documents (89%).
- 0 ambiguous publication places; some of these can be possibly resolved by checking that the synonyme list does not contain multiple versions of the final name (case sensitive).
- 30527 unknown place names These terms do not map to any known place on the synonyme list; either because they require further cleaning or have not yet been encountered in the analyses. Terms that are clearly not place names can be added to stopwords; borderline cases that are not accepted as place names can be added as NA on the synonyme list.
- 5228 discarded place names These terms are potential place names but with a closer check have been explicitly rejected on the synonyme list
- Conversions from the original to the accepted place names
- Unit tests for place names are automatically checked during package build
ESTC Digital

Online version:
- Entire catalogue is browsable at http://estc.bl.ac.uk

Offline version:
- Not publicly available
- One (non-standard) XML file
- 1,629,574,849 bytes
- 1,450,034 lines
- Data entries are discrete/not harmonized
Paper consumption according to book formats in Kungliga and Fennica
Gender

- **Author-gender mappings** in the final data

- 72797 unique male authors

- 1839 unique female authors

- 400414 documents (14.9%) with a male author

- 6776 documents (0.3%) with a female author

- 2157312 documents (80.5%) with **unresolved gender** (including pseudonyms)

- First names identified as **female** in the preprocessed data (including pseudonyms)

- First names identified as **male** in the preprocessed data (including pseudonyms)

- First names with ambiguous gender (both male and female listed in the gender mapping tables) in the preprocessed data (including pseudonyms). To override and resolve ambiguous mappings, gender info can be added to the custom name-gender mappings or the custom author information table

- First names with unknown gender (no gender mapping info available) in the preprocessed data (including pseudonyms). The missing info can be added to the custom name-gender mappings or the custom author information table
Figure 1. Annual relative print area for common book formats.
Jacob Tonson’s (the younger) edition of *King Lear* in 1734. R. Walker is a competing publisher.
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
Interlinked data harmonization

Editions (Ali Ijaz)
Authors (Mark Hill)
Publishers (Ville Vaara)
Places (Iiro Tiihonen)

Data → Standardization → Validation → Integration → Analysis
The Bard, the Bible and Book Formats
Smaller Books and Reading Habits in Early Modern Europe

Introduction
The eighteenth century entailed a change in printing, reading and writing books. Book sizes became smaller and the public gradually switched from reading a few key works (such as the Bible) to an ever-expanding amount of literature. An anonymous observation from Paris in the 1770s (cited from Reinhart Wirmann) concluded that: “Everyone, but women in particular, is carrying a book around in their pockets. People read while riding in carriages or taking walks; they read at the theatre during the interval, in cafés, even when bathing.”

Research questions
- How popular were large and small book formats?
- For which books, when and where did book formats change?

Historical records suggest that the size of books made a difference in how books were read and distributed:
- Small books could be easily transported, carried in a pocket to places where individuals could read in solitude.
- Large books were appropriate for reading out loud to an audience and for marking passages.

...but the share of large books declined in the eighteenth century...
A statistical analysis of changes in book formats show the increasing popularity of small formats in Europe towards the end of the eighteenth century. The development was uneven, however, and varied according to location.

- The Swedish case (SNB) shows a rise in the production of smaller books in the second half of the eighteenth century.
- In the British case (ESTC), a similar trend occurs earlier, but there is also an increase in the ashendrimen format, indicating an overall shift towards smaller books.
- The same trend is repeated in HBP in the whole of Europe. The trend is clearer for German than Spanish cities.

Shakespeare was made big by small books...
While contemporarily published, Shakespeare’s works were printed less frequently in the seventeenth century. In terms of printed books, Shakespeare’s canonization happened in the eighteenth century through smaller book formats. The year 1736 was a turning point; a comparison between the publishers Jacob Tonson the Younger and Robert Walker printing many editions.

...and the Bible gravitated to small formats!
While the Bible was read out loud, and thus suitable to be printed in larger book formats, it was also published in smaller sizes.
- Larger books dominated Bible printing until the mid-eighteenth century when smaller and smaller book sizes overtook.
- Printing and reading the Bible changed especially in the German-speaking part of Europe, and through the hands of two publishinghouses located in Halle.

