Mapping the STC with MoEML and DEEP

Janelle Auriol Jenstad
jenstad@uvic.ca
University of Victoria, Canada

Tye Landels-Gruenewald
tyelandels@gmail.com
Queen's University, Canada

Joseph Takeda
joey.takeda@gmail.com
University of British Columbia, Canada

Introduction

MoEML's gazetteer of 6500 London place name variants invites the mapping of datasets with a geographical component. As a textual editing project with an interest in print culture, we have long hoped to mobilize our GIS tools and gazetteer data in the service of mapping the English book trade. Our ultimate goal is to publish a layer showing the printing and/or retailing locations of the approximately 25,000 books printed in London between 1475 and 1640. Imprint lines in early modern books include highly granular location data, which has meant that book history has traditionally had an implicit geospatial dimension. A typical imprint line tells us that copies of a folio are "Printed by Elizabeth Purslovv, and are to be sold by Nicholas Bourne, at his shop at the south entrance of the Royall Exchange, 1633." Using the information in such imprint lines, Kathleen Pantzer reorganized the items in the Short Title Catalogue under location headings (Pantzer numbers) in Vol. 3 of the catalogue. Her work facilitates questions about the proximity of one printer or bookseller to another, and thereby questions about affiliations, collaborations, and specialization among a key group of early modern cultural agents.

However, considerable processing of Pantzer's printed lists is required to visualize or map STC items. Thus far, digital databases like *Early English Books Online* (EEBO) and the *English Short Title Catalogue* (ESTC) have captured the imprint line without parsing it into discrete data points, thereby leaving Pantzer's formidable interpretive work behind as we move into the era of digital historical bibliography. *The Database of Early English Playbooks* (DEEP) has included and corrected Pantzer numbers, but only for the printed plays, of course. MoEML has attempted to replicate Pantzer's work via datamining the ESTC. After several unsuccessful NER experiments on ESTC data, we are now mobilizing the curatorial work of DEEP and planning to extend their work beyond playbooks. In this paper, we take imprint lines and geospatial information about the book trade as a case study in mining carefully curated data. We explain the long history of this project as it extends back to Pantzer's own work creating the strict vocabulary for the print locations of early modern texts. We then discuss how MoEML has been able to put the STC data onto the Agas map, giving a better sense of the spatial relationship of printed early modern texts. In doing so, our argument centers on the necessity for authority names and strict vocabularies. Invoking Mike Poston's suggestion that we cannot predict the uses of our data, we use our own work on various print and digital databases to show how we can control and scaffold the mining processes to establish links between several pairs of projects in order to mine and ingest data from databases that do not share a common data field with the initial project in the sequence. We conclude with a list of considerations and principles for maximizing future interoperability between literary datasets.

Methodology

Although not strictly based in MySQL technology, our methodology borrows from the work of digital humanists like Harvey Quamen and Jon Bath who use MySQL to design relational databases. Indeed, in order to establish valuable connections across diverse datasets, we must first identify what data points these datasets have in common (either directly or indirectly). For example, suppose that Dataset 1 contains raw data for categories A, B, and C and Database 2 contains raw data for categories X, Y, and C. By identifying common data points in category C between databases 1 and 2, it becomes possible to make further connections among categories A, B, X, and Y. From here, we could identify common data points in a third database that contains raw data for categories E, F, and X. We believe that relational databases provide the best platform to capture this "web of relations" in big data. Quamen and Bath describe relational databases as "a series of interconnected spreadsheets. Each spreadsheet--called a table in database lingo--contains information on a real world entity such as People or Books or Songs or Birds or Rock Concerts or Places. Those tables are then tied together via relationships" (Quamen and Bath, 2016; 146-147). By providing a set of stepping stones or crosswalks between diverse datasets, relational databases enable us to build links between allied projects (i.e., ones that share a common data point) and more remote projects (i.e., ones that do not share a common data point) in order to combine expertise and mobilize already curated data in new environments.

