Text mining Mill

Title: Text Mining Mill: Computationally Detecting Influence in the Writings of John Stuart Mill from Library Records

Helen O'Neill[1], Anne Welsh[2], David Smith[3], Glenn Roe[4], Melissa Terras[5]

Affiliation(s):

[1] Department of English, Royal Holloway University of London

[2] Beginning Cataloguing

[3] Khoury College of Computer Sciences, Northeastern University

[4] Faculty of Letters, Sorbonne University

[5] College of Arts, Humanities and Social Sciences, University of Edinburgh

Digital Scholarship in the Humanities, fqab010, https://doi.org/10.1093/llc/fqab010
Published:
27 February 2021

Text Mining Mill: Computationally Detecting Influence in the Writings of John Stuart Mill from Library Records

Abstract

How can computational methods illuminate the relationship between a leading intellectual, and their lifetime library membership? We report here on an international collaboration that explored the interrelation between the reading record and the publications of the British philosopher and economist John Stuart Mill, focusing on his relationship with the London Library, an independent lending library of which Mill was a member for thirty-two years. Building on detailed archival research of the London Library's lending and book donation records, a digital library of texts borrowed and publications produced was assembled, which enabled natural language processing approaches to detect textual reuse and similarity, establishing the relationship between Mill and the Library. Text mining the books Mill borrowed and donated against his published outputs demonstrates that the collections of the London Library influenced his thought, transferred into his published oeuvre and featured in his role as political commentator and public moralist. We reconceive archival library issue registers as data for triangulating against the growing body of digitized historical texts and the output of leading intellectual figures. We acknowledge, however, that this approach is dependent on the resources and permissions to transcribe extant library registers, and on access to previously digitized sources. Related copyright and privacy restrictions mean our approach is most likely to succeed for other leading 18th and 19th century figures.

Introduction

How can we understand the relationship between the books an author consults in a library, and those they write? How can computational methods be used to trace how one individual

library has affected the work and public interventions of an author? Under what circumstances will this be feasible, possible, or practical? We report here on an international collaboration that aimed to explore these issues via the reading and writings of the British philosopher, economist, and politician John Stuart Mill (1806-1873), focusing on his relationship with the London Library, an independent lending library in London, U.K., which Mill was an engaged member of for thirty-two years. A detailed archival research of the London Library's lending and donation records, followed by an assembly of a digital library of both these texts and the publications Mill produced, enabled text mining and natural language processing (NLP) approaches to detect textual reuse and similarities between passages of writing in the texts, and further close reading to establish relationship, context, and meaning between works borrowed, and works written.

Building on a closely documented analysis of the archival record, and related synthesis of the results (O'Neill 2015, 2016, 2019) it is demonstrated that, in text mining the books John Stuart Mill borrowed from and donated to the London Library against his published outputs, it is shown that the collections of the London Library influenced his thought, transferred into his published oeuvre and featured in his role as political commentator. Intense periods of reading around a common theme can be identified in authorial practice, and books Mill consulted from the London Library can be found referenced extensively in his publications, showing this institution's importance to his work and public life, which had been previously unrecognised (O'Neill 2015, 2016, 2019). This paper concentrates on the computational approach used to underpin these findings.

Our findings will be useful to others wishing to compare and contrast the content and product of libraries regularly consulted by authors: we show that this combined archival and digital

approach is an effective and efficient means to interrogate the historical borrowing record of a leading intellectual figure. We also demonstrate we have extended the remit of research that can be undertaken with authorial libraries, library issue registers, and borrowing records, fundamentally reconceiving library issue records as data which can be used at the start of a digital continuum. We demonstrate how the triangulation of borrowing record, growing access to digitized resources, and the use of computational tools developed for literary study, holds a rich nexus for 19th century author and bibliographic studies. We also show that this approach is one which depends on interdisciplinary researchers working alongside computational methods, rather than researchers depending entirely on automation. However, our approach has various dependencies: it is reliant upon the survival of historical and archival issue records, and on gaining the permissions and resources to consult them fully; it is dependent on full text access to previously digitized resources (only a selection of which may be available due to copyright and other restrictions); the applicability of this method may be limited to other leading 19th Century intellectuals, given ethical concerns and additional complexities of modern privacy legislation.

John Stuart Mill and the London Library

J. S. Mill (1806-1873), the influential 19th Century British philosopher, civil servant, and political theorist, actively contributed to the fields of logic, ethics, philosophy, social theory, and political economy, foregrounding utilitarian and liberal approaches (Packe 1954, Ryan 1970, Hollander 1985, Reeves 2015). Mill shaped 19th Century British and international political discourse with his extensive publication record, including *Principles of Political Economy* (1848), *On Liberty* (1859), *Utilitarianism* (1861), *The Subjection of Women* (1869), and *Three Essays on Religion* (1874), all of which, naturally, cite other sources. Understanding how the books he read fed into his own published output can help us follow

the development of his thought and influences. Somerville College, Oxford, holds the best-known collection of Mill's books (one thousand, six hundred and seventy-four volumes)[i]. Less well known is Mill's membership and use of The London Library[ii], itself a surviving Victorian institution: a subscription lending library, operating in central London since 1841 (Harrison 1907, Nowell-Smith 1958, 1972, Baker 1988, 1990, 1992, Lynn 2006, McIntyre 2006, O'Neill 2015, 2016, 2019). Mill began his relationship with the London Library in 1840 as an expert subject adviser for the acquisition of books on Political Economy, Logic, and French Histories and Memoirs before the Library opened, becoming a founder life-member, and remaining in membership records until his death in 1873 (O'Neill 2016, p. 257; 2019, p.187). Mill consulted books in and donated books to the London Library over thirty-two years, but prior to O'Neill's transcription and subsequent analysis (2016, 2019) his loan record had never been fully understood. His name as an early book donor had been spotted (Baker 1992) and a donation of twenty-four titles had been documented (O'Neill 2019, p.186) but the number and variety of Mill's book donations was unknown as the books had been integrated into the London Library's general holdings. The scale of Mill's prolific use and enrichment of the Library's collections, and the influence that these holdings had on his own outputs, was not known until O'Neill's detailed archival work, combined with computational analysis described here, to determine both the extent and significance of Mill's loan record and book donations. The four hundred and thirty books Mill is now known to have borrowed from, and one hundred and sixty-five titles he donated to the London Library (O'Neill 2015, 2016, 2019) form a substantial bibliographic backdrop to the work of a pre-eminent Victorian thinker and throw light on a singular, but little researched Victorian institution.

This paper describes how computational comparison was undertaken on a digital corpus of books compiled by O'Neill, following the compilation of the list of books in the London Library that Mill was known to consult over his career. Our experimental research developed from online conversations with the Digital Humanities community on how to best approach our problem from a computational angle (Terras 2013). The research question was: how could we computationally compare the texts of the books known to have been consulted by Mill against publications written by Mill, to triangulate an evidential relationship between the books Mill borrowed, donated and wrote? Could bibliometric analysis demonstrate the importance of the London Library to his research and thinking? What methodological issues, opportunities, and limitations does this present, for others contemplating comparing the known reading habits of an author to their written output?

Related Previous Research

As James Raven argues persuasively, "there cannot and should not be one type of history of the book, but many types" (Raven, p. 141) and so notwithstanding general agreement on basic techniques in Bibliography set out by Bowers, and Gaskell, and more recently by Tanselle and Werner, it is always necessary to identify the type of bibliographic methods that are being used in this research. Much has been written about authors' libraries to reveal their authorial habits (Sealts, 1966, and Olsen-Smith, Norberg and Marnon, 2009-2019 on Melville; Reynolds 1981 and 1986 on Hemingway; Capps, 1966 and Miller, 2012 on Dickinson), in particular, the links that can be drawn between the items read and those written through careful scholarly research (Sealts, 1982 and Faflik, 2018 on Melville; Tyler, 1995 and Jungman and Tabor, 2000 on Hemingway; Coleridge, 1980-2001, Jackson, 2005, and Leinwand, 2016 on Coleridge). However, to date, we have found minimal previous research that has attempted to use computational approaches to detect and analyse the

influence of authors' libraries upon published outputs: an analysis of Melville's marginalia used digital text analysis to identify sources in his surviving copies of Homer, Shakespeare, and Milton (Ohge and Olsen-Smith 2018, Ohge et al 2018); and Antonini et al (2020) have explicity suggested marginalia as a loci for work linking between pre-digital and digitised texts to "enhance the digitisation efforts of authorial libraries by producing interoperable and comparable digital sources" (10)). We employ Digital Humanities tools to interrogate transcribed and collated borrowing records (previously held only in issue registers and a range of institutional archival sources) and provenance information for books donated by certain members (previously available only in dispersed paper records and ownership marks on the books themselves). As scholars including Pearson, Oram, and Darnton have highlighted, research into such sources is significant within the history of libraries, and more widely within cultural history. As early as 1939, Keynes asserted that "The mind of Man is recorded in his books, and the catalogues of the great libraries enable the individual to consult the universal mind," (1980, 3): early studies acknowledged the need for assistance from library staff with access to institutional records and expertise in unpicking them (see, for example, Keynes, but also Harding on Thoreau). Previous research on the London Library issue registers has depended on this supported and approved access, but required additional, intense close reading of often difficult to decipher and transcribe content[iii] (Baker 1981, Atkinson 2013), as was the case in O'Neill's foundational work (2015, 2016, 2019), although earlier work did not then rely on digital methods, to extrapolate findings further.

