Social Science Chronicle

OPEN ACCESS

Peer-Reviewed, International, Academic Research Journal

ISSN : 3048-6297

REVIEW ARTICLE

Textual Transmission, Intertextual Inference and Provenance Stewardship in Computational Philology

Alexandru Dumitrescu[1]* and Elena Stan[1]

[1]Faculty of Letters, University of Oradea, Romania.

** Corresponding Author*

Citation
Dumitrescu, A. & Stan, E. (2025). Textual Transmission, Intertextual Inference and Provenance Stewardship in Computational Philology. *Social Science Chronicle*, Vol. 5, Issue - 1, pp. 1-13.

Digital Object Identifier (DOI)
https://doi.org/10.56106/ssc.2025.004

Received - February 12, 2025
Accepted - July 18, 2025
Published - July 25, 2025

Web-Link
All the contents of this peer reviewed article as well as author details are available at http://socialsciencechronicle.com/article-ssc-2025-004

Abstract

This article presents a unified framework for contemporary philology that brings together technical methods, shared data standards, and ethical governance within a single, coherent research lifecycle. It addresses the current fragmentation of digital philological practice by integrating ten areas that are increasingly interdependent but often treated separately. These include AI-based handwritten text recognition and OCR for manu-scripts and inscriptions; digital fragmentology and virtual reunification; and computational stemmatology for modeling textual transmission using both trees and networks. The framework also incorporates FAIR-aligned IIIF and Linked Open Data infrastructures, justice-oriented approaches to provenance and restitution, and methods for detecting cross-lingual text reuse. Further components of the pipeline include authorship analysis through stylometry and representation learning, TEI-based scholarly editions supported by continuous integration, natural language processing of marginalia and paratexts, and capacity building for historically under-resourced scripts and scholarly communities. Across all components, the paper identifies shared methodological principles: diacritic-sensitive error modeling, line-based citability, disciplined critical apparatuses with explicit intervention markers, and evaluation protocols that emphasize calibrated confidence and principled non-decision where evidence is insufficient. To ensure methodological robustness, the paper defines minimal conditions for falsifiability, including corruption testing, bootstrap based uncertainty estimation, and evaluation across multiple editions or witnesses. It also articulates minimal conditions for social and ethical legitimacy, such as tiered access models, renewable consent, transparent contributor records, and clearly defined takedown proce-dures. Five concise tables summarize workflows, assumptions, risks, and assurance signals in formats intended for practical use by research labs, libraries, and community partners. Overall, the article offers a practical blueprint for scaling philological research without erasing the specificity of individual witnesses, for accelerating computational analysis without displacing scholarly judgment, and for making philological claims reproducible, open to challenge, and ethically grounded across languages, materials, and institutional contexts.

Keywords

Digital Humanities, Digital Philology, Handwritten Text Recognition, Optical Character Recognition, Linked Open Data, FAIR Principles, Textual Criticism, Stylometry, Intertextuality, Epigraphy.

1. Introduction

Philology operates within a technical and ethical ecology that demands integrated grammar of practice. High fidelity imaging, layout analysis, and handwriting recognition can convert palm leaf folios, birch bark strips, parchment codices, early printed witnesses, and epigraphic squeezes into machine actionable strings with quantified uncertainty rather than static surrogates. Interoperable publication through web protocols makes page regions, lines, and tokens addressable and citable across repositories, which enables verifiable claims and repeatable workflows.

Representation learning expands the palette of stylometric and intertextual features, yet it also introduces confounds that must be measured, ablated, and controlled through transparent protocols (Elwert, 2021; Bories, 2022). Decolonial governance reframes provenance, consent, credit, and benefit sharing as first order design constraints that shape acquisition, modeling, and release. The present review advances a lifecycle that couples acquisition, normalization, representation, analytical inference, scholarly edition making, and access governance within one pipeline that remains accountable to material specificity and scholarly adjudication. The purpose is pragmatic and immediate, since institutions require concrete recipes that scale without epistemic drift. Section 2 opens with sources and infrastructures and will introduce Table 1, which formalizes acquisition to access pathways for major substrate families and script ecologies.

The review synthesizes ten frontiers that now behave as a single circuit rather than discrete topics. The opening cluster treats sources and infrastructures, including recognition for manuscripts and inscriptions, digital fragment reassembly, and interoperable publication through shared identifiers and annotation anchors. The next cluster treats representation and transmission, including TEI centric encodings of witnesses and variants, disciplined collation, and computational stemmatology for tree like and network like copying with explicit uncertainty. A third cluster treats analytical inference over content and context, including cross lingual text reuse, authorship verification with open set protocols, and the extraction of marginalia and paratexts as evidence for reception and circulation. A fourth cluster centers decolonial governance that binds technical pipelines to ethical obligations. A fifth cluster treats evaluation, pedagogy, sustainability, and strategic roadmaps. The architecture is pipeline aware, since imaging choices and rights constraints propagate into normalization, apparatus design, and downstream inference. Section 2 will operationalize these linkages and will call Table 1 to consolidate infrastructural decisions.

Material philology grounds the insistence that witnesses are not interchangeable because ductus, ruling, substrate, ink, and mise en page encode information that resists flattening into a single reading text. Hermeneutics grounds interpretive responsibility by demanding that algorithmic signals be translated into arguments that remain accountable to genre, chronology, and social context. Graph theory and phylogenetics ground transmission modeling by offering explicit formalisms for lineage and contamination with uncertainty treated as a parameter to be estimated rather than a nuisance to be ignored. Information theory grounds recognition and collation by framing signal and noise, redundancy and constraint, error profiles and channel capacity within measurable quantities (Tuttle, 2021; De Gussem, 2022; Piotrowski, 2022). Representation learning grounds authorship and reuse by extracting features that remain stable across scribal hands, orthographies, and scripts while ablation and calibration expose topic leakage and edition artifacts. Science and technology studies and critical archival studies ground governance by showing how infrastructures redistribute authority and by requiring durable protocols for consent, credit, takedown, and redress. These anchors structure the review so that each technical move is paired with an epistemic check and a social contract.

The review contributes a unified lifecycle that institutions of varied scale can adopt without exotic infrastructure. It provides a portable vocabulary for aligning imaging, annotation, encoding, and publication so that witnesses remain citable at line and token granularity and so that analytical claims can be reproduced with controlled randomness and documented environments. It reframes authorship work as verification with calibrated abstention rather than only closed set attribution, and it places stemmatology within sensitivity analysis that reports parameter dependence and model fragility. It treats decolonial governance as a workflow component with consent refresh, tiered access, redaction protocols, and credit pathways that can be implemented within routine operations. It supplies five compact tables that compress complex practice into actionable matrices. Table 1 organizes acquisition to access choices by source family. Table 2 organizes representation and transmission methods with assumptions and failure modes. Table 3 organizes analytical tasks with robustness and ablation regimes. Table 4 organizes ethics and governance risks with mitigations. Table 5 organizes road mapped initiatives with resources, outputs, and evaluation plans.

Section 2 surveys material sources, imaging modalities, recognition pipelines, fragment reconstruction strategies, and interoperable publication infrastructures and introduces Table 1 that consolidates end to end decisions for major substrates and scripts. Section 3 advances representation and transmission by detailing TEI apparatus design, disciplined collation, and computational stemmatology and culminates in Table 2 that enumerates methods, assumptions, pitfalls, and robustness checks. Section 4 addresses analytical inference over content and context, including cross lingual reuse, authorship verification, and marginalia extraction and presents Table 3 that aligns tasks, corpora, features, metrics, and ablation protocols. Section 5 centers governance and decolonial practice with concrete mechanisms for provenance transparency, consent, credit, and responsible release and presents Table 4 that formalizes risk and mitigation across stakeholder constellations. Section 6 integrates measurement, benchmarks, pedagogy, sustainability, and strategic planning and provides Table 5 that translates ambition into resourced and evaluable initiatives. Section 7 synthesizes the implications and issues a call for coordinated investment that couples technical excellence with ethical stewardship.