Past Work

In 2014, MoEML research assistant Tye Landels-Gruenewald undertook a directed study course with director Janelle Jenstad with the aim of geocoding the *English Short Title Catalogue* (ESTC) from 1475 through 1666. With the generous assistance of David Eichmann and Blaine Greteman of the Shakeosphere project (based at the University of Iowa), we were able to extract toponyms from transcribed imprints in the ESTC catalogue using natural language processing (NLP) technology. We had intended on using named entity recognition to find matches between the extracted ESTC toponyms and our own gazetteer of early modern London locations; however, the toponyms themselves included too many errors or extra text to make this feasible. As Grover, Givon, Tobin, and Ball note in their white paper on "Named Entity Recognition for Digitised Historical Texts," there is still much work be done in order to teach named entity recognition software to recognize early modern English (Grover et al., 2008).

Concomitantly, Jenstad was manually compiling a spreadsheet of Pantzer numbers and cross-referencing them to MoEML location identifiers. Pantzer numbers are an alphanumeric string consisting of a letter and an integer. The letter indicates a general location. All the Pantzer numbers beginning with the letter O indicate locations in, near, or "against" the Royal Exchange. The numbers offer more granularity. For example, O.2 designates a location "at the north side of the Royal Exchange." Key challenges in matching Pantzer numbers with MoEML IDs were (1) different controlled vocabularies, and (2) the different levels of granularity inherent in the projects. Pantzer's authority names came from the imprint line wording; MoEML authority names are standardized spellings of the official or most common toponym variant (determined by set of critical rules we codified in order to build our gazetteer). Granularity differences emerged from the different interests of the two projects. Book historians map the bookstalls in the Royal Exchange, a location for which MoEML considered as a single entity (ROYA1); MoEML finer granularity emerges elsewhere, in our mapping of conduits, landings, and the many other precise locations that John Stow mentions in his Survey of London. A full crosswalk between Pantzer and MoEML would require the addition of sublocations to MoEML's placeography, a goal we will likely realize via the development of MoEML microsites for the Royal Exchange and Paul's Churchyard. In the meantime, we lose some of the granularity of Pantzer's data by assigning the same MoEML id to two or more Pantzer numbers.

Current Work

These past-attempts at establishing interoperability between datasets illustrate the challenges in attempting to traverse projects that only weakly share common data points. Between MoEML and the ESTC are a number of assumptions, potential errors, and remediations that weaken the link between the two respective datasets. To get to our larger project of mapping the STC, we must take smaller steps.

Our current work relates the playbook data collected by Zachary Lesser and Alan B. Farmer at *The Database of Early English Playbooks* (DEEP) to our own toponymic data, relying on Pantzer's vocabulary as a shared data-point. Jenstad's spreadsheet was transformed into a TEI-conformant XML table, which we ran across DEEP's openly available XML data. Doing so allows us to integrate DEEP numbers into the site, linking outwards to DEEP's newly static and predictable URLs.

The DEEP data and Pantzer-MoEML table can be related, but we recognize that this relationship is not immutable. In other words, both datasets are "living" databases insofar as the data can—and should be—curated and edited. Once Jenstad's spreadsheet was converted into TEI, Landels-Gruenewald was tasked with editing and refining Jenstad's initial findings to reflect the the growth of MoEML's gazetteer over the past two years (the MoEML team tagged nearly 2000 more toponyms between July 25, 2014– the last day Jenstad worked on the spreadsheet– and October 31, 2016, from 11,259 to 16,120). Lesser and Farmer have also recognized the need to amend Pantzer's findings in their data.

Future Work

The experiment with DEEP data has given us a stronger link to the ESTC. Now that we know Pantzer numbers are relatable to MoEML toponym IDs, we can now mobilize the data from Pantzer's appendix to connect MoEML with the ESTC. We plan to convert Pantzer's printed aggregations of STC numbers to digital files via OCR. With some curation, we will then have a list of all the STC numbers at each Pantzer number; using our crosswalk between Pantzer numbers and MoEML IDs, we will have a list of STC numbers (and therefore of unique print editions and issues) associated with MoEML locations. From the ESTC, we can obtain a crosswalk relating STC numbers to ESTC numbers. We add the caveat that Pantzer's locations will need to be corrected as book historians like Lesser and Farmer bring their knowledge to bear on her interpretation of STC data; every crosswalk dependent on her data will need to be refreshed and all the data maps remade. We can display these STC numbers as lists on MoEML location pages, much as Pantzer's print database does; in the digital environment, we can make dynamic links to DEEP or ESTC open-access pages for the book. We can also map these numbers on our open-layers Agas map platform as a layer of imprints associated with locations, eventually in combination with other tags (such as genre, now being added to EEBO by other scholars) or with other metadata fields harvested from the ESTC. All this data will pivot on the STC-MoEML data crosswalk that we are producing via Pantzer, following DEEP's initial work.

