click the image to launch the slideshow

Forms of Attention: Digital Humanities Beyond Representation

John Unsworth

A paper delivered at "The Face of Text: Computer-Assisted Text Analysis in the Humanities," the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004.

Frank Kermode's Wellek Library Lectures, published in 1985 as the book titled *Forms of Attention* (Chicago: University of Chicago Press), are prefaced with this opening remark:

> By what means do we attribute value to works of art, and how do our valuations affect our ways of attending to them?

Kermode is interested in canon-formation, and in order to understand how quality and reputation are related, he focuses on

> the history of those unusual objects which vanish from the "canon,'"vanish, indeed, from view, but are restored after long periods of neglect. (xiii)

The title I've chosen alludes to Kermode's book, and I am interested, as he was, in the way we value and the way we attend to works of art, but more precisely, in what follows I'm interested in the different forms of attention we pay to those works, in humanities computing, and why those forms of attention change over time. Kermode's focus was on the institutional and social forces, and sometimes the sheer accidents, that shape accepted notions of what is good and what is bad, and what is beneath notice. My interest tends more toward a different set of social forces—the technical and the legal—that shape our digital engagement with the works that have survived the process Kermode describes.

For as long as there has been humanities computing, humanities computing has been about the representation of primary source materials, and about building tools that either help us create those representations or help us manipulate and analyze them. Source materials might be—and have often been—texts, but they might also be drawings, paintings, architectural or archaeological sites, music, maps, textiles, sculptures, performances, or any of the other forms in which the record of human culture and human creativity can be found. Under that description, it might seem that we've been working steadily at more or less the same thing, for quite a long time now—more than half a century. But it would be a mistake to make that assumption, and a disservice to impute mere stasis or repetition to the field. In fact, over the half-century leading up to now, there have been significant accomplishments and significant changes. As a result of them, and (even more) as a result of changes in the world around us, we are, I think, on the verge of what seems to me the third major phase in humanities computing, which has moved from tools in the 50s, 60s, and 70s, to primary sources in the 80s and 90s, and now seems to be moving back to tools, in a sequence that suggests an oscillation that may repeat itself in the future. But whether or not the pattern ultimately repeats, I think we are arriving at a moment when the form of the attention that we pay to primary source materials is shifting from digitizing to analyzing, from artifacts to aggregates, and from representation to abstraction.

In an essay published in 1996 in *Victorian Studies* (and online the year before that), Jerome McGann wrote that "To date, computerization in humanities has been mainly located in work closely associated with the library, and with library-oriented research projects" ("Radiant Textuality"). That's a very durable generalization, all the more so because the library represents not only a collection of primary source materials, but also tools for use with those materials. Things like reference works, dictionaries, encyclopedias, concordances—or, for that matter, critical editions—are texts-as-tools, texts that are composed as tools to be used with other texts (which they sometimes contain, as in the case of a critical edition). The analysis of texts by means of other texts is one of the oldest methods of humanities scholarship, dating back to the earliest days and oldest traditions in every textual culture, and my point here is not to suggest that the interest in tools and the interest in primary source materials are mutually exclusive: on the contrary, the two are complementary—but one of the two is usually foregrounded as the primary object of attention.

[new slide: texts as tools] Most fields cannot point to a single progenitor, much less a divine one, but humanities computing has Father Busa, who began working (with IBM) in the late 1940s on a concordance of the complete works of Thomas Aquinas. Shown here with a Pope he has subsequently survived, the 91-year-old Jesuit has seen this project through generations of computer technology, from punchcards to magnetic tape and ultimately to CD-ROM (with a 56-volume print edition appearing along the way). The *Index Thomisticus* and other early forays into humanities computing (like the Reverend John W. Ellison's 1957 Concordance to the Revised Standard Version of the Bible, or Roy Wisbey's work on 12th-century German) in the late 1950s and early 1960s, were geared toward producing **tools**—albeit tools that could take the form of texts. The effort (and it was enormous effort) devoted to these projects was justified on the grounds that source material was worthy of scholarly attention, and the resulting tool would facilitate the application of that attention. The choice to produce the tool with computer technology was justified on the grounds that the computer "freed the scholar from the tedious, repetitious clerical work and returned him to his real vocation—scholarly research" (1).

