# Palimpsest: Improving Assisted Curation of Loco-specific Literature

Beatrice Alex†, Claire Grover†, Jon Oberlander†, Ke Zhou†, Uta Hinrichs*

†ILCC, School of Informatics, University of Edinburgh
[balex][grover][jon]@inf.ed.ac.uk, zhouke.nlp@gmail.com[1]

*SACHI, University of St.Andrews
uh3@st-andrews.ac.uk

## 1. Introduction

This paper reports on interdisciplinary work carried out for the Palimpsest project, focusing on mining literary works set in Edinburgh, a UNESCO City of Literature.[2] The project's aim is to use text mining to scour accessible literary works and find those mentioning Edinburgh or places within it. We ground "loco-specific" passages of text by identifying their latitudes and longitudes, so that both scholars and the public can geographically explore their fictional city. Palimpsest is a collaboration between literary scholars studying the use of place in literature and computer scientists working on text mining and information visualisation. Through a range of maps and accessible visualisations, users are able to explore the spatial relations of the literary city at particular times in its history, in the works of specific authors, or across eras and writers.

We present an overview of the project workflow and describe the assisted curation process adopted. It involves automatic retrieval and ranking of accessible literature according to its loco-specificity followed by manual selection of ranked documents, resulting in a set of literary works identified as set in Edinburgh. We report on the fine-tuning of the retrieval and ranking prototype based on literary scholar annotators' feedback.

## 2. Palimpsest

Fig. 1. shows the Palimpsest workflow. The input data is made of five literary document collections amounting to approximately 380,000 works, most of which are out of copyright, as well as a small set of modern books from authors which are well known for their literature being set in Edinburgh (incl. Irvine Welsh, Alexander McCall Smith and Muriel Spark). The out-of-copyright collections are varied in content and contain literary fiction and nonfiction genres. The data is first indexed using Indri 5.6[3] and ranked using a set of 1,633 Edinburgh place name queries.[4] We use the Indri inference network language model based ranking approach (Strohman et al., 2015). The ranking score of a document is increased given certain meta data information (including a set of favoured Library of Congress codes and subject terms) or down-weighted for ambiguous Edinburgh place names. We combine the score for genre in the meta data with the location query retrieved from the content of the book.  The output of the document retrieval component is a set of ranked Edinburgh-specific candidate documents per collection.

[1] Zhou has recently started working for Yahoo Labs London.

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

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

[4] This includes 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).