Materials and methods
These include:
- Finnish National Bibliography (SNB)
- Swedish National Bibliographic (SNB)
- English Short Title Catalogue (ESTC)
- Hessian of the Printed Book database (HBPG), which is a compilation of 488 bibliographic sources, primarily 18th-century books, pamphlets, and is more uneven than the others.

The bibliography:
- Grass 3.68 million entries from the incunabula period.
- Precise, good coverage of the publication records.
- Include: books, libri, pamphlets, newspapers, journals, publishers, languages, publication places, publication year, book formats and other features of printed documents.

Methods:
- Automatic harmonization of selected sociology fields.
- Custom data science workflows in R and Python.

Size matters, or at least the authors thought so...
From the early eighteenth century book formats already carried cultural connotations with regard to prestige and status. In a satirical text Joseph Addison’s The Bespectacled (1712) described authors of books in different formats (and sizes):
- “I have observed that the Author of a Folio, in all Companies and Concoctions, sets himself above the Author of a Quarto above the Author of an Octavo; and so on, by a gradual Decent and Subordination, to an Author in Twenty Folios. This Distinction is so well observed, that in an Assembly of the Learned, I have seen a Folio Writer place himself in an Elbow-Chair, when the Author of a Duo-decimo has, out of a just Deference to his Superior Quality, seated himself upon a Squab. In a word, Authors are usually ranked in Company after the same manner as their Works are upon a Shelf.”

Poster board: 54
Paper ID: 596
Example of bias that is particular to data:
The 5-year theory with respect to ESTC catalogue
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
The rise of Octavo: paper consumption

A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828
Mikko Tolonen, Leo Lehti, Hege Roiha, & Jani Marjanen
Title count share for books in Latin (primary language)
Share of large books declined in the eighteenth century

**Fig. 1:** Annual relative print area for common book formats.
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)
Size matters, or at least the authors thought so...

From the early eighteenth century book formats already carried cultural connotations with regard to prestige and status. In a satirical text Joseph Addison (The Spectator, 6 November 1712) described authors of books in different formats (and sizes):

“I have observed that the Author of a Folio, in all Companies and Conversations, sets himself above the Author of a Quarto; the Author of a Quarto above the Author of an Octavo; and so on, by a gradual Descent and Subordination, to an Author in Twenty Fours. This Distinction is so well observed, that in an Assembly of the Learned, I have seen a Folio Writer place himself in an Elbow-Chair, when the Author of a Duo-decimo has, out of a just Deference to his superior Quality, seated himself upon a Squabb. In a word, Authors are usually ranged in Company after the same manner as their Works are upon a Shelf.”
ESTC as historical network data

- The current outputs result in a network of 72,066 nodes connected by 328,996 edges.
- Split into overlapping subsets which only contain actors active or living during a given period.

- **Good**
  - Actual historical records, not curated data.
  - Geographically centralized in London.
  - Hand-operated printing press required individual actors and relationships.

- **Bad**
  - Not historically comprehensive: records “obviously include only surviving and recorded publications, and also fail to identify publications in which the bookseller had shares (if indeed given on the original imprint)...” (Raven 2007: 406-407).
William Shakespeare’s posthumous ego networks

Ego networks weighted by number of connections.

Red nodes are historically noted publishers.
Publishing and reprint patterns by publisher role in the printing sequence.
Publisher subject topic specialization and canon share.
Actors (authors and booktrade)
Authors and Actor Fields (100, 110, 700, 710)

- 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.
AN ANSWER
TO THE RIGHT HON. EDMUND BURKE'S
REFLECTIONS ON THE REVOLUTION IN FRANCE,
WITH SOME
REMARKS ON THE PRESENT STATE,
of the
IRISH CONSTITUTION.

BY AN IRISHMAN.

When I see that a generous Nature has been suffer'd to take her own Way to Perfection,—When I consider how profuse this has been to Us, I feel all the Pride of Power sunk, and all the Vanity in the Wield of Human Contrivance, melt and die away within me.—My Rigour relents.—I pardon Something to the Spirit of Freedom.