That choice—to use computer technology to automate repetitious clerical work—was visionary, at the time. Even if you lived through the changes, it is actually difficult to remember how recently computers and their peripherals acquired some of the features that make them useable for the work of daily life. The first personal computer—meaning one that would fit on your desk—wasn't commercially available until 1973. The first computer with a mouse (the Xerox Alto) came along in 1974. The first laser printer was introduced by Xerox in 1978: it cost half a million dollars. Optical storage wasn't available until 1980, and it wasn't until 1981 that the IBM PC was introduced, with its brand-new MS-DOS operating system and a 4.77MHz Intel processor. Proportional fonts for display and printing don't become available until the first Macintosh computer is introduced by Apple, in 1984. Color monitors on personal computers become available first with the Apple II (1977), and on IBM PCs in the 80s (4 colors in 1981, 16 in 1984—only in 1990, do we get 800x600 pixel resolution in "true color" (16.8 million colors) and 1,024x768 resolution in 65,536 colors). And, of course, all of this applies to largely stand-alone computing: computer networks are invented in the 1970s, but not widely used even in research universities until the 1980s, and not used by the general public until the 1990s.

My point is that humanities computing projects in the 50s and 60s, before all this innovation, faced and were shaped by some very basic challenges at the level of representation, as Susan Hockey notes in her chapter on the history of humanities computing, in the new Blackwell's Companion to Digital Humanities:

> At this time much attention was paid to the limitations of the technology. Data to be analysed was either texts or numbers. It was input laboriously by hand either on punch cards with each card holding up to eighty characters or one line of text (uppercase letters only), or on paper tape where lower case letters were perhaps possible but which could not be read in any way at all by a human being. . . . Character-set representation was soon recognized as a substantial problem. . . . Various methods were devised to represent upper and lower case letters on punched cards, most often by inserting an asterisk or similar character before a true upper case letter. Accents and other non-standard characters had to be treated in a similar way and non-roman alphabets were represented entirely in transliteration.

It is worth reflecting for a moment on the limitations being described here, and on their significance in the humanities. Accurate transcription might be accomplished, but only if characters didn't vary too much from the teletype keyboard, and even then, it would be necessary to use a variety of idiosyncratic, non-interoperable hacks to note things as simple as capital letters or accents. There were similarly basic limitations in other areas as well. Hockey continues:

> Most large-scale datasets were stored on magnetic tape, which can only be processed serially. It took about four minutes for a full-size tape to wind from one end to the other and so software was designed to minimize the amount of tape movement. Random access to data such as happens on a disk was not possible. Data had therefore to be stored in a serial fashion. This was not so problematic for textual data, but for historical material it could mean the simplification of data, which represented several aspects of one object (forming several tables in relational database technology), into a single linear stream. This in itself was enough to deter historians from embarking on computer-based projects.

These are only a couple of examples, but they may serve to remind us that not only was the nature of the tools proscribed by the limitations of computer technology, but the nature and the features of the source material accessible to tools were similarly proscribed. In other words, computers were only good for paying certain very limited kinds of attention to certain very limited features of certain very limited subsets of primary source materials in the humanities. Texts in Latin were a good starting point, because computers used the roman alphabet, and if you were concerned with aspects of the text that you could abstract from the artifact itself by transcription, then you could use the computer's ability to store and retrieve and order and count as an aid in producing text-tools like indexes and concordances. On the other hand, if you were interested in information that resisted representation in this medium, for example the relational social structures that interest historians, then computers, in the 50s and 60s, wouldn't have seemed especially useful as tools for paying attention to important aspects of the human record.

[new slide: batch tools] Naturally, software tools developed during this period shared the characteristics of other computer technology of the time. Take as an example TUSTEP (TUebingen System of TExt-processing Programs), a suite of tools that could be used for text-analysis and for the creation of critical editions, which was developed by Wilhelm Ott and others, at Tuebingen University, over three decades, beginning in the 1970s. In the TUSTEP web site's general introduction, we learn that work on TUSTEP

> ...began in 1966 when we first designed a series of functions and subroutines for character and string handling in FORTRAN . . . and implemented them on the mainframe of the Computing Center of the University of Tuebingen. This made programming easier for projects such as the Metrical Analysis of Latin Hexameter Poetry, the Concordance to the Vulgate, or the edition and indexes to the works of Heinrich Kaufringer. Proceeding from the experiences gained from those projects, the next step in supporting projects was to no longer rely on programming in FORTRAN or other "high level" languages, but to provide a toolbox consisting of several program modules, each covering one "basic operation" required for processing textual data. The function of each program is controlled by user-supplied parameters; the programs themselves may be combined in a variety of ways, thereby allowing the user to accomplish tasks of the most diversified kind. It was in 1978 when these programs got the name TUSTEP.

