Artificial Intelligence and the Recovery of Women's Intellectual History: Digital Methods for Rediscovering Forgotten Women riters

Ms. Swapnil Sharma

Research Scholar, History, Sharda School of Humanities and Social Sciences, Sharda University, Greater Noida, Uttar Pradesh

DOI: https://doi.org/10.47772/IJRISS.2026.102000013

Received: 15 October 2026; Accepted: 20 October 2026; Published: 19 November 2026

ABSTRACT

The silence of women is not a natural phenomenon; it is a historical phenomenon. In the eighteenth, nineteenth, and twentieth centuries, identifiable mechanisms of anonymous and pseudonymous publication, the re-imagining of the literary canon, and the selective activities of libraries and archives systematically closed off opportunities for women to be thinkers and writers.

This paper understands the recovery of women's intellectual history as a problem of the history of knowledge and structures the analysis of this history around the four “moments” of historical silences identified by Michel-Rolph Trouillot. It claims that at each of these times, the historian can intervene, using AI and computational methods enabling him to read what was previously unreadable (handwritten text recognition [HTR] and optical character recognition [OCR]); test the authorship of anonymous and pseudonymous texts (stylometric authorship attribution); and reconstruct intellectual relationships that have been lost from traditional historical narratives (entity recognition, network analysis, and semantic markup).

The paper, which relies on a purposive review of key digital-humanities projects and quantitative analysis of the data on women's authorship in the historical record, demonstrates how digital approaches can uncover and address marginalisation of women in the historical record. It proposes, however, that there is a value dimension to these techniques: machines that create digital corpora make good use of archival material influenced by patriarchy and colonialism, reproducing them. The paper thus argues that the biggest hurdle to the recovery via AI is in a multilingual and non-Western context, and looks at this via the colonial and post-colonial South Asian archive, with shorter case studies in Bengali and Marathi from India. In the spirit of Indian languages and knowledge traditions, which have been emphasized by the National Education Policy 2020, computational history can be a powerful tool for a more complete and fair history of women's contribution to the production of knowledge, when used wisely.

Keywords: women’s history; digital history; archives and silence; authorship attribution; colonial archive; algorithmic bias; Indian regional-language archives; handwritten text recognition

INTRODUCTION: A PROBLEM IN THE HISTORY OF KNOWLEDGE

History, to a great extent, is the story of the survivals that have been remembered. The transmission of past knowledge in libraries, catalogues, anthologies, and critical traditions does not just preserve knowledge, it also constructs it. They have a pattern of inclusion and exclusion and a pattern of neglect, whose thought ends up being included, whose thought ends up being excluded, whose thought ends up being neglected.

The recovery of women's intellectual history is adopted as a problem in the history of knowledge and how artificial intelligence (AI) and computational techniques may transform it is explored. Makes three arguments.

First, the marginalisation of women writers is a historical product: it is possible to identify practices, in the

process of publishing and in the process of canon-building and archiving, which marginalise women's writing and that have developed over the years.

Second, AI can be a means through which to intervene at every stage of the process, whether it be "reading" illegible documents, or reconstructing erased intellectual networks, opening up new ways of feminist recovery.

Third, these tools are not neutral – they are created from pre-existing archives that have their own set of perspectives, and may reproduce and, indeed, reinforce the very silences they seek to disrupt and breach.

The paper does not imply an opposition to the use of computation, but rather considers how such tools can be designed to make recovery a transformative process of historical justice, rather than an automatic process of exclusion.

The analysis is organised through an influential account by Michel-Rolph Trouillot, on the role of silences in the construction of history. According to Trouillot (1995) silence is present in four ways: the creation of a fact (the making of sources), the assembly of a fact (the making of archives), the retrieval of a fact (the making of narratives), and retrospective significance (the making of history in the final instance).

RESEARCH METHODOLOGY AND SOURCE SELECTION

This paper is interpretive and integrative in nature and will be qualitative. It is not experimental research, since it does not test or train new computational models, rather it is a conceptual and critical review that relates to the specific mechanisms through which women have been silenced. The four moments of silencing are used as the framework for analysis, meaning that each method presented below is discussed in the context of the specific silencing it can and cannot explain. Purposive sampling was used and the sources are divided into three groups. The first consists of theoretical and historiographical texts related to the concepts of silence, archives and gender (Trouillot 1995; Spender 1986; Risam 2018; D'Ignazio and Klein 2020).