Gribben's call in 1986 for collaboration between historians, computer scientists and librarians has been realized in many digital author library projects, including Melville's Marginalia Online[iv], The Freud Library (Davies and Fichnter), The Gladstone Reading Database[v], and the multi-national and crowdsourced RED, The Reading Experience Database[vi]. However,

digital research upon author's libraries "is currently limited to the provision of either digital catalogs that make library metadata available … or of simple digital copies, which are offered in a viewer and/or as a PDF download" (Busch et al 2019, although they tackle this by developing a prototype visualization of Theodor Fontane's Library, in particular to identify patterns in marginalia). A recently funded project, Books and Borrowing 1750-1830: An Analysis of Scottish Borrowers' Registers[vii] (2020-23), aims to reveal hidden histories of book use, knowledge dissemination and participation in literate culture, but is yet to report. The advances in this paper move beyond digital catalogue or visualization, demonstrating the affordances of sequence alignment techniques to identify textual matches between items borrowed and items written by an author at scale.

The identification of similarities and relationships between passages in large collections of historical texts - including direct quotations, commonplace expressions, plagiarisms, and other forms of borrowings - is of great interest to a variety of humanities scholars, as it can advance our understanding of influence, writing habits, and ethical approaches in a writer's work, while "placing it in a larger intellectual and cultural context" (Olsen et al 2011). Relationships between texts are complex and often multi-faceted, ranging from directly attributed quotations to influences and allusions, and a key approach in humanistic study is tracing these relationships (Jardine and Grafton 1990). Intertextuality is a rich area of technical and theoretical research development, requiring collaboration between computer science and the digital humanities to build upon and utilise the growing number of digitised texts available to researchers via mass digitisation sourced from either commercial (Google Books, Microsoft, Gale Cengage), government (Library of Congress, Bibliothèque nationale de France) or non-profit (Internet Archive, HathiTrust, Project Gutenberg) providers (Smith et al 2019a, Olsen 2009). Machine-assisted reading can be used to identify intertextuality,

particularly when "faced with the intricacies of text recycling in historical and literary works, along with the frequently degraded status in which these texts are currently made available" (Olsen et al 2011), although it has been argued that this is an "undertheorized" practice in the Humanities (Underwood 2014).

Here, we are concerned with Text Recycling, or local text reuse, which identifies small regions of similarity, ignoring large amounts of difference, predicated on the pairwise comparison of many documents to identify typically infrequent instances (Seo and Croft 2008)[viii]. There are many NLP approaches that can be used to do so (see Graham 2019, and Smith et al 2019b for overviews). Sequence alignment, which "divides the source and target strings into overlapping sets of consecutive words… called 'shingles' or 'n-grams'" (Graham 2019, p. 122) is widely used in bioinformatics, and as the basis for many plagiarism detection algorithms (Lyon et al 2001, Bourdaillet and Ganascia 2007). However, it has also been used by humanities scholars to detect sources, influence and allusion in historical texts: in Classical Latin poetry (Bamman and Crane 2008), in the 18th-century *Encyclopédie* of Denis Diderot and Jean d'Alembert (Edelstein et al 2013, Roe 2018), in detecting reuse of Homeric epics across 15 million words of Greek and 10 million words of Latin (Büchler et al 2012). Coffee et al examined allusions to Vergil's *Aeneid* in the first book of Lucan's *Civil War* (2012). Büchler et al extract relationships between different English editions of the Holy Bible (2014). Franzini detected similarities in English translations of the Polish romantic epic *Pan Tadeusz* by Adam Mickiewicz (2016).

Sequence alignment algorithms vary in complexity and resulting computational tractability. An alternative approach using n-gram matching, for example, is presented in Ganascia et al (2014). Smith et al (2013) detect clusters of reused texts to analyse the culture of reprinting in

newspapers in the United States before the American Civil War, refining the n-gram shingling approach to optimize effectiveness and efficiency by employing hashing for space-efficient indexing or repetition and local alignment techniques to find compact passages with the highest probability of matching. This approach was also used to trace the flow of policy ideas in legislation (Wilkerson et al 2015, Funk and Mullen 2018). Recent developments in this method have also included visualization of results to support interpretation (Abdul-Rahman et al 2017). However, although computational detection of textual reuse is becoming an established method in humanistic study, we have uncovered no previous application of this approach to authors' libraries or borrowing records.

Method

Our research consisted of four distinct stages. Firstly, O'Neill compiled the list of Mill's borrowing record of books held within the London Library, which was foundational archival research on both loans and donations records (2015, 2016, 2019). Secondly, O'Neill compiled a digital corpus of books written by Mill, from extant online sources (O'Neill 2016, 2019). Thirdly, sequence alignment NLP approaches to align subsequences and then cluster common passages were used to identify commonalities in texts between the books Mill wrote, and those he read. Fourthly, analysis of the results of text mining enabled understanding of the relationship between Mill's London Library borrowing record and his published output (O'Neill 2016, 2019). We detail our approach, its successes, and its shortcomings, here.

This research was given ethical approval from UCL Department of Information Studies. Given the timescale of the author records in question, there are no concerns regarding the

General Data Protection Regulation, or the need to obtain permissions from the individuals involved.

Library Record compilation

This research depended on the time-consuming, detailed archival work with the library's extant loan records (see figure 1) undertaken by O'Neill, and the permission from the London Library to do so. The challenges of such archival work are presented in O'Neill (2016, 258; 2019, 190). Additionally, identifying books donated by Mill within the collection required forensic and extensive consulting of thirty-four years of internal Library administrative records, catalogues, and supplements (O'Neill 2016, 260; 2019, 190). The extracted loans data presents a unique corpus of four hundred and thirty books consulted by Mill, albeit for a finite period from the early part of his membership (1842-1849 and 1856-1857), given the extant London Library issue registers: he is therefore likely to have consulted far more over his membership. This may be a topic for future research with these methods: estimating the probability that Mill quotes from books within the London Library, using their cataloguing and accession records to compile a wider corpus, and detecting matches in his output. In addition, Mill's donations were marked by three significant deposits over three decades, totalling one hundred and sixty-five titles (see O'Neill 2016, 269-276 or 2019, 379-390 for a complete listing). The records of the books loaned and donated were transcribed and entered into an Excel spreadsheet in order to enable further analysis.