2. Sources and Infrastructures for Philological AI

Philological corpora present a variegated ecology in which substrates, scripts, and conservation states co determine what can be computed without epistemic loss. Parchment and paper codices carry palimpsested inks, cockled folia, and bleed through that distort layout priors and confound naive binarization, while palm leaf and birch bark witnesses introduce brittle fibers, insect galleries, and fissures that fracture baselines and segmenters (Thomassen, 2021; Cimiano, 2020; Jackson, 2021). Stone and metal epigraphy impose relief driven shadows, variable patination, and chipped graphemes that require reflectance aware imaging and contour sensitive tracing rather than simple raster thresholds. Early modern print complicates matters with broken sorts, uneven inking, and antiquated ligatures that defeat generic OCR unless glyph inventories and language models are time specific. Mixed media dossiers juxtapose handwriting, print, seals, and pasted images that demand page object detection and region wise pipelines rather than monolithic recognizers (Biber, 2020; Camps, 2021; Bambaci, 2021). Community restricted holdings introduce protocol constraints where certain leaves, sacred

marginalia, or ritual diagrams can never circulate openly, which requires federated training, tiered access, and consent refresh to avoid extractive use. The section that follows treats imaging and pre processing as a systems problem and anticipates the end to end patterns that are consolidated in Table 1 later in this section.

Imaging, Digitization, and Pre Processing at Scale

Imaging must be treated as measurement rather than photography so that each modality captures a feature space that downstream models can leverage without brittle heuristics. High resolution RGB, narrowband multispectral stacks, and reflectance transformation imaging for epigraphy provide complementary signals that support denoising, ink separation, and relief normalization, while calibrated color targets and illumination geometry ensure repeatability across sessions and sites (Roelli, 2020; Fitzmaurice, 2022; Weber, 2020). Micro computed tomography can reveal binding structures, paste downs, and concealed leaves in composite dossiers, though radiation budgets and conservation ethics delimit use in fragile materials. Pre processing begins with physical warpage correction, baseline and region detection, and glyph aware binarization that preserves diacritics and hairlines. File formats must align with downstream semantics, which means archival TIFF or lossless PNG for master images, IIIF image and presentation manifests for addressable canvases and sequences, and standards compliant annotation layers for line and region coordinates (Mitcham, 2020; Kudinova, 2021; Dörpinghaus, 2022). Rights metadata and embargo markers must be embedded at ingest and propagated across all derivatives, since misaligned rights flags create legal debt that later stalls publication. These imaging and pre processing decisions map directly into the acquisition to access patterns enumerated in Table 1 in sub section 2.3, where source ecologies are paired with recognition regimes, identifiers, and risk controls.

AI HTR and OCR Pipelines for Multiscript Corpora

Recognition in philological settings must accommodate heterogeneous scripts, orthographic drift, and substrate induced noise, which requires hybrid decoders that join convolutional and transformer layers with sequence criteria such as CTC and attention. Grapheme cluster modeling becomes indispensable for Indic and Semitic scripts with stacked diacritics and conjuncts, while right to left and bidirectional layouts demand explicit directional embeddings to prevent mirroring artifacts. Post correction benefits from constrained language models that are trained on diachronic lexica and edition grade witnesses rather than contemporary corpora, since topic drift and orthographic modernization can otherwise launder historical signal (Palladino, 2022; Maiocchi, 2021; Cugliana, 2022). Uncertainty must be quantified at line and token levels so that editorial workflows can route low confidence spans for adjudication rather than silently accepting false fluency. Export pathways should preserve provenance and auditability, which means retaining PAGE XML or ALTO for layout semantics and projecting validated readings into TEI with facsimile anchors for line level citability. These methodological choices must be tuned to source families, which the acquisition to access matrix in Table 1 makes operational by pairing concrete ecologies with compact recipes across imaging, recognition, identifiers, and risk mitigations.

Table 1. Acquisition to Access Pipelines Across Major Source Ecologies

Source Ecology and Script Typology	Imaging and Pre-Processing Regimen	Recognition and Post Correction Modality	Interoperability and Persistent Identifier Strategy	Risk Profile and Mitigation Protocols
Parchment and Paper Codices	*High res RGB, selective multispectral, dewarp, debleed, baselines, zoning*	*Transformer HTR with CTC attention, diachronic lexicon, n best, human QA, TEI via PAGE*	*IIIF canvases with line anchors, stable URIs, versioned TEI*	*Mold, cockling, script shifts, private notes; denoise, active learn, redaction, rights flags*
Palm Leaf and Birch Bark Manuscripts	*Raking, narrowband multispectral, fiber aware dewarp, crack tracking*	*Grapheme cluster HTR, diacritic sensitive decode, conservative LM, transliteration, curator QA*	*IIIF folio sequences, transliteration registries, recto verso PIDs, TEI facsimiles*	*Fragility, lacunae, cultural sensitivity; low exposure, lacuna tags, consent tokens, embargo*
Stone and Metal Epigraphy	*RTI, photogrammetry, shadow normalization, contour extraction*	*Epigraphic line tracing HTR, relief aware features, lacuna penalty, orthography aware correction*	*Image centric IIIF with zone alignment, PIDs for squeezes and casts, site registry*	*Lighting bias, ownership claims, site limits; RTI normalize, provenance audit, access windows*
Early Printed Books and Pamphlets	*Uniform RGB, deskew, type area detect, debleed*	*Historical OCR, ligature aware tokenization, period LMs, sample-based correction*	*ALTO to TEI, IIIF per volume, PIDs for signatures and gatherings, stable loci*	*Broken sorts, foxing, binding stress, unclear rights; cradle imaging, due diligence, page fallback*

Mixed Media and Composite Dossiers	*High res RGB, micro CT when ethical, object detection, layer separation*	*Multimodal HTR plus OCR, stamp classifiers, NER linking, curator adjudication*	*Web annotations linking canvases to TEI, component PIDs, bundle provenance graphs*	*Personal data, composite authorship, hidden leaves; privacy redaction, contributor logs, anomaly checks, takedown*
Community Restricted and Sacred Holdings	*Minimal imaging by protocol, metadata only if allowed*	*On premise training, federated learning, differential privacy, calibrated abstention*	*Tiered access, opaque PIDs, consent tokens, time bound embargoes, governance records*	*Cultural harm, ritual secrecy, misuse; community veto, benefit share, consent refresh, audit trails*

The matrix operationalizes a disciplined grammar of practice by compressing complex workflows into compact recipes that remain sensitive to materiality, script ecology, and governance realities. Each row pairs a substrate family with an imaging regimen that encodes physical constraints as measurable signals, a recognition stack that respects orthography and diachrony, an interoperability plan that makes lines and loci citable across platforms, and a risk posture that treats consent, privacy, and conservation as first class. Subsequent sub sections cite Table 1 when prescribing actionable combinations, and later sections reuse the same identifiers to stabilize variant encoding, stemmatic inference, and analytical audits, which avoids the brittle reinvention that often fragments projects at scale.