The latter consists of computational and digital humanities research and projects which either quantify women's authorship or show how to recover it (Underwood, Bamman, and Lee 2018; the Orlando Project; the Women Writers Project; and documentation of tools like Tesseract and the R package "stylo"). The third group includes scholarship and archives on South Asia which provide the case material (Tharu and Lalita 1991; O'Hanlon 1994; Burton 2003; British Library's Endangered Archives Programme). Works were considered to be related to at least one of Trouillot's four moments and included if they were scholarly publications or primary documentation of a work. The result is that the distribution of sources by decade and category is as shown in Figure 2. All the methods were evaluated on the following three points: what type of silence does it refer to, and what has been found in the literature about how effective and successful it is, and which limitations and biases does it have, especially in the case of non-Latin scripts and archives outside the West?

The results are summarised in Tables 1-3 and Figure 3. There are constraints to the approach which must be explicitly identified. The list is not definitive, and is biased towards English language scholarship. The quantitative evidence is secondary: the data on women's authorship are drawn from Underwood, Bamman and Lee (2018); the data are used to illustrate their reported trend in Figure 1, rather than to be analysed anew. The Indian case studies provided here are not an exhaustive survey, but indicative and could be further explored through pilot experiments with corpora based on regional languages, as suggested in the research directions for the future.

A HISTORICALLY PRODUCED SILENCE: HOW WOMEN WERE WRITTEN OUT

Dale Spender had already been trying to rewrite the history of the English novel, displacing many women whose contributions to the novel's beginnings were ignored, long before the computational turn (Spender

1986). This erasure has been made measurable with computational evidence since then. Ted Underwood, David Bamman, and Sabrina Lee's analysis is of a very large sample of fiction in the English language, covering almost two centuries. Were they able to conclude that the proportion of published fiction written by women did not continuously increase over time, but instead dropped precipitously, from about half of English-language fiction around 1850 to about a quarter by the mid-twentieth century (Underwood, Bamman, and Lee 2018; see Figure 1)? This is a reversion to a well-known story of incremental steps.

This marginalisation of women writers thus was not just a relic of a bygone era which they were slowly emancipated from, but a historical process that in particular times became more pronounced. As literacy rates and the number of women writers increased, publishing, reviewing and canon-making in the period 1850-1950 resulted in a displacement of women's writing from print. To demonstrate that such a decline does exist and to date it accurately, is just the type of claim historical scholarship is designed to challenge. Only with the advent of large digital corpora, however, was such change possible, and even more so to date (Underwood 2019). The trend is summarized in Figure 1, and the partial recovery following approximately 1970 indicates that the trend could be reversed, thereby proving that it was historical and not natural.

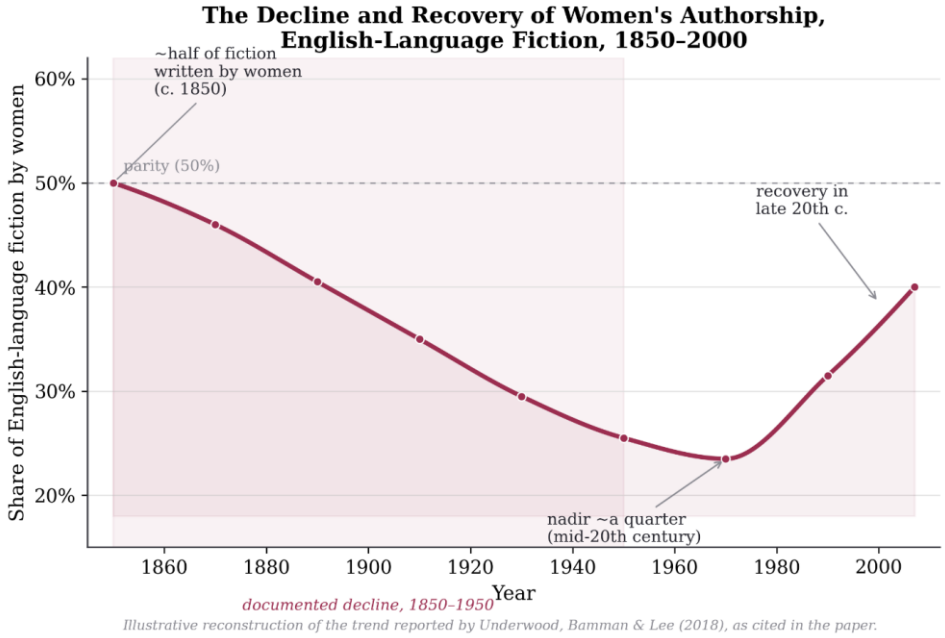

Figure 1. The decline and recovery of women's authorship in English-language fiction, 1850–2000

Source: Illustrative redrawing of the trend reported in Underwood, Bamman, and Lee (2018). Values are approximate and show the shape of the trend; they are not the original data.