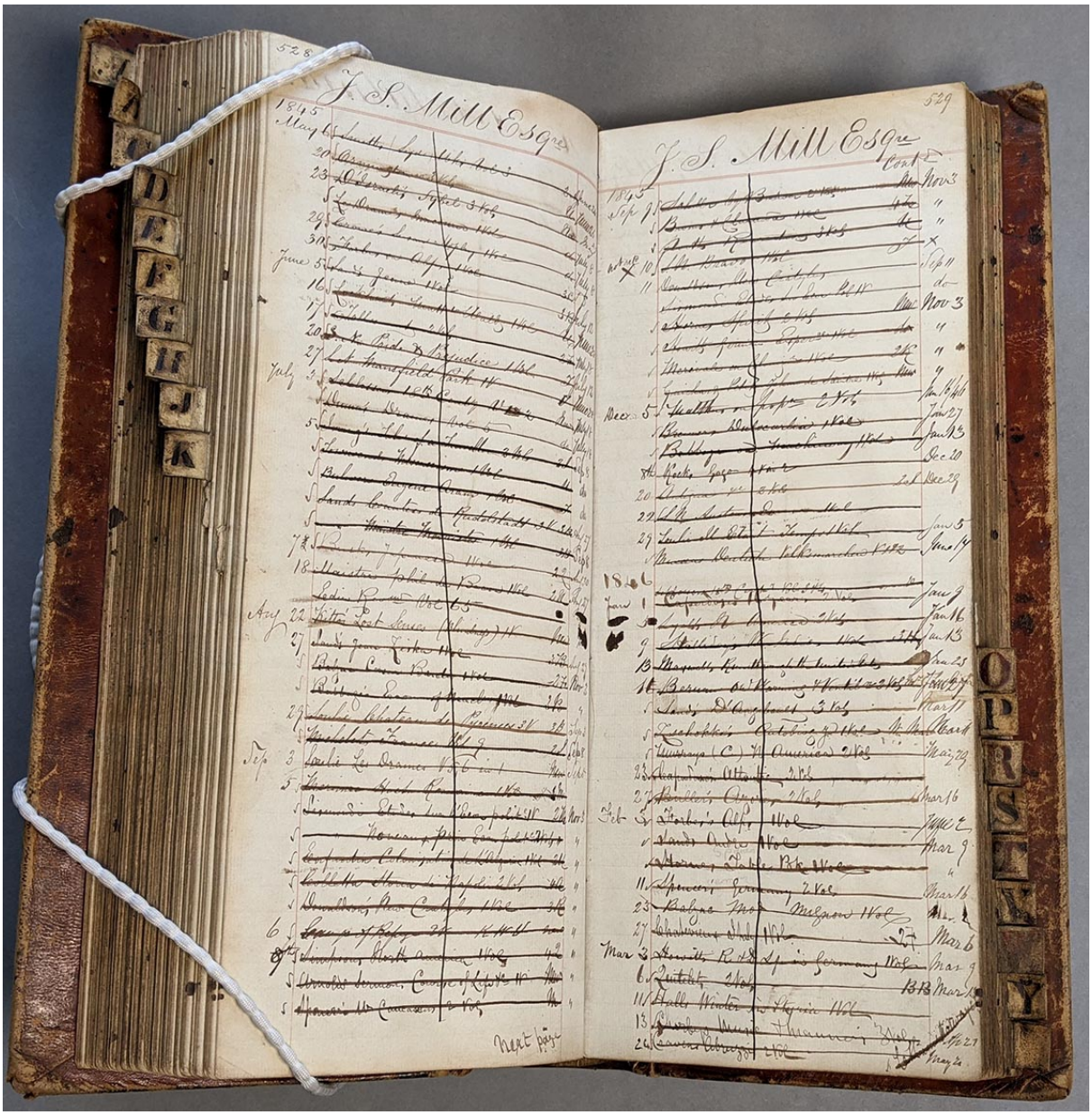


Figure 1: London Library Issue Book No. 3 showing Mill's intensive borrowing record during 1845, London Library Issue Book 3, 529. The horizontal lines indicate the return of individual books. The vertical lines indicate that all the books listed on the page have been returned. This is representative of the type of library issue record that required transcription and identification from Mill's loan record. Image reproduced with the kind permission of the London Library. © The London Library.

Assembling the Virtual Library

O'Neill attempted to source digitized, machine-searchable text of the five hundred and ninety-five book titles Mill was shown to have consulted within the collections of the London Library, being careful to identify exact editions required where possible, from the Internet Archive[ix]. While the digitisation of texts "afford opportunities for more extensive, data-rich and quantitative approaches to literary historical scholarship" (Bode 2012, p. 1) it was unfortunately not possible to locate previously digitized versions of all of the titles. Two hundred and fifty-five of the four hundred and thirty books Mill borrowed (59%), and ninety-one of the one hundred and sixty-five books he donated (55%) were obtained in machine-processable format. A limitation to this research approach is the still patchy digitisation landscape (Nauta et al 2017). We also acknowledge the limitations that poor OCR of digitized texts can inject into this process, and that depending on the quality of previously digitized content can affect research outputs in unknown ways (Cordell 2017).

Given Mill heavily revised certain monographs, we were dependent on the scholarly edition of Mill's *Collected Works* (Mill, ed. Robson 1963-91, henceforth referred to as *CW*), available in an accessible format in the Online Library of Liberty[x], which facilitated and accelerated close reading of textual matching. Our choice of subject matter is only suitable because of this detailed prior work on Mill's publishing history.

Text mining approaches

We benefited from two text different text mining approaches previously designed for the detection of textual alignment, with one being a computationally light-weight approach, the other which involves significantly more resource, in the hope that we would be able to identify matches. The intention was not to compare these tools *per se*, however, using two

available systems which differ in approach and execution allows insight into where these tools may benefit others.

We first used TextPAIR[xi] (Pairwise Alignment for Intertextual Relations), an open-source software package developed by Roe and colleagues as part of the ARTFL Project at the University of Chicago for text reuse discovery in digitized text collections, originally implemented in 2009, and rewritten in 2018[xii]. TextPAIR is an implementation of a very general sequence alignment algorithm for humanities text analysis that supports one-against many comparisons using a generalized Python module. Sequence Alignment respects order in documents, and can align similar passages directly, dealing with variations in similar passages such as insertions, deletions, spelling, OCR errors, etc. TextPAIR identifies regions of similarity shared by strings using word or k-tuple heuristics in order to balance efficiency and completeness while identifying occurrences of the same word sequences shared between documents (see Roe 2012 for further documentation on TextPAIR's approach, and use of TextPAIR in Olsen et al 2011, Edelstein et al 2013, Kokkinakis and Malm 2015). A benefit of using TextPAIR is that the results are stored in individual files associated with each source document, sorted chronologically by year of document publication, where the start of the matched passage is highlighted. This makes for a quick way to see borrowings and quotations, and for a researcher to return to results and identify linked passages. Parameters can be adjusted to loosen or tighten the degree of similarity. In our case, searches were run by Roe on a dedicated server at the Oxford e-Research Centre running the associated PhiloLogic search and retrieval software also developed by ARTFL. Our source and target datasets were indexed and loaded into PhiloLogic before using TextPAIR to pre-process the texts into overlapping tri-grams, or the three-word shingles used for identifying similar passages between corpora. We settled on matching parameters of 10 or more words occurring within a

sliding window of matched n-grams to avoid over-fitting of many banal expressions, which output an appropriate number of results for a researcher to return to for analysis.

Using the same set of source and target texts, we then used Passim[xiii], implemented by Smith in 2012 at Northeastern University and since continually improved (Smith 2019), which uses probabilistic approaches to text-reuse analysis to successfully detect alignments between noisy OCR sources. The software performs an initial filtering stage using n-gram shingling and then implements the Smith-Waterman algorithm (Smith and Waterman, 1981) with an "affine gap penalty", which encourages inserted/deleted passages to be more compact. For this corpus of books, we treated each page as an independent document and had passim return a maximally aligned subsequence in each pair of pages. As with TextPAIR above, we pruned away aligned passages with fewer than ten words. (See Smith et al 2019c for a full overview of the implementation.) Openly available under the Eclipse Public License, Passim has been successfully used in a variety of studies and projects (see Smith 2013, Wilkerson et al 2015, Vesanto et al 2017, Smith 2019). Passim detects pairs worth aligning where textual variation and OCR errors mean that more straightforward approaches are less robust, but it is therefore more computationally expensive in both memory and time than pure shingling n-gram methods (although the code can be parallelised via Apache Spark, either on a single machine or a cluster). The results are a set of aligned text passages, highlighting matches between source and target texts, providing reference to document name and page number, allowing the researcher to return to both to undertake close scrutiny. Running Passim on the corpus of books required less than thirty minutes on a fourteen-node cluster at Northeastern. It is imperative to note that this research is not an attempt at distant reading (Underwood 2017), or purely quantitative literary analysis making "a false claim to absolute knowledge and objective truth" (Bode 2012, p. 10). The results from our searches required extensive

close reading and synthesis from the researcher, Helen O'Neill, in a process that combines "digital and computational methods with traditional modes of literary analysis" (Rosen 2011). The distant reading approaches used here are a "supplement to traditional close reading practices", as an example of how "the invaluable resource of digital archives and the utility of searchable databases can be most rewarding when deployed in concert with close reading, archival research skills, and careful argumentation" that "attend to the complexity and contingency of historical phenomena" (ibid). The computational analysis therefore pinpointed where close reading analysis should occur, allowing us to "quantify without losing the disruptive detail and splitting significations to which we have learned to attend" (Rothberg, 343), as a "productive way of integrating empirical data with the paradigm of humanities knowledge as a critical, analytic and speculative process of enquiry" (Bode, 2012 p. 8). The scale and scope of the texts in question requires computational approaches for the identification of potential matches to be feasible; however, the results from this process require in-depth human synthesis and analysis to understand trends and assign meaning.

