This is a pre-copyedited, author-produced PDF of an article accepted for publication in *Digital Scholarship in the Humanities* following peer review. The version of record Alex B, Grover C, Oberlander J, Thompson T, Anderson M, Loxley J, Hinrichs U & Zhou K (2017) Palimpsest: Improving Assisted Curation of Loco-specific Literature. *Digital Scholarship in the Humanities*, 32 (Supplement 1), pp. 4-16 is available online at: https://doi.org/10.1093/llc/fqw050

Palimpsest: Improving Assisted Curation of Loco-specific Literature

Beatrice Alex†, Claire Grover†, Jon Oberlander†, Tara Thomson•, Miranda Anderson*, James Loxley*, Uta Hinrichs°, Ke Zhou‡

†ILCC, School of Informatics
University of Edinburgh
[balex][grover][jon]@inf.ed.ac.uk

• School of Arts and Creative Industries
Edinburgh Napier University
t.thomson2@napier.ac.uk

* School of Literature, Languages and Cultures
University of Edinburgh
[miranda.anderson][james.loxley]@ed.ac.uk

°SACHI Group, School of Computer Science
University of St Andrews
uh3@st-andrews.ac.uk

‡ School of Computer Science
University of Nottingham[1]
zhouke.nlp@gmail.com

Abstract

Text mining and information visualisation techniques applied to large-scale historical and literary document collections have enabled new types of humanities research. The assumption behind such efforts is often that trends will emerge from the analysis despite errors for individual data points and that noise will be dominated by the signal in the data. However, for some text analysis tasks, the technology is unable to perform as well as domain experts, perhaps because it does not have sufficient world knowledge or metadata available. Yet, the advantage of language processing technology is that it can process at scale, even if not perfectly accurately. Geo-locating literary works is one example where human expert knowledge is invaluable when it comes to distinguishing between candidate works. This was

[1] Both Tara Thomson and Ke Zhou were employed at the University of Edinburgh, when they carried out the work for this project.

the underlying assumption in Palimpsest, an interdisciplinary digital humanities research project on mining literary Edinburgh. From the outset, the project adopted an assisted curation process whereby the automatic processing of large data collections was combined with manual checking to identify literary works set in Edinburgh. In this article, we introduce the assisted curation process and evaluate how the feedback from literary scholars helped to improve the technology, thereby highlighting the importance of placing humanities research at the core of digital humanities projects.

1. Introduction

The following quotation from John Wilson's *The Recreations of Christopher North* (1854) illustrates one of the many ways in which Edinburgh has been used as a literary setting.

> *Edinburgh Castle is a noble rock — so are the Salisbury Craigs noble craigs — and Arthur's Seat a noble lion couchant, who, were he to leap down on Auld Reekie, would break her back-bone and bury her in the Cowgate.*

Edinburgh, the first UNESCO City of Literature, has a rich literary heritage which provides the backdrop for many novels and stories. This paper reports on interdisciplinary work carried out for the Palimpsest project, which focussed on text mining literary works set in Edinburgh.[2] The project's aim was to examine the dimensions of literary Edinburgh through using text mining to scour accessible historical and fictional literary works in order to uncover those which mention Edinburgh or places within it. The term "loco-specific literature" here describes the widespread use of non-fictional place names in literary texts. This reflects an investment in place by these works, which through the use of a place name provide an anchoring mechanism that both enables and constrains the imagination of the author and the reader. We grounded "loco-specific" passages of text by identifying their latitudes and longitudes, so that both scholars and the public can geographically explore the historical and fictional city via the geo-located passages of text. Palimpsest was a collaboration between literary scholars studying the use of place and place names in literature, and computer scientists working on text mining and information visualisation. Through a range of maps and

[2] http://palimpsest.blogs.edina.ac.uk/

visualisations accessible via the LitLong.org site, the product of Palimpsest, users are now able to explore the associations of place names and the spatial relations of places in the literary city at particular periods in its history, in the works of specific authors and works, or across periods and authors. The Palimpsest data is also accessible via the mobile LitLong iPhone app in situ while walking through the city.

In this article we present an overview of the project workflow and describe the assisted curation process adopted. This process involved automatic retrieval and ranking of accessible literature according to its loco-specificity, which was followed by the manual selection of ranked documents, resulting in a set of literary works identified as *set in Edinburgh*. We also report on the fine-tuning of the retrieval and ranking prototype based on literary scholar annotators' feedback and explain how we evaluated it using standard retrieval metrics.

2. Palimpsest

The Palimpsest project was an AHRC funded collaboration between: the University of Edinburgh's School of Literatures, Languages and Cultures and its School of Informatics; EDINA; and the University of St Andrews' SACHI lab. The idea of the project arose from Dr. Miranda Anderson's project called "Palimpsest: The Literary High Street"[3] for which a prototype map interface to literary quotes containing Edinburgh place names was developed. The data presented in this interface is a small set of around 200 quotations from around 100 works crowd-sourced from Anderson's colleagues. Our aim with this project was to scale up this effort by relying on computer-assisted processing for some of the steps involved in collecting the data and geo-locating place name mentions within them. The LitLong interfaces, the final outputs of Palimpsest, link to more than 1,600 locations within Edinburgh mentioned in over 47,000 literary excerpts from around 550 books. They are aimed at scholarly and non-specialist audiences, including tourists exploring the streets of Edinburgh virtually or physically, locals who want to discover how authors described their city 150 years ago and literary scholars who are interested in place and the relations between place and literature.