A lot of this silencing took place at Trouillot's first moment, that of the making of sources. The anonymity and pseudonymity of texts resulted in women's works being published without their names and authorship therefore disappearing at the moment of publication, and not later. Table 1 summarizes the silencing which took place at each of Trouillot's four moments; each of the following sections will examine one of these moments.

Table 1: Historical Silencing of Women's Intellectual History

Stage of Historical Silencing	**Historical Process**	**Effect on Women's Intellectual History**
Source Creation	Anonymous and	Women's authorship

	pseudonymous publication	remained hidden
Archive Formation	Selective preservation and cataloguing	Women's writings excluded from institutional archives
Narrative Construction	Canon formation and literary criticism	Women omitted from mainstream literary history
Retrospective Significance	Historical interpretation and recognition	Women's intellectual contributions undervalued

Source: Author's synthesis, based on Trouillot (1995).

THE ARCHIVE AND ITS SILENCES

The first important silence is created in the process of making the sources; the second in the process of making the archives. The archive as a institution has come under intense scrutiny by historians who have been more and more interested in the values and selection practices of the archive. Feminists and empire scholars have further developed Trouillot's discussion of "making of archives" to critique the types of "preserving and forgetting" it entails (Trouillot 1995; Burton 2003). In addition, the Indian case is also a testimony to the extent of labour recovery required prior to digital methods. Susie Tharu and K. Lalita collected women's writing from over the last two millennia and across a variety of languages for their groundbreaking anthology. They had recognized how hard it was to even find such texts in the library, as some copies were nearly unreadable, and biographical information had to be gleaned from letters, memories of family and friends, and works that were preserved in private collections or in the oral tradition (Tharu and Lalita 1991).

THE DIGITIZED TURN: NEW INSTRUMENTS FOR THE HISTORIAN

In Trouillot's terms, the third moment of silencing, fact retrieval or the making of narratives, occurs when the historian selects some facts and leaves out others in telling a story about the past. The situation has been completely transformed by digitisation. Mass digitization and full text searching revolutionized the way historians search for information: as Lara Putnam has explained in the American Historical Review in 2016, it has "significantly diminished the significance of 'place-specific knowledge' in the discipline" (Putnam 2016). The turn that's 'digitised' is particularly important for the marginal: a person who only has a short mention in a provincial newspaper, and who would have been otherwise inaccessible to any researcher not already immersed in the archive.

Not all collections will, however, be digitised evenly the quality of what is digitised and made searchable is not necessarily representative of the past, but is determined by institutional priorities, funding, and copyright, and search that is seamless across collections can mask this lack of evenness (Putnam 2016). This warning is in line with larger historiographical discussion. The theory of "distant reading" developed by Franco Moretti was based on the premise that patterns of a literary system could only be discerned from the study of thousands of texts, rather than only a few (Jockers 2013; Moretti 2013).

Katherine Bode warned however, that digital collections are frequently viewed as windows into the past, and these windows are seldom complete, and in some cases, are 'accidents of survival and digitisation' (Bode 2018). Intellectual genealogy of the authors sources can be seen in figure 2. Of the 22 dated works cited, 11 are computational or digital-humanities studies, nine of which appeared in the 2010s (earlier, 4 works, between 1986 and 2003, are works of feminist or archival scholarship which frame these works). The pattern is similar to the paper's argument: Recent computational methods are examined here in the light of earlier feminist and archival critiques. It also reveals a weakness, in so far as the work's sources are limited to the recent and the writing in English.

