Editing and Curating Online

Beginning Again

Jerome McGann

Abstract

The complexity of natural language works, especially transinformational works, cannot be adequately represented by what has become the institutional standard for DH editorial projects, TEI. In that respect book technologies remain far superior to current digital tools in sustaining their reciprocal communicative action. Recent developments in graph database platforms suggest ways to accommodate the n-dimensionality of such work to the disambiguating inertia of digital tools.

I.

> Beginning again and again is a natural thing even when there is a series. Beginning again and again and again explaining composition and time is a natural thing.
>
> —Gertrude Stein, *Composition as Explanation*

Online projects over the past twenty-five years have built an impressive record in making cultural legacy available across the globe, often with facsimiles of extraordinary excellence. In the United States, *The Women Writers Project*, *The William Blake Archive*, and *The Walt Whitman Archive* are exemplary of broad achievements that have come in many countries and in every discipline. At the same time they have all but institutionalized models for the digital transmission of literary works that are in important ways profoundly misguided. Moved by the significant, even epochal, opportunities for knowledge and communication that have come with digital technology, we continue to forget or ignore how and why

Textual Cultures 15.1 (2022): 53–62. DOI 10.14434/tc.v15i1.34497

natural language technologies are so much more powerful than the DH designs we have adopted and keep producing.

Recent developments in the data sciences point to ways we might begin to change course and organize digital platforms that operate more like books. In the final part of this essay I will discuss in brief and general ways how some of these will be implemented in the editorial venture *Jaime de Angulo's Old Time Stories: Voice, Text, Image*. But first I must revisit the problems that DH projects keep perpetuating.

The problems were forecast, indirectly but decisively, in Claude Shannon's seminal 1948 essay "A Mathematical Theory of Communication". Recall its opening sentence: "The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point" (379). For the purposes of electronic transfer, if messages carried by "analogue" waves were reconceived as discrete "bits" of "information", "signal noise" could be removed as the message made its way through its sending/receiving channels.

This mathematicized approach to communication recalls problems well-known to scholars working with documents coded in natural languages (oral, graphical, textual). Transmission over time accumulates what Charles Lamb lamented as a "mingled mass" of "strange defeatures" that curators and editors are missioned to clear, as best they can.[1] But in that frame of reference, "clearing the text" can be parsed to mean either "clearing up" the mingled mass or "clearing away" what are judged its accumulated errors. Shannon's achievement implicitly appeals to the latter: the pursuit of an authoritative, uncorrupted original, what Thomas Tanselle has named the "ideal text".[2]

The problem is that natural language communications are always transactional (dialogic) even when the goal is transmission. Built into them are the histories, only ever partially recoverable, of how they have been handed back and forth. Even the self-identical Word of God, that ultimate ideal text, is only a word known to all men only in their different ways of knowing/communicating. So traditional oral and textual works operate a model of communication very different from Shannon's. By exploiting the many transinformational (rhetorical and poetic) devices of natural languages, they multiply the redundant and ambidexter features of the channels, hypertexting ambiguities, obscurities, and all manner of apparently incidental expressive devices. In our digital age, Randall McLeod's various

1. See Lamb's splendid poem "In my own Album".
2. See Tanselle 1995.

investigations into what he came to call "Information about Information" delivered especially dramatic proofs of the importance of these transinformational functions of communicative exchange — those *signal* features that natural languages exploit and that readers use and scholars are called to notice.[3]

Those features show why the issues here are broadly human and social, not just scholarly or humanist. Scholars are professionally involved because we have a vocation to preserve and study the world's natural languages and their works. We do this for two reasons: first, because ordinary human beings manipulate them all the time in our day by day lives, including our lives played out in virtual spaces; and second, because some not-so-ordinary human beings work hard at arranging them into instructive and pleasurable show-and-tell public demonstrations: Ovid for instance, or (they don't have to be poets) Montaigne; Hannah Arendt or Martin Luther King (they don't have to be long gone). We do this because we know that we all get by in language only with a little help from our many friends (and enemies, though the Beatles neglected to tell us that) who exchange language with us.

