Analytical determination of editions from bibliographic metadata

Ali Zeeshan Ijaz (University of Turku), Mikko Tolonen (University of Turku), Leo Lahti (University of Helsinki) and Iiro Tiihonen (University of Turku)

Abstract

Analytical bibliography aims to understand the production of books. Systematic methods can be used to determine an overall view of the publication history. In this paper, we present the state of the art analytical approach towards the determination of editions using the ESTC meta data. The preliminary results illustrate that metadata cleanup and analysis can provide opportunities for edition determination. This would significantly help projects aiming to do large scale text mining.

1 Introduction

Analytical bibliography studies books as material objects and aims to understand how they were produced (Tanselle, 1977). Large scale library catalogues, where millions of documents have been cataloged, can be utilized to develop an automated framework that can highlight book publication and production (Tolonen et al., 2015; Lahti et al., 2019). Systematic methods can provide a more thorough view of the publishing history (Eliot and Rose, 2009;Tolonen et al. 2018). At the early times of book printing, publishers could reprint the same book several times based on the public reception to the literary work. Additionally, the expiration of the Licensing Act of 1695 and the resultant increase in book production meant that some publishers resorted to duplicating their own issues. The term "edition" was then used for successive issues as means to keep on selling the same work (Todd, 1951). Furthermore, by the 18th century, the octavo format became the most popular for books, paving the way for cheaper production (Lahti et al., 2019). For these reasons, the number of editions and their ordering are integral to understanding book production (Howsam, 2014).

Contemporary text mining approaches generally ignore edition level information, or provide generic solutions that may omit important details. Projects such as "Commonplace Cultures" (Morrissey, 2016) have performed large-scale text mining of the Eighteenth Century Collections Online (ECCO), a collection of books printed in the United Kingdom during the 18th century. "BookSampo", a semantic portal which uses the FRBRoo ontology (Riva et al., 2008), covers metadata on Finnish fiction literature, though only at the work level (Mäkelä et al., 2011). Furthermore, standard algorithms used in this field, such as latent Dirichlet allocation are time agnostic (Blei et al., 2003), and while later algorithms have become time aware, they only focus on topics. Hence, such methods may fall short in contextualizing historical developments in book printing and publishing.

2 Background

According to Joseph Dane, printed books are not individual objects but are in fact, instances of an historical process that produces similar members of a defined collection (Dane, 2016). Hence contextualizing their production chronologically and in relation to their close relatives is more important than just viewing each book in isolation.

Considering how fundamentally important editions are for this research work, it is essential to determine what is meant by an edition of a book and the concepts related to it. Fredson Bowers defined editions as the combined number of books printed from the same type-pages and therefore includes both the issues and variants. In this manner, issues were a form of editions put on sale by the publisher at planned times, while variants represented alterations in a book but with no change in the title page (Bowers, 1994).

Libraries have begun moving from the traditional print oriented model to having digitization projects being conducted under them. For many digital humanities projects, the library's source material serves as the groundwork on which research is conducted (Cunningham, 2010). This is especially true for catalogs and bibliographic metadata that various libraries hold, which become essential tools in information science and provide avenues for significant research work (Lahti et al., 2019).

Historically, the hand press era, circa 1450 to 1825, had a starkly different means of production of books as compared to the more recent publishing eras. Technologically, this time period saw little change in the method of book production. However, the more specific requirements of cataloging books from this era necessitated a modification of cataloging rules developed mainly for modern publication of books. This required much more sophisticated record keeping, where identification of an edition is no small task. Nonetheless, several new editions were discovered when more comprehensive bibliographic metadata were utilized (Snyder and Hutchinson, 2014).

Analysis of large textual datasets such as ECCO conducted under projects such as Commonplace cultures project tends to be incomplete as only the earliest edition information is utilized. This significantly hampers the ability to properly contextualize and analyze book production, as the patterns in the historical production of printed books is as important as the written text itself. Furthermore, this resulted in a significant reduction in usable data, as later editions were discarded. Coverage wise, ECCO consists of around 200,000 volumes of texts from 18th century, while bibliographic metadata such as ESTC provides information for around 480,000 titles, from 14th to 18th century.

Hence bibliographic metadata provides a great opportunity for pattern discovery, the ability to determine editions and works, and exploratory analysis to complement earlier studies in the history of knowledge production. Unfortunately, metadata cannot be readily used for research due to remarkable shortcomings in the raw data quality, and has to be first properly harmonized. The quality and overall value of metadata can be significantly improved by transforming it into a more standardized format. Bibliographic data science is a field that aims to produce high quality, consistent and improved metadata for further analysis (Lahti et al., 2019). Given that edition information is crucial for determining patterns in historical book publication and production, we have developed an automated pipeline that can process and clean up the raw metadata into a more usable format in order to determine edition level information for various works of authors. This provides the foundation for exploratory analysis and elucidating significant patterns in book production.