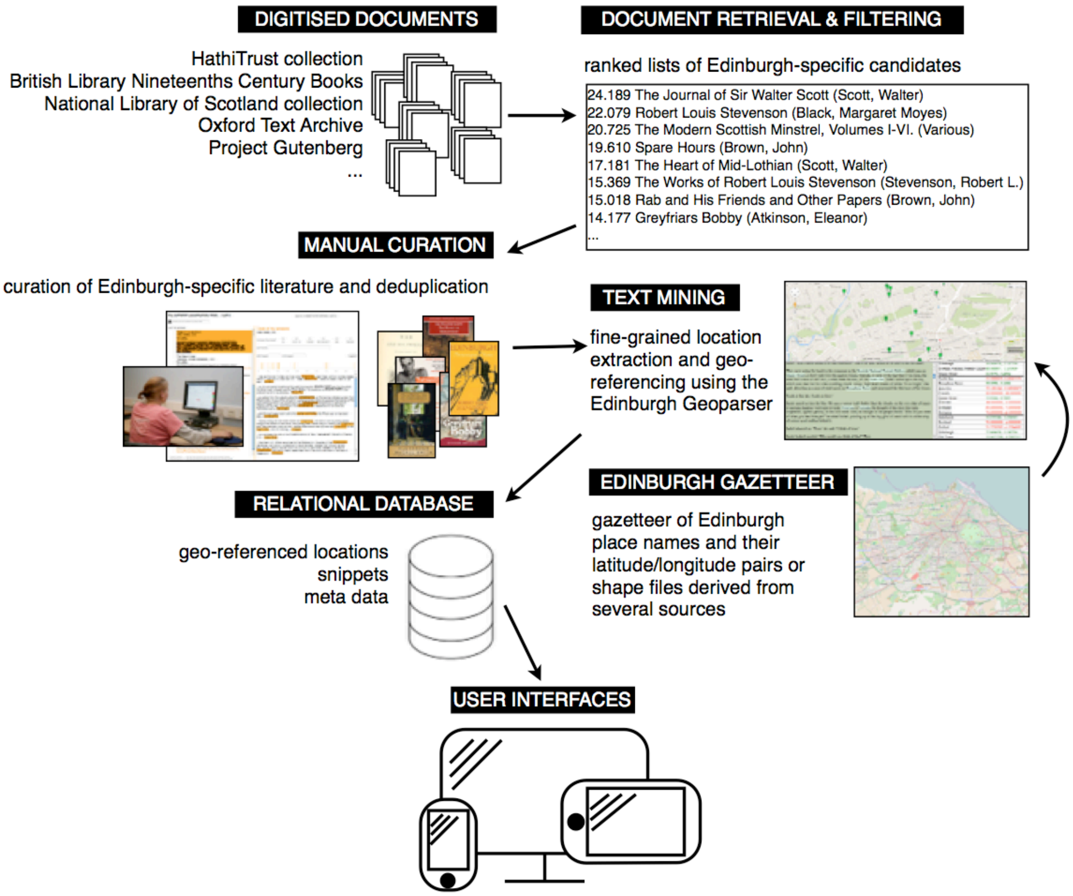

DIGITISED DOCUMENTS
HathiTrust collection
British Library Nineteenths Century Books
National Library of Scotland collection
Oxford Text Archive
Project Gutenberg
...
DOCUMENT RETRIEVAL & FILTERING
ranked lists of Edinburgh-specific candidates
24.189 The Journal of Sir Walter Scott (Scott, Walter)
22.079 Robert Louis Stevenson (Black, Margaret Moyes)
20.725 The Modern Scottish Minstrel, Volumes I-VI. (Various)
19.610 Spare Hours (Brown, John)
17.181 The Heart of Mid-Lothian (Scott, Walter)
15.369 The Works of Robert Louis Stevenson (Stevenson, Robert L.)
15.018 Rab and His Friends and Other Papers (Brown, John)
14.177 Greyfriars Bobby (Atkinson, Eleanor)
...
MANUAL CURATION
curation of Edinburgh-specific literature and deduplication
TEXT MINING
fine-grained location extraction and geo-referencing using the Edinburgh Geoparser
EDINBURGH GAZETTEER
gazetteer of Edinburgh place names and their latitude/longitude pairs or shape files derived from several sources
RELATIONAL DATABASE
geo-referenced locations
snippets
meta data
USER INTERFACES

**Figure 1: Palimpsest workflow.**

This data was loaded into a web-based annotation tool for manual curation. All Edinburgh place names occurring in the document and snippets surrounding them were displayed to aid the annotators', three literary scholars from the School of Literature at the University of Edinburgh, decision-making. The sub-set of works which were manually curated as Edinburgh-specific are further processed by text mining which geo-references place names by grounding them to their latitude/longitude coordinates using the Edinburgh Geoparser (Grover et al., 2010)[5] and in particular the Edinburgh gazetteer which is being developed in Palimpsest. The output (geo-referenced location mentions and snippets) is stored in the Palimpsest database which is accessible via web-based visualisations.

## 3. Assisted Curation

By assisted curation we refer to the process of semi-automatically curating a set of Edinburgh-specific literature from all accessible literature. Related endeavours have relied on the collection of titles, or passages, by a few individuals or via crowd sourcing (e.g. Edinburgh Reads[6] run by Edinburgh Libraries or Global Bookmap[7]). The idea for Palimpsest arose out of an initial prototype which visualises a small set of extracts manually collected by literary scholars at the University of Edinburgh.[8] Such an approach results in high-quality data with the disadvantage of