DUBLIN:
PRINTED FOR JAMES MOORE, NO. 45, COLLEGE-GREEN.
M DC XXI.
End results: tables of distinct booktrade actors & links between printed objects

<table>
<thead>
<tr>
<th>actor_id</th>
<th>is_organization</th>
<th>name_unified</th>
<th>name_variant</th>
<th>year_birth</th>
<th>year_death</th>
<th>year_pub_first_estc</th>
<th>year_pub_last_estc</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://bbtibodleianox.ac.uk/details/?traderid=55412&amp;printer_friendly=true">http://bbtibodleianox.ac.uk/details/?traderid=55412&amp;printer_friendly=true</a></td>
<td>False</td>
<td>POTTS, J</td>
<td>J. Potts; J. Potts Jun; J. POTT</td>
<td>1791</td>
<td>1791</td>
<td>1781</td>
<td>1796</td>
</tr>
<tr>
<td><a href="http://bbtibodleianox.ac.uk/details/?traderid=52779&amp;printer_friendly=true">http://bbtibodleianox.ac.uk/details/?traderid=52779&amp;printer_friendly=true</a></td>
<td>False</td>
<td>PARTRIDG E, J</td>
<td>J. Partridge</td>
<td>1798</td>
<td>1798</td>
<td>1798</td>
<td>1798</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>curves</th>
<th>actor_id</th>
<th>source_tags</th>
<th>actor_name_p</th>
<th>actor_roles</th>
<th>actor_addresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CU-Riv ES)T113 973</td>
<td>306130621</td>
<td>100; 260</td>
<td>Gadesby, Richard.</td>
<td>author; publisher</td>
<td></td>
</tr>
<tr>
<td>(CU-Riv ES)T113 973</td>
<td>229360957</td>
<td>260</td>
<td>S. Bladon</td>
<td>bookseller</td>
<td>{@} {i:NA} No 13, Paternoster Row</td>
</tr>
</tbody>
</table>
Researching publishing in the 18th century

- Previous attempts at quantifying publishing have been painstakingly done by hand.
  - -> take one limited geographic region and time, and start counting
  - -> labour intensive, methods difficult to reapply to create comparative results

- Catalogues have been utilized
  - but only really in their raw format

---

Raven: Bookscape (2014)
publisher harmonization workflow
Step 1: NER

- Named Entity Recognition (Stanford Parser)
- Machine learning based method
  - Needs teaching for the language processing algorithm material (similar to earlier exercise)
- Testing, evaluation, retraining, ...

by Christopher Barkar dwelling in Powles Churchyard at the signe of the Tygres head,

- **PER** --- Christopher Barkar
- **LOC** --- Powles Churchyard
- **LOC** --- signe of the Tygres head
Step 2: Name variant unification

- Correct and enrich names
  - Iohn becomes John; VWoodcocke becomes Woodcock

- Using town, address, matching initials and name, name combinations, years of activity, etc, harmonized and expand on existing named entities.
  - I.e., initials to full names: I. Newbury becomes John Newbery
Step 3: Pairing and grouping

- External databases (BBTI, VIAF)
- Name variations (fuzzy name matching)
  - eg.: J. Walley; John Walley; Iohn Walley; Iohn VValley; Jhon Walley
- Internal duplicates in ESTC
  - Same actor in multiple fields
Step 4: Finished data, evaluation, manual correction

- Automated step finished
- Still not perfect though, so the important data points need to be hand checked and corrected
- This kind of stuff never perfect, so a threshold of accuracy needs to be decided on
Validation: page count (ECCO vs. ESTC)

Clean up messy entries

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

### [1] 142
Timeline of Shakespeare’s publications included in canon. The point size indicates the share of the publisher with most prints of the indicated work (rows) per decade (columns).
Actor numbers and role counts

- Unique actors (all): 144,399
- Links: 1,107,777

*primary role == role most commonly associated with the actor. "unknown" only if no other roles found.
Authors

