Analytical edition detection in bibliographic metadata

The aim of analytical bibliography is to understand books and other printed objects as artefacts, and, most importantly, how they were produced (Tanselle, 1977). Systematic methods can unlock patterns otherwise hidden and provide an overall view of the publishing history (Eliot and Rose, 2009). Hence, bibliographic metadata can represent important historical trends (Tolonen et al., 2015) as well as resolve long standing issues, such as the ordering of editions and impressions of books (Hawsam, 2014).

In this paper we present the state of the art analytical approach for determining editions and their ordering. This enhances the applicability of metadata towards bibliographical analysis, and provides a systematic quantitative perspective on early modern publishing. Furthermore, it will be a great aid for projects aiming to do large data set oriented text mining, by providing harmonized data and information on historical developments in book production. Such analysis will be essential for recognizing how books and other textual artifacts have organically evolved over time, and for delivering a broader context for full text mining and interpretation.

State-of-the-art

Contemporary text mining approaches typically ignore edition level information, or provide very generic solutions that omit many details. Related earlier work includes the “Commonplace Cultures” project (Morrissey, 2016), where large-scale text mining of the ECCO dataset was carried out, though with only the earliest edition information present. Other projects include BookSampo, a semantic portal which draws from FRBRoo ontology (Riva et al., 2008). It covers metadata on Finnish fiction literature, however at the work level and does not have complete edition information (Mäkelä et al., 2011). Additionally, commonly used analysis algorithms such as Latent dirichlet allocation are inherently time agnostic (Blei et al., 2003), and although newer approaches such as Topics over Time can include time spans (Wang and McCallum, 2006), they are not applicable to the problem of edition detection, due to their focus on topics. Hence these methods may not be able to contextualize historical developments in book printing and publishing in a chronological fashion. Effectively, these projects are limited in their scope by providing a simplistic and static view into the nature of book production.

Data

We have demonstrated these ideas based on The English Short Title Catalogue (ESTC), which provides a wealth of knowledge with regard to the books, their publication and editions. However, it follows the Machine Readable Cataloging (MARC, 1999) standard, which is unsuitable for research in its raw form. This is a common characteristic of the metadata in general (Nilsson, 2010: 1).

To overcome this limitation, we have developed dedicated and semi-automated harmonization techniques that convert free-form textual information into more coherent and consistent entries that are readily amenable for statistical analysis. This required robust handling of differences in title texts, spellings, and more, hence going beyond simple textual comparisons.

The harmonization approach

Developed in the popular statistical environment R (https://www.r-project.org), our approach begins with cleaning up edition information present in the edition statement field, MARC Field 250, provided by ESTC. Unfortunately, this information is unavailable for the majority of entities and hence the publication date, MARC Field 260, is used to provide a starting point for editions. Next, a new work field is constructed to collect titles into a small collection or work, by using title uniform, which is MARC Field 240, or a cleaned up title of the book.

Considering the variety of titles in the ESTC, the spelling variations, and different styles of writing, we developed an algorithm to handle these issues. It iteratively builds up a title for the collection using the initial work field and collects similar books into the collection. The algorithm performs sub string matching using a variety of methods such as grep, fuzzy or exact matching, on a word by word basis. Alongside manual corrections, the algorithm allows for gaps in the matching and uses a coverage metric to determine if two titles are similar. The benefit of this approach is then realized as similar titles are grouped together despite variations in word spellings, title length and more. At the moment, the algorithm is being worked on to improve its applicability across different genres.

In between automatic and manual

The project has been supported by manual checking and corrections. While we are looking at the whole of ESTC, our priorities are focused on published books. Considering the scale of the ESTC dataset, with the number of documents going beyond 460,000 (Tolonen et al., 2015), a smaller sample of the works of 7 authors were selected, keeping in mind the diversity present in ESTC as well as the popularity of these authors. The list of authors consisted of William Shakespeare, David Hume, Jonathan Swift, John Locke, Isaac Watts, Alexander Pope and Daniel Defoe. The dataset sample was then used for development of the cleanup techniques and algorithms. Finally, the harmonized entries were manually checked to determine the corrections needed for the cleanup.

In our experience, different techniques may be required for different genres. Such as the case with works concerning poetry or religious sermons, which contrast significantly with the works consisting of popular books. The issue is further complicated by the fact that different spellings have been used for the same words, titles have been written in different manner across different genres, and therefore may have no universal clear pattern.

Validation

Validation is an essential requirement to determine whether an algorithm based approach is performing as expected. It provides a description of its performance and its ability to generate correct data. In the context of this work, it is imperative that such algorithmic techniques are supported by a humanities based approach in order to develop a correct view of the underlying data, in order to obtain reliable conclusions as well as to assist more concretely in the development of techniques and algorithms. Therefore, in the interest of reproducibility, a gold standard is developed for the purpose of validating the harmonization process.

We sampled a total of 250 authors, each with 5 to 50 publications with unique titles in the ESTC. Then we manually evaluated these samples to determine those titles, which should be grouped together as the same work based on the non-material content of the entries.

Each of the records are considered as a distinct impression or edition. This way we did not have to bother ourselves with publication years or page counts signifying reprints: the chronological ordering of the entries computationally is straightforward once the correct grouping is known.

Construction of the gold standard for validation

We created several edition layers for the gold standard. The first one was a simple straight-forward layer, with all the works with exactly the same content but with spelling mistakes or obvious occasional word replacements in the title defined as a single work. The second layer combined the first print with new revised editions and those with added content as a singular work. Also we connected the multi-volume works with each volume annotated separately together with the works containing all the volumes in the same entry. On the third layer we regarded the same content with differing time periods as one work. This type of layer was required for including calendars designed and marketed for a specific area as the same work, as well as music performance handouts for different dates.

We also added yet another level signifying that the work is a collection of other works. Seeking out collections allows to research which part of an author's curriculum actually was revered at a given time. Additionally, we made a rudimentary classification of the genre, so that automating exclusion from subset based on the record type would be possible, even if the genre field had not been annotated for the record. For example, formally structured documents, such as proclamations, court case reports, meeting minutes or dictionaries, in subsets designed for word embeddings would skew the outcome.

Open science

While the development is ongoing, our overarching aim has been to provide a reproducible ecosystem for the harmonization and analysis of the ESTC data collection. This project complements the overall analysis by investigating the harmonization of the edition field, and by providing the first harmonized version of the data. A quantitative perspective on early modern publishing would be greatly improved by combining the edition level information with publisher data. Accurate description of the publishing network and the various changes it had undergone in the eighteenth century would then become available. Additionally, combining the edition level information from ESTC with text mining of large datasets such as ECCO would provide a finer description of what was the first edition of a book as well as the subsequent changes between it and the later editions. This would be supplemented by text reuse approaches, enabling a more detailed account of the evolution of the written text during the early modern period and hence can serve as the foundation for more descriptive analysis.

Bibliography

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of*

machine Learning research, *3*(Jan), pp.993-1022.

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.

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.

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.diva-portal.org/smash/get/diva2:369527/FULLTEXT02.pdf (Accessed 27 November 2018).

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

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

Wang, X. and McCallum, A. (2006). August. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 424-433). ACM.

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.

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*.