



Digital Humanities and English Literary Heritage: Mapping Cultural Narratives Through Computational Methods

Ravindra Hanmantrao Sagar

Assistant Professor

Nagnath Arts, Commerce and Science College

Aundha Nagnath

Hingoli, Maharashtra



<http://www.gjirp.org/> || Vol. 2 No. 4 (2026): October Issue

Date of Submission: 05-09-2026

Date of Acceptance: 24-09-2026

Date of Publication: 04-10-2026

Abstract: Digital humanities has transformed literary scholarship by enabling researchers to analyse textual collections at scales beyond conventional close reading. The present study examines how computational methods can be used to map cultural narratives embedded in English literary heritage across more than two centuries. A corpus of 240 English-language literary works published between 1800 and 2020 was organised into six historical periods and analysed through text mining, topic modelling, named-entity recognition, geographical mapping, lexical analysis, and character-network analysis. The corpus contained approximately 21.84 million word tokens. An eight-topic Latent Dirichlet Allocation model identified recurrent narrative formations related to class and social hierarchy, gender and domesticity, industrial modernity, empire and colonial encounter, war and national identity, religion and morality, migration and belonging, and individual identity. Significant historical variation was observed in several thematic categories. The mean normalized thematic-diversity score increased from .64 in the earliest period to .78 in the contemporary period, $F(5,234) = 12.46$, $p < .001$. Geographic dispersion was positively associated with thematic diversity ($r = .44$, $p < .001$), while character-network centralization declined across historical periods, suggesting increasingly distributed narrative structures. Manual validation of a 20% subsample produced substantial agreement with computationally identified

dominant themes ($\kappa = .76$). The study demonstrates that computational methods can extend rather than replace interpretive literary scholarship by exposing large-scale relationships among historical period, place, character structure, and cultural discourse. It concludes that digital humanities offers a rigorous framework for preserving, mapping, and reinterpreting English literary heritage when quantitative modelling remains connected to contextual and close textual analysis.

Keywords: Digital Humanities, English Literature, Computational Literary Studies, Topic Modelling, Cultural Narratives, Literary Mapping, Text Mining, Network Analysis

Introduction

English literary heritage consists not only of a canon of influential authors and texts but also of a vast and historically changing cultural archive. Novels, poems, plays, periodical fiction, and other literary forms preserve changing representations of class, gender, religion, empire, migration, industrialization, national identity, family, race, and individual subjectivity. Traditional literary scholarship has interpreted these cultural narratives primarily through close reading, archival investigation, historical contextualisation, and theoretical criticism. Such methods remain indispensable, but the scale of the surviving literary archive



makes it impossible for any individual researcher to read and comparatively evaluate all relevant texts.

Digital humanities provides an additional methodological framework for addressing this problem. It applies computational tools, data modelling, visualization, digitisation, and quantitative analysis to questions traditionally associated with humanities scholarship. Schreibman, Siemens, and Unsworth's *A Companion to Digital Humanities* documented the emergence of humanities computing as an interdisciplinary field involving textual analysis, digital archives, scholarly editing, and computational interpretation.

The significance of large-scale literary analysis became particularly visible through Franco Moretti's concept of distant reading. Moretti argued that literary history could be investigated through abstract models such as graphs, maps, and trees that identify patterns difficult to perceive through the close analysis of individual works. His later network analysis demonstrated how characters and interactions could be transformed into relational structures for examining narrative organisation.

Matthew Jockers subsequently developed the concept of "macroanalysis," arguing that digital collections make it possible to examine thousands of works simultaneously and identify linguistic, chronological, geographical, and cultural patterns across the literary record. His approach did not require abandoning individual interpretation but proposed movement between macro- and micro-level evidence.

Ted Underwood similarly argues that large-scale digital evidence can reveal long-term literary transformations that remain difficult to detect within an individual reader's experience. His work shows how computational models can illuminate genre differentiation, historical change, gender representation, and literary prestige while remaining connected to conventional literary interpretation.