- 369109 **unique authors** These final names capture all name variants from the custom author synonyme table, and exclude known pseudonyms (see below). If multiple names for the same author are still observed on this list, they should be added on the author synonyme table.
- 2011002 documents have unambiguous author information (75%).
- 1082 **unique pseudonyms** are recognized based on custom pseudonyme lists.
- 484 **discarded author names** This list should not include any real authors (if it does, please send a note to the admin). The stopword lists are considered when discarding names.
- **Author name conversions** Non-trivial conversions from the original raw data to final names.
Language

- 266 unique languages
- 206 unique primary languages
- 2573145 single-language documents (95.99%)
- 107482 multilingual documents (4.01%)
- Conversions from raw to preprocessed language entries
- 226436 documents (8.45%) with unrecognized language

Language codes are from MARC; new custom abbreviations can be added in this table.

Title count per language (including multi-language documents):

![Bar chart showing language distribution](image)

Top languages

Number of documents assigned with each language (top-10). For a complete list, see accepted languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Documents (n)</th>
<th>Fraction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td>723610</td>
<td>27</td>
</tr>
<tr>
<td>German</td>
<td>463161</td>
<td>17.3</td>
</tr>
<tr>
<td>French</td>
<td>454052</td>
<td>16.9</td>
</tr>
<tr>
<td>English</td>
<td>313745</td>
<td>11.7</td>
</tr>
<tr>
<td>Undetermined</td>
<td>225162</td>
<td>8.4</td>
</tr>
<tr>
<td>Dutch</td>
<td>114609</td>
<td>4.3</td>
</tr>
<tr>
<td>Italian</td>
<td>77547</td>
<td>2.9</td>
</tr>
<tr>
<td>Swedish</td>
<td>65572</td>
<td>2.4</td>
</tr>
<tr>
<td>Spanish</td>
<td>57428</td>
<td>2.1</td>
</tr>
<tr>
<td>Latin;German</td>
<td>13446</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Publication places

- 31939 unique publication places; available for 2372974 documents (89%).
- 0 ambiguous publication places; some of these can be possibly resolved by checking that the the synonyme list does not contain multiple versions of the final name (case sensitive).
- 30527 unknown place names These terms do not map to any known place on the synonyme list; either because they require further cleaning or have not yet been encountered in the analyses. Terms that are clearly not place names can be added to stopwords; borderline cases that are not accepted as place names can be added as NA on the synonyme list.
- 5228 discarded place names These terms are potential place names but with a closer check have been explicitly rejected on the synonyme list.
- Conversions from the original to the accepted place names
- Unit tests for place names are automatically checked during package build

Top-20 publication places are shown together with the number of documents.

Publication countries

- 59 unique publication countries; available for 2062853 documents (77%).
- 30950 places with unknown publication country (97% of the unique places; can be added to country mappings)
- 0 potentially ambiguous region-country mappings (these may occur in the data in various synonyms and the country is not always clear when multiple countries have a similar place name; the default country is listed first). NOTE: possible improvements should not be done in this output summary but instead in the country mapping file.

<table>
<thead>
<tr>
<th>Country</th>
<th>Documents (n)</th>
<th>Fraction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>588582</td>
<td>22.0</td>
</tr>
<tr>
<td>France</td>
<td>438723</td>
<td>16.4</td>
</tr>
<tr>
<td>England</td>
<td>311857</td>
<td>11.6</td>
</tr>
<tr>
<td>Netherlands</td>
<td>199860</td>
<td>7.5</td>
</tr>
<tr>
<td>Italy</td>
<td>132050</td>
<td>4.9</td>
</tr>
<tr>
<td>Sweden</td>
<td>66048</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Why does this metadata harmonization matter?
Combining harmonized metadata information in research

- “Data-Driven Canon” looking at top c. 1000 works published in 1470-1800 based on longevity and publishing frequency
- Crucial to have the actor information & workfield information so we are also able to study publishing
- We have also implemented a hand-curated genre identification to the DD canon based on Dewey classification
- Our Dewey classification has been further extrapolated based on matches in the existing subject_topics information in ESTC
Publishers

- 792622 unique publishers

- 2260169 documents have unambiguous publisher information (84.3%). This includes documents identified as self-published; the author name is used as the publisher in those cases (if known).