[3] http://palimpsest-eng.appspot.com

2.1 Palimpsest Workflow

Figure 1 shows the workflow adopted in Palimpsest. The input data was made up of five literary document collections amounting to over 380,000 works, most of which are out of copyright, as well as a small set of modern books from authors who are well known for their literature being set in Edinburgh (including, for example, Irvine Welsh, Alexander McCall Smith and Muriel Spark). The out-of-copyright collections are varied in content and quality and contain literary fiction and nonfiction genres mostly in English but also in other languages. They include the world public domain subset of the HathiTrust data (253,350 documents)[4], the British Library Nineteenth Century Books collection (65,235 documents)[5], the Project Gutenberg data (64,047 documents)[6], a collection of documents from the National Library of Scotland (3,007 documents)[7] and the Oxford Text Archive TEI text data (2,729 documents)[8]. Unfortunately, our data collection and preparation work was carried out before the EEBO-TCP and ECCO-TCP data sets were made available for research. In future iterations of Palimpsest, these collections should also be considered.

In Palimpsest, our analysis was limited to English language documents. If the information on the language of a text was not already present in the metadata for the document then we computed it automatically using the TextCat language identification tool which works very reliably even when given just a few sentences of text.[9]

In our workflow the input data was first converted to one common format necessary for the document retrieval component where it was first indexed. Edinburgh-specific candidate documents were then retrieved automatically from this index. This component outputs a set of ranked Edinburgh-specific candidate documents per collection.

[4] https://www.hathitrust.org
[5] http://labs.bl.uk/Digital+Collections+-+Books+and+Text
[6] https://www.gutenberg.org
[7] http://www.nls.uk
[8] http://ota.ox.ac.uk/catalogue/index.html
[9] http://odur.let.rug.nl/~vannoord/TextCat/

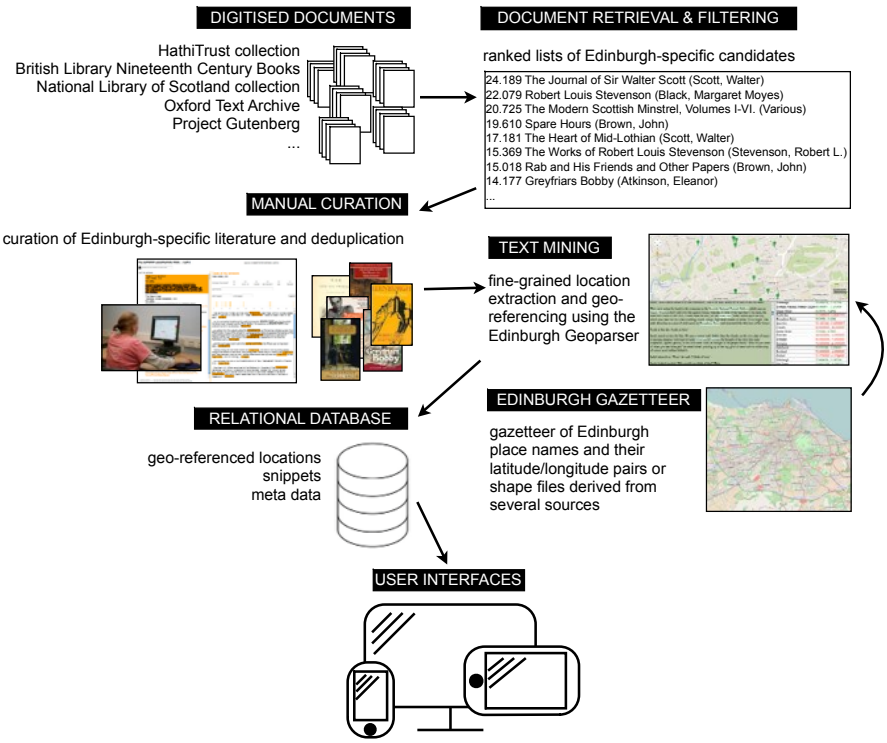

Figure 1: Palimpsest workflow.

The ranked output was loaded into a web-based annotation tool for manual curation. All Edinburgh place names occurring in the document along with the snippets of text surrounding the place name mentions were displayed to aid the decision making of the annotators, three literary scholars from the English Literature Department at the University of Edinburgh. The curators considered candidate literary works as *set in Edinburgh* if the city featured as a setting but not necessarily as the primary setting. For example, a book containing a chapter with dense mentions of Edinburgh place names was considered as being sufficiently set in Edinburgh for inclusion, particularly since in the LitLong interfaces excerpts are abstracted away from the their source works.

While going through the ranked list of documents, the curators decided to stop curating after going through the top 10% of the ranked documents per collection. Aside from project time constraints, this decision was taken because, as the likelihood of a document being a true Edinburgh-specific candidate decreases with a decreasing ranking, it took longer and longer to find real true data points in the ranked list. Moreover, the annotators also added any

documents which were not already selected as Edinburgh-specific candidates as part of the assisted curation process, but which were in the pool of documents crowd-sourced manually for the Palimpsest prototype.