Textual alignment methods have a bias, as mentioned above, towards high-precision, surface-level matches. Other research projects to apply text-reuse methods to literary influence—such as the Tesserae project at Buffalo (Coffee et al 2012) or the eTRAP project at Göttingen (Franzini 2016)—have focused significant effort on improving recall, e.g., by parameterizing textual variation with synonym dictionaries and part-of-speech substitution rules. However, alignment methods are useful as null models of textual influence. Since each mutation of a text in passing from source to destination is equally likely, we can establish a baseline for future investigations that account in a more nuanced way for authors' transmutation of their sources. In a similar way, null models of gene drift establish a baseline against which certain genes may be deemed adaptive. Textual alignment focuses our attention on the most likely

matches. While we can evaluate the (generally high) precision of these methods, it is impractical to perform an exhaustive evaluation of recall. The more one relaxes the matching parameters - to attempt to capture allusion, or "indefinite or diffused source" (Altick 1975, 94) for example - the more noise is introduced into the system, which can often overwhelm the signal of re-uses.

For this project, the use of metadata in triangulating with ownership and borrowing records allows us to check model output with independent observations to some extent, we are well aware that subtle allusions, unconscious borrowings, and lapses of memory may pass unrecorded (and may be better served by alternative methods such as topic modelling, or stylometric analysis). This uncertainty at the level of individual instances of text reuse, however, can be mitigated by aggregating our analysis at the level of books—which also happens to be the level of our bibliographic analysis. While books whose only contribution to Mill's writing was indirect may thus escape detection, we are more confident in finding the books that made some direct textual contribution. State-of-the-art language models exhibit a growing sensitivity to a range of genres and long-distance dependencies among lexical choices. While these capabilities have to date been fine-tuned on fairly shallow paraphrase, translation, and question-answering tasks, we expect that future directions in text-reuse research will focus on systems that combine search in dense contextual embedding spaces with models of text mutation trained on collections of documents and their sources.

Results

Understanding the Borrowing Record

The compilation of Mill's borrowing record is fascinating in itself as a snapshot of the zeitgeist of his age. From the works of leading European economists, philosophers, and historians, to children's books, it reveals Mill's lifelong interest in and affection for all things French; his active engagement with European culture; his attentiveness to women's writing, actions and opinions; and his focus on the economic, political, social and cultural developments in countries, colonies and continents across the globe. For a complete analysis of Mill's loans and donations by title and theme, see O'Neill (2015, 2016, 2019).

Results from Text Mining

Of interest here is the textual matching between loan and publication, where text mining has highlighted important further influence on Mill's thinking, that may have gone undetected by keyword searching. This allows us to establish a direct link between books Mill borrowed from the Library, his published oeuvre, his political interventions and his public profile, which would not have been possible without using computational methods due to the scale and extent of the task.

Seven hundred and ninety-five text matches of strings at least ten words long were found between the "source" (Loans, Donations) files and the "target" (Publications) via the TextPAIR approach, and one thousand eight hundred and sixty-three were found via the Passim system. The difference in these numbers can be explained by the difference in tolerances between each system's algorithmic approach to matching, and correction for errors in OCR. There were many false positives, or more accurately, matches of texts within the digital files that are not necessarily useful to our cause. Some of these are due to the nature,

form and content of the digitized texts, and artefacts from the digitization process the systems were comparing, such as “the borrower will be charged an overdue fee if this book is not returned to the library on or before the last date stamped below”, or “please do not move cards or slips from this pocket” which made it into the OCR-generated text! Some of these matches reflected use of common aphorisms or phrases: “an eye for an eye and a tooth for a tooth” was found in nine matches between source and target, “either to the right or to the left and that” was found in eight. The overestimate of the significance of such common phrases results, in part, from the focused input collection. A larger corpus would help both models of text reuse better infer the relative frequency of these phrases or, as in the case of the biblical quotation above, explain away their co-occurrence in two books as arising from a common third book.

Both systems matched the same five hundred substantive matches within source and target texts, often between multiple editions of Mill’s works. These became of immediate relevance for comparison with the *CW* to establish the relationship between the published item and source, and to determine what these matches told us about Mill. Smith’s Passim system identified significantly more potential matches where the OCR transcriptions were poor, which then required returning to the source texts, and manually checking details, for corroboration. The outputs of the text mining therefore provide a starting point for the detailed analysis, which requires much human synthesis to rationalize and establish relevance and conclusion, rather than a fully automated process.

A major result of the text mining and close reading analysis was that Mill clearly cited his sources: we do not identify any significant uncited or newly discovered influence. However, this computational approach (as postulated by Olsen et al 2011) significantly improved “the

manner in which these relationships are linked from text to text. Rather than parsing a reference and link using citation data or outside references schemes – which can be highly variable, inconsistent, and typically keyed on page number of other rather arbitrary attributes" we identified and contextualized links and relationships in an efficient manner (ibid). Doing this manually would have been possible given Mill's use of citation for these specific sources, but would have been a life's work. Our major finding is to show how the results of sequence alignment can indicate important influence of particular individuals, particular texts, references from particular genres of text (such as French literature), and around specific historical events (such as the Great Famine in Ireland).

Our results give an indication of importance of influence, firstly by the number of times specific authors or their texts are referred to by Mill. The speeches of the statesman and philosopher Edmund Burke, who was staunchly opposed to the 1789 French Revolution, were borrowed by Mill in May 1848 (Burke 1826-7): there are forty-six references to Burke in Mill's publications, showing his importance on Mill's argumentation. During the 1865 election in which Mill stood for Westminster, he quoted Burke on the hustings (Mill *CW*, Vol. XXVIII, 45) and recognised in him the significance of principled action:

> What was it which made Edmund Burke, with all his errors …soar so immeasurably above the vulgar orators, and still more vulgar statesmen of his day? What, except that he was a man of general principles? (Mill *CW*, Vol. IV, 115).

Mill's respect for Burke is evidenced through the repeated quotation throughout his publications.

Secondly, the text mining allows us to identify parallels between significant books read, and significant outputs. The text mining results from both systems generated key textual matches

which identified Mill's germinal work *The Principles of Political Economy with some of their Applications to Social Philosophy* (*PPE)* (1848) as a target work of significance (understandable given that Mill would have been reading extensively for this in the period covered by the transcribed loans records). For example, a ground-breaking work on political economy appears in Mill's loan record: *Economy of Machinery and Manufactures* (Babbage 1835) which established Charles Babbage's reputation as a European authority on factories and workshops in England and on the continent. Babbage's work is cited frequently in *PPE* and eleven substantial portions quoted: indeed, at one point, Mill stresses "I still quote Mr. Babbage" (Mill *CW*, Vol. II p.136). In discussing the association between labourers and capitalists, Mill used payments to crews of whaling ships as an example, further synthesizing Babbage's arguments to add to his own:

> Mr. Babbage, who also gives an account of this system, observes that the payment to the crews of whaling ships is governed by a similar principle; and that "the profits arising from fishing with nets on the south coast of England are thus divided: one-half the produce belongs to the owner of the boat and net; the other half is divided in equal portions between the persons using it, who are also bound to assist in repairing the net when required." Babbage has the great merit of having pointed out the practicability, and the advantage, of extending the principle to manufacturing industry generally (Mill *CW*, Vol. III, 1013).

This long quotation from Babbage, which appeared in the first and second editions (1848, 1849), disappears from the 3rd (1852), indicating that Mill continued to revise his reasoning. However, text mining reveals here that Babbage's *Economy of Machinery and Manufactures,* consulted by Mill from the collection of the London Library, operated as a crucial and central influence when Mill was writing *PPE*.