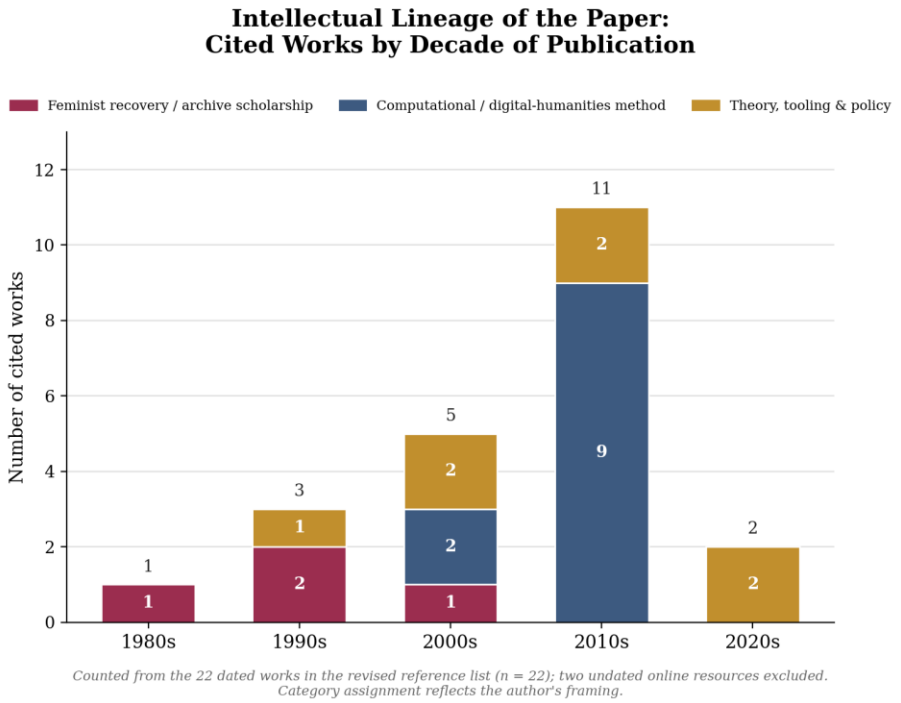

Figure 2. Works cited in this study, by decade of publication and category (22 dated works).

Source: Author's count from the reference list of this paper; two undated online resources are excluded.

METHODS OF RECOVERY

To date, a set of particular techniques provides a means for historians to intervene in each of these times (see Figure 3) in the history of women's thought. Their discussion here is not of them as ends in themselves, but for their role as means to an argument in the history of ideas.

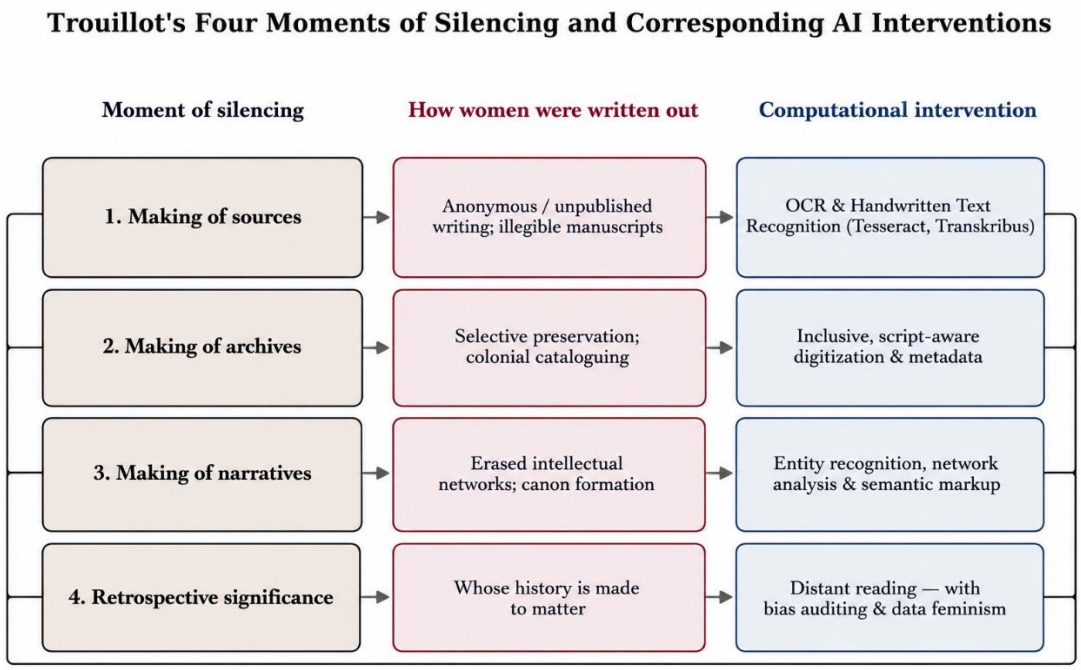

Figure 3. Trouillot's four moments of historical silencing and the corresponding AI/digital intervention

Source: Author's framework, adapted from Trouillot (1995). Stylometry, discussed below, also acts at the first moment by contesting anonymity.

Making Sources Legible: OCR and Handwritten Text Recognition

The first step in recovery is to make sources legible. For print, optical character recognition (OCR) software such as Tesseract transforms images of text into searchable text, but it performs less well on historical typefaces, damaged pages, and non-Latin scripts (Smith 2007).

Manuscript handwritten text recognition platforms (e.g. Transkribus) obtain handwritten text recognition models from a set of samples of handwritten pages from a specific hand or collection of hands; therefore, the accuracy of the models is dependent on the quality and quantity of the training samples (Muehlberger et al. 2019). Such restrictions are more stringent with Indian languages.

OCR problems are different for Indic scripts than for Latin scripts (Govindaraju and Setlur 2009) and handwritten and historical Indic material is still a research challenge, especially for handwritten material due to limited training data.