The sub-set of 546 works which were identified as Edinburgh-specific in this way were then further processed by our text mining pipeline which geo-references place names by grounding them to their latitude/longitude coordinates using the Edinburgh Geoparser (Grover et al. 2010, Alex et al. 2015).[10] The geoparser is set up to work by default with GeoNames[11], a global gazetteer. We adapted it to work with the fine-grained Edinburgh gazetteer, which was aggregated and cleaned during the Palimpsest project. The text-mined output (geo-referenced location mentions, snippets etc.) was stored in the Palimpsest database, which is accessible via our web-based LitLong visualisations (see Figure 2), an iPhone app[12] and via a search API hosted by EDINA.[13]

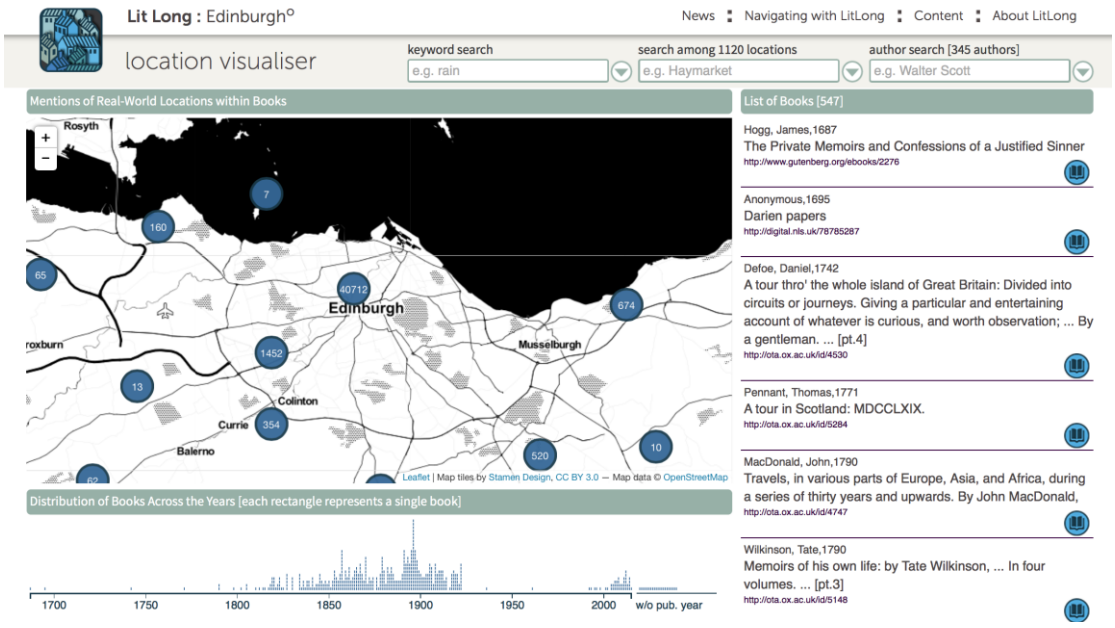

Figure 2: The LitLong web interface accessible at litlong.org

[10] https://www.ltg.ed.ac.uk/software/geoparser
[11] http://www.geonames.org
[12] http://litlong.org/navigating-with-litlong/download-our-app/
[13] http://litlong.edina.ac.uk/search/

2.2 Document Retrieval and Ranking

The aim of Palimpsest was to be able to discover and make available for exploration a broad spectrum of books, including forgotten gems which are not part of the established canon of Edinburgh literature. The main literary research question was: is the characteristic of literary setting, and the detailed ways in which this is narratively established, sufficiently amenable to machine reading to allow us to work automatically across large scale collections of digitised texts? This involved finding ways to define when a book qualifies as being Edinburgh-specific, exploring the literary use of place names and their utility as a marker of setting (Anderson and Loxley, 2016), and developing a document retrieval and ranking tool to sift potential candidates out of the pool of literature to which we had access.

In order to retrieve candidates of Edinburgh-specific literature, the literary data was first indexed using Indri 5.6[14] and ranked using a set of 1,633 Edinburgh place name queries.[15] We used the Indri inference network language model based ranking approach (Strohman et al., 2015). It combines the language modeling (Ponte and Croft, 1998) and inference network (Turtle and Croft, 1991) methods to information retrieval. The resulting model allows structured queries similar to those used in Inquery (Callan et al., 1992) to be evaluated using language modeling estimates within the network. Indri is a research oriented information retrieval framework, which supports various effective ranking functions. The reason why we chose it over other search engines is that its structured query representation allows us to enforce constraints for ranking.

The document retrieval and ranking prototype was developed using the HathiTrust data only, the largest of our document collection. Metadata and ambiguity weightings were taken into account when computing the ranking.

The ranking score of a document was computed by combining the score for the location query retrieved from the content of a book with a score based on information in the metadata

[14] http://sourceforge.net/projects/lemur/files/lemur/indri-5.6/

[15] They included entries appearing in at least three of five resources used to construct the Edinburgh gazetteer (OpenStreetMap, OSLocator, Royal Commission for Ancient Historic Monuments of Scotland, Historic Scotland, QuatroShapes of Edinburgh areas).

of the book.[16] For example, the ranking was increased given the presence of a set of favoured subclasses from the Library of Congress classification system, including PR (English literature), DA (description and travel), PZ (fiction and juvenile belles lettres), PN (literature (general)) and PS (American literature), or given a list of relevant subject terms (Edinburgh, Scotland, literature, fiction, novel, poetry, poem, story, stories, drama, novella, English, biography, ballads, ballad, Scottish).[17] This was to ensure documents with such metadata information appeared higher in the ranking.

At the same time, the ranking score was down-weighted for ambiguity of Edinburgh place names in order to push documents which mention place names most likely not referring to a location within Edinburgh down the list. There are various kinds of ambiguous place names, for instance: common place names which occur in other towns or cities (for example, 'Market Street'); place names which are derived from person names or which describe a person (for example, 'the Town Guard of Edinburgh') or place names which are also common nouns (for example, 'Trinity'). The weight for a location was determined by means of its frequency in GeoNames[18] so that more frequently occurring place names are considered less likely to be locations occurring within Edinburgh.

