BABY ElEPHãT - Building an Analytical Bibliography for a Prosopography in Early English Imprint Data

Nushrat J. Khan[1], Terhi Nurmikko-Fuller[2], Kevin R. Page[2]
[1]University of Illinois at Urbana-Champaign
[2]University of Oxford e-Research Centre

Abstract
Prosopography of the people involved with publishing and selling of Early English books can be useful for the investigation of diachronic change in that sector. In this study, we developed an analytical bibliography from the metadata available from 25,000 texts published by the Early English Books Online Text Creation Partnership (EEBO-TCP), focusing exclusively on information captured in the 'Publisher' field of the original TEI-XML Header. From there, we extracted the named entities associated with "printed by", "printed for" and "sold by" relationships to generate and analyze a bibliographical network. We extended the EEBOO ontology to accommodate these relationships and generated RDF from the resulting structured metadata, enriching existing triples capturing other information within the dataset. This work is challenging because of the ambiguity and inconsistency in data and therefore can be of interest for further investigation by those with domain specific knowledge.

Keywords: Historical data; digital humanities; metadata; entity extraction; linked data
doi: 10.9776/16588

Acknowledgements: David M. Weigl for assistance with coding; the Oxford Illinois Digital Libraries Placement Programme (OIDLPP) including the University of Oxford e-Research Centre, Bodleian Libraries, and University of Illinois iSchool, for the academic study visit that supported this work.
Research Data: All the source codes and resulting data are available at - https://github.com/oerc-elephat/BABY-Elephant
Contact: njkhan2@illinois.edu, terhi.nurmikko-fuller@oerc.ox.ac.uk, kevin.page@oerc.ox.ac.uk.

1 Introduction

Detailed analysis of the metadata records for historical texts can reveal previously inaccessible, implicit information. The Early English Books Online Text Creation Partnership (EEBO-TCP) consists of a dataset of some 25,000 texts published between 1473-1700 [4]. A collaboration between the University of Oxford and the University of Michigan, EEBO-TCP started in 1999, and in 2015, the first part of the database was made available in the public domain. For our project, we analyzed metadata extracted from these extensively curated TEI-XML records, focusing on information regarding titles, authors, dates, publication places, and publishers.

In this paper we report on an extension to the earlier ElEPHãT (Early English Print in HathiTrust) [3] project, which created a Linked Semantics Worksets Prototype [9][13] connecting and combining the collections of the EEBO-TCP and the HathiTrust. The BABY ElEPHãT project – a smaller, second-generation project, benefitting from heuristics of earlier data extraction and cleaning, and a reiteration of the ElEPHãT workflow – focused on the largely unexamined information from the publisher field, in particular those entities associated with the terms 'printed by', 'printed for', and 'sold by'. This information was re-published as linked data (as RDF, in the Turtle format), allowing for a new analysis of the prosopography of Early English publishers.

In the following sections we discuss the data preprocessing, extraction, and linked data generation processes, and illustrate how these can be used to analyze data and to further research. We also discuss the direction and implementation of this work in the context of future study.

2 Project Activities

The metadata extracted from 24,925 original TEI-XML files was examined to identify the collectible types of information available, and was found to contain many instances where the cataloger had a degree of uncertainty regarding particular statements (largely due to the publication processes for Early English literature, when naming conventions and publishing industry paradigms differed from those in place today). Many special characters, such as '?', '[ ]', '.', containing specific meaning, appear in the data. We thus began with data preprocessing, proceeding to named entity extraction, followed by ontology selection, and culminating in RDF generation. These stages are discussed in greater detail below.

2.1 Data Preprocessing

We began by removing all superfluous punctuation from the metadata records. We then proceeded to remove 4,199 records, which contained an unknown entity, denoted as 's.n.', in the publisher field. This left us with a final total of 20,726 records.

2.2 Entity Extraction

The most computationally expensive task for the project was the extraction of the named entities based on their relationship within the sentences, particularly those preceded by 'printed by', 'printed for' and 'sold by'. The efficiency of NLTK Entity Extractor [2], ReVerb [10] and Open Calais [8] was tested, and it was found that in case of Early English names and short phrases, NLTK outperformed the other two.