I mention Ovid because the problem was clearly illustrated in the history of Willard McCarty's visionary *Onomasticon* project, which he began in the late 1980s and finally had to shut down in 2004 because of a "core failure" in the research design. His 2017 report — the essay is collected here — explains his decision:

> For a computational tool, design [. . .] is its responsiveness to the interpretative moves of the user-designer, moment by moment; constraint is provided by the imperative of complete explicitness and absolute consistency. Hence the core failure of analytical markup lies not in its rigidity but in the lack of responsiveness which that rigidity entails, its propositional rather than subjunctive, as-if form. Unlike human language, in which fossilized metaphors can come back to life at the poet's touch, [the computational tool's] units of expression remain fossils from the moment of utterance (this issue, p. 51).

Because TEI had not yet established itself when he began, McCarty devised his own imbedded markup for a relational database of named textual things. But as the tagged texts kept turning up duck-rabbits, the "propositional"

3. See McLeod [McLoud] 1991. See also Kirschenbaum 2014 and 2021, especially pp. 69–73.

design simply ran aground, frustrating the interpretive "responsiveness" of the human user and, reciprocally, making the "bewildering combinatorial complexity" of Ovid's language computationally inaccessible.

As McCarty pursued the project through the 1990s he was "assured by no less than Michael Sperberg-McQueen that my metalanguage could be algorithmically translated into TEI". But by 2004, when TEI had become an orthodox "encoding methodology", McCarty saw that both the *Onomasticon* and TEI were DH "dead ends". No matter how carefully and thoroughly executed, such designs translate natural language's asymmetric duplicities into static "content objects" to be handed over for relational database retranslation as self-identical data neatly arranged on a grid of columns and rows.[4] That is a fundamental misrepresentation of natural language. While imbedded text markup (TEI) *cum* relational database can and do efficiently organize, at least for some purposes, the study of traditional documentary information, they cannot compute the (trans)information from which that information is extracted and abstracted.[5]

Though he wasn't thinking of the DH dead-end McCarty's project exposed, Paul Eggert recently clarified how the problem might be practically addressed. A properly theorized edition of any kind, Eggert pointed out, "implicitly builds the reader into itself" (2019, 7). The observation is relevant to scholarly editing because natural language operates an immediate and recursive codependent relation between the agents Eggert named "The Work and the Reader". Simply, a "Reader" understands and actively reciprocates a "Work"'s languages: not just its linguistics, but all the expressive forms that textual artifacts put into play. While book makers, authors, and their allies, have developed various specialized linguistic, graphical,

4. Ted Nelson succinctly identified the problem, though his point of view was slightly different from McCarty's: "We greatly need a general structure to represent all forms of interconnection and structure, and changes in both content and structure; and to visualize and re-use variants and alternatives, comparing them in context in order to understand and choose"; see Nelson 1997. See also my debate with Allen Renear at the 1999 Humanities Conference at University of Virginia, "What is text? A debate on the philosophical and epistemological nature of text in the light of humanities computing research": http://www2.iath.virginia.edu/ach-allc.99/proceedings/hockey-renear2.html.

5. Stand-off markup was an interesting attempt to break through that problem: see Berrie et al. 2006; see also Eggert 1994, especially the discussions at 16–18. While a stand-off approach can cite sets of overlapping textual features, it cannot compute their dynamic interoperations, which is the core of their asymmetrical (nonhierarchized) character.

and bibliographical codes, they are not, as are digital machines, coded in the universal language of mathematics but in a nonuniform array of expressive features spawned and shared by different social groups. You declare a scholastic allegiance if you see that situation as either the Tower of Babel or the Library of Alexandria.