[5] https://www.ltg.ed.ac.uk/software/geoparser

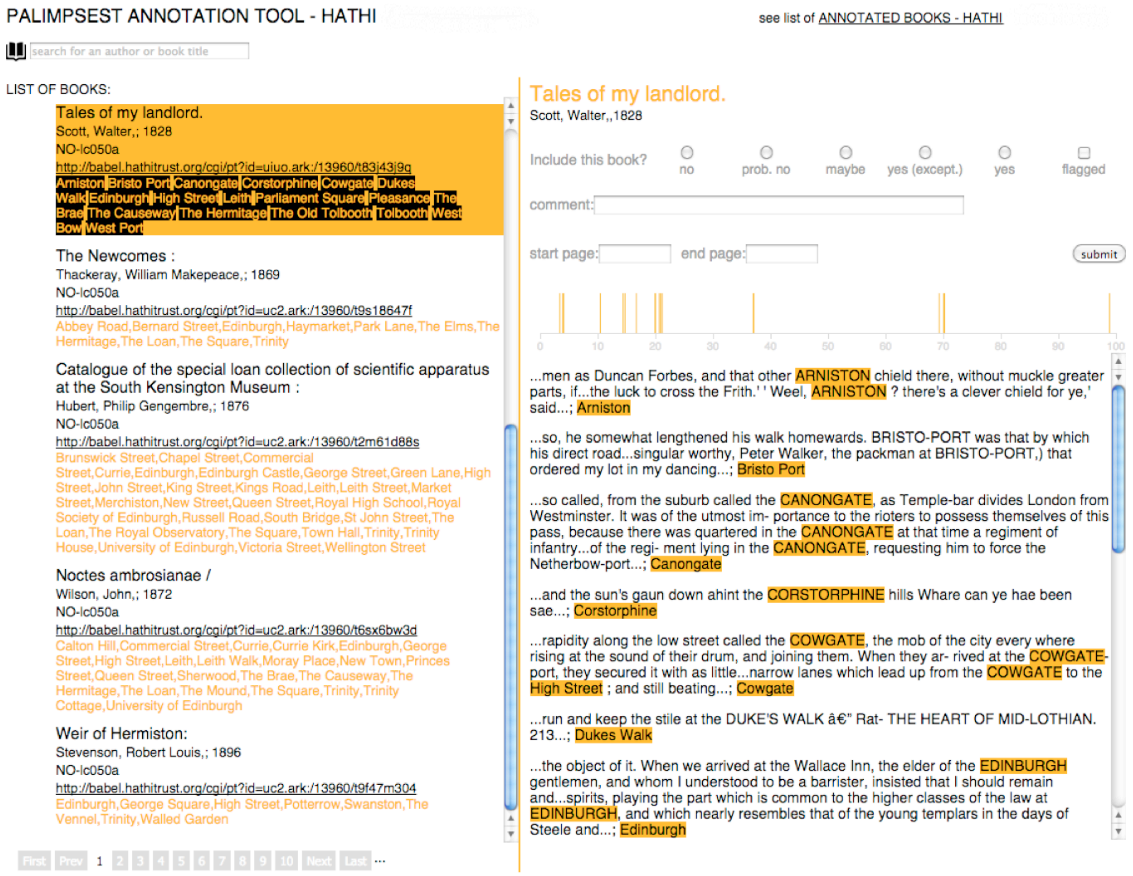
[6] http://yourlibrary.edinburgh.gov.uk/fictionmap

[7] http://www.mappit.net/bookmap/

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

missing less well-known but potentially interesting works. In Palimpsest we consider the entire pool of accessible literature accessible to determine a sub-set of highly ranked Edinburgh-specific candidates automatically using location-based document retrieval. The aim is to uncover a large range of Edinburgh-specific literature, so not only famous and well-read titles. 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). We combine text mining and information retrieval for assisted curation and show how user feedback can improve the technical stages to this process.

The manual annotation of the ranked candidates to select actual Edinburgh-specific literature was done using the annotation tool displayed in Fig. 2. 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, 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 can determine a work as being Edinburgh-specific or not, enter further comments and identify the start and end content pages of a document. When clicking the submit button, a document annotation is saved to the database and disappears from the panel on the left.



PALIMPSEST ANNOTATION TOOL - HATHI

see list of ANNOTATED BOOKS - HATHI

search for an author or book title

LIST OF BOOKS:

Tales of my landlord.
Scott, Walter,; 1828
NO-lc050a
http://babel.hathitrust.org/cgi/pt?id=uiuo.ark:/13960/t83j43j9q
Arniston Bristo Port Canongate Corstorphine Cowgate Dukes Walk Edinburgh High Street Leith Parliament Square Pleasance The Brae The Causeway The Hermitage The Old Tolbooth Tolbooth West Bow West Port

The Newcomes :
Thackeray, William Makepeace,; 1869
NO-lc050a
http://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t9s18647f
Abbey Road,Bernard Street,Edinburgh,Haymarket,Park Lane,The Elms,The Hermitage,The Loan,The Square,Trinity