Due to the data structure of NLTK Entity Extractor it was comparatively difficult to locate the entities based on the preceding relationship. We took a lexical approach and initially extracted the names followed by the prepositions 'by' and 'for'. This however resulted in overlapping data for 'printed by', 'sold by' and 'printed and sold by' fields in the extracted named entities. Further de-duplication and revision was thus required.

2.3 De-duplication

For the de-duplication process we separated the named entities immediately following the phrase 'sold by', using regular expressions. We then compared the output of the first extraction of named entities followed the second iteration, which yielded 326 duplicates in total. There were however additional records that contained information for entities with a dual-role of 'printed and sold by'. These records were subset of both 'printed by' and 'sold by' set and there were 142 of such records.

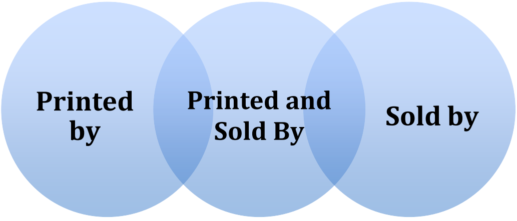


Figure 1. Data types

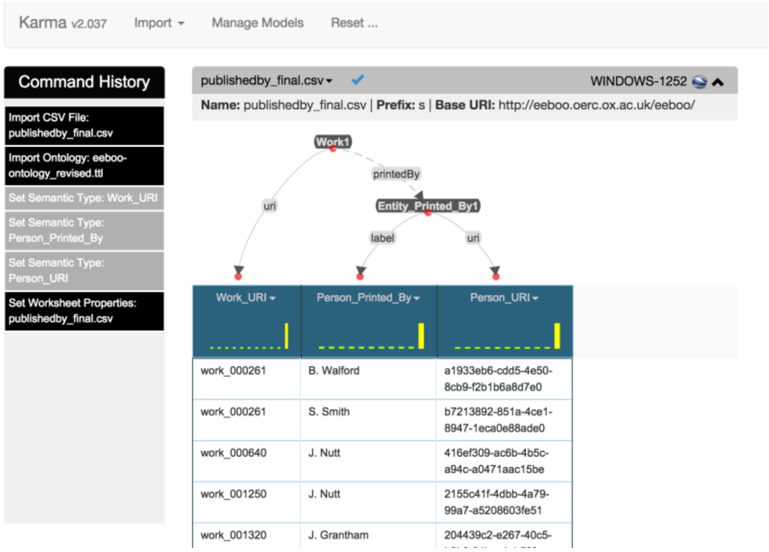
We kept those 142 records in the 'printed by' field and removed the other 184 duplicate records that overlapped with the 'sold by' field. The de-duplication was mainly executed using Microsoft Excel's Advanced Filter.

2.4 Ontology Extension

The next step was to select the appropriate ontology for the mapping of inter-entity relationships. We examined a number of existing ontologies with a bibliographical metadata-focus, such as MODS/MADS RDF [7], Bibframe [1] and FRBRoo [12]. Since the publishing field in the Early English period was different and more complex than it currently is, we were unable to directly map these historical relationships to these modern ontologies. Instead, we extended the EEBOO ontology to include these three relationships, maintaining their unique underlying meanings, and enabling further study in future.

2.5 RDF Generation

To generate RDF from the existing data, we used an open-source data-integration tool developed by the University of Southern California (Web Karma) [5]. We provided the tool with the extracted data in CSV format and the EEBOO ontology, producing RDF for the 'printed by', 'printed for' and 'sold by' relationships. Below is a sample figure of the procedure.