Distant Future Work

A longer-term goal is to harvest from the ESTC's XML files the strings of characters transcribed in the imprint line metadata field. Since we will already know from the STC-MoEML crosswalk which location is described in the imprint line, we can sort the imprint lines by locations and do rapid human scans for outliers, which may be a quick way of correcting Pantzer's data. We can also wrap TEI tags around the toponyms in the imprint lines, thereby increasing the number of toponymic variants in the MoEML gazetteer. The more variants in the gazetteer, the more accurate any future NER or geoparsing of large corpora will be. Given that we already search the EEBO-TCP corpus manually for references to place, we aspire to run our gazetteer against the entire TCP corpus to find and then map toponyms.

Principles and Practices of Curation for Future Mining and Interoperability

Acknowledging that the most interesting future uses of a project's data have not yet been imagined (Poston, 2011), how can we maximize the opportunities for other people to do things with that data? We suggest the following principles and practices as a starting point for discussion:

1. Make your data free to the world, preferably in easily downloadable and manipulable formats (in .json or .xml files, for example).
2. Be clear about how you compiled your data.
3. If you are aware of limitations in your data, tell the world.
4. As you correct and refine your data, communicate regularly about data updates.
5. If you are using other people's data in your own applications, check back regularly and rebuild the data crosswalks.
6. Know the weak link(s) in your data crosswalks.
7. Plan for corrections as other projects improve their data.
8. Be mindful of the potential for error to compound. Errors in my data, combined with errors in your data, have the potential to lead scholars to false conclusions.
9. Test your data crosswalks in a variety of ways. Take a small subset of the data and compare NLP results to hand curated results, for example.

Conclusion

Pantzer died in 2005, the year before MoEML was published at a public URL, but we like to think that she would have welcome the digital recreation, correction, curation, and connection of her data. She used the capacities of print to create a map and dense cross-references. Having "o'erleapt" Pantzer's curatorial work in building our digital catalogues, we now need to capture her formidable scholarship of interpreting and relating disparate types of data. We began with the goal of relating MoEML toponyms to ESTC numbers, but discovered that Pantzer's hand-curated data was more reliable than the results of NER and NLP. Our new question then became: "What sort of steps, processes, principles, and practices are necessary in doing this sort of work?" Handcrafted data, in conjunction with computer processing, allows for greater interoperability between projects and begins to achieve the possibilities of the data not conceived by Pantzer.

Bibliography

British Library Board. (n.d.) *English Short Title Catalogue (ESTC).* Available at: http://estc.bl.uk/ [Accessed 20 March 2017].

Farmer, A. and Lesser, Z., eds. (2007). *DEEP: Database of Early English Playbooks*. Available at: http://deep.sas.upenn.edu/ [Accessed 20 March 2017].

Farmer, A., and Lesser, Z. (2008). Early Modern Digital Scholarship and DEEP: Database of Early English Playbooks. *Literature Compass* 5(6), pp. 1139-1153.

Grover, C., Givon, S. Tobin, R., and Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association, n. pag. Available at: http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf [Accessed 20 March 2017].

Jenstad, J., ed. *The Map of Early Modern London (MoEML)*. Available at: http://mapoflondon.u vic.ca/ [Accessed 20 March 2017].

Pantzer, K. and Rider, P. (1991). *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland and of English Books Printed Abroad, 1475-1640*. Begun by A. Pollard and G. Redgrave. Vol. 3. London: Bibliographical Society.

Poston, M. (2011). *The most interesting use of our data will not be what we think it is.* [Blog] The Collation*.* Available at: http://collation.folger.edu/2011/12/the-most-interesting-use-of-our-data-will-not-be-what-we-think-it-is/ [Accessed 20 March 2017].

ProQuest LLC. (2003-) *EEBO: Early English Books Online*. Available at: http://eebo.chadwyck.com/home [Accessed 20 March 2017].

Quamen, H., and Bath, J. (2016). Databases. In: C. Crompton, R. Lane, and R. Siemens, eds., *Doing Digital Humanities: Practice, Training, Research*. London and New York: Routledge, pp. 145-162.

Stow, J. (1598). *A Survey of London.* London: Printed by John Windet for John Wolfe. STC 23341.