Catalogue of the special loan collection of scientific apparatus at the South Kensington Museum :
Hubert, Philip Gengembre,; 1876
NO-lc050a
http://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t2m61d88s
Brunswick Street,Chapel Street,Commercial Street,Currie,Edinburgh,Edinburgh Castle,George Street,Green Lane,High Street,John Street,King Street,Kings Road,Leith,Leith Street,Market Street,Merchiston,New Street,Queen Street,Royal High School,Royal Society of Edinburgh,Russell Road,South Bridge,St John Street,The Loan,The Royal Observatory,The Square,Town Hall,Trinity,Trinity House,University of Edinburgh,Victoria Street,Wellington Street

Noctes ambrosianae /
Wilson, John,; 1872
NO-lc050a
http://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t6sx6bw3d
Calton Hill,Commercial Street,Currie,Currie Kirk,Edinburgh,George Street,High Street,Leith,Leith Walk,Moray Place,New Town,Princes Street,Queen Street,Sherwood,The Brae,The Causeway,The Hermitage,The Loan,The Mound,The Square,Trinity,Trinity Cottage,University of Edinburgh

Weir of Hermiston:
Stevenson, Robert Louis,; 1896
NO-lc050a
http://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t9f47m304
Edinburgh,George Square,High Street,Potterrow,Swanston,The Vennel,Trinity,Walled Garden

Tales of my landlord.
Scott, Walter,,1828

Include this book? ○ no ○ prob. no ○ maybe ○ yes (except.) ○ yes □ flagged

comment:

start page: end page: submit

...men as Duncan Forbes, and that other ARNISTON chield there, without muckle greater parts, if...the luck to cross the Frith.' ' Weel, ARNISTON ? there's a clever chield for ye,' said...; Arniston

...so, he somewhat lengthened his walk homewards. BRISTO-PORT was that by which his direct road...singular worthy, Peter Walker, the packman at BRISTO-PORT,) that ordered my lot in my dancing...; Bristo Port

...so called, from the suburb called the CANONGATE, as Temple-bar divides London from Westminster. It was of the utmost im- portance to the rioters to possess themselves of this pass, because there was quartered in the CANONGATE at that time a regiment of infantry...of the regi- ment lying in the CANONGATE, requesting him to force the Netherbow-port...; Canongate

...and the sun's gaun down ahint the CORSTORPHINE hills Whare can ye hae been sae...; Corstorphine

...rapidity along the low street called the COWGATE, the mob of the city every where rising at the sound of their drum, and joining them. When they ar- rived at the COWGATE-port, they secured it with as little...narrow lanes which lead up from the COWGATE to the High Street ; and still beating...; Cowgate

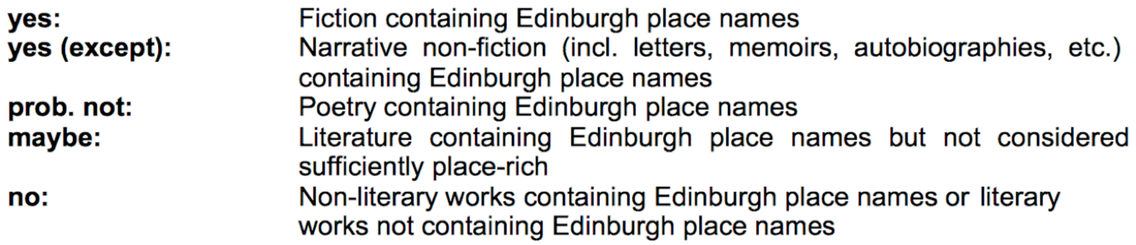
...run and keep the stile at the DUKE'S WALK â€" Rat- THE HEART OF MID-LOTHIAN. 213...; Dukes Walk

...the object of it. When we arrived at the Wallace Inn, the elder of the EDINBURGH gentlemen, and whom I understood to be a barrister, insisted that I should remain and...spirits, playing the part which is common to the higher classes of the law at EDINBURGH, and which nearly resembles that of the young templars in the days of Steele and...; Edinburgh

**Figure 2: Palimpsest annotation tool.**

An item can be annotated using the annotation scheme shown in Fig. 3. We consider documents annotated as *yes* or *yes (except)* as Edinburgh-specific within Palimpsest.[9] The scheme was developed by the annotators while working on an initial ranking of HathiTrust documents.[10]

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



| | |
|---|---|
| **yes:** | Fiction containing Edinburgh place names |
| **yes (except):** | Narrative non-fiction (incl. letters, memoirs, autobiographies, etc.) containing Edinburgh place names |
| **prob. 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 |

**Figure 3: Annotation scheme.**

We used the HathiTrust collection (253,350 documents) to develop the retrieval and ranking component. This resulted in 20,542 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.