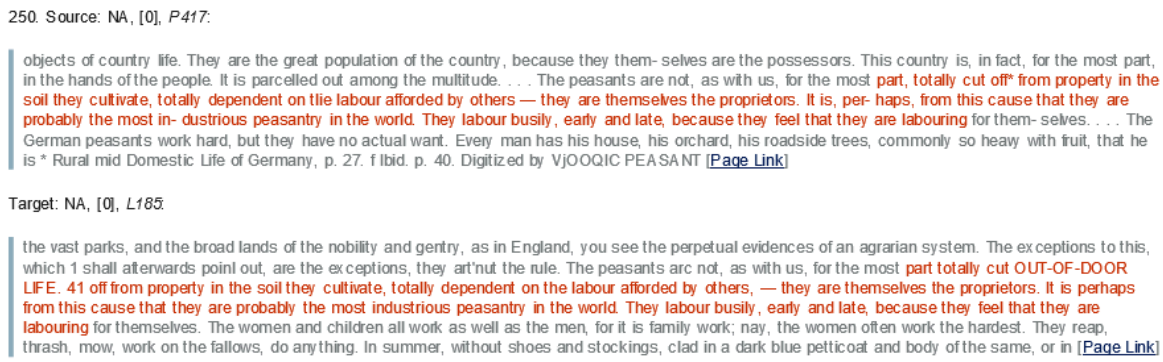
Thirdly, text mining allows us to reinforce the importance of Mill's international influences, such as French literature and history, subjects on which he was particularly knowledgeable. Mill's active consumption of the French novel, revealed by his borrowing record, shows a heightened engagement with political, philosophical and economic thinking in the dominant cultural medium of the day, much read by the London intelligentsia (Atkinson 2013). In his essay on the writing of Alfred de Vigny Mill contrasted Sue's "Literature of Despair" with de Vigny's "touching and beautifully told stories, founded on fact" (Mill *CW*, Vol. I, 488). Three significant French historians appear in Mill's loans record: Jules Michelet; François Mignet, and the French speaking Genevan, Simonde de Sismondi: all three were extensively reviewed by Mill. Between 1826 and 1849 Mill reviewed Mignet's *French Revolution* (1826); *Scott's Life of Napoleon* (1828); *Alison's History of the French Revolution* (1833); *Carlyle's French Revolution* (1837); *Michelet's History of France* (1844); Guizot's *Essays and Lectures on History* (1845); Duveyrier's *Political Views of French Affairs* (1846) and wrote impassioned essays on *Armand Carrel* (1837) and *A Vindication of the French Revolution of 1848* (1849). Two French economists and one French speaking Genevan appear in Mill's loans record, all of whom are directly quoted in *The Principles of Political Economy with some of their Applications to Social Philosophy* (*PPE)*: Charles Dunoyer, M.H. Passy and Simonde de Sismondi. Passy's work *Des Systemes de Culture, et de leur influence sur L'Economie Sociale* (Passy 1846) is referred to fifteen times in in relation to peasant and capitalist farming. Sismodi's *Nouveaux Principes d'Économie Politique* (1819) and *Etudes sur L'Économie Politique* are referred to over fifteen times and are also cited in Mill's articles on the condition of Ireland during the Great Famine. *De la Liberté du Travail* (Dunoyer 1845) by Dunoyer is particularly praised by Mill in *PPE:*

> In M. Dunoyer's work will be found, what we in general vainly seek from political economists, a clear view of the relation between the political economy of any society

> and its state of general intelligence & of moral and social improvement; nor am I aware of any other work in which this important relation is traced out in anything like similar detail (Mill *CW*, Vol. II, 111n).

Strong parallels can therefore be drawn between Mill's consultation of the Topography collections in the holdings of the London Library, and his publications.

Fourthly, citation practices can be shown to cluster around specific texts and topics (O'Neill 2016, 2019). A direct link between books borrowed and published can be found in the citations within *PPE*, and also in Mill's forty-three *Morning Chronicle* articles (1846 -1847) on "The Condition of Ireland" which were precipitated by the Great Famine, when he was also writing *PPE* (Mill *CW*, Vol. 1, 879). *Rural and Domestic Life in Germany* by William Howitt (Howitt 1842) was borrowed by Mill in March 1846 and is quoted by him in *PPE* as "Evidence Respecting Peasant Properties in Germany" (Mill *CW*, Vol. II, 262).


250. Source: NA, [0], *P417*:

> objects of country life. They are the great population of the country, because they them- selves are the possessors. This country is, in fact, for the most part, in the hands of the people. It is parcelled out among the multitude. . . . The peasants are not, as with us, for the most part, totally cut off* from property in the soil they cultivate, totally dependent on tlie labour afforded by others — they are themselves the proprietors. It is, per- haps, from this cause that they are probably the most in- dustrious peasantry in the world. They labour busily, early and late, because they feel that they are labouring for them- selves. . . . The German peasants work hard, but they have no actual want. Every man has his house, his orchard, his roadside trees, commonly so heavy with fruit, that he is * Rural mid Domestic Life of Germany, p. 27. f Ibid. p. 40. Digitized by VjOOQIC PEASANT [Page Link]

Target: NA, [0], *L185*:

> the vast parks, and the broad lands of the nobility and gentry, as in England, you see the perpetual evidences of an agrarian system. The exceptions to this, which 1 shall afterwards poinl out, are the exceptions, they art'nut the rule. The peasants arc not, as with us, for the most part totally cut OUT-OF-DOOR LIFE. 41 off from property in the soil they cultivate, totally dependent on the labour afforded by others, — they are themselves the proprietors. It is perhaps from this cause that they are probably the most industrious peasantry in the world. They labour busily, early and late, because they feel that they are labouring for themselves. The women and children all work as well as the men, for it is family work; nay, the women often work the hardest. They reap, thrash, mow, work on the fallows, do anything. In summer, without shoes and stockings, clad in a dark blue petticoat and body of the same, or in [Page Link]

Figure 2: Mill's use of *Rural and Domestic Life in Germany* by William Howitt (1842) quoted in *Principles of Political Economy* (1849), as noted by TextPAIR. There are various differences in spellings and in OCR quality, but the match is still detected.

This passage is quoted (and cited) again as testimony for introducing peasant properties in Ireland in "The Condition of Ireland [24]" which appeared in the *Morning Chronicle* on 30 Nov 1846 (Mill *CW* Vol. XXIV, 984).

Mill borrowed David Henry Inglis's work *Switzerland the South of France and the Pyrenees* (Inglis 1827) in November 1846 and July 1847, which he quotes in *PPE*. He also quotes Arthur Young, *Travels during the Years 1787, 1788, and 1789 in France* (Young 1792), which he donated to the Library in April 1841, as evidence in support of his argument for the introduction of peasant properties in Ireland, both in "The Condition of Ireland [29]" which appeared in the *Morning Chronicle* on 9 December 1846 (Mill *CW*, Vol. XXIV, 984), and in the densely referenced second volume of *PPE* (Mill *CW*, Vol. II, 256-8, 273). Mill quotes from Arthur Young's *Travels in France* throughout *PPE* in relation to the productivity of peasant proprietors (Mill *CW,* Vol. II, 273, 274-5, 276, 278, 291, 298, 301-5) and also quotes him in five of "The Condition of Ireland" articles (Mill *CW,* Vol. XXIV, 957-8, 968, 985, 1004, 1018, 1049, 1061), using Young's statement "The magic of property turns sand into gold" four times:

> An authority on this point, not to be disputed, is Arthur Young…travelling over nearly the whole of France in 1787, 1788, and 1789, when he finds remarkable excellence of cultivation, never hesitates to ascribe it to peasant property…Young notes 'The magic of property turns sand to gold. Give a man the secure possession of a bleak rock, and he will turn it into a garden; give him a nine years' lease of a garden and he will turn it into a desert (Mill *CW,* Vol. II, 273-4).

Further examples of multiple matches found between books loaned and donated, *PPE*, and the "Condition of Ireland" series of articles are detailed in O'Neill (2016, 2019). In this way, she demonstrated a close link between the books Mill borrowed from the London Library and his published output.

Discussion

The methodology successfully employed here suggests it would be possible to analyse the reading records of other Victorian intellectuals held in the London Library issue registers, to identify how their use of the library influenced their outputs, or indeed, how Mill's donated books feature in their outputs. Given the numbers of politicians, civil servants, academics, clerics, writers, and prime ministers who were Victorian members of the London Library, Mill's donations introduced ideas from other countries and on contentious issues onto the bookshelves of some of the most significant power brokers in Victorian London (in a way which was undetectable given the absence of Mill's personal book label or signature, perhaps avoiding his identification as a radical influence (O'Neill 2016, p. 277). Currently, work on the corpus continues, as a case study for advanced matching algorithms. It would be logical to extend this work, including: incorporating the titles from Mill's private library held at Somerville College, Oxford into our source texts; returning to the list of required books and using a wider variety of online sources to see if these were now available[xiv] in digitised format for inclusion in our target texts; extending beyond the mining of Mill's monographs to also using this technique to compare Mill's correspondence, speeches, and articles in the *CW* against his London Library loans and donations; and applying statistical models to see if the quotations present in Mill's writing temporally align with his borrowing record.