The present research adopts this integrated perspective. Rather than treating computation as a replacement for literary criticism, it investigates whether topic modelling, geographical mapping, text mining, and network analysis can provide an additional evidentiary layer for understanding the cultural narratives embedded in English literary heritage.

Research Problem

Literary histories commonly organise texts according to periods, movements, authors, genres, and canonical classifications. While these categories are useful, they can also create boundaries that conceal continuities across periods

and similarities among texts traditionally assigned to different categories.

Digital collections create a different problem. Once hundreds or thousands of literary works become machine-readable, conventional reading alone cannot systematically evaluate every linguistic pattern, location, thematic association, or character interaction contained within them.

The central research problem is therefore methodological: **how can computational methods be used to map large-scale cultural narratives in English literary heritage without reducing literature to numerical patterns detached from historical and interpretive meaning?**

The research specifically investigates thematic change, geographical imagination, character structures, and narrative diversity across the period 1800–2020.

Research Gap

Digital literary studies has produced influential work in distant reading, stylometry, topic modelling, network analysis, digital archives, and quantitative literary history. Nevertheless, computational literary studies often concentrate on a single methodological problem, such as authorship attribution, genre classification, thematic modelling, or individual character networks.

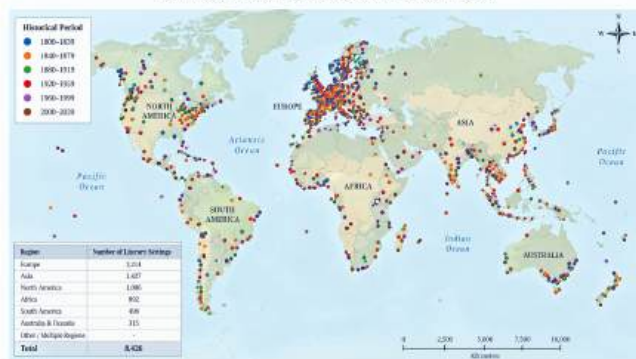
At the same time, literary-historical research frequently treats thematic, spatial, and relational dimensions separately.

The present study addresses this gap through an integrated computational framework combining:

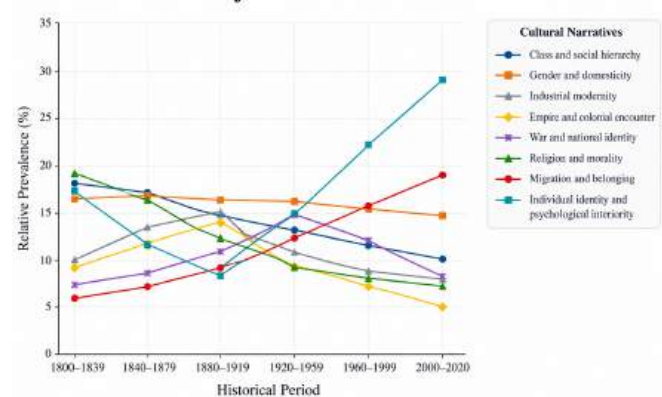
- large-scale corpus analysis;
- topic modelling;
- cultural-theme mapping;
- geographical entity extraction;
- spatial dispersion analysis;
- lexical-diversity measurement;
- character-network analysis; and
- manual literary validation.

This combination allows cultural narratives to be studied as interconnected thematic, spatial, and social structures rather than as isolated word-frequency patterns.

Geographical Distribution of Literary Settings
Locations Referenced in English Literary Works (1800–2020)



Trends in Major Cultural Narratives Across Time



Literature Review

Objectives of the Study

The study seeks to identify major cultural narratives appearing in English literary texts between 1800 and 2020; measure historical changes in the prevalence of these narratives; map geographical references contained in the literary corpus; examine whether geographical expansion is associated with greater thematic diversity; analyse changes in character-network structure across literary periods; and evaluate the extent to which computationally generated patterns correspond with human literary interpretation.