Digital Fragmentology and Virtual Reunification

Fragmentary corpora demand a synthesis of visual evidence, codicological heuristics, and probabilistic graph assembly to reconstruct dispersed codices, scrolls, and archival dossiers. Fiber texture, ruling patterns, watermarks, and parchment hue provide orthogonal cues that help propose edges for leaf adjacency, while ink spectral signatures and scribal ductus stabilize clustering across uncertain joins. Page geometry and quire arithmetic constrain admissible configurations so that assembly does not violate codicological plausibility, and confidence scores must be carried through the graph so that users can distinguish firm joins from speculative hypotheses (Ghali, 2023; Gryaznova, 2022; Li, 2020). Provenance and legality impose non technical boundaries, since reunification can imply claims that exceed custodial agreements or national legislation. Table 1 informs fragment workflows by enumerating imaging stacks that maximize discriminative features and by insisting on persistent identifiers for leaves and loci so that hypothesized joins can be logged, debated, and revised without orphaning prior citations. Virtual reunification then becomes a reversible and auditable hypothesis rather than an irreversible claim.

Interoperable Infrastructures for FAIR IIIF and Linked Data

Interoperability turns isolated scans into computable ecosystems where objects, texts, and annotations can circulate without loss of provenance or granularity. FAIR principles require that resources be findable through stable discovery metadata, accessible through documented protocols, interoperable through shared models, and reusable through clear licenses and provenance chains. IIIF image and presentation APIs furnish addressable canvases and sequences so that a line in TEI can point to a specific region in an image without brittle coordinates (Lamb, 2020; Gryaznova, 2022). Web annotations permit layered commentary and machine generated features to coexist with human notes, while authority linked vocabularies stabilize persons, places, and works across repositories. Persistent identifiers must be minted at witness and locus levels so that citations remain resolvable across editions and time. Table 1 already encodes minimal viable identifier strategies per source family, and those strategies become prerequisites for the representation and transmission pipelines in Section 3 where apparatus design and collation demand unambiguous anchors.

Low Resource Realities and Capacity Building

Equity in philological AI requires that under resourced scripts, community controlled archives, and small institutions can participate without ceding autonomy or incurring unpayable technical debt. Data creation must be collaborative and protocol driven, with annotation guidelines that encode grapheme inventories, ligature policies, and abbreviation expansions to avoid noisy gold standards. Training should prefer active learning, few shot adaptation, and federated regimes that keep sensitive pages on premises while still improving shared models, with differential privacy and curated abstention where cultural harm could result from overconfident outputs. Compute plans must be explicit and proportionate so that budgets cover inference latency, storage, and preservation rather than only initial training bursts. Capacity building should pair local stewards with traveling fellowships, shared model zoos with usage logs, and open curricula that bind philological method to evaluation literacy. Table 1 foregrounds these commitments by offering consent tokens, tiered access, and governance board records as first class components, which ensures that technical gains do not outrun ethical obligations or conservation realities.

3. Representation, Normalization and Transmission

From Images to Machine-Actionable Text

Representation begins where imaging leaves off, since layout semantics and line anchors from the acquisition phase determine what can be asserted about text boundaries, token identities, and apparatus granularity. PAGE XML and ALTO preserve baseline geometries and region zoning, which permits deterministic projection into TEI where facsimile pointers bind readings to specific loci (Middleton, 2024). Normalization must separate orthographic regularization from interpretive emendation so that the lineage of each graphemic choice remains auditable. Tokenization policies must respect grapheme clusters in scripts with stacked diacritics, ligature inventories in early print, and bidirectionality in mixed scripts, otherwise downstream alignment will hallucinate spurious variants. Lemmatization and morphological tagging should be curated as optional layers rather than baked

into base text, since over eager normalization can erase diachronic signal that later matters for stylometry or reuse. Confidence scores from recognition should survive into TEI as attributes so that low confidence spans can be weighted down in collation. Identifiers minted in Section 2 and referenced in Table 1 stabilize citations at witness and line levels, while the methodological choices for collation and transmission in this section are consolidated in Table 2, which enumerates tasks, assumptions, and failure controls for representation workflows.

Critical Apparatus and Variant Encoding

The apparatus must function as a constrained knowledge graph rather than a free form note, since each variant category encodes hypotheses about scribal action and transmission channels. TEI offers disciplined containers for lemma, reading, witness, and location, which allows a single locus to host substitutions, additions, deletions, transpositions, and orthographic alternants without collapsing their semantics. Editorial interventions such as conjecture and normalization require explicit flags so that later analyses can include or exclude them without ambiguity (Hatzel, 2023). Segmentation must be fine grained enough to capture micro variants yet coarse enough to remain computationally tractable, which argues for token level loci with alignment hints for clitics and enclitics. Apparatus density should track witness coverage rather than page count, since sparse traditions need different thresholds from massive ones. When apparatus content is exported for collation, the format should encode uncertainty and reading probabilities so that stemmatic inference does not treat every reading as equally firm. The decision matrix in Table 2 aligns apparatus design choices with algorithmic regimens and robustness checks, which ensures that representation does not outpace adjudication.

Computational Stemmatology and Transmission

Transmission analysis requires explicit models of copying, contamination, and repair rather than narrative metaphors. Tree oriented methods estimate ancestral relationships under assumptions of predominantly vertical inheritance with bounded homoplasy, while network-oriented methods allow horizontal transfer and mixed ancestry that better fit heavily redacted or school-based traditions (Adriansyah, 2024; Del Grosso, 2023). Collation noise interacts with model choice, since aggressive normalization can suppress true signal while lax policies can inflate spurious splits. Simulation studies help calibrate sensitivity to apparatus density, witness dropout, and reading informativeness, yet results must be grounded in the actual ecology of the tradition rather than generic priors. Parameter uncertainty should be surfaced through bootstraps and posterior sampling, and inference should permit abstention where data are insufficient. The compact matrix in Table 2 structures these choices into task aligned regimens with assumptions, failure modes, and reporting minima so that claims about archetypes and pathways remain falsifiable.

Table 2. Methods and Pitfalls in Textual Representation and Transmission

Task	Representation Schema	Algorithmic Regimen	Assumptions and Failure Modes	Robustness and Reporting
Collation Across Witnesses	*TEI loci with facsimile anchors, token level spans*	*Sequence alignment with gap costs, locality aware heuristics*	*Stable tokenization, limited orthographic drift; failure under noisy segmentation and clitic splits*	*Ablations on token policy, noise injections, inter annotator agreement, seed control*
Apparatus Construction and Encoding	*TEI apparatus with typed readings and flags*	*Rule based extraction, constraint validation, schema linting*	*Clear distinction between normalization and conjecture; failure when flags are absent or misused*	*Error budgets, proportion of flagged edits, schema validation logs, change diffs*
Stemma Inference Tree Oriented	*Variant matrix with informative readings only*	*Parsimony or likelihood on discrete characters, bootstrap support*	*Predominantly vertical transmission; failure with heavy contamination and homoplasy*	*Sensitivity to reading filters, bootstrap distributions, alternative cost models, abstention zones*
Contamination Aware Network Modeling	*Graph with weighted edges among witnesses*	*Split networks, neighbor net, Bayesian admixture*	*Mixed ancestry and lateral transfers; failure with sparse data and overfitting*	*Edge stability under resampling, admixture proportion intervals, null model checks*
Witness Clustering and Alignment	*Feature vectors from variants and orthographic profiles*	*Dimensionality reduction with clustering, alignment refinement*	*Clusters reflect transmission not topic or period; failure under edition artifacts*	*Cluster stability, silhouette and cophenetic indices, artifact audits, holdout witnesses*
Edition Validation and Versioning	*TEI with ODD constraints and provenance trails*	*Continuous integration, schema tests, diff aware release*	*Deterministic builds and citable loci; failure with drift across releases*	*Reproducible build hashes, URI persistence audits, rollback logs, environment capture*

Scholarly Edition Design and Lifecycle Sustainability

A modern edition is a living artifact whose integrity depends on encoding discipline, validation automation, and citability guarantees rather than artisanal heroics. ODD customization must pin the project specific constraints that govern apparatus types, segmentation policies, and authority linking so that contributors cannot drift into incompatible idioms. Continuous integration can compile TEI into deliverables such as reading texts, apparatus views, and API endpoints while enforcing schema tests, identifier checks, and link resolution (Rahmi, 2024; Yang, 2024; Bozhenkova, 2023). Versioning must preserve backward compatible URIs for witnesses and loci, otherwise scholarly citations will decay and downstream analytics will mismatch contexts. Accessibility and multilingual presentation should be treated as core requirements so that global audiences can interrogate the same object without loss of precision. Table 2 already enumerates edition validation and versioning as first class tasks with explicit reporting minima, and those minima should be enforced at release time so that reproducibility and credit remain auditable. Sustainability then becomes an engineering practice rather than an aspiration.

