Text Analysis and the Dynamic Edition? A Working Paper, Briefly Articulating Some Concerns with an Algorithmic Approach to the Electronic Scholarly Edition

Ray Siemens, with the TAPoR Community
University of Victoria
siemens@uvic.ca

Abstract

Beginning with the assumption that text analysis and 'algorithmic' approaches to textual navigation provide a foundation for the dynamic scholarly edition, in 2002 I began a number of conversations of varying degrees of formality chiefly, but not exclusively, with researchers associated with the Text Analysis Portal for Research (TAPoR) initiative. My working paper is intended to recount, generically, the salient points of these discussions, as they document an important stage in our thinking about the future of the scholarly edition.

KEYWORDS: Electronic editing, text analysis, TAPoR, dynamic edition.

Introduction & Context

Those who work in my specific area of inquiry, English Renaissance literature, and those who worked in literary text analysis in the early 1990s, may recall a largish review article from that time by Whitney Bolton, published in the journal Computers and the Humanities. Entitled "The Bard in Bits: Electronic Editions of Shakespeare and Programs to Analyse Them" (Bolton, 1990) it was in many ways a tour-de-force for the community that was interested in literary studies of the period and textual analysis; the article documented what was then (as it is now) a very fertile area, rife with discussion of a good number of encoded texts -- most derived from accepted scholarly editions or prepared to accepted standards for scholarly editing and transcription -- and good mention of what, then, was felt to be the natural companion to such electronic texts: useful, integrated text analysis software.

Bolton's discussion is still, today, exemplary in illuminating one of two dominant perspectives on the electronic scholarly edition. I mean,

here, the notion of the "dynamic text," which consists of an electronic text and integrated advanced textual analysis software; the dynamic text presents, in essence, a text that indexes and concords itself, allowing the reader to interact with it in a dynamic fashion, enacting text analysis procedures upon it as it is read (see Lancashire, 1989). The widespread adoption of hypertext, because of its facility for organising electronic objects, was evident at the same time as Bolton wrote, but hypertext had not yet reached its apex in the Internet's adoption of HTML; but it would soon give rise to the "hypertextual edition," an edition exploiting the ability of encoded hypertextual organisation to facilitate a reader's interaction with a text, the apparatus (textual, critical, and otherwise) that traditionally accompanies scholarly editions, and relevant external textual and graphical resources, critical materials, and so forth (Faulhaber, 1991). Bolton's review champions the dynamic text, though were the review to have been written even a few short years after its 1990 publication date, the hypertextual edition would have evolved such that it could receive at least the same level of attention as the dynamic text that predominantly occupies Bolton's attentions.

Such was the state of things in the early 1990s; and, for all intents and purposes, some might say that such is the state of things today. Many, though, will know the finer subtleties of our current situation... of where we've been since Bolton's exemplary review article, since Lancashire succinctly documented the notion of the "dynamic text," and since Faulhaber rightly anticipated so many elements of the hypertextual edition. What the last decade has seen, as my colleague and collaborator Susan Schreibman has noted, is that computer-assisted approaches to the textual edition had such impact on the work of the textual scholar that it quickly became commonplace, and well-justified, to assert that textual scholarship had been released "from the spatial restrictions of the codex form." The "focus of the textual scholar's work," Schreibman notes, had changed considerably "[t]hroughout the 1990s; rather than synthesising [textual materials], the textual scholar accumulated" (Schreibman 2002). I might restate this assertion as follows: rather than exploring the ways in which textual materials related to one another -- textually, critically, analytically, and so forth – the main focus of computing-oriented textual scholars of the past decade has been upon digitising textual and extra-textual information, encoding it to accepted scholarly standards (at the same time as those standards have evolved), and associating it via hypertextual links.[1]

This activity represented a significant shift in focus from trends in

the days of the late 1980s and early 1990s, before the rise of hypertext and its now near-intuitive embrace resultant from our use of the World Wide Web. It is uncontroversial, I think, to state that what is different about the 90s from the decade that preceded it is that, alongside this accumulation of textual materials, for the past decade textual scholars have not paid the same attention to the potential of computer-assisted techniques for analysis of those materials, nor to the relation of the electronic scholarly edition and text analysis which appears, in Whitney Bolton's survey, to be assumed.

But, before getting ahead of myself, I must offer a few caveats. Firstly, though most of the textual studies community did not fix on computer-assisted textual analysis, those in other areas of computer-assisted, text-oriented computing have retained a clear focus on computer-assisted analysis. Secondly, exemplary programs -- such as TACT: Text Analysis Computing Tools -- were published in the 90s, though it is fair to note that these were based predominantly upon work begun before the rise of the World Wide Web, in the climate that Bolton documents. Lastly, encoding texts and establishing hypertextual relations are, indeed, important analytical activities, and they are necessary to programs of computer-assisted textual analysis; the field traditionally associates them with textual bibliography more so than critical-theoretical processes they embody.[2]

Returning to the matter at hand: what we have as the predominant products of electronic textual scholarship of the 1990s, then, are two related things: highly-encoded electronic texts and related extra-textual digital objects, at times gathered in archives, and electronic scholarly editions that, while based on such highly-encoded texts (and themselves rife with information for computer-assisted analysis), make exemplary use of hypertextual facilitation for the navigation of textual materials, but do not yet offer text-analysis features commensurate with expectations even at the beginning of the last decade.