Initially, the annotators reacted enthusiastically to the annotation and discovered several works set in Edinburgh which they did not know (e.g. *John and Betty's Scotch History Visit* or *Noctes Ambrosianae*). As they worked through the documents, however, they lost trust in the ranking. They noticed relevant documents appearing far down the list and sometimes had to go through many documents to find a positive example. They also recorded a list of ambiguous place names (*High Street* or *Trinity*) mostly referring to other locations as well as a list of words in titles suggesting non-literary content (*catalogue* or *dictionary*). Finally, they observed that most Edinburgh-specific documents contain a reference to *Edinburgh* or a variant.

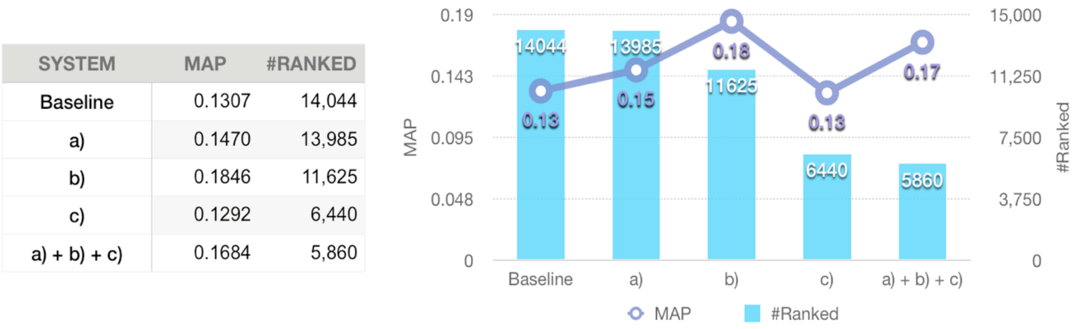
## 4. Improving the Ranking

Based on this feedback, we then fine-tuned the retrieval component. We used the set of 1,710 annotated works as an evaluation set to determine the effect of a modification. 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), the following three measures and their combination.

a) Down-weighting ambiguous place names identified by the annotators.
b) Removing documents containing non-literary title words (*catalogue*, *dictionary*, etc).
c) Ensuring that *Edinburgh* or one of its variants (*Embra*, *Edinburrie*, etc.) occurs in the work.

Fig. 4 shows that down-weighting of ambiguous place names (a) resulted in a small improvement in average precision (MAP) (Baeza-Yates and Ribeiro-Neto, 1999). Filtering documents with non-literary title words (b) had the highest increase in MAP. The condition of Edinburgh or a variant to appear in the document (c) decreased MAP slightly. However, it resulted in a large decrease in the number of ranked documents reducing the workload of the annotators significantly. We therefore consider measure (c) to be beneficial as well. When combining all three measures, the retrieval component yielded an improved MAP score of 0.1684 (compared to the baseline MAP of 0.1307), and the workload of documents to be curated was reduced by 60%.

[10] HathiTrust: http://www.hathitrust.org



| SYSTEM | MAP | #RANKED |
|---|---|---|
| Baseline | 0.1307 | 14,044 |
| a) | 0.1470 | 13,985 |
| b) | 0.1846 | 11,625 |
| c) | 0.1292 | 6,440 |
| a) + b) + c) | 0.1684 | 5,860 |

**Figure 4: Mean average precision (MAP) and number of ranked documents (#RANKED) per retrieval method.**

## 5. Conclusion

The assisted curation process undertaken in Palimpsest attempts to keep the user in the loop during iterative technical development. We received useful feedback from the literary scholars on issues that appeared as they curated documents and considered their suggestions in changing the underlying methods for ranking Edinburgh-specific literature. Our results show that document retrieval performance improved and curation workload was reduced as a result. The improved method was subsequently applied to all document collections which resulted in very positive feedback from the curators reporting that the ranking improved considerably.

## 6. Acknowledgements

Palimpsest is funded by the AHRC (Digital Transformations in the Arts and Humanities - Big Data, PI: Prof. Loxley). We would like to thank Dr. Anderson, Dr. Otty and Dr. Thomson who were in charge of the curation, and Dr. Harris-Birtill who helped with the data processing for the annotation tool.

## 7. References

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. Proceedings of the Pacific Symposium on Biocomputing, pp.556-567.

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

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.

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. Cambridge University Press.

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.