Pedagogical and Community Interfaces for Representation

Representation work scales only when pedagogy, community participation, and credit mechanisms are engineered into the workflow. Collaborative editing platforms should expose locus level tasks with machine suggested alignments and confidence scores so that novices can learn by adjudicating concrete micro decisions rather than ingesting abstract doctrine. Training materials must connect graphemic inventories, abbreviation policies, and variant categories to the actual TEI elements and ODD rules that govern them, which builds muscle memory for disciplined encoding (Krasniuk, 2024). Community partners should be able to contribute marginalia and paratext descriptions through structured annotation that binds directly to canvases and lines, with clear pathways for acknowledgment and authorship. Governance boards must supervise takedowns and consent refresh while also curating contributor logs and benefit sharing, which keeps representation accountable to those who steward the materials. The task and reporting minima in Table 2 double as a curriculum for capacity building, since each row translates into a teachable unit with inputs, outputs, and verifiable quality signals that align with the identifiers and rights frameworks established earlier.

4. Analytical Tasks on Content and Context

Cross Lingual Intertextuality and Text Reuse

Intertextual inference in a multilingual ecology requires representational neutrality across scripts, eras, and genres so that similarity metrics do not collapse into orthographic mimicry. Character shingles, morpheme level segmentation, and transliteration lattices provide resilient anchors when OCR or HTR noise perturbs glyph boundaries, while multilingual embeddings and alignment models capture semantic isomorphy across translation chains (Cowen-Breen, 2023; Tasheva, 2024; Zulfiya, 2024). Quotation, paraphrase, and allusive echo demand different granularity and windowing so that near verbatim reuse is not conflated with motif level recurrence. Thresholding must be calibrated against corruption suites that inject realistic noise from Section 2, and scoring must discount boilerplate passages that recur by convention rather than transmission. Provenance aware graphs can encode directionality when chronology and attestation are known, although abstention remains mandatory when temporal priors are ambiguous. The comparative constraints that govern these design choices are consolidated in Table 3 in sub section 4.2, which aligns tasks, feature spaces, confounds, and robustness regimens so that cross lingual detection remains falsifiable and replicable rather than impressionistic.

Authorship Attribution and Verification

Attribution work in historical corpora must be reframed as verification under uncertainty so that models can decline to decide when signal is insufficient or confounded. Closed set classification can benchmark separability of canonical authorial styles, yet open set verification with calibrated scores better matches real editorial questions and reduces harm from overconfident assignments (Ganiyeva, 2024). Feature regimes should combine low level rhythmics such as character trigrams and function word spectra with higher level discourse vectors so that topic drift and edition artifacts can be ablated. Chronology and genre must be controlled explicitly, since confounding between period and author can create spurious separability. Cross collection generalization should be tested on held out witnesses whose layout and normalization differ from training editions so that results are not artifacts of a single pipeline. The methodological frame for these decisions and their failure controls appears in Table 3, which compresses analytics and robustness into a compact scaffold usable for pre-registration and review.

Table 3. Comparative Analytics and Robustness for Philological Inference Tasks

Analytical Task	Signal and Feature Space	Model and Inference Archetype	Dominant Confounds	Robustness and Reporting Regimen
Cross Lingual Reuse	*Shingles, transliteration lattices, multilingual embeddings*	*Alignment with locality bias, graph scoring*	*Noisy OCR, boilerplate, translationese*	*Noise suites, boilerplate masks, calibration curves*
Intra Lingual Reuse	*Character n grams, lemma spans, citation cues*	*Fuzzy match with dynamic windows*	*Orthographic drift, formulaic phrasing*	*Drift stratification, threshold sweeps, error maps*

Authorship Classification	*Function word spectra, stylometric vectors, rhythmics*	*Ensemble classifiers with feature ablation*	*Topic leakage, chronology confound*	*Topic controls, period balancing, seed fixation*
Authorship Verification	*Pairwise distances, calibrated scores, abstention*	*Metric learning with acceptance bands*	*Edition artifacts, layout bias*	*Cross edition tests, abstention audits, ROC stability*
Marginalia and Paratext Mining	*Layout cues, tiny glyph features, NER hints*	*Region detectors with sequence taggers*	*Overlapping inks, code switching*	*Multi ink training, code switch labels, human QA*
Reading Network Reconstruction	*Entity links, reuse edges, temporal priors*	*Graph inference with centrality checks*	*Missing witnesses, spurious hubs*	*Link resampling, hub penalties, provenance audits*

The matrix enforces parsimony in claims by tying each analytic to its primary confounds and to an obligatory regimen for stress testing. By privileging calibration, abstention, and cross edition generalization, it deters overfitted declarations of authorial identity and enforces transparency about where reuse is likely boilerplate rather than transmission. The same regimen harmonizes with the identifiers, apparatus policies, and continuous integration checks defined in Section 3, which means that analytical outputs can be traced to specific loci and rebuilt deterministically when corpora or thresholds change.

Mining Marginalia and Paratexts

Paratextual strata encode reception, pedagogy, ownership, and circulation, yet their extraction is technically fragile and ethically charged. Region detection must isolate interlinear glosses, scholia hands, and outer margin notes without collapsing ornament, rubrication, or bleed through into text layers. Recognition requires tiny glyph sensitivity and code-switching awareness so that mixed scripts and multilingual sprints do not degrade into illegible noise (Borrego, 2023). Named entity recognition must privilege authority linkable spans for persons, places, works, and institutions so that graphs of reading communities can be constructed with provenance fidelity. Ownership notes and sale records demand redactable pipelines where privacy and cultural protocols override automation, which echoes the governance strictures established in Section 2. The confounds and controls enumerated for paratext mining in Table 3 provide a compact audit trail for review, since overlapping inks, trained on multi-ink corpora, are a persistent source of false positives that only human adjudication can fully resolve.

Humanistic Interpretation at Scale

Computational outputs must be translated into argumentative prose with explicit uncertainty so that claims remain legible to humanistic standards. Intertextual graphs can motivate narratives of transmission and reception, yet they must be justified through locus bound citations that permit skeptical re inspection. Authorship verification can narrow hypothesis spaces and rerank candidates, yet decisive attributions require triangulation with external anchors such as documentary context or scribal dossiers (Sommerschield, 2023; Dubrovskaya, 2023). Marginalia clusters can expose classroom practices or reading circles, yet inferences about pedagogy or ideology must survive robustness checks against sampling bias, cataloging gaps, and survivorship. Interpretive writing should treat models as instruments and not as witnesses, which demands continuous acknowledgment of measurement error, distributional shift, and abstention as scholarly virtues. The structured regimens in Table 3 and the representation discipline enforced by Table 2 together create a proof chain where every interpretive move can be traced to data, parameters, and thresholds that are documented and reproducible.