Recognition errors are not random noise: they might systematically fail to recognize infrequent names, women's names, and dialect forms and thus re-silence the very sources they are recovering. If the output is to be used for analysis, the OCR or HTR output should always be accompanied by a reported error rate, on hand checked transcriptions.

Contesting Anonymity: Stylometry and Authorship Attribution

Numerous women's works were published anonymously, under the name of their husbands or under a pseudonym. The statistical analysis of measurable features of style (stylometry) provides a principled approach to testing attributions.

The methods have now been introduced to historians without specialist expertise, e.g. by the R package “stylo” (Eder, Rybicki, and Kestemont 2016). Stylometry could be employed in recovery work to determine if a text is probably written by the specific woman writer or if letters that have been assigned to a male family member are actually written by a woman.

However, stylometric results should be carefully validated. Attribution is an educated guess, rather than proof. Good practice involves forming a reference set of securely attributed texts of similar genre, date, and length; checking to see if the technique can correctly identify the author for texts, for which the author is already known; comparing a number of techniques, not just one; and reporting results as a measure of confidence based on historical evidence, not as a verdict (Koppel, Schler, and Argamon 2009).

The results are weakest with short texts, with texts that have been edited, translated, or co-written, and for languages which lack many digitised comparison texts, all of which are prevalent in women's writing. Attributing a work to an unknown author solely on stylistic grounds is particularly problematic as this can result in the perpetuation of gender stereotypes, and is more defensible than attributing it to a particular candidate.

Reconstructing Erased Networks: Entity Recognition and Semantic Markup

But it is not sufficient to find an individual; the reconstruction of the intellectual relationships, the networks with which women made their work possible- correspondents, collaborators, publishers, reviewers- is just as important, since women's influence often works through networks that are not covered in the conventional narrative.

Named-entity recognition identifies entities like people, places, and institutions in digital text; network analysis identifies links between named entities and provides a map of the relationships between them (see Figure 4).

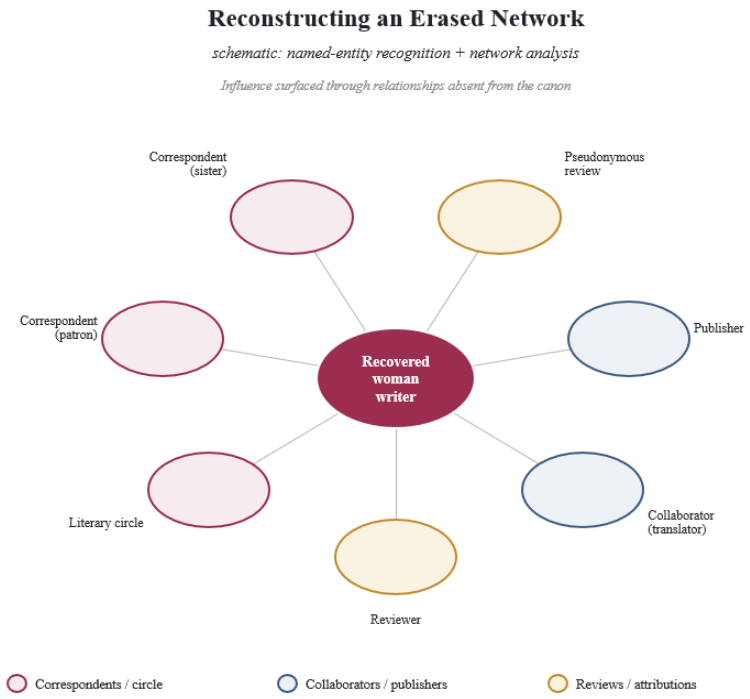

Figure 4. Schematic reconstruction of an erased intellectual network using named-entity recognition and network analysis

Source: Author's schematic. It illustrates the kinds of ties that entity recognition and network analysis can extract; it is not drawn from a real dataset.

The Orlando Project was one of the first projects on women's writing to develop a "text-base", in which semantic layering of the texts enables intertextual influences, reception, and biographical connections to be traced across the whole collection rather than within single entries; its creators frame it as a feminist experiment in encoding relationships within literary history (Brown, Clements, and Grundy n.d.). The Women Writers Project likewise demonstrates the possibilities of encoding pre-Victorian women writers' texts to common scholarly standards and preserving them for the long term, and serves as a model for data curation in the humanities (Flanders et al. 2017).

THE COLONIAL ARCHIVE AND THE MULTILINGUAL FRONTIER

The strategies discussed above are mostly from the Anglophone and European sphere and they have to be considered in the context of a global, just recovery, as they pose fundamental questions. These are no simple questions. Digital archives are a product of colonial practices, and so are the processes of what was collected, catalogued, and preserved, and what was not. Digital humanities, therefore, must not be a digital version of colonial practices and violence (Risam 2018).