This approach should be feasible for other authors, provided access to library issue registers and institutional archival records were possible, and the resources are available to gather the required data upon which to build analysis. It was not our intention to compare available software for text alignment in this study, but we would recommend using TextPAIR to identify core matches between texts and deploying Passim when the OCR is known to be more problematic, or where the volume of texts to be searched would benefit from

parallelisation, given the differences in computational requirements between the two tools. Both TextPAIR and Passim provided very useful results for this study, and are effective and advantageous computational tools, available for others to employ.

However, there are a variety of generic limitations to this research, which is dependent on a body of existing digitised content, including Mill's oeuvre, and the texts he read (even though we could not get access to digital surrogates of all titles required). Not all authorial figures have their writings digitised so completely, and therefore this method could be most successfully applied to authors whose outputs have benefited from prior digitisation, building upon known biases within the historical digital canon which may have consequences for our understanding of the past (Putnam 2016, Hauswedell et al 2020). Digitisation of cultural heritage content remains incomplete and uneven (Nauta et al 2017) and it is difficult to understand how different our results would be with access to a full set of digitised texts, and we have to provide methodological explanation to continually grapple with incomplete corpora and representativeness (Bauer and Aarts 2000). We are at the mercy of prior digitisation activities, including quality control for generation of high enough quality OCR transcripts to allow even advanced NLP algorithms to identify potential matches, and little information is provided to researchers about the digitisation process and how this may affect text-mining approaches (Cordell 2017, Hauswedell et al 2020). Researchers operating within this space should therefore do so in a critical manner, to understand how the digitisation process may be shaping their findings.

Furthermore, there is a legal component to both the affordances of this methodology. Copyright remains a driving force of digitisation practices and the "the nineteenth century is particularly well represented in digital archives, owing perhaps to its 'goldilocks' (or just

right) conservation-copyright status" than specific academic rationales (Hauswedell et al 2020). Influences on the writings of other Victorian figures may be successfully analysed using our method, but this is not the case for more recent authors, due to the "20th century black hole" in our digitised cultural heritage (Fallon and Uceda Gomez 2015). It is also unlikely that researchers will be able to access the borrowing records of modern writers without explicit consent, due to changes in privacy legislation and the resulting appropriate responses from the library sector (Bowers 2006, Dowling 2017, Bailey 2018): it is unlikely that modern reading records will survive to enable this type of research. We therefore suggest that this method is applicable to the reading and writing of authors beyond Mill, but is most likely to succeed, or even only be possible, for other leading figures professionally active from the mid 18th to early 20th centuries.

Conclusion

Text mining the books John Stuart Mill borrowed from and donated to the London Library against his published outputs has shown that the collections of the London Library influenced his thought, transferred into his published oeuvre and featured in his role as political commentator and public moralist. This research has moved discourse about the impact of the London Library onto an evidential footing, and also provides a proven methodological approach from which to approach future case studies involving understanding and mining the reading records of other 19th century intellectual figures, in order to detect and analyse influence in their published oeuvre. Identifying and showing these links benefited from interleaving computational matching (or "distant reading"), and detailed, or "close reading" undertaken on both archival registers and authorial outputs. We therefore believe we have demonstrated a virtuous relationship between archival research, computational analysis, and close textual scholarship, which will be a fruitful triangulation for others to explore in author

library studies. Opportunities in this area will continue to advance in the future, as the digital resources that are the result of mass digitisation of historical texts grow. However, given the current digitisation landscape, and complexities associated with privacy legislation, this method is most likely to be successful for other leading 18th and 19th century figures, particularly where prior digitisation of their works has already been undertaken, given the dependencies identified.

Acknowledgements

This PhD research was undertaken by Dr Helen A. O'Neill at UCL (O'Neill 2019). We would like to thank the former Librarian, Inez Lynn and the Trustees of the London Library for allowing access to the Library's historic institutional records. Development of the Passim software was supported in part by NEH Digital Humanities Start-Up Grant #HD-51728-13 and by a grant from the Andrew W. Mellon Foundation's Scholarly Communications and Information Technology program. Any views, findings, conclusions, or recommendations expressed do not necessarily reflect those of the NEH or Mellon.

References

Abdul-Rahman, A, Roe, G, Olsen, M, Gladstone, C, Whaling, R, Cronk, N, Morrissey, R and Chen, M 2017 Constructive visual analytics for text similarity detection. *Computer Graphics Forum,* 36(1): 237-248.

Altick, 1975 *The Art of Literary Research*. Revised Edition. London: W. W. Norton and Co.

Antonini, A, Benatti, F, and Blackburn-Daniels, S, 2020 On Links To Be: Exercises in Style #2. In: *31st ACM Conference on Hypertext and Social Media* (HT'20), 13-15 Jul 2020, http://oro.open.ac.uk/70781/1/on%20links%20to%20be.pdf

Atkinson, J 2013 The London Library and the circulation of French fiction in the 1840s. *Information & Culture*, 48(4): 391-418.

Babbage, C 1835 *Economy of machinery and manufactures*. London: Charles Knight.

Bailey, J 2018 Data protection in UK library and information services: Are we ready for GDPR?. *Legal Information Management*, 18(1): 28-34.

Baker, W 1981 The London Library Borrowings of Thomas Carlyle, 1841-1844 *Library Review* 1 February 1981, 30(2): 89-95.

Baker, W 1988 The London Library: A Study of its Early Rules and Regulations. Library Review 37(2), 33-41.

Baker, W 1990 J.G. Cochrane and the London Library at Pall Mall. Library History 8(6), 171- 179.

Baker, W 1992 The Early History of the London Library. New York; Lampeter: Edwin Mellen Lewiston.

Bamman, D and Crane, G 2008 The logic and discovery of textual allusion. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008). Language Technology for Cultural Heritage Data 2008, Marrakech, June 1st 2008. http://www.perseus.tufts.edu/~amahoney/latech2008.pdf

Bauer, M W and Aarts, B 2000 Corpus construction: A principle for qualitative data collection. In Bauer, M A and Gaskwll, G *Qualitative researching with text, image and sound: A practical handbook*. London: Sagem pp.19-37.

Bode, K 2012 *Reading by numbers: Recalibrating the literary field*. London: Anthem Press.

Bourdaillet, J and Ganascia, J G 2006 MEDITE: A unilingual textual aligner. In International Conference on Natural Language Processing, Finland, 23-24 August, pp. 458-469.

Bourdaillet, J and Ganascia, J G 2011 "Alignment of noisy unstructured text data." In IJCAI 2007: Workshop on Analytics for Noisy Unstructured Text Data, India, 8 January.

Bowers, F 2005 *Principles of bibliographical description*. New Castle, Delaware: Oak Knoll.

Bowers, S L 2006 Privacy and library records. *The Journal of Academic Librarianship*, 32(4): 377-383.

Bremer, F 1843a *The diary with strife and peace*. London: John Wright & Co.

Bremer, F 1843b *The home, or family laws and family joys*. Trans. London: John Wright & Co.

Bremer, F 1845 *Life in Dalecarlia: The Parsonage of Mora*. London: Chapman and Hall.

Bremer, F 1848 *Brothers and Sisters*. London: Henry Colburn.

Bremer, F 1856 *Hertha*. London: Arthur Hall, Virtue.

Büchler, M, Crane, G, Moritz, M and Babeu, A 2012 Increasing recall for text re-use in historical documents to support research in the Humanities. In International Conference on Theory and Practice of Digital Libraries, Cyprus, 23-27 September, pp. 95-100.

Büchler, M, Burns, P R, Müller, M, Franzini, E and Franzini, G 2014 Towards a historical text re-use detection. In Biemann, C and Mehler, A (eds.) *Text Mining*. Cham: Springer. pp. 221-238.

Burke, U R and Staples, R 1886 *Business and pleasure in Brazil*. London: Field & Tuer.

Busch, A, Bludau, M-J, Brüggerman, V, Genzel, K, Seifert, S, Trilcke, P, 2019 Scalable exploration: prototype study for the visualization of an author's library on the example of 'Theodor Fontane's Library'. In Digital Humanities 2019, Utrecht, 8-12 July, https://dev.clariah.nl/files/dh2019/boa/0490.html

Capps, J L 1966 *Emily Dickinson's reading*, 1836-1886. Cambridge, Massachussets: Harvard University Press.

Coffee, N, Koenig, J P, Poornima, S, Forstall, C W, Ossewaarde, R and Jacobson, S L 2012 The Tesserae Project: Intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, 28(2): 221-228.