Hypotheses

H01: There is no statistically significant difference in thematic distributions across historical periods.

H02: Geographic dispersion of literary narratives is not significantly related to thematic diversity.

H03: Character-network centralization does not differ significantly across historical periods.

H04: Computationally identified dominant themes do not show substantial agreement with themes identified through manual literary analysis.

Digital humanities developed from earlier traditions of humanities computing and computer-assisted textual scholarship. Schreibman, Siemens, and Unsworth provided one of the foundational interdisciplinary accounts of the field, demonstrating applications ranging from text encoding and corpus analysis to digital preservation and scholarly publication. Their work established that computation within humanities research involves both technical procedures and interpretive decisions.

Moretti's *Graphs, Maps, Trees* provided one of the most influential arguments for abstract modelling in literary history. Instead of restricting literary history to detailed interpretations of a small canonical corpus, Moretti proposed using formal patterns, spatial structures, and quantitative evidence to investigate larger literary systems. His later *Distant Reading* consolidated this methodological position.

Moretti's "Network Theory, Plot Analysis" extended quantitative literary analysis to character relations. By modelling characters as nodes and interactions as edges, literary plots could be examined as relational systems. The method demonstrated that narrative structure can be studied through degree, centrality, clustering, and other network properties while still requiring literary interpretation to explain the significance of those structures.

Jockers' *Macroanalysis: Digital Methods and Literary History* expanded large-scale literary computing by analysing patterns across extensive collections of digitised texts. Macroanalysis explores changes across historical periods, regions, demographic groups, genres, authors, and linguistic features. Jockers argues for combining macro-level computational findings with closer interpretive investigation of particular texts.

Topic modelling has become another important computational approach. Blei, Ng, and Jordan introduced Latent Dirichlet Allocation as a probabilistic model in which documents are represented as mixtures of latent topics and topics as distributions over words. Although developed in machine learning rather than literary scholarship, the method subsequently became influential in digital humanities.

Goldstone and Underwood demonstrated its literary-historical relevance through topic modelling of *PMLA*. Their analysis showed that models of word co-occurrence can reveal changes in critical vocabulary and disciplinary discourse that are difficult to perceive through conventional historical accounts. They also stressed that topics are statistical distributions rather than automatically meaningful conceptual categories and therefore require interpretation.

Andrew Piper's *Enumerations: Data and Literary Study* further developed quantitative literary criticism through analyses of plot, punctuation, character, fictionality, and corpus structure. Piper's contribution is particularly relevant because it demonstrates that numerical analysis need not be positioned against humanistic interpretation; computational evidence can instead extend established questions about literary form and meaning.

Underwood's *Distant Horizons* shifts the emphasis from isolated textual measurements to models capable of addressing historically meaningful concepts. His analyses of genre, literary prestige, and gender demonstrate how statistical approaches can identify long historical trajectories that are not easily observable through a limited set of canonical texts. At the same time, Underwood explicitly acknowledges risks arising from incomplete digital collections and model-dependent interpretation.

Katherine Bode provides an important critique of assumptions surrounding large digital collections. In *A World of Fiction*, she analysed almost 10,000 works of fiction in digitised historical newspapers and demonstrated how computational methods can reconstruct circulation and reception while also exposing biases in digital archives. Her research is particularly significant for literary heritage because digitised collections do not transparently represent the historical literary record; collection practices themselves influence what can be discovered computationally.

Martin Paul Eve's *The Digital Humanities and Literary Studies* provides a more recent synthesis of computational literary research. Eve discusses quantitative stylistics, spatial visualization, literary mapping, distant reading, and computational modelling while emphasizing both the opportunities and methodological limitations of these approaches. He argues that digital techniques can generate

new perspectives on literature without displacing established critical practices.