The output of the document retrieval component is a set of ranked Edinburgh-specific candidate documents per collection as depicted in Figure 1. This data was then loaded in order of ranking into the curation tool.

3. Assisted Curation

The term *assisted curation* refers to the process of semi-automatically curating a set of Edinburgh-specific literature from all accessible literature. This means that the results of the retrieval and ranking process were checked manually by literary curators. In the case of Edinburgh, related endeavours to geo-locate literature have relied on the collection of titles,

[16] By metadata we refer to traditional library cataloguing record data.
[17] In this article we also refer to this information as genre and subject, respectively. By subject information we refer to MARC 21 bibliographic data information, in particular subject access field information stored under code 600 (Personal Name), 650 (Topical Term) or 651 (Geographic Name).
[18] The ambiguity-related weight was computed by dividing 1 by the sum of the frequency of the place name in GeoNames and 1.

or passages, by a few individuals or via crowd sourcing (e.g. Edinburgh Reads[19] run by Edinburgh Libraries or Global Bookmap[20]). The Book Navigator, a web-based tool and mobile app interface which allows the users to manually geo-locate place name mentions in literary data directly in eBooks (Hinze et al., 2015), could be used for such crowd-sourcing endeavours.

As mentioned previously, the idea for Palimpsest arose out of an initial prototype which visualises a small set of around 200 extracts manually collected by literary scholars at the University of Edinburgh. Such an approach results in high-quality data with the disadvantage of missing less well-known but potentially interesting works. In this further iteration of Palimpsest we considered the entire pool of literature accessible to us in order to determine a sub-set of highly ranked Edinburgh-specific candidates automatically using location-based document retrieval. The aim was to reveal a wider range of Edinburgh-specific literature, by uncovering now obscure and neglected literary works, and juxtaposing them alongside the more famous and well-known works.

Assisted curation by means of text mining alone has shown encouraging results in other domains (e.g. Kristjansson et al., 2004 and Alex et al., 2008). In Palimpsest specifically, we combined text mining and information retrieval for assisted curation and studied how user feedback can improve the technical stages of this process. Extending the same model to other collections for identifying Edinburgh-specific place names, using the same parameters and setup, is straightforward in terms of running the text processing tools. The main effort would be in converting the data to the same input format required by the information retrieval and text mining pipeline. Some decisions made in Palimpsest and some resources used as part of this processing are very specific to Edinburgh. This means that applying the model to a different city of literature would involve more work on the technical side, including the development of a place name gazetteer and some thinking about what constitutes a place name specific to the new location.

[19] http://yourlibrary.edinburgh.gov.uk/fictionmap
[20] http://www.mappit.net/bookmap/

3.1 Curation Tool

The manual annotation of the ranked candidates to select actual Edinburgh-specific literature was done using the web-based annotation tool displayed in Figure 3. All ranked documents are displayed on the left-hand panel, listing the title of each work, the author and publication date if available, a link to the original source document and a list of location mentions identified within the book. When clicking on a title and thereby selecting a document, additional information appears in the right-hand panel, including a graph showing occurrences of place names within a document and snippets containing Edinburgh place names. Based on this information and by following the link to the original source, the annotators were able to determine a work as being Edinburgh-specific or not, enter further comments and identify the start and end content pages of a document. The latter was useful to avoid identifying Edinburgh place names in the paratext of a literary work. When clicking the submit button, a document annotation is saved to the database and disappears from the panel on the left. However, the annotators were also able to access all previous annotations by clicking on the link on the top right corner ("see list of ANNOTATED BOOKS"). The tool also allows users to search for an author or book title in the list of ranked document using the search box in the top left corner.

The annotation tool was developed specifically for Palimpsest as the curation process involved a specific set of aforementioned requirements (including rating, commenting, linking to original documents, annotated documents disappearing from view, highlighted location entities in context, searching, marking start/end content pages and graph of location occurrences across the document). We were aware of existing tools supporting such features but not one supporting them all.