While I emphatically endorse Eggert's practical point about the codependent relation between "The Work and the Reader", I would revise his comment to read something like this: "a properly *designed* scholarly edition should *explicitly* build the reader into itself". That requirement underscores why traditional literary works remain so important. They explicitly realize a reciprocity of "The Work and the Reader" that is absent from digital machines as they are currently designed for, and used by, "the Reader", whether "Common" or scholastic. The back-end of a platform — middleware, software, and hardware — is *terra incognita* for nearly all users, and developers themselves commonly have expertise in narrowly defined specializations. Codependence operates then, if it operates at all, as the machine is being designed and built by its developers. Afterwards, because users — even those who are more-or-less DH savvy — are specifically set apart from the computational design, reciprocity with the machine has turned implicit.

That happens because "the Reader and the Work", the user and the device, are engaged at the interface, where natural language and its visible extensions have been algorithmically reshaped as an ordered hierarchy of self-identical content objects.[6]

Natural language artifacts are completely different. First of all, they have become naturalized — or perhaps one should say "second-naturalized" — over millennia of human practice and in all the particular textualities — from alphabets to quipu — that human beings have devised for themselves: not just the textual language (words, syntax, usage, and punctuation) but all the artifactual extensions that are the purview of codicology, bibliography, typography, and graphic design. In addition, like spoken language, they are amazingly tolerant of every kind of human or nonhuman devi-

6. The computational process is designed to validate the self-identity of the information across the entire channel, not to promote exchanges between the various sending and receiving agents. It's true that if one were to read the interface texts as one reads traditional artifactual texts explicit reciprocity would obtain at least at the interface level. But even there the inertia of algorithmic design, its commitments to speed, generality, and mathematicized precision, distorts the flexible character of natural language works, necessarily short-circuiting interaction.

ance or intervention, deliberated or otherwise. Indeed, the many sorts of ruptures and defacements they regularly suffer also regularly become a significant part of the documentary meaning and design.

Natural language works are the mother of double-tongued inventions. They go bravely in fear and trembling of the information they ask us all to handle, operating under the sign that Dante laid down centuries ago for poets and artists: "Ma la natura la dà sempre scema, / similmente operando a l'artista/ che l'abito de l'arte ha man che trema" (*Paradiso* XIII. 76–78). They cultivate their arts, and their artifacts, with a trembling hand.

The Textual Condition is explicitly reciprocal throughout.

II.

Cultural memory is preserved in multiple media — oral, performative, textual, graphical — and the different vehicular forms regularly interact with each other in complex ways over time. Sometimes a single artist — Dante Gabriel Rossetti, Richard Wagner — operates simultaneously in more than one medium, but even their work is multiply-mediated by multiple agents across its composition and transmission phases. When electronic devices proliferated through the twentieth-century, multiple-(re)mediation became even more widespread and sundry. Yeats saw in phonography and radio the promise of oral poetic performance, a rebirth of "The Living Voice".

De Angulo's *Old Time Stories* (*OTS*), a notable offspring of those conditions, presents editorial challenges more demanding than even Ovid or Rossetti, being an interlaced network of oral, textual, and graphical works preserved in both traditional and electronic documents. At once complete and unfinished, *OTS* is a long narrative prose poem in an American English that he extruded from his two Western mother tongues (Spanish and French) and then decisively reshaped to echo and imitate the pre-Modern indigenous languages, ritual and music-based, that were the focus of de Angulo's ethnolinguistic research. Conceived originally (1928) as a set of "Indian Tales for a little boy and girl" to be recited to his children, it shape-shifted over the next twenty years: first as a massively illustrated textual narrative in several instantiations, finally as a series of radio performances — daily in fifteen minute episodes — that extended over a year (from April 1949 to March 1950) in at least two distinct versions, one complete (in thirteen hours), the other open-ended (it comprised some twenty-two hours). After de Angulo's death in 1950, the work was editorially reinvented yet again in multiple textual, oral, illustrated, and electronic modes.