- 177 documents are identified as self-published (0%).

- Discarded publisher entries

- Conversions from original to final names (only non-trivial conversions shown)

The 20 most common publishers are shown with the number of documents.

![Top publishers chart]

- Müller
- Luctus
- Mullerus
- Meyor
- Dieterich
- Killus
- kungl tryckeriet
- [Kongl. tryckeriet]
- Langenheim
- Gotta
- Handel
- Endter
- Henckellus
- Reusner
- Nilsus
- Richter
- [G. W. Lonticer]
## Default sheet sizes

<table>
<thead>
<tr>
<th>format</th>
<th>gatherings</th>
<th>width</th>
<th>height</th>
<th>area</th>
</tr>
</thead>
<tbody>
<tr>
<td>sheet</td>
<td>1to</td>
<td>60</td>
<td>90</td>
<td>5760</td>
</tr>
<tr>
<td>broadside</td>
<td>bs</td>
<td>60</td>
<td>64</td>
<td>3840</td>
</tr>
<tr>
<td>folio-large</td>
<td>2long</td>
<td>30</td>
<td>53</td>
<td>1749</td>
</tr>
<tr>
<td>folio</td>
<td>2to</td>
<td>30</td>
<td>45</td>
<td>1350</td>
</tr>
<tr>
<td>folio-small</td>
<td>2small</td>
<td>25</td>
<td>38</td>
<td>950</td>
</tr>
<tr>
<td>quarto-long</td>
<td>4long</td>
<td>27</td>
<td>35</td>
<td>945</td>
</tr>
<tr>
<td>quarto</td>
<td>4to</td>
<td>22</td>
<td>28</td>
<td>616</td>
</tr>
</tbody>
</table>
Page counts

- Page count available for 2665032 documents in total (99.4%), including both readily available and estimated page counts.

- Page count readily available for 2161345 documents (80.6%).

- Page count estimated for 503687 documents (18.8%).

- Page count missing and could not be estimated for 15578 documents (0.6%).

- Page count updated for 281559 documents in the validation phase.

- Conversions from raw data to final page count estimates

  - Augmented pagecounts For these cases the page count is missing (or discarded) in the original data, and estimated based on median page counts for single volume, multi-volume and issues, calculated from those documents where page count info was available.

  - Automated unit tests for page count conversions - these are used to control that the page count conversions remain correct when changes are made to the cleanup routines
Titles

- 2223108 unique titles
- 2680279 documents (100%) have a title
- Discarded titles

Top-20 titles are shown together with the number of documents.

![Bar chart showing top 20 titles]

- Publicatio
- Opera
- Poems
- Gedichte
- Oeuvres
- Resoluto
- Edito
- Biblia
- Oeuvres complètes
- Arrest du Conseil d'État du roy
- Opera omnia
- Arrest du Conseil d'État du roi
- Programma funebre
- Die Bibel, oder die ganze Heilige Schrift des alten und neuen Testaments
- Notifications
- By the King
- Theses juridicae inaugurales
- Poems on several occasions
- Goethe's Werke

Documents
Data sources

The name-gender mappings were collected from the following sources using this script:

- U.S. Social Security Administration baby name data as implemented in the babynames and gender R packages. For each year from 1880 to 2013, the number of children of each sex given each name. All names with more than 5 uses are given.
- The U.S. Census data in the Integrated Public Use Microdata Series as implemented in the genderdata R package.
- The Kantrowitz corpus of male and female names as implemented in the genderdata R package.
- The genderdata R package mappings for Canada, UK, Germany, Iceland, Norway, and Sweden.
- Multilingual database (Prenoms.txt)
- French first names
- German first names
- Pseudonymes provided by the authors of the bibliographica R package.
- Custom name-gender mappings constructed manually by the authors of this R package.
- Custom author information constructed manually by the authors of this R package.

The name-gender mappings from different years and regions are combined. When the sources give conflicting gender mappings, the gender is marked to be ambiguous. Afterwards, our custom name-gender mappings and custom author information tables are used to augment this information. The genderizeR R package could also be useful but the genderizer.io API has a limit of 1000 queries a day, hence omitted for now.