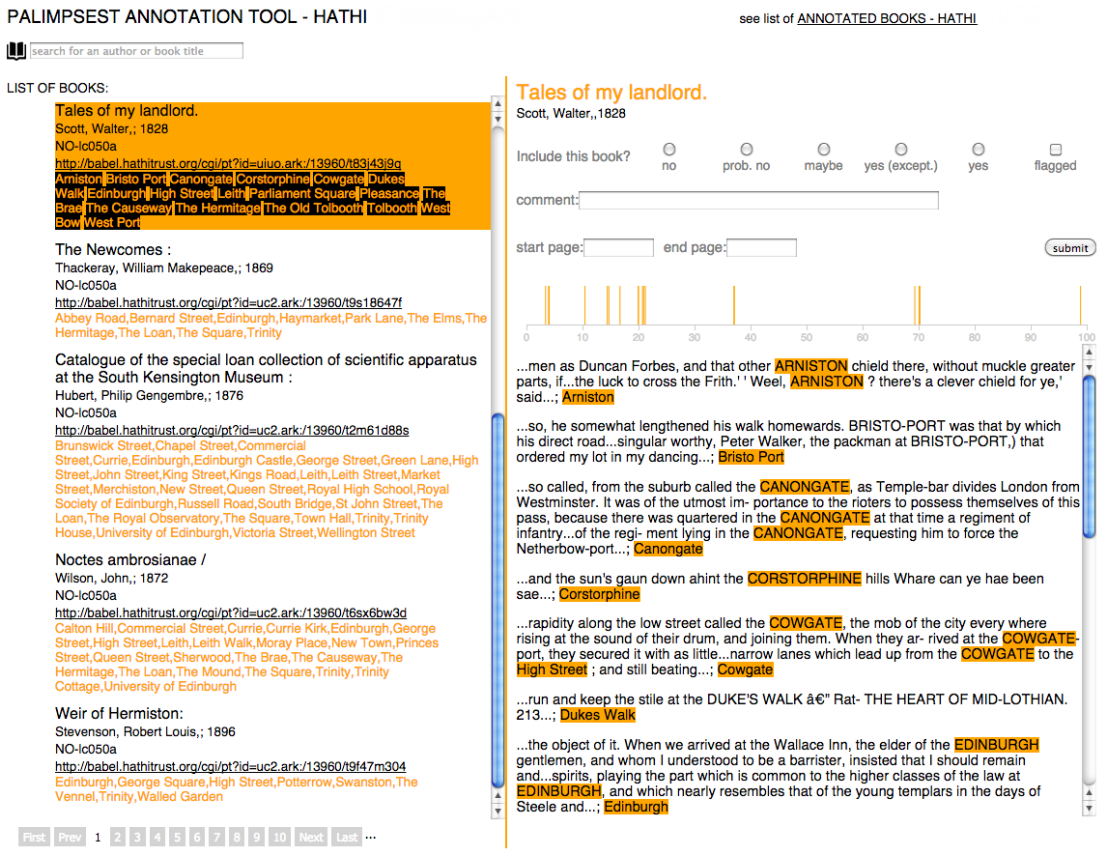

Figure 3: Palimpsest annotation tool.

3.2 Annotation Scheme

Items were annotated using the annotation scheme shown in Table 1. We consider documents annotated as *yes* or *yes (except)* as Edinburgh-specific within Palimpsest.[21] The scheme was developed by the annotators while working on an initial ranking of HathiTrust documents.

Label	Explanation
yes	Fiction containing Edinburgh place names
yes (except)	Narrative non-fiction (incl. letters, memoirs, autobiographies, etc.) containing Edinburgh place names
probably not	Poetry containing Edinburgh place names
maybe	Literature containing Edinburgh place names but not considered sufficiently place-rich
no	Non-literary works containing Edinburgh place names or literary works not containing Edinburgh place names

Table 1: Annotation Scheme

The decision to include certain non-fictional works within the database, such as letters, memoirs, autobiographies, biographies and travels journals, reflects the widening of the literary canon more generally to include such non-fictional works, as well the inclusion of

[21] We excluded poetry but we annotated it (*probably not*) to be able to work on it in future.

such works in the Palimpsest prototype, where literary scholars had suggested passages from both fictional and non-fictional works for inclusion. However, this added to the complexity of creating a reliable automated ranking system, as discussed further in later sections. Poetry was reluctantly excluded in this iteration of Palimpsest due to the text mining challenges created by its form as well as its textual presentation.[22]

No distinction was made between non-literary works containing Edinburgh place-names (e.g. gazetteers, directories etc.) and literary works not containing Edinburgh place-names. Both types were annotated using the "no" label. However, the annotators frequently added comments on the type of work in question which could help to make this distinction.

3.3 Experiment Data

We used the world public domain HathiTrust collection (253,350 documents) to develop the retrieval and ranking component. For setting up the prototype, we started with all HathiTrust documents with available genre information in the metadata in the form of codes from the Library of Congress classification system. We found that 142,680 (56.3%) of the works in that collection had genre information in their metadata records, i.e. Library of Congress classes and subclasses.[23]

Applying the document retrieval and ranking prototype to this data yielded 14,044 ranked candidate documents containing one or more Edinburgh place names. Over a period of two weeks, the annotators curated the ranked documents in order. This resulted in 1,710 annotated documents, of which 200 were considered Edinburgh-specific literature. We considered this to be our gold standard data which we used later to test different document retrieval and ranking component modifications.

[22] Poetry set in Edinburgh was excluded from the Palimpsest database accessible via LitLong but was annotated during the manual curation as "probably not" for processing in possible future iterations of this work. The text processing stages would need to be tailored to poetry as they are currently developed for running text containing capitalisation and punctuation, conventions which are often not adhered to in poetry.

[23] 106,962 (42.2%) documents had subject information stored in their metadata records.

3.3 Initial Feedback

Initially, the annotators reacted enthusiastically to the annotation and discovered numerous works set in Edinburgh of which they had not been aware, such as Margaret Williamson's *John and Betty's Scotch History Visit* (1912), a history and travel guide in the guise of a fictional story about two school children travelling Scotland, and Professor John Wilson's *Noctes Ambrosianae*, a collection of popular political and editorial columns originally published in Blackwood's Edinburgh magazine between 1822 and 1835. They were also pleased to discover a large number of travel memoirs from the mid- to late-nineteenth century, most written by Americans travelling in Scotland.

As they worked through the documents, however, they lost trust in the ranking system. They noticed relevant documents appearing far down the list and sometimes had to go through many documents to find a positive example. Given the sheer volume of the results, and the amount of time it was taking the annotators to examine each text for relevance, it was apparent that they would be unable to manually curate the full set of results. As such, improving the ranking system would be of paramount importance in curating a strong, relevant collection of works for the final database.