The material conditions of such a work, so abundantly nonuniform, would present serious problems for our traditional Western models of editing which hold in view the delivery of an "ideal" or "authoritative" text. The *OTS* stands aside those traditions because its approach to language has been so decisively shaped by de Angulo's ethnolinguistic study of the performative language cultures of native California. In that respect the *OTS* might seem such an outlier case as to be useless for thinking in normative terms about scholarly editing. But it seems to me the exceptional case that proves a key rule about The Textual Condition of all works of natural language: that they emerge and then regenerate through the many agents involved in lending them life over time.

Works of natural language are explicitly reciprocal machines because their messages are handed back and forth in shared codings. Some agents — authors and readers for example — hold offices that are usually more consequential than, for example, typesetters or editors (commercial or scholarly). But that is only a general rule. Typesetting and design (the Gutenberg Bible; *Leaves of Grass*) and editing (*King Lear*; *The Waste Land*) are often as consequential as a work's linguistic codes. The other general rule is that all the agents are more or less realized duck-rabbits in their acts of exchange (classically, the author is also the reader and the reader is also the author).

Editing natural language works with digital tools calls for platforms that explicitly operate with the originals' full range of expressive reciprocities. The *OTS*'s complex expressive materials answers to the range requirement. Even more important, however, is the need for a platform that maintains reciprocal exchanges between the machinic computations and human recalculations. Graph database technologies offer practical ways of responding to that need. The project is necessarily experimental since graph databasing has to date been applied almost exclusively to data-centric commercial enterprises. Working with the open source graph database Neo4j, we've spent two years designing the platform that will deliver a functioning model for editorial investigations of the most complex features of natural language works.

Neo4j responds to the reciprocal operations of natural language by an initial translational move that brackets out all localized presence of natural language: it atomizes any designated set of natural language units (e.g., words) to abstract arithmetic coordinates (specifically, to time stops in the audio files and pixel coordinates in the textual files). Each is a unit in an array of nodes with n-dimensional edges (discrete objects with relational values). This elementary move in the platform's design hands over

to the user a complete but explicitly unfinished computational product of documentary materials. Specific functional relationships are left to be defined by the user and then fed back for serial recomputation/recalculation exchanges.[7]

The platform reflects the presence of natural language by setting in place a number of finished "works" (e.g, recognized editions) whose contents have been historically defined at a second order. Nonetheless, the platform abstracts and atomizes all the elementary units of those second order works, leaving them open to further research and interpretive change. The platform remains fundamentally organized to help users produce declarative investigations of the primary natural language materials in the three interpretive genres of traditional philology: annotations in an open-ended set of "Notes and Queries"; annotations organized to frame an interpretive argument (synthetic or deconstructive); annotations organized to propose an "edition" that embraces some or even all of the materials in a finished interpretive proposal.

The computational design returns these declarative investigations to the database when a particular line of investigation — a research project — declares itself closed. Until that terminal point an investigation is maintained as a user-defined, stand-alone ("sandbox") editorial inquiry into the master graph dataset. Once "closed", however, the result becomes an interpretive action that interoperates with all of the platform's second order philological entities, including the initial historically defined set. The editorial research is thus directed toward discovering both new second order philological perspectives as well as fresh insights into how they emerge from relationships that get exposed at more primitive levels. As to the latter, we expect to find word vectors an especially useful tool. We also expect — this has yet to be tested — that our graph approach will allow us to define the graph's nodes much more broadly/usefully to include syntactic and rhythmic units.