Command History

- Import CSV File: publishedby_final.csv
- Import Ontology: eeboo-ontology_revised.ttl
- Set Semantic Type: Work_URI
- Set Semantic Type: Person_Printed_By
- Set Semantic Type: Person_URI
- Set Worksheet Properties: publishedby_final.csv

publishedby_final.csv WINDOWS-1252

Name: publishedby_final.csv | **Prefix:** s | **Base URI:** http://eeboo.oerc.ox.ac.uk/eeboo/

Work_URI	Person_Printed_By	Person_URI
work_000261	B. Walford	a1933eb6-cdd5-4e50-8cb9-f2b1b6a8d7e0
work_000261	S. Smith	b7213892-851a-4ce1-8947-1eca0e88ade0
work_000640	J. Nutt	416ef309-ac6b-4b5c-a94c-a0471aac15be
work_001250	J. Nutt	2155c41f-4dbb-4a79-99a7-a5208603fe51
work_001320	J. Grantham	204439c2-e267-40c5-

Figure 2. RDF Generation Using Web-Karma

3 Results/Outcome

The final output consisted of 10,083 named entities for ‘printed for’; 9,000 named entities for ‘printed by’; and 1,446 named entities for ‘sold by’ from 24,925 records. Publishing the results as linked data opens it up to potential bridging across other external Web resources, with possibilities for future enriching of the dataset, and making information more accessible.

The linked data can also be searched and analyzed using SPARQL. Our preliminary analysis of the top 20 people for each of the three fields found sellers who worked with top publisher (Henri Hills). The query showed four works by Henri Hills, sold by Will Larner, Jane Underhill and Francis Smith. In a similar way, the relationships that existed amongst early publishers can be analyzed, further contributing to the generation of new knowledge.

4 Conclusions and Further Work

This project explored the possibility of analyzing historical texts with a prosopographical perspective on Early English imprint data by extracting named entities in the Early English publishing field based on their specific roles, and by generating and publishing it as linked data to make it accessible for further reuse by scholars in this area. The extracted information needs to be further examined, overcoming the limitations of the NLTK Entity Extractor – for example, NLTK does not detect initials such as ‘A. B.’ as name, so it needs to be specifically trained for this dataset to avoid such limitations. There are many other future directions in which this project could be developed, ranging from extracting specific place names from the publisher field and linking them with the existing data-stream, to alternative analyses and visualizations of the data. Domain specialists may also use the newly generated RDF to examine diachronic change to name expressions, or to authoritatively identify the authors using sources such as the London Book Trades Index [6] and the British Book Trade Index [11].

5 References

1. Bibliographic Framework Initiative. (n.d.). Retrieved December 17, 2015, from http://www.loc.gov/bibframe/
2. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.
3. Early English Print in the HathiTrust (ElEPHãT). (n.d.). Retrieved September 21, 2015, from http://www.oerc.ox.ac.uk/projects/elephant
4. EEBO-TCP. (n.d.). Retrieved September 21, 2015, from http://www.bodleian.ox.ac.uk/eebotcp/
5. Karma. (n.d.). Retrieved September 21, 2015, from http://usc-isi-i2.github.io/karma/
6. London Book Trades Database. (n.d.). Retrieved September 21, 2015, from http://www.oxbibsoc.org.uk/resources/london-book-trades-database
7. MODS RDF Ontology. (2012, September 12). Retrieved December 17, 2015, from https://www.loc.gov/standards/mods/modsrdf/

8. Open Calais. (n.d.). Retrieved September 21, 2015, from http://new.opencalais.com/
9. Page, K., & Willcox, P. (2015). ElEPHãT: Early English Print in the HathiTrust, a Linked Semantic Worksets Prototype.
10. ReVerb. (n.d.). Retrieved September 21, 2015, from http://reverb.cs.washington.edu/
    1. http://www.bbti.bham.ac.uk/
    2. The CIDOC CRM. (2009, February 23). Retrieved December 17, 2015, from http://www.cidoc-crm.org/frbr_inro.html
11. The British Book Trade Index. (n.d.). Retrieved September 21, 2015, from
12. Workset Creation for Scholarly Analysis. (n.d.). Retrieved September 21, 2015, from http://worksets.htrc.illinois.edu/worksets/