The annotators identified two main issues with the ranking system, which had resulted in a substantial number of false hits appearing higher on the list than other more relevant results. The most urgent issue arose from the inclusion of ambiguous place names in the search gazetteer. These were of two varieties: Edinburgh location names that also appeared with great frequency in other cities (for example, 'Commercial Street'), and place names not specific to particular locations (for example, 'Town Hall').

In the first category, there were a great number of texts appearing high in the ranking that were not set in Edinburgh, but instead in other British and North American cities, particularly London and Boston. Most of those texts set in North American cities were histories of those regions, which had been given primacy in the ranking due to the high density of shared place names. The annotators observed that most of the Edinburgh-specific texts also included reference to the name 'Edinburgh', or one of its variants, such as 'Edinboro', 'Edina', or

‘Embro’, among others, whereas it naturally did not feature as often in those texts set in other cities, especially in North America.

In the second category, the place names resulting in false hits were largely general, rather than loco-specific, names, including ‘Police Station’, ‘the Square’, ‘Main Street’, ‘Town Hall’, ‘Medical School’, ‘Great Hall’, and others. Many of the texts ranked high in our results included a high density of these general types of places, making the texts appear to be dense in Edinburgh locations. There was also a high instance of names such as ‘Trinity’, ‘the Loan’, or ‘the Murrays’, which name not only places but historical, social, or religious concepts, or even people. The annotators compiled a list of such ambiguous place names to feed back to the natural language processing developers in order to improve the document retrieval and ranking prototype.

Another problem with the ranking was the frequent occurrence of non-narrative non-fiction works, which the project team had not planned to include in the database, such as regional and family histories, encyclopaedias, dictionaries, catalogues, and county registers. One especially amusing result that appeared high in the ranking was *A Record of Unfashionable Crosses in Shorthorn Cattle Pedigrees* (1883), by F.P. and O.M Healy, which dealt with cattle breeds in Ohio that descended from imported British stock. It was apparent that non-fiction, in general, would pose an issue for ranking, as place names appear to be used with much greater frequency in non-fiction writing than in fiction. However, since some types of non-fiction were going to be included in the database, such as memoirs and literary correspondence, non-fiction as a general category could not be entirely excluded. In hopes of limiting manual curation of non-fiction works, the annotators observed a series of titular words that always marked a non-relevant text, including (but not limited to) ‘Record(s)’, ‘Register’, ‘Catalogue’, ‘Dictionary’, ‘Encyclopedia/Encyclopaedia’, ‘Topography’, and ‘Index’. The annotators then fed this list back to the language technology team for deletion from the ranking. Where codes from the Library of Congress classification system were available, the annotators also suggested that giving literary categories a higher ranking than

non-fiction categories may help bring fiction higher in the ranking, despite its often minimal density of place names.

4. Improving the Ranking

In summary, the annotators recorded a list of ambiguous place names mostly referring to other locations and a list of words in titles suggesting non-literary content. They also observed that most Edinburgh-specific documents contain at least one reference to Edinburgh or one of its variants.

Based on this feedback, we then fine-tuned the retrieval component. There is a body of research on using relevance judgments for improving information retrieval, a good summary of which is provided by Manning et al. 2008. We tested the initial ranking (baseline) as well as the following three measures and their combination:

1. Down-weighting further ambiguous place names identified by the annotators.
2. Removing documents containing non-literary titular words identified by the annotators (‘catalogue’, ‘dictionary’, etc.).
3. Ensuring that Edinburgh or one of its variants (‘Auld Reekie’, ‘Edinboro,’ ‘Edinbra’, ‘Edinburg’, ‘Edinbrughe’, ‘Edinburrie’, ‘Embra’ and ‘Embro’) occurs in the work.

The latter step also meant that the query gazetteer was increased from 1,633 to 1,641 place names in order to include the various name variants of Edinburgh.

4.1 Results

As document retrieval systems produce ranked output, they are most standardly evaluated by means of the mean average precision (MAP) metric which results in one single figure measuring the quality across all recall levels (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al. 2008). The set of 1,710 annotated HathiTrust works was used as an evaluation set to determine the effect of each modification. MAP scores are computed by comparing the system ranking to the ground truth of ratings of the same data created by the annotators.

Our baseline, the document retrieval and ranking prototype before the modifications were made, performed at a MAP score of 0.13 when retrieving 14,044 documents from the

HathiTrust collection. Figure 4 shows that down-weighting of ambiguous place names (see System 1) resulted in a small improvement in the MAP score. Filtering documents with non-literary title words (System 2) had the highest increase in the MAP score and also lead to a sizeable reduction in the number of ranked document by 17.2% compared to the baseline system. The condition of Edinburgh or one of its variants to appear in the document (see System 3) decreased the MAP score slightly which is unsurprising since a small number of Edinburgh-specific documents do not refer to the city itself; the feedback from the annotators was that in the majority of cases (but not all) the city name is mentioned. However, this measure resulted in a large decrease in the number of ranked documents reducing the workload of the annotators significantly (by 54.1%). We therefore consider measure 3 to be beneficial as well. When combining all three measures, the retrieval component yielded a small improvement in the MAP score of 0.17 (compared to the baseline MAP of 0.13), and the workload of documents to be curated was reduced considerably by 58.4%.