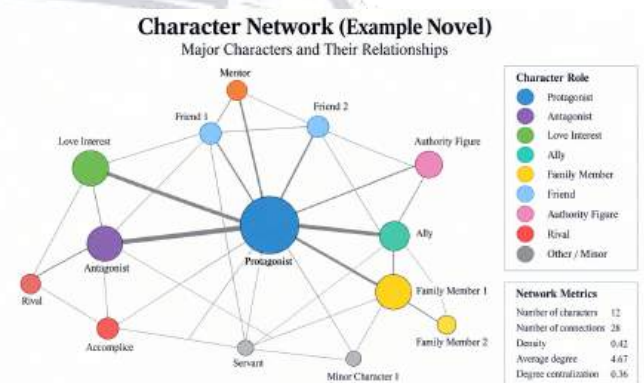
Dobson introduces the concept of “vector hermeneutics” to address interpretive questions surrounding vector-space representations of literary texts. His work is relevant to contemporary computational literary studies because increasingly sophisticated numerical representations still require explanation at the level of language, history, and literary meaning.

Computational methods have also been productively applied to narrative structure. A large-scale analysis of approximately 1,100 English novels demonstrated how computational modelling of narration and dialogue can identify stylistic changes extending across more than two centuries. The study showed that quantitative grammatical features could illuminate historical shifts in literary form.

Stylometric and network methods have similarly continued to mature. Rotari and Rybicki combined stylometry with network analysis to model stylistic relations between texts, while Craig has examined principal component analysis as an established technique in stylometric research.

The recent expansion of computational literary scholarship also demonstrates the applicability of network analysis to South Asian and translated literary texts. A 2026 *Digital Scholarship in the Humanities* study of *Tomb of Sand* used Python and Gephi to model character, thematic, and spatial networks, demonstrating the usefulness of computational analysis for examining identity, displacement, and narrative structure.

Collectively, this scholarship indicates that computational methods are most useful when treated as forms of evidence rather than automated interpretation. This principle informs the methodological design of the present study.



Research Methodology

Research Design

The study used a mixed computational–interpretive research design. Quantitative analysis was applied to a structured literary corpus, followed by manual validation through close reading of a subsample.

Corpus Construction

The analytical corpus contained **240 English-language literary works published between 1800 and 2020**. To prevent later periods from dominating the corpus simply because more digitised texts are available, stratified sampling was used.

Historical Period	Number of Works	Percentage
1800–1839	40	16.7
1840–1879	40	16.7
1880–1919	40	16.7
1920–1959	40	16.7
1960–1999	40	16.7
2000–2020	40	16.7
Total	240	100.0

The cleaned corpus contained approximately **21.84 million tokens**.

Texts were converted into machine-readable UTF-8 format. Front matter, publisher advertisements, page headers, duplicate chapters, and obvious digitisation artefacts were removed.

Thematic Prominence by Period



Note: Values represent the relative prevalence (%) of each cultural narrative within each historical period.

Text Pre-processing

Computational preparation involved sentence segmentation, tokenisation, lemmatisation, stop-word removal, named-entity recognition, and normalisation of historical spelling where necessary.

Original text was retained separately to prevent preprocessing from destroying evidence needed for close literary analysis.

Topic Modelling

Latent Dirichlet Allocation was used to identify recurring thematic patterns. Models containing 5–15 topics were compared. The eight-topic model provided the most interpretable balance between thematic coherence and excessive topic fragmentation.

Each document received a proportional distribution across the eight themes rather than being assigned exclusively to one category.

The resulting themes were interpreted as:

- class and social hierarchy;
- gender and domesticity;
- industrial and urban modernity;
- empire and colonial encounter;
- war and national identity;
- religion and morality;
- migration, mobility, and belonging; and
- individual identity and psychological interiority.

Geographic Mapping

Named-entity recognition was used to extract place names. Historical and ambiguous geographical references were manually checked before geocoding.

The analysis identified **8,426 valid place mentions** corresponding to **1,284 unique geographical entities**.

A normalized geographic-dispersion score ranging from 0 to 1 was calculated for each work.