TUSTEP functions include automatic collation of different versions of a text; text correction (either with an editor or in batch mode, using instructions prepared beforehand); decomposing texts into elements (e.g. word forms) according to rules provided by the user; building and sorting logical entities (e.g. bibliographic records); preparing indexes by building entries from the sorted elements; processing textual data by selecting records or elements, by replacing strings or text parts, by rearranging, completing, compressing and comparing text parts on the basis of rules and conditions provided by the user, by retrieving numerical values which are already given in the text (like calendar-dates) or which can be derived from it (such as the number of words in a paragraph); and transforming textual data from TUSTEP files into file formats used by other systems (e.g. for statistical analysis or for electronic publication). And because Professor Ott and his colleagues continually updated TUSTEP, it acquired the ability to output Unicode and XML, and continued developing well into the age of the Web, CGI, Web Services, SOAP, and the like.

All of TUSTEP's features are quite useful if you are preparing an index, concordance, critical edition in print, bibliography, or dictionary—but they all fall within the realm of attention that one can pay to the alphanumeric elements of source material, and the fundamental paradigm on which TUSTEP is based is the batch process, as Chaim Milikowsky pointed out in an email to Humanist in 1996:

> To the best of my knowledge — I received my tutorial in TUSTEP in December 1994— there is absolutely no interactivity in TUSTEP at all. Now some level of interactivity would probably be relatively trivial to create in TUSTEP, but my sense is that the basic algorithms are deeply geared to a batch mode of operating. This doesn't bother me that much, but somehow I doubt that any such set of programs could become popular in this day and age.

Tools, too, are subject to fashion, and in 1996, TUSTEP seemed out of step with the times—and yet some of the principles on which it was founded are quite timely, still: modularity, professionality (meaning that it should go beyond off-the-shelf office products to meet the needs of research users), integration, and portability. These are things we still look for in our tools, but we also look for interactivity, networkability, and internal data representations that use open standards.

[new slide: interactive tools] The next generation of tools for digital humanities might be represented by TACT (Text Analysis Computing Tools), which was developed by John Bradley, and released in 1989. Subsequently, TACT (originally a DOS-based program) was ported to the Web, as TACTWeb. TACT follows a more recent query-based paradigm of computing that grows out of database software: indeed, the user installs the software on a personal computer and builds a TACT database of the text or texts that are of interest. TACT queries return keyword in context (with user-configurable parameters for what constitutes the context) or they return statistical information about the frequency and distribution of lexical elements in the database (words in a text). But in a 1992 paper that Bradley wrote with Geoff Rockwell, the authors note that, in working with TACT on their own research problems,

> (a) computing work was most efficiently done by combining small computing "primitives" to perform larger tasks. To this end, we used existing software wherever possible, including various Unix tools such as Awk, Grep, and Sort, with larger pieces of software such as Tact and Simca. However, we were still obliged to write some bridging software in an ad hoc fashion. This brings us to our second observation: that (b) our computing interest required us to continually add new processes to answer new questions. Some of the new software used results of Tact searches, some read the Tact database directly, some worked on other intermediate data formats. As mentioned above, we were in a unique position to reuse and reassemble parts of Tact for a range of different purposes. However, it quickly became clear that it would be virtually impossible for anyone else to use the Tact program code to assemble on-the-fly specialized processing as we could, even if they had access to all of it. [2]

There are two other conclusions in the article, and I'll return to one of them later, but for now, I just want to note that this passage makes it clear that TACT was really a very different kind of tool from TUSTEP. To begin with, it was much smaller—not a comprehensive suite of tools, but something that you might use alongside grep and sort, for a fairly narrowly defined purpose. As the authors tried to use it for more comprehensive investigation of text, they found they needed to mix it with other tools. In the TUSTEP world, this would have been difficult at best—the TUSTEP documentation actually points out that one virtue of the system is that your text can live inside of TUSTEP for all phases of the project. What changes, between TUSTEP and TACT, is the rise of the database paradigm: TACT assumes queries submitted to the database return results which are then piped to various external programs, which might in turn send output to other programs. TUSTEP was fundamentally about "textdata processing" (a term invented to distinguish its focus from word-processing software), but TACT is fundamentally a much more narrowly focused tool designed for analyzing the statistical properties of texts. TUSTEP mentions as one of its features the preparation of texts for statistical analysis, but the documentation doesn't suggest that TUSTEP would actually do that analysis itself.