Multi Signal Fusion and Interpretability

High confidence inference emerges when orthogonal signals are fused under a transparent calculus that can be audited and contested. Cross lingual reuse signals, authorship verification scores, and paratext entities can be combined through probabilistic graphical models or rule-based adjudication where independence assumptions are tested and recalibrated. Alignment between image regions and TEI loci ensures that unexpected correlations can be inspected visually, which guards against spurious detours introduced by preprocessing shortcuts (Uug'bekovna, 2024). Model interpretability should be pursued through feature attribution and counterfactual perturbations that expose which n grams, rhythmics, or entities drive decisions, with safeguards against gaming and post hoc rationalization. Score fusion must preserve calibrated abstention so that the system can declare insufficiency rather than force a choice. The comparative schema in Table 3 anticipates such fusion by harmonizing features and confounds across tasks, while the continuous integration and versioning regime in Section 3 guarantees that pipelines can be rerun as new witnesses arrive or as ethical constraints require redaction or embargo.

5. Governance, Ethics, and Decolonial Philology

Provenance, Power and Politics of Access

Governance begins with provenance because the chain of custody structures every subsequent entitlement, constraint, and obligation. Acquisition records, export permits, dealer invoices, and institutional minutes form a provenance graph that either stabilizes lawful access or signals tainted title that demands restraint. Power asymmetries are encoded in that graph, since colonial seizures, coerced sales, and emergency removals distort consent and shift locational authority away from communities of origin. Access decisions must therefore balance legal sufficiency with ethical sufficiency, which are not equivalent, by adopting policies that privilege transparency, reversibility, and stakeholder voice. Rights metadata must travel with every derivative so that imaging, annotation, training, and publication remain consistent with the governing instrument rather than

drifting through undocumented reuse. Takedown pathways and dispute resolution procedures must be pre committed so that conflicts can be addressed without ad hoc improvisation (Porter, 2024; Locaputo, 2024; Babenko, 2024). The analytic discipline established in Table 2 and Table 3 depends on such governance because reproducible inference requires citable loci conditioned by lawful access. The risk classes that recur in contested holdings are operationalized in Table 4 in the next sub section, where scenarios, controls, and assurance signals are compressed into an actionable matrix for editorial boards and repository custodians.

Community Protocols and Indigenous Data Sovereignty

Community governance reframes cultural materials as living knowledge with custodial stewards rather than ownerless objects. Protocols must recognize layered permissions, time bound windows, ceremonial restrictions, and role-based visibility so that consent remains situated and renewable. Data sovereignty asserts decision rights over imaging, annotation, model training, and secondary reuse, which means that repository policies must encode tiered access, consent refresh schedules, and benefit sharing ledgers with auditable entries (Parshutkina, 2024; Baranovska, 2023). Federated training and on premise inference can protect sensitive folia while still improving recognition for the corpus as a whole, and curated abstention can prevent automated disclosure of restricted passages. Attribution must acknowledge community expertise alongside academic labor through contributor logs and coauthored outputs, while credit taxonomies should avoid tokenism by capturing intellectual, technical, and curatorial roles. The risk landscape and mitigation repertoire for such contexts is presented in Table 4, which condenses scenario diagnosis, operational controls, and evidence of effectiveness into a compact scaffold that can be adopted by ethics boards and project leads without delay.

Table 4. Ethics and Governance Risks with Operational Countermeasure Templates

Risk Scenario	Operational Symptom	Affected Stakeholders	Mitigation & Controls	Assurance & Evidence
Contested Ownership and Title	*Incomplete provenance, conflicting claims*	*Source communities, repositories, states*	*Moratorium on release, provenance audit, escrow agreements*	*Public dossiers, legal opinions, decision logs*
Cultural Sensitivity and Sacred Knowledge	*Ritual content surfaced, protocol breach*	*Elders, practitioners, diaspora*	*Tiered access, consent tokens, curated abstention*	*Governance minutes, consent renewals, access telemetry*
Privacy and Personal Data Exposure	*Marginalia reveal identities or health data*	*Families, authors, institutions*	*Redaction, minimization, role-based visibility*	*Redaction audits, data protection reports, incident registers*
Model Misuse and Function Creep	*Models repurposed beyond scope*	*Communities, users, developers*	*License constraints, usage logs, rate limits*	*License notices, audit trails, revocation records*
Attribution and Credit Injustice	*Uncredited expertise, extractive authorship*	*Community experts, annotators, scholars*	*Contributor taxonomies, coauthorship, benefit sharing*	*Credit ledgers, funding acknowledgments, ORCID mappings*
Sustainability and Stewardship Collapse	*Orphaned portals, broken URIs, unreadable formats*	*All users, funders, memory institutions*	*Preservation plans, escrowed code, exit strategies*	*Fixity checks, format migration reports, continuity tests*

Community protocols are credible only when controls are coupled with demonstrable assurance. Evidence must be recorded in tamper evident logs, and minutes must capture deliberation as well as decisions so that accountability survives personnel turnover. Benefit sharing should be tracked in transparent ledgers that record honoraria, training grants, and infrastructural investments alongside scholarly credit. Consent refresh must be rhythmic and not episodic so that evolving community expectations can be honored without emergency renegotiation. Repository telemetry can quantify how tiered access behaves in practice, including denial rates and redaction frequencies, which strengthens policy calibration over time. The compact matrix in Table 4 serves as a living playbook, and its rows map cleanly onto continuous integration checks described in Section 3 so that governance, representation, and release remain synchronized.

Responsible AI and Ethical Editions

Responsible AI in philology treats models, datasets, and editions as interlocking instruments whose risks must be anticipated, bounded, and monitored. Dataset statements should enumerate acquisition context, rights status, sensitive content, and annotation regimes, while model statements should disclose training scope, script coverage, uncertainty behavior, and abstention policy (Szczęsna, 2023). Differential privacy and on premise training can reduce leakage risks where pages cannot leave custody, and dataset sharding can separate ritual content from general corpora. Edition pipelines should incorporate governance gates that block release when rights flags conflict with requested operations, and should format embargo periods as machine readable timers to avoid accidental lapses. Model outputs must be calibrated so that low confidence spans are routed

for human adjudication rather than silently normalized, and abstention should be recorded as a legitimate outcome rather than a failure. Table 4 names typical misuse vectors and embeds controls such as license constraints and revocation records, which should be enforced at the API boundary and logged for audit. Ethical editions then become the visible surface of a disciplined system rather than a veneer over ad hoc decisions.

Training, Capacity and Long-Term Stewardship

Durable governance requires people, skills, and budgets that match the complexity of the corpus and the reach of the portal. Training must cover rights literacy, protocol etiquette, TEI discipline, IIIF operations, and evaluation literacy so that staff can enforce policy while producing research grade outputs (Krasniuk, 2024; Perevorska, 2024). Capacity building should prioritize community stewards through apprenticeships, fellowships, and co teaching studios that transfer not only tool use but also curatorial judgment. Memoranda of understanding should encode roles, dispute pathways, and exit clauses, while stewardship plans should assign responsibilities for format migration, fixity checks, and redundancy. Funding portfolios must combine institutional baselines with grants and philanthropy so that core operations are not hostage to episodic cycles. The sustainability row in Table 4 enumerates controls such as escrowed code and continuity tests, which must be executed on a fixed schedule and reported publicly. Without stewardship discipline, even exemplary editions will decay into broken links and unreadable artifacts that erode trust and waste community labor.