The chance is great. Antoinette Burton (2003) demonstrates that women's memoirs, fiction, and family histories, like Janaki Majumdar's or Attia Hosain's, can be viewed as archives on their own, and that such a material is appropriate for encoding and linking as the Orlando Project has done for British women writers. Likewise, the anthology of Tharu and Lalita (which includes authors in a number of Indian languages) could be digitised, encoded and analysed to yield something similar to the Orlando Project for the British Isles (Tharu and Lalita 1991).

Indian Case Studies: Regional-Language Archives and Women Writers

Three clusters of material provide insights into what recovery is in practice in India. The first one is in Bengali. Amber Jiban (1876) by Rassundari Devi is considered as one of the earliest autobiographies written by an Indian woman, and can be found in Tharu and Lalita's (1991) anthology. Periodicals like Bamabodhini

Patrika, a Bengali journal for educated women readers, are partially preserved in the wider print culture in which Bengali women read and write; volumes from the 1870s to the 1900s were digitised as part of the British Library's Endangered Archives Programme (EAP921) (British Library n.d.).

These types of page images are not text that can be searched. Without applying OCR or HTR systems that are trained on Bengali script, NER and network analysis cannot be used for the women who wrote in and about these journals. Second case is Marathi. In response to a newspaper article, Stri Purush Tulana (A Comparison between Women and Men) by Tarabai Shinde was printed in Pune in 1882 and is a strong condemnation of the gender relations (O'Hanlon 1994).

It elicited hostile responses, and was largely forgotten and rediscovered in the 1970s. A text can be suppressed at the archive, as evidenced by its history: it can be in print and not be part of the scholarly record.

It also reveals that recovery is a process, digitisation is just part of it. Finding the answers, comments and companion pamphlets that accompanied the text would necessitate OCR of Marathi (Devanagari) degraded newsprint and entity recognition for Marathi names and places. Research and tools to aid these tasks are still inadequate for Indian languages (Govindaraju and Setlur 2009; Gala et al. 2023).

The third cluster is an additional and different form of writing, one which is not part of the literary archive. In the unpublished family memoir of Janaki Majumdar, the memoirs of Cornelia Sorabji and Attia Hosain's novel Sunlight on a Broken Column, Burton (2003) sees a home, family and partition which are written from a women's perspective as counter-archives of late colonial India. This kind of writing is domestic, autobiographical, and in Majumdar's case unpublished, and is the type of writing that is poorly catalogued in libraries. Semantic markup might be useful here, for instance by linking people, places and households across texts; and this might be achieved with the help of computational methods, once the manuscripts have been digitised and transcribed.

These cases are compared in Table 2. The challenges that arise at each step are markedly different: printed Bengali and Marathi books and newspapers are hard to locate and read in non-Latin scripts; digitised periodical has for its printed page images to be transcribed before they can be searched; memoirs and novels are domestic and unpublished writing, outside of typical cataloguing. None of these cases is due to a lack of algorithms; it is a combination of script, condition of the source and archival neglect.

Table 2: Indian Case Studies: Sources, Archival Conditions, and Computational Opportunities

Writer and source	**Language and archival condition**	**Computational opportunity and caution**
Rassundari Devi, *Amar Jiban* (1876)	Bengali; printed autobiography, available in English translation (Tharu and Lalita 1991)	Bengali OCR and entity recognition could locate her in other sources; working only from English translations would add a colonial-language filter
Tarabai Shinde, *Stri Purush Tulana* (1882)	Marathi; printed pamphlet, largely unknown until republished in the 1970s (O'Hanlon 1994)	OCR for Marathi (Devanagari) newspapers and pamphlets to trace replies and reception; Marathi named-entity tools are needed
Bamabodhini Patrika	Volumes digitised as page images by the British	Bengali OCR or HTR to make the text searchable;

(Bengali periodical)	Library's Endangered Archives Programme (EAP921)	network analysis of contributors and readers once transcribed
Majumdar, Sorabji, and Hosain (Burton 2003)	English; an unpublished family memoir, memoirs, and a novel	Digitisation and transcription first; then semantic markup to link people, places, and households in an Orlando-style text-base

Source: Author's synthesis, based on Tharu and Lalita (1991), O'Hanlon (1994), Burton (2003), and the British Library Endangered Archives Programme record for EAP921.

