

A Poetics of the Rebuilt Archive

A Statement on Research Methods

Brian Kim Stefans — Department of English, UCLA

Los Angeles · 20 July 2026, 5:31 p.m.

PREFACE

For much of this summer, most of my scholarly and creative energy has gone into a single form: the *reading edition* — a clean, designed, editorially framed text reconstructed from a large, degraded, or unstructured digital archive by means of a computational pipeline. My two principal sites are the mid-century little magazine *Trace* (edited by James Boyer May) and the SUNY-Buffalo Poetics listserv. The method, in this automated form, is new; the editorial impulse behind it is not, and I set out its longer history below.

The *Trace* work belongs to a larger, ongoing archival project on Los Angeles poetry; more journals are to come — *Coastlines*, *Bachy*, and others whose only surviving copies sit in library storage. The Poetics list is the eastern counterweight to that West-Coast effort, and the more technically demanding of the two.

Part of what drives the work is a plain fact about scholarly publishing: the books that most need to exist are often the least able to pay their way. My own critical study, *Word Toys: Poetry and Technics* (University of Alabama Press, 2017), lists at \$49.95 — and yet appeared, almost at once and through no doing of mine, as a pirated file online. The audience was there; the price was the obstacle. A freely distributed, well-made edition is one answer to that mismatch, and it is the tradition I came up in.

What follows names the techniques the method borrows from other fields, situates the work among its humanities neighbors, and gives an account of the Poetics project and the episodes that have most shaped my thinking about it.

I. THE METHODS AND THEIR VOCABULARY

The problem is always the same: find what matters inside a large, noisy corpus, then render it into something a person can actually read. Literary study did not invent the tools for this; law, journalism, and evidence-based medicine did, and their vocabulary is precise enough to be worth borrowing.

The umbrella term, from litigation, is **technology-assisted review (TAR)**: software classifies documents in a vast collection from an expert's coding of a sample, so that human attention can be aimed where it will count. Its engine is **predictive coding** — the reviewer marks a sample

relevant or not, and a statistical model ranks the remainder by likelihood of relevance. The legal profession's turning point was Magistrate Judge Andrew Peck's 2012 opinion in *Da Silva Moore v. Publicis Groupe*, which blessed this kind of concept-driven, statistical searching over brute keyword matching. I am not litigating anything, but the underlying gesture — rank the unread by probable relevance, then read down the ranking — is exactly what selecting from two decades of list traffic demands.

A handful of further terms name operations my pipeline performs daily. **Optical character recognition (OCR)** turns page images into machine-readable text — the first and most error-prone step whenever a project begins, as mine do, in scans. An **extraction pipeline** is the chain that follows: OCR, cleanup, normalization, indexing — the machinery that converts a heap of PDFs and raw email into something searchable at all.

Email threading reconstructs the reply-structure of a correspondence so that a scattered exchange can be read as the conversation it once was — precisely the operation required to turn a listserv dump back into legible discussion. **Near-duplicate detection** flags documents that are almost but not exactly alike, indispensable when the same text survives in several corrupted or partial copies. **Concept search** — clustering by meaning rather than by shared keywords — earns its place too, since a mailing list quickly develops a private idiom that literal keyword matching sails straight past.

The whole enterprise is judged, across all these fields, by two measures worth importing into editorial work: **precision** (how much of what you retrieved is actually relevant) and **recall** (how much of the relevant material you actually found). Their real value is that they turn "have I found everything?" from an article of faith into a question with an answer.

One adjacent field deserves a passing nod, because it runs the same machinery for the highest stakes: evidence-based medicine, whose systematic reviews screen tens of thousands of study abstracts, and where a missed document is measured in patient outcomes.

II. THE HUMANITIES NEIGHBORS

The nearest disciplinary home for this work is **digital scholarly editing**, and its own verdict on itself is instructive. For forty years, by the field's candid account, it has been "fundamentally conservative" — producing machine-readable surrogates of printed originals, reproducing the old apparatus of annotation and gloss in a new medium without rethinking what an edition is, or dares to be. Mine departs on two counts: it treats born-digital ephemera as worth editing at all, and it ends not in a database but in a finished, designed object made to be read.