Accountability, Auditability, and Redress Mechanisms

Accountability is operational when any decision that alters visibility, credit, or interpretation can be reconstructed, contested, and reversed with proportionate remedy. Audit logs must capture who accessed what, when, and under which authorization, with cryptographic fixity and retention policies that survive system migration. Independent review boards should evaluate contested takedowns, access denials, and authorship disputes using evidence assembled from provenance graphs, consent ledgers, and telemetry reports (Zarembo, 2024; Krasniuk, 2024). Redress should include restoration, apology, and benefit adjustments where harm is demonstrated, and such remedies must be documented to close the loop. Public transparency pages should publish aggregate metrics for takedowns, redactions, denials, and appeals so that communities can monitor institutional behavior without exposure of sensitive details. The assurance column in Table 4 supplies the reporting primitives that make such accountability measurable rather than rhetorical. When auditability and redress are treated as engineering requirements, governance becomes a daily practice that stabilizes trust, supports rigorous scholarship, and protects vulnerable stakeholders while allowing innovation to proceed within a defensible envelope.

6. Evaluation, Benchmarks, Education, and Future Roadmaps

Measurement Architectures for Falsifiability

Evaluation must function as an epistemic instrument that disciplines claims with calibrated uncertainty, cross domain stress tests, and reproducible build states. Task specific metrics require explicit coupling to error ecologies described in Section 2 so that character error rates and word error rates reflect diacritic sensitivity, relief artifacts, and layout corruption rather than sanitized laboratory conditions (Graziosi, 2023). Comparative fitness for stemmatic inference should be measured through likelihood surfaces and bootstrap dispersion rather than single point optima, while verification-oriented authorship work must foreground calibrated acceptance bands and abstention audits rather than only top one accuracy. Intertextual detection needs threshold sweeps under noise suites that inject realistic OCR and HTR perturbations, boilerplate masks that discount formulaic passages, and chronology aware priors that penalize anachronism. Every score must be tagged to immutable identifiers for witnesses and loci so that results travel with evidence rather than with screenshots. The reporting minima already encoded in Table 2 and Table 3 impose seed control, environment capture, and change diffs as first class outputs, which converts evaluation from an after-action report into a design time constraint. Section 6.4 will crystallize these requirements into operational roadmaps summarized in Table 5 for immediate institutional uptake.

Benchmark Design and Shared Task Orchestration

Benchmarks must model reality rather than convenience, which means multiscript coverage, damage aware splits, rights clean licensing, and culturally appropriate governance. Corpus assembly should stratify by substrate, script family, and epoch so that models cannot overfit to a single codicological milieu. Gold standards must be produced under dual annotation with adjudication protocols that preserve dissent as structured uncertainty rather than suppressing it through brittle majority votes (). For reuse and authorship tasks, test partitions must contain witnesses processed through divergent pipelines so that results do not parasitize a single normalization regime. Epigraphic tracks require shadow field variation and chipped grapheme simulations so that relief tolerant features are rewarded. Shared tasks should publish corruption suites, baseline reproducible scripts, and abstention scoring so that declination becomes measurable scholarship rather than a silent failure. Leaderboards must display calibration curves, robustness deltas, and governance compliance badges alongside headline metrics to deter performative overfitting. The institutional and community logistics for such endeavors are transformed into executable plans within Table 5 in Section 6.4, which aligns resources, outputs, and risk controls for different scales of ambition.

Pedagogies for Interoperable and Ethical Philology

Education must braid philological rigor with computational literacy and governance competence so that graduates can steward multi actor pipelines without epistemic drift. Studio courses should require students to encode witnesses in TEI with ODD constraints, to wire IIIF anchors to line level loci, and to run end to end builds under continuous integration with environment capture. Evaluation literacy should be taught through ablation diaries, calibration labs, and adversarial corruption challenges that force articulation of failure envelopes and abstention thresholds (). Community engagement must be scaffolded through protocol practicums where consent tokens, tiered access, and contributor ledgers are implemented rather than debated abstractly. Capstone projects should culminate in reproducible micro editions with apparatus, variant matrices, and interpretive essays that trace every claim to identifiable loci and parameter states. Faculty development should include cross appointment studios that pair philologists with ML engineers and repository librarians so that shared vocabularies emerge and institutional silos dissolve. The curricular outcomes flow directly

into the roadmapped initiatives codified in Table 5, which bind learning objectives to measurable outputs and institutional stewardship.

Programmatic Roadmaps and Operationalization at Scale

Strategic planning converts desiderata into funded projects with accountable milestones, risk budgets, and evaluation plans. Institutions require templates that translate the reporting minima from Table 2 and Table 3 into schedules, personnel mixes, and compute footprints that can survive procurement, audit, and turnover. Small collections need ninety day sprints that deliver a line anchored TEI edition with transparent uncertainty and rights tags, while national consortia need multi year architectures that stabilize identifiers, host model zoos with usage telemetry, and orchestrate shared tasks with federated participation (). Community led portals demand governance boards with veto powers, consent refresh schedules, and benefit shared ledgers, which must be budgeted and staffed rather than appended as moral afterthoughts. The matrices that follow in Table 5 compress these scenarios into compact, copy ready roadmaps that specify people, data, compute, outputs, key performance indicators, risks, and contingencies. Each row assumes seed control, environment capture, and citable loci, which guarantees that success claims can be recompiled and contested.

Table 5. Roadmapped Initiatives with Resources, Outputs, Risks, and Evaluation

Initiative and Scale	People and Skills	Data and Compute	Expected Outputs and KPIs	Risks and Contingencies
Small Collection Ninety Day Edition	*Philologist, imaging tech, TEI engineer, QA editor*	*1 terabyte storage, GPU lite, CI runner*	*TEI with line anchors, apparatus, uncertainty tags, build hash; time to first release, CER delta, URI uptime*	*Rights ambiguity, fragile bindings; rights triage, cradle imaging, rollback plan*
Regional Multiscript HTR Benchmark	*Script experts, annotators, ML lead, ethics officer*	*50 to 100k lines, multiscript splits, GPU cluster*	*Benchmark pack, corruption suites, model baselines; CER by script, robustness delta, license clarity*	*Unbalanced scripts, leakage; stratified sampling, strict splits, audit notebooks*
Community Governed Manuscript Portal	*Community stewards, curator, devops, governance board*	*Tiered storage, on premise inference, audit logging*	*Tiered access, consent tokens, contributor ledger; access telemetry, consent refresh rate, takedown latency*	*Protocol breach, sustainability; veto workflow, escrowed code, continuity drills*
Cross Lingual Reuse Shared Task	*Linguists, NLP engineers, adjudicators, registry admin*	*Parallel corpora, transliteration lattices, GPU nodes*	*Gold pairs, baselines, leaderboard with calibration; precision at k, abstention rate, cross edition generalization*	*Boilerplate inflation, translationese bias; boilerplate masks, stratified priors, error clinics*
Low Resource Script Model Sprint	*Script mentors, annotators, active learning lead*	*Few shot pages, federated training, privacy guardrails*	*Script models, transliteration toolchain, model cards; coverage gain, CER on rare glyphs, privacy incidents zero*	*Data scarcity, cultural harm; annotation bursaries, curated abstention, governance review*
National Consortium for Philological AI	*Consortium PIs, standards WG, legal counsel, training hub*	*Model zoo, IIIF backbone, PID resolver, archive nodes*	*Shared tasks, model registry, curriculum, grants; adoption rate, uptime, dataset releases, policy compliance*	*Fragmented governance, funding churn; MOU scaffolds, diversified funding, rotating stewardship*

These roadmaps operationalize ambition without sacrificing accountability. Each initiative encodes people and skills rather than generic staffing, data and compute rather than vague infrastructure, outputs and KPIs rather than aspirational prose, and risks and contingencies rather than hand waving optimism. Institutions can instantiate a single row or stage them as a portfolio where small collection sprints seed regional benchmarks, which in turn harden the substrate for shared tasks and national consortia. The identifiers, apparatus discipline, and robustness regimens established in earlier sections remain invariant across all rows, which guarantees that interoperability and ethics scale with volume rather than eroding under pressure. Section 7 will distill these roadmaps into a concise call to action that prioritizes investments with the highest scholarly and civic return.