3 Material and Methods

The English Short Title Catalogue (ESTC) provides a wealth of knowledge concerning the books published in the early modern period. Nonetheless, it uses the Machine Readable Cataloging standard (MARC, 1999), where the raw data is unsuitable for research, due to the presence of spurious and erroneous information, as well as differences in standards and languages (Nilsson, 2010; Lahti et al., 2019). The ESTC contains metadata for more than 460,000 documents, covering the hand press era (1470-1800). It contains various fields that can be used to determine each book uniquely, while providing rich details on the editions.

For this work in progress, our aim has been to develop a pipeline that can be used to determine the chronological ordering of editions of various books in the ESTC metadata. Before such ordering is determined, it is imperative to describe the “work” under which these editions of books are found. Considering the difficulty in determining the definition of work that is widely applicable, we opted for a collection oriented approach. Here, books of similar titles are collected in a collection, henceforth known as the work field. Discriminating information such as publication date, edition number (if available), publisher details, and more can be used to determine more precise edition information, as well as to further subset the collection as per requirement of the end users. The harmonization process begins with selecting the edition field, and other supporting metadata fields that are needed during the harmonization. The “edition statement”, for instance, provides the edition number as well as fields related to book title, publisher information, date of publication and more. Additionally, this work is a part of a larger project, which includes several collaborators who use the harmonized ESTC metadata for various research purposes (Lahti et al., 2019; Tolonen et al., 2018).

The raw data is harmonized in an iterative and progressive manner, with various processing steps performed to handle certain aspects of the harmonization process. This is done on a per author basis to ensure consistency and minimize any errors, by linking titles to their respective authors. Harmonization is then performed for each author individually. Figure 1 illustrates the flow diagram for the harmonization process, while the step by step details are described below.

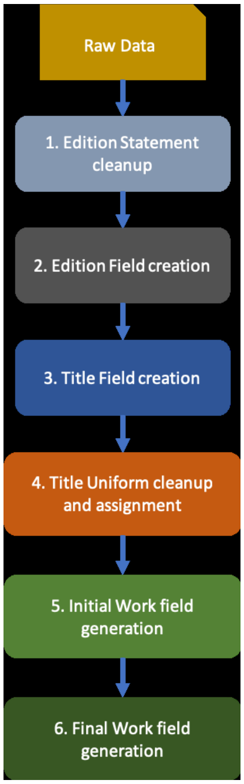
Raw Data
1. Edition Statement cleanup
2. Edition Field creation
3. Title Field creation
4. Title Uniform cleanup and assignment
5. Initial Work field generation
6. Final Work field generation

Fig. 1. Flow diagram for the harmonization process.

The process begins with the "edition statement" field[1], which is processed to determine the edition number of the book. Currently, only a small subset of the whole of ESTC metadata contains information in this field. Hence the publication date is combined with the edition number, if available, to create a new edition field. This provided the ability to distinguish various editions on chronological basis.

The book titles[2] are cleaned up, removing any unwanted characters and non-important words. Furthermore, the "title uniform" field[3], which provides a representative title for the work undergoes a similar cleanup as well. An initial work

[1] MARC Field 250a

[2] MARC Fields 245a and 245b

[3] MARC Field 240a

field is then generated by combining the author identifier with the harmonized title. This provided a unique work field identifier for each book. A final work field is then created by combining similar titles into a collection. We have created custom algorithms in order to combine similar titles, by assigning a represetative title of the collection, alongside the author identifier. Source code will be accessible via the COMHIS homepage.

At the moment, the algorithms employed in the harmonization process as well as the creation of collections are being worked upon to improve their applicability across different genres in ESTC metadata. Considering the scale of ESTC, during the initial development and tuning of the pipeline, a small subset of works of 7 popular authors was selected. The list of authors included William Shakespeare, David Hume, Jonathan Swift, John Locke, Isaac Watts, Alexander Pope and Daniel Defoe. The harmonized entries were manually checked to determine if the titles were being properly assigned the correct work field. Corrective steps were then taken to improve the performance of the harmonization process, by tuning the various methods and parameters of the underlying algorithms.