The institutional analogs closest to *Trace* are the little-magazine digitization projects. The **Modernist Journals Project** (Brown and Tulsa, from 1995) has built searchable editions of more than thirty modernist magazines — *Poetry*, *The Little Review*, *Blast* — with introductions and

biographies; it began, fittingly, with a graduate student and a slow flatbed scanner in a converted storage closet. Its Princeton cousin, the **Blue Mountain Project**, extends the same ambition to the international avant-garde periodical. Both, though, largely stop at the facsimile-plus-metadata; the reset, redesigned reading text is rarer, and rarer still for born-digital material.

The most spectacular version of these methods lives outside the academy, in investigative journalism. The ICIJ's Panama and Paradise Papers ran to millions of files — one reporter at a minute per document would have needed twenty-six years for a single leak — and the answer was the same pipeline at industrial scale: extract, index, search, then train a classifier on a hand-coded sample to sort what no one had yet opened. The lesson I take is not technical but ethical: the machinery is worth building only because it *concentrates* human judgment rather than replacing it.

Closer to home is a colleague's work. Daniel Scott Snelson, who teaches in English and Design Media Arts here at UCLA, mapped this whole territory before I set foot in it, under the name of the *little database*. In *The Little Database: A Poetics of Media Formats* (University of Minnesota Press, 2025) he makes the case for small, handmade online archives — PennSound, Eclipse, Textz, UbuWeb, and, not least, the Electronic Poetry Center that housed the Poetics list — as a deliberate counterpoint to the vast aggregations of Google and Facebook, and develops contingent, often playful methods for reading the ordinary digital objects (TXT, JPG, MP3, MOV) such collections are made of.

One further precedent is closer still — it is my own. From roughly 2000 to 2003 I ran a small PDF-publishing practice: [the Arras Media PDFs](#), self-made editions of avant-garde poetry that I typeset in QuarkXPress under my own imprint, designing the covers and doing the copyediting myself — electronic reprints of Jennifer Moxley's *The Impercipient* and Robert Fitterman's *Object 9*, and *Poli Sci*, which collected the political-science writings that the Language poet Bruce Andrews had scattered across journals and pamphlets over the academic career he made as a professor of the field, among others.

Out of that work I initiated and edited UbuWeb's first PDF series, /ubu Editions — later widened into *Publishing the Unpublishable* — and wrote the introductions to its first two series. I have since gathered [those PDFs and their introductions on my own site](#). The editions themselves remain on UbuWeb, but my introductions no longer appear there, and current descriptions of the series attribute its editorship otherwise; those introductions survive in the Internet Archive's 2012 captures ([ubu_first.html](#), [ubu_second.html](#)).

It is the direct ancestor of what I am doing now — the same conviction that recovering neglected material, with real editorial and typographic care, is itself a creative act. What I once did by hand in Quark and Photoshop, twenty-odd years ago, the pipeline now does in a fraction of the time.

III. THE UB POETICS READING-EDITIONS PROJECT

The listserv

The University at Buffalo Poetics listserv — the Poetics List, UBPOETICS — opened on 8 December 1993 under Charles Bernstein's sponsorship and became one of the best-known mailing lists in contemporary poetry: an outgrowth of the Poetics Program that Bernstein and Robert Creeley had founded at Buffalo in 1991, and a sibling of the Electronic Poetry Center (Loss Pequeño Glazier and Bernstein, 1994), among the first poetry sites on the Web. For twenty years it was the nervous system of a whole tendency in American writing; its subscribers included Marjorie Perloff and much of the world of Language poetry and its successors.