Alongside this, advances in computing -- and computing's further advance into disciplines that include textual editing -- have, over the same past decade, made it clear that electronic scholarly editions can incorporate dynamic interaction with the text and its related materials and, at the same time, also reap the benefits of the fixed hypertextual links that typify the standard relation of materials we find in most editions of this sort. Indeed, contemporary scholarly consensus is that the level of dynamic interaction in an electronic edition itself -- if facilitated via text analysis in the style of the dynamic text -- could replace much of the interaction that one typically

has with a text and its accompanying materials via explicit hypertextual links in a hypertextual edition. That said, at the moment, there is no extant exemplary implementation of this new "dynamic edition," an edition that transfers the principles of interaction allowed by a dynamic text to the realm of the full edition, comprising that text and all its extra-textual materials (textual apparatus and commentary, and beyond).

Such is the observation that led me into a series of discussions with a number of those involved in humanities computing and textual editing, beginning with these questions: Why have we not yet properly melded the computer-assisted textual analysis of the dynamic text and the multimedia navigational facility of the hypertextual edition? Why have we not yet developed a dynamic edition? What might that edition look like? And how might we go about building it?

Articulating Some Concerns with an Algorithmic Approach to the Electronic Scholarly Edition

Beginning with the assumption that text analysis and 'algorithmic' approaches to textual navigation provide a foundation for the dynamic edition, in 2002 I began a number of conversations of varying degrees of formality with parties who might be willing to accept such an assumption, and might be willing to consider exploring the questions. Chiefly, but not exclusively, these discussions were carried out with researchers associated with the Text Analysis Portal for Research (TAPoR) initiative. What follows, then, is by no means the results of a scientific study, nor the results of extensive work towards establishing consensus on the future of the electronic scholarly edition. Rather, my working paper is intended to recount, generically, the salient points of these discussions, as they document an important stage in our thinking about the future of the scholarly edition. At the very least, my efforts do just this, even if in heavily condensed form; at best, perhaps, in these observations lie the basic parameters of thinking surrounding the future of one type of electronic scholarly edition from the perspective of those who use them and those who are most likely to create them.

Concerns relating to an algorithmic approach to the development of the electronic scholarly edition fell into roughly four areas of need: meeting community needs, expectations, and expertise or familiarity levels; repurposing existing tools and developing new tools; the seamless integration of those tools with one another; and development of an

interface to the edition that is at least as seemingly-intuitive as the model provided by the World Wide Web and its browsers.

We noted that the first step toward meeting the needs of the community who might best make use of a dynamic scholarly edition is in understanding the expertise and familiarity levels of that community in the first instance with print-based scholarly editions, secondly electronic editions, and text analysis processes. Hypertext was adopted quickly because it appeared to be intuitive; it formalised an existing, longstanding understanding of the relation between one thing and the next; on the other hand, most automated and semi-automated text analysis processes are significantly less familiar to textual editors and literary scholars – with notable exceptions to this being those processes associated with traditionally-accepted tasks involving basic searching, concording, and collation. In addition to having the need for searching, concording, and collation to be handled by the electronic edition, we identified the need for allowing simple navigational strategies (those associated with linear reading) as well as strategies that are more complex – requiring organisation of textual materials, included in the matter of the edition, beyond relating one thing and the next via fixed hypertextual links, as well as the possibility of using visualisation techniques beyond linear representation of the text. An ideal dynamic edition would also feature the ability to work with large texts and textual corpora, in varied encoding formats and across networks and systems, as well as with non-textual material (image and sound, moving image, &c.) and their metadata. As well, those working with such an edition should have the ability to relate those text-centred activities associating 'lower' critical functions (applied bibliography) and 'higher' critical practices (which are informed by critical theory).

In facilitating all this, we articulated a premise that readers should be able to expect a straight-forward functionality and ease of operation, a seamless integration of analysis procedures with reading practices – rather than analysis being a separate act, involving imposition on the reading process – and some degree of alignment of the textual analysis approaches with dominant interpretive practices, perhaps via tools that have an expert system component.

We noted that there are valuable tools already available in the text analysis community that could be repurposed and adapted to this end, among them those associated with the TAPoR group such as TACT and TACTWeb – with additional extant coding, into critical modules such as those suggested by Rockwell and Bradley (1998) – and HyperPo (Sin-

clair); further, we noted TAPoR's general mandate to provide an open text analysis portal, as well as Sinclair's work within that mandate, exploring adaptable open source text analysis tools. Even so, there are much-needed tools that remain to be developed, among them intuitive and dynamic text collation software that work with texts in various formats and encoded-states,[3] complex search tools that draw both upon a text's conformance to extant encoding systems and upon algorithmic locational, relational, and retrieval processes – incorporating content and context analysis techniques, having the ability to operate within the semantic web – as well as a number of display and visualisation tools that draw upon innovative work in the area but also provide their results in ways historically recognisable by those in a number of fields.