SYSTEM	MAP	#RANKED
Baseline	0.1307	14,044
1.	0.1470	13,985
2.	0.1846	11,625
3.	0.1292	6,440
1.-3.	0.1684	5,860

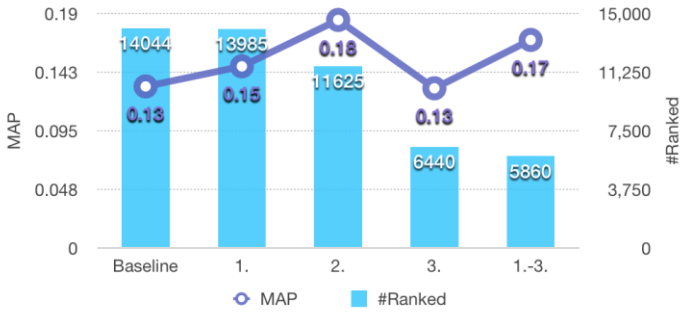

Figure 4: Performance of the document retrieval baseline and various modifications. We report mean average precision (MAP) and number of ranked documents (#RANKED) per retrieval method.

The improved document retrieval and ranking component (when applying all three measures) was re-applied to the as yet un-annotated HathiTrust collection (with/without metadata information) and was also run on the four remaining, out-of-copyright data collections listed in Section 2.1. The different ranking outputs (one per collection) was then presented to the annotators in the annotation tool for further manual curation.

4.2 Feedback from the Curators

The annotators reported the new ranking to be significantly improved after their feedback had been taken into consideration when making the modifications to the prototype. While the annotation process was slowed a great deal, this was due to the increased instances of relevant documents demanding closer scrutiny. After down-weighting ambiguous place names and applying the 'Edinburgh +' criteria, the annotators found more Edinburgh-specific works rising to the top of the list, and fewer instances of works set in other cities. The occurrence of texts set in Boston and other American cities almost entirely disappeared from the top 10% of results, although many texts remained that were set in London. The substantial overlap of place names in London and Edinburgh, coupled with cultural connections that lead to mentions of Edinburgh in these texts (for instance, minor references to a person from or a trip to Edinburgh), would make these results challenging for the automated process to differentiate between. However, the annotators noted particular place names that were more often associated with London than Edinburgh (such as 'Haymarket'), which became red flags for the annotators, speeding up the curation process slightly. Adding variants of 'Edinburgh' to the 'Edinburgh +' criteria yielded positive results that would not have been identified in the early ranking phase, highlighting the value of including historical variants. For instance, William Beatty's novel *The Secretar* (1897) is set in Edinburgh and written partly in Scots dialect, so 'Edinburgh' does not appear in the text but 'Embro' does.

The exclusion of documents with non-relevant title words also made the curation process much more manageable, as stated, but non-fiction in general still dominated the higher ranked documents and the ranking system remained unable to make finer genre distinctions, especially where metadata was incomplete. Fiction still ranked lower due to the lower density of place names; however, as annotators were not spending as much time sifting through obvious false hits, they were able to find fictional texts as they moved deeper into the ranked documents. This improvement in the ranking system seemed to be somewhat undermined by either limited or incorrect metadata in the results from other collections, in particular the National Library of Scotland's digital collection and additional results from HathiTrust without relevant metadata information, which led to non-relevant non-fiction (particularly

family histories) rising in the ranking. However, these were smaller sets of results so the higher number of false hits was more manageable than with the main batch of results from the HathiTrust collection.

Improvements in the ranking system enabled annotators to discover more relevant documents than they found initially, although it remained apparent that the ranking system would not be able to make reliable distinctions between imaginative descriptions of place and references to real-world locations. A telling example was the appearance of Sir James Matthew Barrie's novel *Quality Street* (1920) fairly high in the ranking. Barrie is a Scottish author, most famous for being the creator of Peter Pan. The front matter of the book contains a mention of Edinburgh in one of the other works he is the author of (An Edinburgh Eleven), so the text met the 'Edinburgh +' criteria. Other place name mentions within the content of the book included 'Quality Street' and 'the Causeway', actual locations in Edinburgh; however, the novel is not set in Edinburgh, but in a fictional small town. Within the Palimpsest project's workflow and its scope of resources, works such as this could not be resolved through the improved ranking, only through human curation. This is a clear example of why domain expert knowledge within technology-assisted projects such as Palimpsest is essential.

5. Discussion and Conclusion

In the Palimpsest project we have explored how to combine computational approaches (document retrieval and ranking, text mining and information visualisation) to facilitate literary research. The technical partners in the project built on their know-how already acquired in the Trading Consequences project[24], a digital humanities collaboration involving environmental historians as domain experts (Hinrichs et al., 2015). From this past experience, it was clear to the team even at the stage of writing the proposal that the involvement of domain experts, the literary scholars in the case of Palimpsest, was fundamental to the success of this interdisciplinary digital humanities project. As a result, the

[24] http://tradingconsequences.blogs.edina.ac.uk

assisted curation process undertaken in Palimpsest, which we described and evaluated in this article, was planned right from the project outset.

This process attempted to keep the user in the loop during the iterative technical development. We received very useful feedback from the literary scholars on issues that appeared as they curated documents and considered their suggestions in changing the underlying methods for retrieving and ranking Edinburgh-specific literature. Our results show that system performance improved slightly and that curation workload was reduced substantially as a result. The improved method was subsequently applied to all document collections, which resulted in mostly positive feedback from the curators reporting that the ranking revealed more relevant documents.