Table 3: Challenges and AI-Based Solutions for Recovering Women's Intellectual History

Challenge	Impact	AI/Digital Solution
Anonymous publications	Unknown authorship	Stylometric authorship attribution
Poor manuscript quality	Difficult transcription	Deep learning-based HTR
Archive bias	Missing historical evidence	Inclusive digitization and metadata expansion
Multilingual manuscripts	Low OCR accuracy	Script-specific AI models
Algorithmic bias	Reinforces historical exclusion	Bias auditing and Data Feminism framework

Source: Author's synthesis, drawing on Smith (2007), Eder, Rybicki, and Kestemont (2016), Risam (2018), and D'Ignazio and Klein (2020).

Multilingual Language Models and Their Limits

One possible way is to go the multilingual language model way. According to its authors the IndicTrans2 machine-translation system is the first system that allows for the translation of all 22 scheduled Indian languages, and models of this type have the potential to enable cross-lingual search and cataloguing of regional-language archives. To historians their limits are as significant. These models are usually trained and tested on modern text and cannot be guaranteed to perform well on the orthography, dialect and vocabulary of the nineteenth century and must be evaluated. Translation can also flatten the linguistic particularity that recovery attempts to maintain; systems that go through the English filter are liable to recreate the colonial-language filter mentioned above. Their outputs should therefore only be used as finding aids to be checked against the source text and by readers of the language concerned and not as a substitute.

BIAS, ETHICS, AND THE MAKING OF RETROSPECTIVE SIGNIFICANCE

The fourth and final moment in Trouillot's scheme is retrospective significance, when the past is given significance, and some histories come to the surface, while others go under. Computational recovery comes in here, too and here it has the highest risk. Catherine D'Ignazio and Lauren Klein have developed an approach they call "data feminism" which states that data is never neutral but always a reflection of power dynamics, and that best practice is to reflect on and question those power dynamics, not assume the objectivity of quantification (2020). Regarding recovery work, this poses concrete questions such as who

will form the training data, who will not, and who will be responsible for deciding the categories for which we will mark up? Such bias can be perpetuated and reinforced by algorithmic systems.

Safiya Noble's research into the racialisation and genderisation of search technologies resonated in the public sphere with the possibility of prejudicial technologies (Noble 2018). More recently the focus has shifted to large language models. They can play a supporting part in speeding up transcription and cataloguing, but can also cause false readings and translate from their training set onto material that is incompatible with it. In fact, if the task is to recapture what dominant traditions have been left out, it is counterproductive to simply accept models that have been trained on dominant traditions. Generative tools should be used as tools to assist; not as oracles (see Figure 5). Verification is essentially a comparison of the model output with the source image or text, recording the tool and version used to create the model and ensuring that readers of the language of the model verify it.

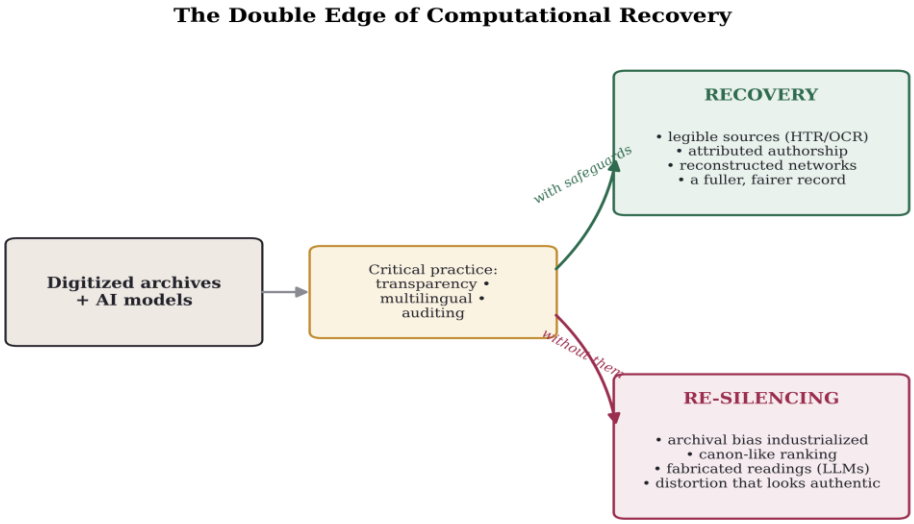

Figure 5. The double edge of computational recovery: the same archives and models enable recovery with critical safeguards and risk re-silencing without them.

Source: Author's conceptual synthesis, drawing on Noble (2018) and D'Ignazio and Klein (2020).

FUTURE RESEARCH DIRECTIONS

Based on the analysis there are three directions that follow. The first is pilot projects of OCR and HTR on Indian regional language corpora, like transcribing a sample of pages in a Bengali periodical like Bamabodhini Patrika, and Marathi newspapers of the 1880s, and reporting the character/world error rates by script and print quality, and checking whether named-entity recognition can recover the women who appear in them.