Bernstein closed it in January 2014, once social media had absorbed its function, but kept the archive. That archive — the daily intellectual life of a poetic community, in a register the finished essay never preserves — is my corpus. In raw form it is close to unreadable: undifferentiated text dumps, broken threads, quoting noise, header cruft, and, as I discovered, outright mislabeling. What the raw files flatten is exactly what made the list matter — the velocity of the argument, the feuds and reconciliations, a poetics being improvised in public at the speed of email. To read it as the record it is, you first have to rebuild it.

What the edition is — and is not

The reading edition is not meant to replace the archive it draws on; that record has been open to scholars for years. The problem was never access but tractability. At Buffalo the list simply sat on a server as a heap of undifferentiated .txt files that, realistically, no one was ever going to read through. With *Trace*, the issues existed as library copies at UCLA, which I scanned myself, page by page, and put through my own OCR before any automated work began.

OCR is worth a caveat here. The recognition Acrobat performs on old, unevenly printed pages is badly error-prone, and getting genuinely clean text out of them means bringing a range of tools to bear and checking the result by hand — laborious, but far easier than repairing the mess that automated OCR leaves behind.

I take some care that nothing is elided. Editors of collected and selected volumes are perennially, and sometimes justly, accused of silencing voices — of letting selection quietly do the work of judgment. My aim is the reverse: the source remains whole and obtainable, and the edition is a lens laid over it, not a substitute for it.

The workflow has three stages. First, the *source* — the raw material, easily obtained by any researcher, whether a book or a folder of text files. Second, *reduction through AI* — the model ingests the entire corpus and begins to make editorial decisions across it at a scale no single reader could match. Third, *presentation*, which is where the human editorial work actually happens: I review what the AI has proposed and send it back revisions, corrections, and second thoughts, much as one would work with a sharp but over-eager editorial assistant. Often, when I

catch it having passed over the contributions of a figure of real importance, I send it back to comb the corpus afresh.

Working this closely with a single model — in my case Claude — one learns to recognize its semantic register: the characteristic grain of its prose and its judgment. The strengths I have come to rely on are specific ones: a genuine capacity for logic processing, following an argument and holding a distinction rather than merely matching words; synthesis across a very large body of text held in view at once; fine control of register and tone, so that a passage can be pitched as commentary or as plain description on request; consistency and tirelessness across thousands of entries; and a readiness to be corrected. These are also, precisely, an editor's tools — which is why the collaboration works, and why it must be supervised.

The method and its run

The apparatus itself is a Node.js and Python pipeline — chiefly `build.js` and `hyphenate.py` — that ingests the raw archives, performs the editorial equivalents of threading and near-duplicate detection, and typesets the result as a designed reading edition in Garamond, in a recurring oxblood-teal-gold livery that gives the series a face. The design is not decoration: a listserv post asks to be skimmed and deleted, and setting it in a considered book face is already an argument that it deserves to be kept and read again.

Onto the bare text I have added the furniture of a proper edition: an introduction to each issue, grounded in that month's actual arguments; *Reader's Guides* and navigational tools that let a reader either move quickly through a month or settle into a single thread; biographical notes on the participants; and a layer of commentary that draws out what is at stake in a given exchange. This editorial matter — introductions and commentary alike — is produced in the first instance through the Kristen M. Fabians persona, in dialogue with the AI, and then reviewed and approved by me: I write none of it from scratch, but I answer for all of it. The formatting is built to serve both close reading and skimming, the two ways anyone actually uses an archive like this.

The monthly run is now effectively complete: a reading edition for every month from the list's first posts in December 1993 to its last in January 2014, on the order of 240 issues. Alongside it I have made special editions of two kinds — single-author collected editions, gathering one writer's contributions across the whole archive (the volume of Kevin Killian's posts, assembled in about two hours by reusing procedures I'd developed for the monthly editions, is a companion to Semiotext(e)'s *Selected Amazon Reviews*), and thematic editions assembling the list's engagement with a single event or quarrel: its long and unlovely argument over Ezra Pound and fascism, and *It's 8:23 in New York*, which gathers the list's response, in real time, to the attacks of September 11, 2001. A general reading edition of the list and a critical introduction to the series round out the whole.¹