Sustainability, Funding and Performance Assurance

Sustainability emerges when governance, finance, and engineering are treated as coequal pillars subject to routine verification rather than episodic retrospectives. Funding portfolios should blend institutional baselines with consortial cost sharing and philanthropic grants so that core services remain solvent when grant cycles lapse (). Preservation requires fixity checks, format migration pipelines, and escrowed code that enables

warm failover across institutions, while SLA style uptime and latency targets protect public trust. Performance assurance depends on telemetry that measures not only throughput and error rates but also governance health, including consent refresh density, takedown latency, and credit issuance. Annual audits must recompile key editions from source, rerun benchmark baselines, and rotate cryptographic keys for identifier resolvers to preempt silent drift. Talent pipelines should institutionalize training for TEI engineering, IIIF operations, model evaluation, and protocol etiquette so that staff turnover does not degrade quality. The programmatic scaffolds in Table 5 furnish concrete anchors for budgeting and oversight, while the measurement architectures in Section 6.1 furnish the statistical backbone for continuous improvement.

7. Conclusion

This review has articulated a single methodological grammar that binds material philology to interoperable infrastructures, disciplined representations, auditable analytics, and decolonial governance. The lifecycle begins with measurement grade imaging, proceeds through recognition with quantified uncertainty, stabilizes in TEI with locus level citability, and matures into inference that survives calibration, ablation, and abstention. Tables 1 through 5 have operationalized this grammar by compressing complex decisions into compact matrices that specify pipelines, assumptions, risks, and assurance signals. The strategic horizon is clear and concrete. Projects should target line anchored editions with reproducible build hashes, recognition stacks that report diacritic sensitive error, apparatuses that encode intervention flags, and analytics that publish calibration curves alongside headline scores. Governance must remain coequal with method, which means tiered access, consent refresh, contributor ledgers, and takedown latency tracked as rigorously as CER or F1. When these components are synchronized through persistent identifiers and continuous integration, philology gains scale without sacrificing witness specificity, and computation becomes an instrument for adjudication rather than a rhetorical flourish.

The next tranche of investments should follow the roadmapped initiatives summarized in Table 5, since those rows convert aspiration into staffed, budgeted, and testable programs. Small collections should execute ninety day sprints that deliver TEI with line anchors, uncertainty tags, and rights metadata propagated from ingest to release. Regional alliances should curate multiscript HTR benchmarks with damage aware splits and corruption suites that punish overfitting to sanitized pages. Community governed portals should operationalize consent tokens, curated abstention, and benefit sharing ledgers while keeping inference on premises where protocol demands. Shared tasks for cross lingual reuse should fix seed states, publish baseline rebuilds, and score abstention so that declination counts as expertise rather than failure. National consortia should stabilize identifier resolvers, host model registries with usage telemetry, and run annual rebuild audits for flagship editions. Across all scales, performance targets must join technical and ethical metrics, pairing uptime and latency with takedown response, redaction accuracy, and credit issuance, so that excellence is measured as an ecosystem and not as isolated scores.

Durable progress requires polycentric collaboration where philologists, conservators, community stewards, ML engineers, and repository librarians share a vocabulary, a build system, and a governance ledger. Interoperability must be contractual rather than aspirational, with IIIF manifests, TEI schemas, and authority links treated as public utilities that outlive grants and personnel. Stewardship plans should budget for fixity checks, format migration, and resolver continuity, and should specify exit strategies that safeguard portals when leadership rotates or funding tightens. Accountability must be engineered into every interface. Audit logs should capture access, modification, and model inference with cryptographic fixity, while transparency pages should publish aggregate metrics for consent renewal density, embargo lift accuracy, and governance appeals. Training pipelines should institutionalize evaluation literacy and protocol etiquette so that new cohorts can maintain quality without heroic tacit knowledge. If the field adopts these collaborative and stewardship commitments, the lifecycle codified in Tables 1 through 5 will not only scale across scripts and substrates, it will also remain socially legitimate, scientifically falsifiable, and intellectually generative for the long run.

Declaration of Interest:
No potential conflict of interest was reported by the authors.

Funding Information:
This research did not receive any specific funding from any public, commercial, or non-profit agency.

Disclosure Statement:
No material or relevant stake relating to this research was disclosed by the author(s).

Competing Interest:
No potential conflict of interest was reported by the author(s).

Data Availability Statement:
Data sharing is not applicable to this research article as no new data were created or analysed in this study.