The latter is the validation of stylometric techniques in Indian languages the construction of securely-attributed reference corpora and the assessment of the transferability of techniques developed for English to Bengali, Marathi, Hindi and Urdu. The third is collaborative and ethical practice: collaborative and co-designed metadata and selection criteria with regional archives, libraries and scholars of the languages concerned; audit models for gender, caste, language bias; documenting all at every step so that results can be checked forward. Comparative evaluation is the natural next step, and the main empirical gap in the present paper, in which several tools are tested on the same material.

CONCLUSION

It has always been difficult to recover intellectual women's history, as there is often a silence in the archive. It has been fought writer by writer until very recently, and what has been lost – most of which – was not

even remarked. What each form of artificial intelligence changes is the subject of Trouillot's account of how silences are made: handwritten, the recognition of text and optical; stylometry, the anonymity and the attribution of the text; entity recognition and semantic markup, the erased networks of ideas; and distant reading, the combination of distance reading with bias auditing, the marginalisation of women itself measured and contested. All of these techniques greatly increase the historian's scope of reach.

The Indian examples mentioned above reflect the possibilities and yet to be accomplished. There are texts printed and/or available as page images, but have not yet been easily searchable, and/or analysed computationally, such as Assamese texts like Amar Jiban by Rassundari Devi or Marathi texts like Stri Purush Tulana by Tarabai Shinde, and periodicals like Bamabodhini Patrika, which are harder to find than the English equivalents, and less developed in terms of tools for Indic scripts, handwriting and multilingual content. Hence, a recovery that is just and that builds on the priority of the National Education Policy 2020 for Indian languages and knowledge traditions needs investments in digitisation of regional languages, validated tools and critical reading. The historian's reach is extended by the help of AI, but only the attention of language-aware and critical practice will expand it towards the people who the archives excluded.

REFERENCES

1. Bode, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History.* Ann Arbor: University of Michigan Press.
2. British Library. n.d. Endangered Archives Programme, project EAP921 (Sadharan Brahmo Samaj collection): *Bamabodhini Patrika* volumes (Bengali). https://eap.bl.uk/archive-file/EAP921-7-2-467.
3. Brown, Susan, Kim Martin, and Asen Ivanov. "Linking Out: The Long Now of Digital Humanities Infrastructures." *Future Horizons: Canadian Digital Humanities* (2023): 315.
4. Burrows, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–287.
5. Burton, Antoinette. 2003. *Dwelling in the Archive: Women Writing House, Home, and History in Late Colonial India.* New York: Oxford University Press.
6. D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism.* Cambridge, MA: MIT Press.
7. Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8 (1): 107–121.
8. Flanders, Julia, Sarah Connell, Nicole Infanta Keller, Elizabeth Polcha, and William Reed Quinn. 2017. "Learning from the Past: The Women Writers Project and Thirty Years of Humanities Text Encoding." *Magnificat Cultura i Literatura Medievals* 4: 1–18.
9. Gala, Jay, Pranjal A. Chitale, A. K. Raghavan, et al. 2023. "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages." *Transactions on Machine Learning Research.*
10. Govindaraju, Venu, and Srirangaraj Setlur, eds. 2009. *Guide to OCR for Indic Scripts: Document Recognition and Retrieval.* London: Springer.
11. Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History.* Urbana: University of Illinois Press.
12. Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. "Computational Methods in Authorship Attribution." *Journal of the American Society for Information Science and Technology* 60 (1): 9–26.
13. Moretti, Franco. 2013. *Distant Reading.* London: Verso.
14. Muehlberger, Guenter, et al. 2019. "Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study." *Journal of Documentation* 75 (5): 954–976.
15. Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.
16. O'Hanlon, Rosalind, ed. and trans. 1994. *A Comparison between Women and Men: Tarabai Shinde and the Critique of Gender Relations in Colonial India.* Madras: Oxford University Press.
17. Putnam, Lara. 2016. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast." *The American Historical Review* 121 (2): 377–402.

18. Risam, Roopika. 2018. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy.* Evanston: Northwestern University Press.
19. Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine." In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007),* 629–633. Curitiba: IEEE.
20. Spender, Dale. 1986. *Mothers of the Novel: 100 Good Women Writers before Jane Austen.* London: Pandora Press.
21. Tharu, Susie, and K. Lalita, eds. 1991. *Women Writing in India: 600 B.C. to the Present. Volume I: 600 B.C. to the Early Twentieth Century.* New York: Feminist Press.
22. Trouillot, Michel-Rolph. 1995. *Silencing the Past: Power and the Production of History.* Boston: Beacon Press.
23. Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change.* Chicago: University of Chicago Press.
24. Underwood, Ted, David Bamman, and Sabrina Lee. 2018. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics* 3 (2).