The corrupted archive

The decisive episode was discovering that the publicly hosted (Penn-mirrored) archive is corrupted in a way that quietly falsifies the record. The file served as 1995_10 .txt — supposedly October 1995 — in fact holds posts from May 1998. This is exactly what near-duplicate detection and stubborn provenance-checking exist to catch, and it is invisible to anyone who trusts the archive's own labels. Working from my own faithfully kept copy, I reconstructed the true October 1995 file by locating its real posts elsewhere in the archive and reassembling them.

I dwell on this because it changes what the project is. Correcting a mislabeled file is, in the language of textual scholarship, an *editorial argument* about the constitution of the record, not a chore of data cleaning — the born-digital equivalent of a copy-text decision. The archive as delivered is not the archive as it was, and only a method that treats provenance as a first-class problem can tell the two apart. It is a small correction with a large moral: the historian of the recent past inherits archives that lie, and inherits them silently.

Kristen M. Fabians, and the ethics of an AI writer

The editorial introductions, the commentary, and much of the related criticism appear under the byline **Kristen M. Fabians** — a developed persona (a scholar besotted with the avant-garde and its forgotten figures, who treats my own work as that of "an obscure New Jersey Modernist") under which I publish writing made in collaboration with a large language model. The device is not a disguise but a frame, and its ethics are part of the method.

A named persona has deep precedent in exactly this tradition — Pessoa's heteronyms, pseudonymous editorship, the authorship games of conceptual writing; literature has always known that a made name can be an honest instrument rather than a fraud. What is new is that some of Fabians's prose is generated in dialogue with a machine, and that raises three questions I try to answer in practice rather than evade.

The first is transparency: Fabians is kept an openly acknowledged collaborative fiction, never a claim that some independent scholar exists and vouches for the work.

The second is reliability. Models fabricate, and in archival work a single invented citation would corrupt the very record I mean to repair; so every fact under her name is checked against the primary archive and held to the standard I would hold my own. The corrupted-file discovery is the governing example — the method's whole worth lies in refusing to trust an unverified source, and that discipline extends to the machine collaborator.

The third is credit: the persona launders nothing; the judgment, and the responsibility for any error, remain mine. That this experiment sits inside a project about an earlier community's own negotiations of voice, authorship, and electronic mediation is not incidental — the Poetics list is the right laboratory for it.

Significance

Together the editions make an argument in practice: that the born-digital ephemera of a literary community are primary sources deserving the full apparatus of scholarly editing; that the methods refined in law, journalism, and medicine can be turned to humanistic ends without surrendering interpretation to the machine; and that recovery, correction, and re-presentation — the making of a *reading edition* from a broken record — are forms of scholarship in their own right.

If the method has a thesis, it is that an archive is not a given but a made thing, and that remaking it well is itself a form of reading. The work is ongoing, each edition extending the archive's legibility and testing the method against another month of the past.

NOTE

¹ Beyond the monthly run — a reading edition for each month from December 1993 to January 2014 — the special and collected editions completed to date are: *It's 8:23 in New York — Poetry and the Wars After 9/11 on the UB Poetics List, 2001–2011*; *Jack and I Were Both Fluent in Martian — The UB Poetics List Writings of Kevin Killian, 1995–2014*; *poetics@ — Poetry, Argument, and the Early Years of the UB Poetics List* (edited by Joel Kuszai in 1999); *"There Are the Alps, Fools!" — Pounding Fascism on the UB Poetics List, 1995–2008*; *WARNING ~ POETRY AREA: Publics Under Construction — The UB Poetics List Writings of Charles Bernstein, 1994–2014*; *The Man Who Sold the Listserv — Every Sighting of David Bowie on the UB Poetics List, 1996–2013*; and *The Only New Device — Appropriation, Procedure, and the New on the UB Poetics List, 1994–2013*.