Coda

What I've called "declarative investigations" are what Peirce discussed as "abductions" (or interpretive "inferences" and "guesses"). As Erik Larson's

7. DH scholars have only just begun to give serious attention to this technology; see Spadini et al. 2021, in particular essays by Prosser and Schloen and Neill and Schmidt.

recent book explains, "Computers Can't Think the Way we Do" because AI's (computational) inferences are radically different from the "world knowledge" (2021, 53) that is entailed in Peircean abduction.[8]

What Larson calls "normal intelligence" — "Inferences from particular observations to particular explanations" (2021, 162) — differs from mathematical intelligence because, operating within the orbit of natural language exchange, its users have to realize that their deductions and inductions are driven by abductions, which are always simultaneously feasible and defeasible. Normal intelligence therefore runs a science of particulars: particular exchanges between particular persons occurring at particular times and under particular conditions that natural language holds open to other (former or further, known or unknown) "conjectural inferences" (2021, 163).

For Larson, Peircean "abductive inference" lies at "the core mystery of human intelligence"; the problem for him is that "we don't know how to program it" (2021, 190). But that would be a less serious problem if, like natural language, our computer designs did not set out to program that mystery but to set its agents free to think and, if drawn to make an inductive or deductive throw of the dice, to see once again how they will never abolish chance.

University of Virginia

Works Cited

Berrie, Phillip, Paul Eggert, Chris Tiffin, and Graham Barwell. 2006. "Authenticating electronic editions". In *Electronic Textual Editing*, edited by L. Burnard, K. O'Brien O'Keeffe, and J. Unsworth, 269–76. New York: The Modern Language Association of America. https://ro.uow.edu.au/artspapers/527.

Eggert, Paul. 1994. "Document and Text: The 'Life' of the Literary Work and the Capacities of Editing". *TEXT,* 7: 1–24

———. 2019. *The Work and the Reader in Literary Studies: Scholarly Editing and Book History*. Cambridge: Cambridge University Press.

Kirschenbaum, Matthew. 2014. "Operating Systems of the Mind: Bibliography After Word Processing (The Example of Updike)". *The Papers of the Bibliographical Society of America,* 105.4: 380–412.

———. 2021. *Bitstreams: The Future of Digital Literary Heritage*. Philadelphia: University of Pennsylvania Press.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence. Why Computers Can't Think the Way We Do*. Cambridge: Belknap Press, Harvard University Press.

8. See Peirce 1929.

McGann, Jerome J. and Allen Renear. 1999. "What is text? A debate on the philosophical and epistemological nature of text in the light of humanities computing research". http://www2.iath.virginia.edu/ach-allc.99/proceedings/hockey-renear2.html.

McLeod, Randall [McLoud, Randall]. 1991. "Information about Information". *Text,* 5 (1991): 241–86.

Neill, Iian and Desmond Schmidt. 2021. "SPEEDy: A practical editor for texts annotated with standoff markup". In Spadini et. al., 45–54.

Nelson, Ted. 1997. "Embedded Markup Considered Harmful". https://www.xml.com/pub/a/w3j/s3.nelson.html.

Peirce, Charles S. 1929. "Guessing". *Hound and Horn,* 2: 282–85.

Prosser, Miller C. and Sandra R. Schloen, 2021. "The Power of OCHRE's Highly Atomic Graph Database Model for the Creation and Curation of Digital Text Editions". In Spadini et. al., 55–72.

Shannon, Claude. 1948. "A Mathematical Theory of Communication". *Bell System Technical Journal,* 27 (July, October): 379–423, 623–56.

Spadini, Elena, Francesca Tomasi, and Georg Vogeler. 2021. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing.* Norderstedt: BoD.

Stein, Gertrude. 1926. *Composition as Explanation*. London: Hogarth Press.

Talfourd, Thomas Noon, ed. 1879. Charles Lamb, *The Complete Works of Charles Lamb: Containing His Letters, Essays, Poems, Etc., with a Sketch of His Life*. Philadelphia: W. T. Amies.

Tanselle, G. Thomas. 1995. "The Varieties of Scholarly Editing". In *Scholarly Editing. A Guide to Research,* edited by D. C. Greetham, 9–32. New York: The Modern Language Association of America.

Yeats, W. B. 1906. "Literature and the Living Voice". *Samhain,* 6: 4–14.