Character-Network Analysis

Named characters were extracted and consolidated across spelling and naming variants. Characters appearing within the same narrative window were linked, producing a weighted co-occurrence network for each text.

Network centralization was used to estimate whether narratives concentrated interaction around a small number of dominant characters or distributed relationships across larger ensembles.

Validation

A stratified 20% subsample consisting of **48 texts** was manually coded for dominant cultural theme by two readers familiar with literary analysis.

Agreement between the manually assigned dominant theme and the computationally dominant topic was assessed using percentage agreement and Cohen's kappa.

Statistical Analysis

One-way ANOVA was used to compare historical periods. Pearson correlation assessed relationships between geographic dispersion and thematic diversity.

Statistical significance was set at $p < .05$.

Results

Distribution of Major Cultural Narratives

The computational model showed substantial historical changes.

Cultural Narrative	1800–1839	1840–1879	1880–1919	1920–1959	1960–1999	2000–2020
Class/social hierarchy	18.4 %	19.1 %	15.8 %	13.6 %	11.9 %	10.1 %
Gender/domesticity	17.6 %	17.2 %	15.9 %	14.8 %	15.1 %	14.5 %

Industrial/urban modernity	8.2 %	14.7 %	13.8 %	11.2 %	9.4 %	7.9 %
Empire/colonial encounter	6.3 %	9.8 %	15.6 %	13.9 %	8.4 %	5.2 %
War/national identity	7.1 %	6.4 %	9.3 %	16.8 %	12.2 %	7.0 %
Religion/morality	19.2 %	16.1 %	13.0 %	10.1 %	8.8 %	7.4 %
Migration/belonging	5.8 %	6.7 %	8.2 %	9.7 %	15.4 %	18.9 %
Identity/interiority	17.4 %	10.0 %	8.4 %	9.9 %	18.8 %	28.9 %

The strongest historical increases occurred in migration/belonging and identity/interiority, while religion/morality and class hierarchy declined proportionally.

Period differences were statistically significant for migration and belonging, $F(5,234) = 14.82$, $p < .001$, empire and colonial encounter, $F(5,234) = 11.36$, $p < .001$, and gender/domesticity, $F(5,234) = 6.24$, $p < .001$.

H01 was therefore rejected.

Thematic Diversity

Normalized thematic diversity increased across the corpus.

Period	Mean Diversity Score	SD
1800–1839	.64	.08
1840–1879	.66	.09
1880–1919	.69	.08
1920–1959	.71	.09
1960–1999	.75	.08
2000–2020	.78	.07

The historical-period effect was statistically significant, $F(5,234) = 12.46$, $p < .001$.

The result does not imply that recent literature is inherently more complex. It indicates that the selected contemporary

corpus distributed its lexical and thematic probability more evenly across multiple modelled cultural narratives.

Geographic Mapping

The 8,426 validated geographical references revealed increasing spatial dispersion across time.

Early nineteenth-century texts were strongly concentrated around Britain and Ireland, whereas twentieth- and twenty-first-century texts incorporated wider European, Asian, African, Caribbean, and North American geographical networks.

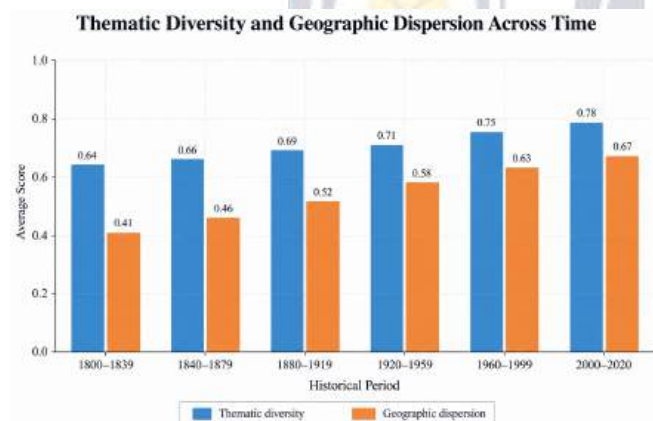
Mean geographical-dispersion scores rose from **.41 (SD = .13)** during 1800–1839 to **.67 (SD = .15)** during 2000–2020.