We discussed that these tools need to be integrated, seamlessly, and offered via an interface that is intuitive or, at least, straightforward. There are two dominant strategies of integration at work in our fields today: one is actual integration, at the level of the software itself, and the other is apparent integration, at the level of operational metaphor and interface. Ideally, the interface to the dynamic edition should be intuitive, or learned easily; it should be robust, or at least not overly-simplified; and should employ operational metaphors drawn from the textual studies community, but understood by, or adaptable to, other interpretive communities in the humanities as well. Central to issues relating both to integration and to interface is the education of the community envisioned to make use of these tools.

Conclusion

It is at this point that our deliberations have stopped, but we anticipate several initiatives to move forward in directions such as those charted above in the near future. Consensus was reached, it is fair to say, in the recognition that the creation of the dynamic edition involves a strong assertion that textual analysis lies at the heart of the electronic scholarly edition, through the repurposing, adaptation, and creation of sophisticated analysis, navigation, and representation software -- such that the tools handle not only text, but pertinent non-textual data as well, toward the end of providing the means by which the highly-encoded texts that currently exist within our textual studies community can be integrated into textual editions and larger text corpora that take fullest advantage of computer-assisted textual analysis.

Notes

[1] The above I've treated in more detail earlier (see Siemens 1998, 2001).
[2] Discussed in Siemens (2002).
[3] Here, several tools were discussed, including Schreibman's Versioning Machine, Peter Robinson's Anastasia, and the edition creation tools (Edition Production Technology [EPT]) being developed by Kevin Kiernan, among others.

Works Cited

Anastasia. 6 Dec. 2004 <http://anastasia.sourceforge.net/index.html>.

Bolton, W. F. (1990). "The Bard in Bits: Electronic Editions of Shakespeare and Programs to Analyse Them." *Computers and the Humanities* 24.4: 275-87.

Faulhaber, Charles B. (1991). "Textual Criticism in the 21st Century." *Romance Philology* 45: 123-148.

Lancashire, D. Ian. "Working with Texts." Paper delivered at the *IBM Academic Computing Conference*, Anaheim, 23 June 1989.

Rockwell, Geoffrey, and John Bradley. (1998). "Eye-ConTact: Towards a New Design for Text-Analysis Tools." *Computing in the Humanities Working Papers* A.4 <http://www.chass.utoronto.ca/epc/chwp/rockwell/>.

Schreibman, Susan. (2002). "Computer Mediated Texts and Textuality: Theory and Practice." *A New Computer Assisted Literary Criticism?* R.G. Siemens, ed. [A special issue of] Computers and the Humanities. 36.3: 283-293.

Siemens, R.G. (2002). "A New Computer-Assisted Literary Criticism?" *An introduction to A New Computer-Assisted Literary Criticism?* R.G. Siemens, ed. [A special issue of] *Computers and the Humanitie*s 36.3: 259-267.

---. (2001). "Unediting and Non-Editions." *In The Theory (and Politics) of Editing*. Anglia 119.3 (2001): 423-455. [Reprint, with additional introduction, of "Shakespearean Apparatus? Explicit Textual Structures and the Implicit Navigation of Accumulated Knowledge." *Text: An Interdisciplinary Annual of Textual Studies* 14. Ann Arbor: U Michigan P, 2002. 209-240. Electronic pre-print published in *Surfaces* 8: 106.1-34. <http://www.pum.umontreal.ca/revues/surfaces/vol8/siemens.pdf>]

---. (1998). "Disparate Structures, Electronic and Otherwise: Conceptions of Textual Organisation in the Electronic Medium, with Reference to Editions of Shakespeare and the Internet." 6.1-29 in Michael Best, ed. *The Internet Shakespeare: Opportunities in a New Medium*. Early Modern Literary Studies 3.3/Special Issue 2. <http://www.shu.ac.uk/emls/03-3/siemshak.html>.

Sinclair, Stéfan. *HyperPo: Text Analysis and Exploration Tools*. <URL: http://huco.ualberta.ca/HyperPo/>.

TACT. Ian Lancashire, in collaboration with J. Bradley, W. McCarty, M. Stairs, and T. R. Wooldridge. (1996). *Using TACT with Electronic Texts: A Guide to Text Analysis Computing Tools*. New York: Modern Language Association.

TACTWeb. 6 Dec. 2004 <http://tactweb.humanities.mcmaster.ca/tactweb/doc/tact.htm>.

Text Analysis Portal for Research(TAPoR). Geoffrey Rockwell, project leader. <http://www.tapor.ca/>.

The Versioning Machine. 6 Dec. 2004 <http://mith2.umd.edu/products/vermach/>.