However, the performance and reliability of the newly designed harmonization and analysis algorithms need to be validated against a known ground truth. Hence, such gold standard was created as follows:

Constructing the Gold Standard

- 250 authors were randomly selected from ESTC, with varying title counts. Additionally, records from seven prolific authors were also selected.
- All records with the same content were carefully inspected and assigned to works.
- The main fields we used for this assignment were Title Uniform (MARC field 240a), Title statement and Remainder of title (245a and 245b). We also consulted the MARC fields for page count, physical extent, edition statement, general note and genre.
- We encountered several issues including spelling mistakes or word replacements in the titles, which we combined into a single work.
- Another issue was the handling of different or multiple volumes of the same work. Although a work could exist in the ESTC as a multi-volume book, each of the volumes could also be listed separately. We resolved this

issue by creating an additional layer, where all the different volumes were combined into a single work.

- A layer was added for handling calendars and music performance handouts.
- A collection oriented layer was added, which described if the book was a collection of other works. In most cases these contain reprints of earlier works; separating them from other material reduces redundancy.
- We also made a crude genre classification for each record. Altogether, 25.3% of the records have an annotated genre.
- Using formally structured documents, such as meeting minutes, dictionaries or court case reports in creating word embeddings could potentially skew the outcomes.

4 Results

The gold standard would serve as the main evaluation method for the harmonization process. Currently, the system performs well on collecting various titles into a single generic collection for some authors. We evaluated the harmonization process for three different authors in the gold standard in terms of precision and recall against the manually constructed ground truth. The results are illustrated in Figure 2. Additionally, the number of titles for each author are listed in Table 1. We define true positives as those titles that were assigned to the correct work field, false positives as those titles that were incorrectly assigned to a work field and lastly, false negatives as those titles who were not assigned to the correct work field and instead had their own work field generated.

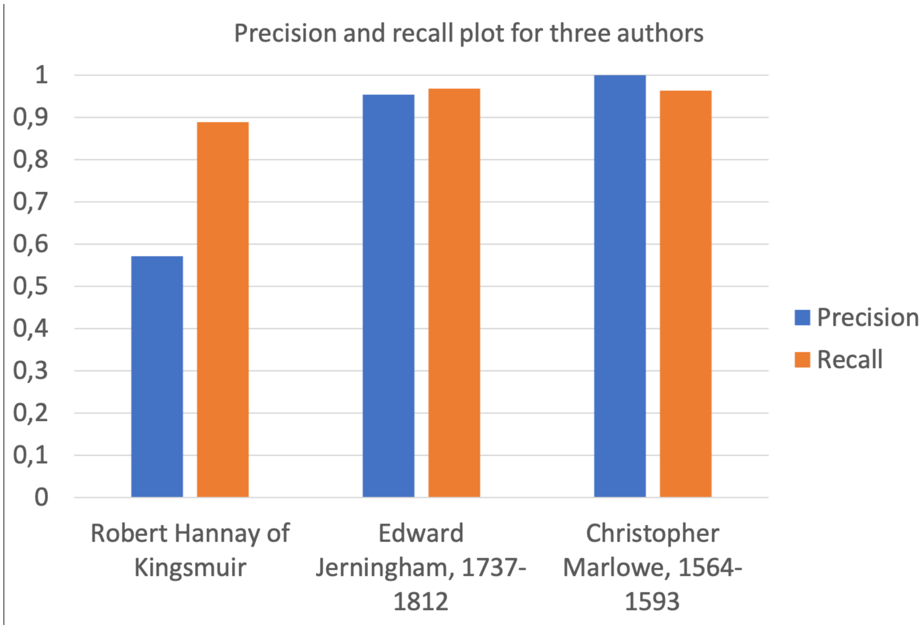

Fig. 2. Precision and recall for three authors.

Table 1. Number of titles for three authors.

Author	Number of Titles
Robert Hannay of Kingsmuir	15
Edward Jerningham, 1737-1812	66
Christopher Marlowe, 1564-1593	55

The results illustrate that the system is capable of assigning titles to the correct work fields in most of the cases, with an overall high recall. For the authors Edward Jerningham and Christopher Marlowe, the precision was at 0.95 and 1 respectively. However, it suffered in the case of Robert Hannay of Kingsmuir, where it decreased to 0.57. Recall was 0.89 for Robert Hannay of Kingsmuir, 0.97 for Edward Jerningham and 0.96 for Christopher Marlowe. In general, titles that differ by a few words increase the chances of false positives. While on the other hand, longer titles tend to reduce this possibility.

In less polished data, both precision and recall would greatly suffer as there are no direct means of assigning each title to a correct work field, unless the titles are exactly the same. This is confounded by the fact that the differences in title length, variations in the title themselves, spelling mistakes and more, complicates the issue further. Hence, developing a work field without such harmonization techniques entails manually assigning each title to a work, a significantly labor-intensive task.