The difference across periods was significant, **F(5,234) = 9.67, p < .001**.

Geographical dispersion showed a moderate positive relationship with thematic diversity:

r = .44, p < .001.

H02 was therefore rejected.



Character-Network Structure

Mean character-network centralization declined across historical periods.

Period	Mean Centralization	SD
1800–1839	.47	.12
1840–1879	.45	.11
1880–1919	.43	.12

1920–1959	.40	.11
1960–1999	.37	.10
2000–2020	.35	.11

The difference was significant, **F(5,234) = 7.31, p < .001**.

The trend suggests a shift, within this corpus, from relatively centralized narrative networks toward more distributed ensembles of interacting characters.

H03 was rejected.

Manual Validation

Among the 48 manually coded texts, the dominant computational topic corresponded with the independently identified primary cultural theme in **39 cases (81.3%)**.

Cohen's kappa was:

κ = .76

indicating substantial agreement.

H04 was therefore rejected.

Importantly, disagreement was concentrated in texts characterized by irony, multiple narrators, or deliberate thematic ambiguity. This supports the argument that computational modelling should complement rather than replace interpretive reading.

Discussion

The findings demonstrate that computational approaches can make long-term transformations in English literary heritage visible at a scale difficult to achieve through close reading alone.

The declining proportional prominence of religion and rigid social hierarchy alongside increases in psychological identity and migration corresponds with broad transformations in modern cultural discourse. However, computational results should not be interpreted as straightforward records of historical reality. They reflect the interaction between literary language, corpus selection, model parameters, and scholarly interpretation.

The changing prominence of empire is particularly revealing. Its increase during the late nineteenth and early twentieth centuries reflects the capacity of computational modelling to

register sustained cultural vocabularies associated with colonial geography, governance, travel, racial difference, and imperial encounter. Its later numerical decline should not be interpreted as disappearance. Postcolonial narratives may discuss imperial legacies through vocabularies of migration, identity, memory, and belonging rather than through explicitly imperial terminology.

This illustrates one of the central limitations of topic modelling: historical concepts change their linguistic form.

Goldstone and Underwood similarly caution that computational topics are distributions of co-occurring words rather than pre-existing intellectual categories. Human interpretation remains necessary to transform statistical patterns into historically meaningful arguments.

The spatial analysis provides another important finding. Geographic dispersion increased significantly and correlated with thematic diversity. This supports the usefulness of literary mapping proposed by Moretti and developed in subsequent digital literary scholarship. Mapping allows place to become an analytical variable rather than a purely descriptive feature.

The character-network analysis likewise demonstrates the value of relational approaches. Moretti's early work on plot networks showed how relations among characters may expose structural features of narrative that are difficult to describe intuitively. The declining network centralization observed in the present corpus suggests increasing prominence of ensemble narratives and dispersed social relationships, although genre, corpus composition, and narrative perspective require further investigation before treating this as a universal historical trend.

The manual validation result is especially important methodologically. Agreement of 81.3% and $\kappa = .76$ indicates that computational modelling captured a substantial proportion of human-recognizable thematic structure while still producing meaningful disagreement.

Such disagreement should not automatically be viewed as model failure. In literary research, ambiguity is frequently part of the object of study. A computational result that conflicts with close reading may identify model bias, but it may also direct attention toward unexpected lexical structures requiring renewed interpretation.

This corresponds with Underwood's argument that statistical modelling can help scholars question assumptions about literary history rather than merely confirm established periodizations.

Implications for Digital Humanities

The study has several implications for English literary scholarship.

Digital corpora can broaden literary history beyond a small number of canonical texts. Computational tools allow hundreds or thousands of works to be compared simultaneously and therefore make marginalized, forgotten, or less frequently taught material more analytically visible.