Coleridge, S T 1980-2001 *The collected works of Samuel Taylor Coleridge*. 12, Marginalia ed. George Whaley and H.J. Jackson. London: Routledge and Kegan Paul.

Cordell, R 2017 "Q i-jtb the Raven": Taking dirty OCR seriously. *Book History*, 20(1): 188-225.

Darnton, R 1990 *The kiss of Lamourette: Reflections in cultural history*. London: Faber.

Davies, J K and Fichtner, G (comp.) 2006 *Freud's library: A comprehensive catalogue = Freud's Bibliothek: Vollständiger Katalog*. London: The Freud Museum; Tübingen: Diskord.

Dowling, T 2017 Paths to protecting patron privacy. International Information & Library Review, 49(1): 31-36.

Dunoyer, C 1845 *De la liberté du travail, ou simple exposé des conditions dans lesquelles les forces humaines s'exercent avec le plus de puissance*. Paris: Guillaumin.

Edelstein, D, Morrissey, R and Roe, G 2013 To quote or not to quote: Citation strategies in the Encyclopédie. *Journal of the History of Ideas*, 74(2): 213-236.

Faflik, D 2018 *Melville and the Question of Meaning*. New York: Routledge.

Fallon, J and Uceda Gomez, P 2015 The missing decades: the 20th century black hole in Europeana. Copyright, *Europeana Pro*. Available at https://pro.europeana.eu/post/the-missing-decades-the-20th-century-black-hole-in-europeana [Last accessed 11 March 2020].

Franzini, G 2016 English translations of Pan Tadeusz: a comparison with TRACER. *eTrap project*. Available at http://www.etrap.eu/english-translations-of-pan-tadeusz-a-comparison-with-tracer/ [Last accessed 11 March 2020].

Funk, K and Mullen, L A 2018 The spine of American law: Digital text analysis and US legal practice. *The American Historical Review*, 123(1): 132-164.

Ganascia, J G, Glaudes, P, and Del Lungo, A 2014 Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, 29(3): 412-421.

Gaskell, P 1995 *A new introduction to bibliography*. Winchester: St Paul's.

Graham, B 2019 Using natural language processing to search for textual references. In Hamidović, D, Clivaz, C, and Bowen Savent, S (eds.) *Ancient manuscripts in digital culture*. Leiden: Brill, pp. 115-132.

Gribben, A 1986 Private libraries of American authors: Dispersal, custody and description. *The Journal of Library History,* 21(2): 300-314.

The Gladstone Reading Database. Available at http://gladcat.cirqahosting.com [Last accessed 2 June 2019].

Harding, W 1957 *Thoreau's library*. Charlottesville: University of Virginia Press.

Harrison, F 1907 *Carlyle and the London Library: An Account of its foundation together with unpublished letters of Thomas Carlyle to W D Christie*. London: Chapman & Hall.

Hauswedell, T, Nyhan, J, Beals, M, and Terras, M 2020. Of global reach yet of situated contexts: An examination of the implicit and explicit selection criteria that shape digital archives of historical newspapers. Accepted, *Archival Science*. Available at https://doi.org/10.1007/s10502-020-09332-1 [Last accessed 1st May 2020].

Hollander, S 1985 *The economics of John Stuart Mill*. Toronto: University of Toronto Press.

Howitt, W 1842 *Rural and domestic life in Germany*. London: Longman, Brown, Green and Longmans.

Inglis, D H 1827 *Switzerland, South of France & Pyrenees*. Edinburgh.

Jackson, H J 2005 *Romantic readers: The evidence of marginalia*. London: Yale University Press.

Jardine, L and Grafton, A 1990. Studied for action: How Gabriel Harvey read his Livy. *Past & Present* (129): 30-78.

Jungman, R and Tabor, C 2000 "A Generation of Leaves": Homeric allusion in chapter five of Hemingway's 'In Our Time'. *The Hemingway Review*, 19(2): 108.

Keynes, G 1980 *The library of Edward Gibbon: A Catalogue*. 2nd ed. Charlottesville: University of Virginia Press.

Kokkinakis, D, and Malm, M 2015 Detecting reuse of Biblical quotes in Swedish 19th Century fiction using sequence alignment. In Corpus-Based Research in the Humanities (CRH), Poland, 10 December, p.79-80.

Leinwand, T 2016 *The Great William: Writers reading Shakespeare*. London: Chicago University Press.

Lynn, I 2006 *Modernising to stay the same: The London Library refurbishment*. Library & Information Update (7-8): 27-30.

Lyon, C, Malcolm, J and Dickerson, B 2001 Detecting short passages of similar text in large document collections. In Conference on Empirical Methods in Natural Language Processing, Pittsburgh, 3-4 June, pp.118-125.

McIntyre, T. 2006 *The Library book: An architectural journey through the London Library 1841-2006*. London: London Library.

Melville's Marginalia Online. Available at http://melvillesmarginalia.org [Last accessed 2 June 2019].

Mill, J S 1848 *Principles of political economy*. London: John W Parker.

Mill, J S 1859 *On liberty*. London: John W Parker & Son.

Mill, J S 1863 *Utilitarianism*. New ed. London: Parker, Son and Bourn.

Mill, J S 1869 *The subjection of women*. London: Longmans, Green, Reader and Dyer.

Mill, J S 1874 *Three essays on religion*. New York: Henry Holt and Co.

Mill, J S (1963-91) *The Collected Works of John Stuart Mill*. Toronto: University of Toronto Press.Available at: https://oll.libertyfund.org/people/john-stuart-mill [Last accessed 11 March 2020].

Miller, C 2012. *Reading in Time: Emily Dickinson in the Nineteenth Century*. Amherst: University of Massachussetts Press.

Nowell-Smith, S 1958 Carlyle and the London Library. In Oldman, C B, Munford, W A, and Nowell-Smith, S *English Libraries 1800-1850: Three lectures delivered at University College London* London: H.K. Lewis & Co., pp. 59-78.

Nauta, G J, van den Heuvel, W, and Teunisse, S 2017 "Europeana DSI 2– Access to Digital Resources of European Heritage: D4.4. Report on ENUMERATE Core Survey 4". *Europeana*. Available at:

https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf [Last accessed 11 March 2020].

Nowell-Smith, S 1958 Carlyle and the London Library. *English Libraries 1800-1850*. London: H. K. Lewis and Co, pp. 59-78.

Nowell-Smith, S 1972 London Library Occasions. *Times Literary Supplement*, (18 February): 187-188.

O'Neill, H 2015 The London Library and the Intelligentsia of Victorian London. *Carlyle Studies Annual*, no. 31: 183-216. www.jstor.org/stable/26594492. [Last accessed 28 May 2020].

O'Neill, H 2016 John Stuart Mill and the London Library: A Victorian book legacy revealed. *Book History*, 19: 256-283.

O'Neill, H 2019 *The role of data analytics in assessing historical library impact: The Victorian intelligentsia and the London Library*. Unpublished thesis (PhD), University College London.

Ohge, C, and Olsen-Smith, S 2018. Introduction: Computation and Digital Text Analysis at Melville's Marginalia Online. *Leviathan*. John Hopkins University Press. Volume 20, Number 2, June 2018: 1-16. 10.1353/lvn.2018.0019.

Ohge, C, Olsen-Smith, S, Barney-Smith, E, Brimhall, A, Howley, B, Shanks, L and Smith, L 2018. At the Axis of Reality: Melville's Marginalia in The Dramatic Works of William Shakespeare.

Leviathan. John Hopkins University Press. Volume 20, Number 2, June 2018: 37-67. 10.1353/lvn.2018.0019.

Olsen, M 2009 Sequence alignment and the discovery of intertextual relations. *ARTFL project*. Available at: https://docs.google.com/presentation/d/1tDH9PEkoYMKrnaqm9PMxVvunnVzI81RtV7S_bCFM2Bg/present?slide=id.i0 [Last accessed 11 March 2020]

Olsen, M, Horton, R and Roe, G 2011 Something borrowed: Sequence alignment and the identification of similar passages in large text collections. *Digital Studies/Le champ numérique*, 2(1): https://www.digitalstudies.org/articles/10.16995/dscn.258/ [Last accessed 11 March 2020]

Olsen-Smith, S, Norberg, P, and Marnon, D C (eds.) 2009-2019 *Melville's Marginalia Online*. Available at: http://www.melvillesmarginalia.org [Last accessed 2 June 2019].