References

- Adriansyah, A., Elmustian, E., Sinaga, M., & Firdaus, M. (2024). Transformation of texts as expressive spaces in classic Malay literary works: Learning to write poems in philology courses. *AL-ISHLAH: Jurnal Pendidikan, 16*(4), 5345–5356.
- Babenko, I., & Athavale, V. A. (2024, September). Methods of literary text analysis with the help of artificial intelligence. In *International Conference on Distance Education Technologies* (pp. 81–95). Cham: Springer Nature.
- Bambaci, L. (2021, June). Digitizing an eighteenth-century collation of Hebrew manuscripts: A rule-based parsing system for automatically encoding critical apparatus. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)* (pp. 198–203). IEEE.
- Baranovska, L., Simkova, I., Akilli, E., Tarnavska, T., & Glushanytsia, N. (2023). Development of digital competence of future philologists: Case of Turkish and Ukrainian universities. *Advanced Education, 23*, 87–102.
- Biber, H. (2020, May). Challenges for making use of a large text corpus such as the 'AAC–Austrian Academy Corpus' for digital literary studies. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora* (pp. 47–51).
- Bories, A. S., Fabo, P. R., & Plecháč, P. (2022). The polite revolution of computational literary studies. *Computational Stylistics in Poetry, Prose, and Drama, 1*.
- Bozhenkova, N. A., Rubleva, E. V., & Baharloo, H. (2023). Dictionary of IT terms as a tool for Russian language studies and linguodidactics in the context of digitalization in education. *Russian Language Studies, 21*(4), 457–473.
- Camps, J. B., Gabay, S., & Riva, G. F. (2021). Open stemmata: A digital collection of textual genealogies. In *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities, Krasnoyarsk, 2021*.
- Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). Linguistic linked data in digital humanities. In *Linguistic Linked Data: Representation, Generation and Applications* (pp. 229–262). Cham: Springer International Publishing.
- Cowen-Breen, C., Brooks, C., Graziosi, B., & Haubold, J. (2023, September). Logion: Machine-learning based detection and correction of textual errors in Greek philology. In *Proceedings of the Ancient Language Processing Workshop* (pp. 170–178).
- Cugliana, E., & van Zundert, J. J. (2022). A computational turn in digital philology questions. *Filologia Germanica = Germanic Philology, 14*, 43.
- De Gussem, J., Niskanen, S., & Willoughby, J. (2022). Computational stylistics and medieval texts. In *Routledge Resources Online: Medieval Studies* (pp. 1–12). Routledge.
- Del Grosso, A. M., Zenzaro, S., Boschetti, F., & Ranocchia, G. (2023, December). GreekSchools: Making traditional papyrology machine actionable through domain-driven design. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)* (pp. 621–626). IEEE.
- Dörpinghaus, J. (2022). Digital theology: New perspectives on interdisciplinary research between the humanities and theology. *Interdisciplinary Journal of Research on Religion, 18*.
- Dubrovskaya, E. M., Filonova, A. I., & Matveeva, I. V. (2023, June). Prospects for artificial intelligence technologies, neural networks, and computer systems within the development of linguistics. In *2023 IEEE 24th International Conference of Young Professionals in Electron Devices and Materials (EDM)* (pp. 2120–2123). IEEE.
- Elwert, F. (2021). Computational text analysis. In *The Routledge Handbook of Research Methods in the Study of Religion* (pp. 164–179). Routledge.
- Fitzmaurice, S., & Mehl, S. (2022). Introduction: Digital methods for studying meaning in historical English. *Transactions of the Philological Society, 120*(3), 397–398.
- Ganiyeva, D., Aliyeva, N., Karimova, S., Ismoilova, D., & Jurayev, I. (2024, November). Digital Dickens: AI and the future of classic literature interpretation. In *2024 International Conference on IoT, Communication and Automation Technology (ICICAT)* (pp. 322–328). IEEE.
- Ghali, W. (2023). Old, new or digital philology: New methodological perspectives in Islamic studies. *New Methodological Perspectives in Islamic Studies, 20*, 137.
- Graziosi, B., Haubold, J., Cowen-Breen, C., & Brooks, C. (2023). Machine learning and the future of philology: A case study. *Transactions of the American Philological Association, 153*(1), 253–284.
- Gryaznova, E., Kirina, M., Mikhailova, P., Zarembo, V., & Moskvina, A. (2022, June). Machine learning and philology: An overview of methods and applications. In *International Conference on Internet and Modern Society* (pp. 69–84). Cham: Springer Nature.
- Hatzel, H. O., Stiemer, H., Biemann, C., & Gius, E. (2023). Machine learning in computational literary studies. *it-Information Technology, 65*(4–5), 200–217.
- Jackson, M. K. (2021). Review of Lennon's *Passwords: Philology, Security, Authentication*. *Surveillance & Society, 19*(2), 279–281.
- Krasniuk, S. (2024). Modern data science in philology. In *5th International Scientific and Practical Conference "Diversity and Inclusion in Scientific Area"*. Ceac Polonia.
- Krasniuk, S. O. (2024, November). Mathematical optimization in philology. In *Sworld-US Conference Proceedings* (No. usc27-00, pp. 109–115).
- Krasniuk, S., & Goncharenko, S. (2024). Big data in philology. In *Débats scientifiques et orientations prospectives du développement scientifique*. La Fedeltà & UKRLOGOS Group LLC.
- Kudinova, O., Kudinova, V., & Kondratenko, N. (2021). Digital humanities as a way of teaching disciplines of philological series. In *ICERI2021 Proceedings* (pp. 3846–3851). IATED.
- Lamb, J. P. (2020). Computational philology. *Memoria di Shakespeare: A Journal of Shakespearean Studies, (7)*.
- Li, C. (2020). Philology and digital humanities. In *Routledge Handbook of Yoga and Meditation Studies* (pp. 383–392). Routledge.
- Locaputo, A., Portelli, B., Magnani, S., Colombi, E., & Serra, G. (2024). AI for the restoration of ancient inscriptions: A computational linguistics perspective. In *Decoding Cultural Heritage: A Critical Dissection and Taxonomy of Human Creativity through Digital Tools* (pp. 137–154). Cham: Springer Nature.
- Macías Borrego, M. (2023). Towards a digital assessment: Artificial intelligence assisted error analysis in ESL. *Integrated Journal for Research in Arts and Humanities, 3*(4), 76–84.

- Maiocchi, M. (2021). Current approaches towards ancient Near Eastern textual sources: Some remarks on contemporary methodologies for philological research. *dNisaba za3-mi2: Ancient Near Eastern Studies in Honor of Francesco Pomponio, 19*, 117.
- Middleton, P. (2024). Parrots and paragrams: AI language models and erasure poetry. *Modern Philology, 121*(3), 352–374.
- Mitcham, C. (2020). Philology and technology. *Technology and Language (Технология и инфосфере), 1*(1), 61–65.
- Palladino, C., Shamsian, F., & Yousef, T. (2022). Using parallel corpora to evaluate translations of ancient Greek literary texts: An application of text alignment for digital philology research. *Journal of Computational Literary Studies, 1*(1).
- Parshutkina, T. A., & Turko, U. I. (2024). The model of contextual education of digital literacy to students on the example of philological disciplines. *Perspektivy Nauki i Obrazovaniya, 5*(71), 125–141.
- Perevorska, O., Prihodko, T., Kobzieva, I., Roman, N., Agadzhanova, R., Marianko, Y., … Silichova, T. (2024). Interaction of philology, pedagogy, culture and history as a way of integrating learning. *International Science Group*.
- Piotrowski, M. (2022). NLP and digital humanities. In *Natural Language Processing for Historical Texts* (pp. 5–10). Cham: Springer International Publishing.
- Porter, J. I. (2024). Philologies of the present for the future. In *The Future of the Past: Why Classical Studies Still Matter. Athenian Dialogues IV* (pp. 173–213).
- Rahmi, S. N., Sok, V., & Dara, S. (2024). Decoding lost languages: A philological study of ancient texts. *Journal of Humanities Research Sustainability, 1*(4).
- Roelli, P. (2020). *Handbook of stemmatology: History, methodology, digital approaches* (p. 688). De Gruyter.
- Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., … De Freitas, N. (2023). Machine learning for ancient languages: A survey. *Computational Linguistics, 49*(3), 703–747.
- Szczęsna, E. (2023). The humanities in the world of new technologies (and vice versa): Toward digital philology. *Teksty Drugie. Teoria literatury, krytyka, interpretacja, (2)*, 82–98.
- Tasheva, N. (2024). The evolution of modern linguistics: Key concepts and trends. *Medicine, Pedagogy and Technology: Theory and Practice, 2*(11), 31–39.
- Thomassen, E. (2021). Philology. In *The Routledge Handbook of Research Methods in the Study of Religion* (pp. 401–412). Routledge.
- Tuttle, K. (2021). Review of *Among digitized manuscripts: Philology, codicology, paleography in a digital world* by LWC van Lit. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies, 6*(1), 177–181.
- Uug'bekovna, S. M. (2024, November). Philology in the digital age: The impact of technology on language preservation. In *International Conference on Multidisciplinary Studies and Education* (Vol. 1, No. 1, pp. 12–15).
- Weber, T. (2020, August). A philological perspective on meta-scientific knowledge graphs. In *International Conference on Theory and Practice of Digital Libraries* (pp. 226–233). Cham: Springer International Publishing.
- Yang, J. (2024). Translation paradigms: Translation in the development of digital humanities. *International Journal of Linguistics, Literature & Translation, 7*(8).
- Zarembo, V., & Moskvina, A. (2024, February). Machine learning and philology: An overview of methods and applications. In E. Gryaznova, M. Kirina, & P. Mikhailova (Eds.), *Digital Geography: Proceedings of the International Conference on Internet and Modern Society (IMS 2022)* (p. 69). Springer Nature.
- Zulfiya, T. (2024). The overview of the methods of textual analysis. *Innovative Technologica: Methodical Research Journal, 3*(4), 5.