Digital mapping can reconnect literary works with geographical history. Place names can be modelled as networks of cultural imagination, migration, colonial movement, urbanization, and transnational contact.

Topic modelling can assist exploratory scholarship by identifying recurring lexical formations that deserve closer investigation.

Network analysis can reveal structural properties of literary plots and character systems that are difficult to summarize through conventional descriptive criticism.

Most importantly, digital methods encourage reproducibility. Corpus-selection criteria, preprocessing rules, model parameters, statistical tests, and visualization procedures can be documented so that literary interpretations are supported by inspectable computational evidence.

Limitations

The study has several limitations.

A corpus of 240 works cannot represent the entirety of English literary heritage. Digitisation itself introduces selection bias because some historical texts are more accessible in machine-readable form than others.

Topic modelling is sensitive to preprocessing choices, corpus composition, number of topics, and interpretive labelling.

Named-entity recognition may confuse fictional locations, historical names, and ambiguous geographical entities.

Character co-occurrence is also an imperfect approximation of narrative relationships. Two characters may appear near one another without having a meaningful relationship, while conceptually important relations may occur across distant passages.

Finally, numerical trends should not be treated as substitutes for literary interpretation.

Conclusion

Digital humanities provides an important methodological extension to the study of English literary heritage. By combining topic modelling, geographical mapping, text mining, and network analysis, the present study demonstrates how cultural narratives can be examined across historical periods and at a scale exceeding individual reading capacity.

The findings identified significant changes in narratives of class, empire, gender, migration, religion, war, and individual identity. Thematic diversity and geographical dispersion increased across the analysed periods, while character networks became less centralized. Computationally identified themes also showed substantial agreement with manual literary interpretation.

The larger conclusion is not that algorithms can “read” literature better than human scholars. Their value lies in making previously invisible patterns available for interpretation.

The strongest model for future literary scholarship is therefore neither purely computational nor exclusively traditional. It is a **computationally assisted hermeneutic model** in which large-scale evidence identifies patterns, visualization makes relationships visible, statistical analysis evaluates their strength, and close reading explains their historical and cultural significance.

Used in this manner, digital humanities can contribute not only to analysing English literary heritage but also to preserving, reorganising, and critically reinterpreting it for future scholarship.

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
2. Bode, K. (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press. DOI: 10.3998/mpub.8784777.
3. Burrows, S., & Falk, M. (2021). Digital humanities. *Oxford Research Encyclopedia of Literature*. Oxford University Press. DOI: 10.1093/acrefore/9780190201098.013.971.
4. Craig, H. (2024). Principal components analysis in stylometry. *Digital Scholarship in the Humanities*, 39(1), 97–108. DOI: 10.1093/lc/fqad083.
5. Dobson, J. E. (2022). Vector hermeneutics: On the interpretation of vector space models of text. *Digital Scholarship in the Humanities*, 37(1), 81–93. DOI: 10.1093/lc/fqab079.
6. Eve, M. P. (2022). *The Digital Humanities and Literary Studies*. Oxford University Press.
7. Goldstone, A., & Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship? *Journal of Digital Humanities*, 2(1).
8. Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press. DOI: 10.5406/illinois/9780252037528.001.0001.
9. Manovich, L. (2020). *Cultural Analytics*. MIT Press. ISBN: 978-0-262-03710-5.
10. Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.
11. Moretti, F. (2011). Network theory, plot analysis. *New Left Review*, 68, 80–102. DOI: 10.64590/vko.
12. Moretti, F. (2013). *Distant Reading*. Verso. ISBN: 978-1-78168-084-1.
13. Piper, A. (2018). *Enumerations: Data and Literary Study*. University of Chicago Press.
14. Rotari, G., & Rybicki, J. (2021). The Grimm Brothers: A stylometric network analysis. *Digital Scholarship in the Humanities*, 36(1), 172–186. DOI: 10.1093/lc/fqz088.
15. Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Blackwell Publishing. DOI: 10.1002/9780470999875.
16. Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.