Oram, R W 2014 Writers' libraries: Historical overview and curatorial considerations. In Oram R W with Nicholson J (eds.) *Collecting, curating, and researching writers' libraries: A handbook*. Lanham, Maryland: Rowman & Littlefield, pp. 1-28.

Packe, M S J 1954 *The life of John Stuart Mill*. London: Secker and Warburg.

Passy, M H 1846 *Des systemes de culture, et de leur influence sur l'economie sociale*. Paris, 1846.

Pearson, D 2019 *Provenance research in book history: A handbook*. New and Rev. ed. New Castle, Delaware: Oak Knoll.

Putnam, L 2016 The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review,* 121(2):377-402. Available at: https://academic.oup.com/ahr/article/121/2/377/2581842 [Last accessed 11 March 2020].

Raven, J 2018 *What is the history of the book?* Cambridge: Polity.

RED, The Reading Experience Database. Available at: http://www.open.ac.uk/Arts/reading/ [Last accessed 2 June 2019].

Reeves, R 2015 *John Stuart Mill: Victorian firebrand*. London: Atlantic Books.

Reynolds, M S 1981 *Hemingway's reading, 1910-1940: An inventory*. Princeton: Princeton University Press.

Reynolds, M S 1986 A supplement to Hemingway's reading, 1910-1940. *Studies in American Fiction* 14(1): 99.

Roe, G H, and The ARTFL Project 2012 Intertextuality and influence in the age of Enlightenment: Sequence alignment applications for Humanities research. In Digital Humanities 2012, Hamburg, 16-20 July, pp. 345-347. Available at: http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/intertextuality-and-influence-in-the-age-of-enlightenment-sequence-alignment-applications-for-humanities-research.1.html [Last accessed 11 March 2020].

Roe, G 2018 A Sheep in Wolff's Clothing: Émilie Du Châtelet and the *Encyclopédie*, *Eighteenth-Century Studies*, 51(2): 179-196.

Rosen, J 2011 Combining close and distant, or the utility of genre analysis: A response to Matthew Wilkens's 'Contemporary fiction by the numbers'. *Post45*, 3 December. Available at: http://post45.research.yale.edu/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers/ [Last accessed 11 March 2020].

Rothberg, M 2010 Quantifying culture?: A response to Eric Slauter. *American Literary History*, 22(2): 341-346.

Ryan, A 1970 *The philosophy of John Stuart Mill*. Amherst: Humanity Books.

Schmidt, D, and Colomb, R 2009 A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67(6): 497-514.

Schmidt, D and Fiormonte, D 2010 Documenti Multiversione: una soluzione per gli artefatti testuali del patrimonio culturale. *Multi-Version Documents: a Digitisation Solution for Textual Cultural Heritage Artefacts*. Available at: https://scienzepolitiche.uniroma3.it/dfiormonte/wp-content/uploads/sites/97/2013/11/intelligenza_artificiale-1.pdf [Last accessed 11 March 2020].

Sealts, M M 1966 *Melville's reading: A check-list of books owned and borrowed*. Madison: University of Wisconsin Press.

Sealts, M M 1982 *Pursuing Melville, 1940-1980: Chapters and essays*. Madison: University of Wisconsin Press.

Seo, J and Croft, W B 2008. Local text reuse detection. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, July, pp. 571-578.

Simonde de Sismondi, J C L 1819 *Nouveaux principes d'economie politique*. 2 vols. Paris: Delaunay.

Simonde de Sismondi, J C L 1836-46 *Histoire des francais, continuée par Renee (à 1776)*. Brussels: H. Dumont.

Simonde de Sismondi, J C L 1837-8. *Etudes sur L'economie politique*. Brussels: H. Dumont.

Smith, D A 2019 Improving optical character recognition and tracking reader annotations in printed books by collating and transcribing multiple exemplars. Funder: National Endowment for the Humanities (NEH), Grant number: HAA-263837-19. https://app.dimensions.ai/details/grant/grant.8385506

Smith, D A, Cordell, R, and Dillon, E M 2013 Infectious texts: Modeling text reuse in nineteenth-century newspapers. In IEEE International Conference on Big Data, Santa Clara, 6-9 October, pp. 86-94.

Smith, D A, Cordell, R, Mullen, A, and Fiztgerald, J D 2019a Mass digitization. In *Going the Rounds: Virality in Nineteenth-Century American Newspapers*. Minneapolis: University of Minnesota Press. Available at: https://manifold.umn.edu/projects/going-the-rounds [Last accessed 11 March 2020].

Smith, D A, Cordell, R, Mullen, A, and Fiztgerald, J D, 2019b What is text, probably? In *Going the Rounds: Virality in Nineteenth-Century American Newspapers*. Minneapolis: University of Minnesota

Press. Available at: https://manifold.umn.edu/projects/going-the-rounds [Last accessed 11 March 2020].

Smith, D A, Cordell, R, Mullen, A, and Fiztgerald, J D, 2019c Text reuse. In *Going the Rounds: Virality in Nineteenth-Century American Newspapers*. Minneapolis: University of Minnesota Press. Available at: https://manifold.umn.edu/projects/going-the-rounds [Last accessed 11 March 2020].

Smith, T F and Waterman, M S 1981 Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Tanselle, G T 2009 *Bibliographical analysis: A historical introduction*. Cambridge: Cambridge University Press.

Terras, M 2013 "Does anyone know if folks have used Turnitin to detect plagiarism in historical texts? Would it work? ie stuff published 1800s?" Tweet 6 March. Available at: https://twitter.com/melissaterras/status/309251138799144960 [Last accessed 11 March 2020].

Tyler, L 1995 Passion and Grief in A Farewell to Arms: Ernest Hemingway's Retelling of Wuthering Heights. *The Hemingway Review* 14(2): 79.

Underwood, T 2014 Theorizing research practices we forgot to theorize twenty years ago. *Representations*, 127(1): 64–72. Available at: https://rep.ucpress.edu/content/127/1/64 [Last accessed 11 March 2020].

Underwood, T 2017 A Genealogy of Distant Reading. Digital Humanities Quarterly, 11(2). Available at http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html [Last accessed 1st May 2020].

Vesanto, A, Nivala, A, Rantala, H, Salakoski, T, Salmi, H, and Ginter, F 2017 Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910. In Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Sweden, 22 May, pp. 54-58.

Werner, S 2019 *Studying early printed books: A practical guide, 1450-1800*. Chichester: Wiley Blackwell.

Wilkerson, J, Smith, D A, and Stramp, N 2015 Tracing the flow of policy ideas on legislatures: A text reuse approach. *American Journal of Political Science*. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12175 [Last accessed 11 March 2020].

Young, A 1792 *Travels in France in 1787, 1788, 1789*. 2 vols. Bury St. Edmunds: Printed by J. Rackham.

---

[i] https://www.some.ox.ac.uk/library-it/special-collections/john-stuart-mill-collection/
[ii] https://www.londonlibrary.co.uk
[iii] Baker, writing on the difficulties of working on with the London Library Issue Books, states "The handwriting in them is often difficult to read, and frequently, even with the help of bibliographical guides such as the London Library Catalogues, it proved at times unreadable for the present writer. Legibility is not helped by heavy black ink lining across the entry, a scoring presumably made to indicate the item's return to the Library… early members complained justifiably "that the library's records of books borrowed were hopefully confused" (Baker 1981, 90, latter quote from Nowell-Smith 1958, 72).
[iv] http://melvillesmarginalia.org
[v] https://www.liverpool.ac.uk/english/research/gladstone-library/
[vi] http://www.open.ac.uk/Arts/RED/publications.htm
[vii] https://gtr.ukri.org/projects?ref=AH%2FT003960%2F1
[viii] An alternative approach, Text Collation, is often used to identify *differences* rather than *similarities* in textual witnesses by aligning like passages and looking for regions of variation, and computational approaches to such "textual genetic criticism" have been implemented (Bourdaillet and Ganascia 2006, Schmidt and Colomb 2009, Schmidt and Fiormonte 2010).
[ix] https://archive.org
[x] https://oll.libertyfund.org/people/john-stuart-mill
[xi] https://code.google.com/archive/p/text-pair/. Since undertaking this research, a revised Python version has been made available at https://github.com/ARTFL-Project/text-pair. Full documentation for each is provided at these sources.
[xii] The 2018 release was re-written in Python, see https://github.com/ARTFL-Project/text-pair.
[xiii] https://github.com/dasmiq/passim
[xiv] Although there are millions of texts now online, there is no single place where these can be searched: this is currently under research via the Global Datasets of Digitised Texts network: https://gddnetwork.arts.gla.ac.uk.