While working with the output data, the literary scholars became increasingly familiar with the strengths and limitations of the document retrieval and ranking technology and used this knowledge to their advantage to speed up their work. Aside from providing valuable feedback for improving the technology, they also understood quickly that the automatic process was there to assist and not replace them. Human curation was particularly vital for cases in which the system had insufficient knowledge or capability to perform a task such as the distinction between fictional and real place names or between different types of genre. Palimpsest is therefore a good case study for illustrating the importance of humanities scholarship at the core of digital humanities research.

The fact that the system struggled to differentiate between genre for works which did not contain this information in the metadata is not surprising. In such cases, the system has to rely mostly on the presence or absence of location mentions in the text. This signals the importance of the availability of document level metadata information. Since our work was completed, Ted Underwood and his collaborators have developed a method to classify genre of HathiTrust documents at the page level using machine learning (Underwood, 2014). Using their code, pages can be labelled with 93.6% accuracy as either paratext (front matter, back matter, ads), prose nonfiction, poetry (narrative and lyric), drama (incl. verse drama), or prose fiction. This shows that certain types of metadata information can be inferred automatically

with relatively high accuracy and that it does not necessarily require laborious manual curation to perform the bulk of such work. If we had had this genre classifier available to us from the outset of Palimpsest, the genre distinctions could also have been considered by our document retrieval and ranking system, making its ranked output more reliable.

The aim to uncover hidden literary gems set in Edinburgh was clearly met by the assisted curation approach taken in the project. The underlying idea of the project was to go from big data (all of the accessible literature) to small data (the Edinburgh-specific documents that finally made it into the Palimpsest corpus). By starting with big data, we did not, as Tim Hitchcock put it rightly, want to 'get away with dirty data' (Hitchcock, 2014). The combination of automatic and manual processing meant that we were able to identify a wide range of literary works set in Edinburgh whilst at the same time ensuring that all documents visualised by the LitLong tools contain Edinburgh place name mentions.

In future iterations of Palimpsest, we would like to include additional collections which have become accessible more recently and would also like to adapt the language processing tools to process Edinburgh-specific poetry already annotated during the manual curation phase.

Funding

This work was supported by the Arts and Humanities Research Council [AHRC AH/L009935/1] via the Digital Transformations in the Arts and Humanities - Big Data programme.

Acknowledgements

We would like to thank Dr. Harris-Birtill at the University of St Andrews who helped with the data processing for the annotation tool. We also thank the data providers (HathiTrust, the British Library Labs, Project Gutenberg, the Oxford Text Archive and the National Library of Scotland) as well as the authors and the publishers of a selected set of contemporary authors who have given us access to the texts for the purposes of Palimpsest.

References

Alex B., Byrne K., Grover C. and Tobin R. (2015). Adapting the Edinburgh Geoparser to Historical Georeferencing. *International Journal for Humanities and Arts Computing, 9(1),* March 2015.

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R. and Wang, X. (2008). Assisted Curation: Does Text Mining Really Help? In: *BIOCOMPUTING 2008*. In *Proceedings of the Pacific Symposium on Biocomputing*, pp. 556-567.

Anderson, M. and Loxley, J. (2016). The Digital Poetics of Place Names. *Literary Mapping in the Digital Age*. Cooper, D., Donaldson, C. and Murrieta-Flores, P. (eds.), Routledge, pp. 47-66.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston: Addison-Wesley Longman.

Barrie, J.M. (1920). *Quality* Street. New York: C. Scribner's sons.

Beatty, W. (1897). *The Secretar.* Paisley & London: Alexander Gardner.

Callan, J.P., Croft, W.B. and Harding, S.M. (1992). The Inquery retrieval system. In: *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, 1992, pp. 347-356.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. and Ball, J. (2010). Use of the Edinburgh Geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925), pp. 3875-3889.

Healy, F.P. and Healy, O.M. (1883). *A Record of Unfashionable Crosses in Shorthorn Cattle Pedigrees*. Bedford, Iowa: The Authors.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. and Coates, C.M. (2015). Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration, to appear in DHS, Special Issue of DH2014.

Hinze, A. M., Littlewood, H. and Bainbridge, D. (2015). Mobile annotation of geo-locations in digital books. In: *Proceedings of the International Conference on Theory and Practice of Digital Libraries*. Poznań, Poland: Springer.

Hitchcock, T. (2014). Big Data, Small Data and Meaning. Historyonics blog post, 09/11/2014 based on his keynote talk at the British Library Labs Symposium on 03/11/2014.

Kristjansson, T.T., Culotta, A., Viola, P. and McCallum, A. (2004). Interactive information extraction with constrained conditional random fields. In: *Proceedings of AAAI*, pp. 412-418.

Manning, C.D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Ponte, J.M. and Croft, W.B. (1998). A language modeling approach to information retrieval. In: *Proceedings of SIGIR*, pp. 275-281.

Strohman, T., Metzler, D., Turtle, H. and Croft, W.B. (2005). Indri: A language-model based search engine for complex queries (extended version), *CIIR Technical Report.*

Turtle, H. and Croft, W.B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), pp. 187-222.

Underwood, T. (2014). Understanding Genre in a Collectio of a Million Volumes. Interim performance report for the Digital Humanities Start-up Grant [HD5178713], University of Illinois, Urbana-Champaign.

Williamson, M. (1912). *John and Betty's Scotch History Visit.* Boston: Lothrop, Lee & Shepard co.

Wilson, J. (1855). *Noctes Ambrosianae*. Ferrier, J.F. (ed.), Edinburgh: William Blackwood & Sons.

Wilson, J. (1854). *The Recreations of Christopher North*. Boston: Phillips, Sampson, and company.