The next steps would include expanding this proof-of-concept study from the few selected authors to the complete ESTC collection.

5 Conclusions

Considering the scope of the ESTC, spelling variations and different styles of writing, we opted for a collection-oriented approach towards the work field. As the work is currently in development, the first version of the pipeline was developed using the dataset of popular authors. Nonetheless, the results illustrate the applicability of this work towards the research goal of determining a correct ordering for editions.

There are various issues that still need to be addressed. The collection work field is only meant to ensure that similar titles are collected into the same collection. This ignores more detailed information such as volumes, or titles that may be belonging to different works, depending on how work is defined. This will necessitate finer grained work field to be generated from the current collection-oriented work field for downstream analysis. Considering that discriminating information is available for each book title, the collection-oriented work field provides a suitable starting point for such analysis.

Future steps will include improving the harmonization process for other types of textual artefacts, such as pamphlets and other document types. Furthermore, generation of sub-work fields, so as to develop more specific datasets required for downstream statistical analysis would also be done. Finally, statistical analysis on the harmonized dataset can be used to determine edition ordering. This would enable determination of editions in a more precise manner for each work in a chronological fashion.

As the development on the harmonization process moves forward, the overarching aim has been to provide an automatic and reproducible system for harmonization and analysis of the ESTC metadata. This research work enhances the overall analysis via investigating the harmonization and cleanup of the edition field.

Large scale text mining projects would significantly benefit from this, as combining the harmonized metadata with text mining of large data sets such as ECCO would lead to a finer description of what was the first edition. Furthermore, changes between different editions can also be explored. Overall, a more descriptive analysis can then be performed.

Researchers interested in charting the book publication over time of certain authors in the early modern era would benefit from this research work, as well as those interested in exploratory analysis concerning historical book production in general. Libraries would benefit too, as this would grant them the ability to perform better information retrieval and provide contextual information on the books present in their catalog.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3, pp.993-1022.

Bowers, F. (1994). Principles of bibliographical description. St. Paul's Bibliographies. Oak Knoll Press.

Cunningham, L. (2010). The librarian as digital humanist: the collaborative role of the research library in digital humanities projects. Faculty of Information Quarterly, 2(1).

Dane, J. A. (2016). Abstractions of evidence in the study of manuscripts and early printed books. Routledge.

Eliot, S. and Rose, J. (eds.) (2009). A Companion to the History of the Book (Vol. 98). John Wiley & Sons.

ESTC. English Short Title Catalogue. http://estc.bl.uk/ (Accessed 27 November 2018).

Howsam, L. (ed.) (2014). The Cambridge companion to the history of the book. Cambridge University Press.

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly, 1-19.

Mäkelä, E., Hypén, K. and Hyvönen, E., (2011), October. BookSampo—lessons learned in creating a semantic portal for fiction literature. In International Semantic Web Conference (pp. 173–188).Springer, Berlin, Heidelberg.

MARC. (1999). MARC 21 Format for Bibliographic Metadata. https://www.loc.gov/marc/bibliographic/ (Accessed 27 November 2018).

Morrissey, R. (2016). Commonplace Cultures: Mining Shared Passages in the 18th Century using Sequence Alignment and Visual Analytics. Retrieved from https://hcommons.org/deposits/item/hc:12365/

Nilsson, M. (2010). 'From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization', Doctoral Thesis, KTH School of Computer Science and Communication. https://www.divaportal.org/smash/get/diva2:369527/FULLTEXT02.pdf (Accessed 27 November 2018).

Riva, P., Doerr, M. and Zumer, M., (2008), August. FRBRoo: enabling a common view of information from memory institutions. In World Library and Information Congress: 74th IFLA General Confrence and Council.

Snyder, H. L., & Hutchinson, H. L. (2014). Cataloging of the hand press: a comparative and analytical study of cataloging rules and formats employed in Europe (Vol. 1). Walter de Gruyter GmbH & Co KG.

Tanselle, G. T. (1977). 'Descriptive Bibliography and Library Cataloguing', Studies in Bibliography, 30: 1-56.

Todd, W. (1951). Bibliography and the Editorial Problem in the Eighteenth Century. Studies in Bibliography, 4, 41-55. Retrieved from http://www.jstor.org/stable/40371090

Tolonen, M., Lahti, L. and Ilomäki, N. (2015). A quantitative study of history in the English shorttitle catalogue (ESTC), 1470-1800. Liber quarterly.

Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2018). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 1–22.