[new slide: math as method] With due respect to TACT, statistical methods in humanities disciplines have had what you might call a limited vogue: perhaps their widest acceptance was in history, in the early 1970s, best exemplified in Stanley Engerman and Robert Fogel's book *Time on the Cross*, which applied econometric statistical methods to quantitative data gathered from communities in the antebellum South, and argued for a reassessment of slavery as "socially benign." That outcome caused a considerable backlash in history, and contributed to the decline in favor, in the United States at least, of this particular form of attention to humanities data. Other humanities disciplines, like English, have found some limited use for statistical analysis in authorship studies and more general kinds of stylometrics, but frankly there are not that many scholars in any branch of the humanities who are avid users of statistical methods, and it's not because the tools aren't there—it's because, at least until recently, the problems those tools could solve were not ones that were of central importance in the humanities. Statistics are all about measuring the measurable, whereas the humanities are all about effing the ineffable.

In fact, the whole model of "problem-solving" is somewhat ill at ease in the humanities. In the sciences, one solves problems in order to move on to other problems—but as Kermode points out, we're not about solving the problem of Shakespeare so we can go on to some other problem and stop paying attention to the Bard. Instead of problem-solving, the rhetorical model in the humanities is appreciation: we believe that by paying attention to an object of interest, we can explore it, find new dimensions within it, notice things about it that have never been noticed before, and increase its value. And when image-based digital scholarship emerged, it was a much better fit with the humanities' native paradigm of appreciation than statistical analysis had been.

Although there were computer-based editions before the 1980s, with very rare exceptions, these editions were realized in print. With the advent of Apple's Hypercard (and the CD, and color graphics) in the 1980s, a new form of humanities computing appeared: the electronic text designed to be delivered and used in electronic form, hypertextual in its interface, and gradually including more graphical content. Some of the very earliest of these texts were new creative works, but from very early on at least a few people were working on electronic scholarly projects. One of the earliest of these was the Perseus Project, founded officially in 1987, originally delivered on hypercard, on CD-ROM (published by Yale UP, in 1992, Mac only, with accompanying videodisc), and providing you with a significant sampling of the record of classical civilization, on your home computer. There was a second release of the Perseus CD in 1996 (and a third in 2000), but from 1995 on, the Perseus Project has focused on the Web as its main mode of delivery.

From the beginning, projects such as Perseus included tools, especially texts-as-tools (like dictionaries), but the focus had shifted: in the foreground now was the artifact. The reason for that was simple: we now had the ability to produce and share digital images in color.

[new slide: authoritative representation] Given the focus on the artifact, it is not surprising that this sort of humanities computing project, casting about for something to describe its scope and its aims, should have settled on "archive" as the generic label. Archivists, once they became aware of this usage, pointed out that a true archive was an unintentional record of some other activity, which disqualified these highly intentional structures from being considered archives—but that distinction, while important to archivists, was less important to humanities scholars, who saw archives as the storehouses of original artifacts, first and foremost. A digital archive, or an electronic archive, or a hypertext archive, or a web archive (no consensus ever emerged on the modifier) was a storehouse of digital representations of original artifacts, and that was a pretty good description of how many of these projects took shape: they produced, collected, organized, and edited digital representations of artifacts on a theme. That theme could be an author (like Whitman, for example), but it could as easily be an event (like the Civil War). The Whitman Archive, whose home page is pictured in this slide, is one example, but we sprouted archives right and left at the Institute for Advanced Technology in the Humanities—none of them, of course, with a long-term plan for preservation.

[new slide: archival sources] This is a screen shot of the navigational apparatus for the Valley of the Shadow project, a civil-war history "archive" that includes church records, maps, images, letters, diaries, tax records, newspapers. The image before you makes it quite clear how literally humanities computing has taken the term "archive," in imagining and organizing its work. And in spite of the obvious misappropriation of the term, there are some other interesting historical and etymological implications behind the word archive, in the context of digital humanities. An archive is a place where important (often unique) documents are kept. In their origins, and still today, archives sometimes hold the very records the authenticity of which gives legitimacy to government, and in fact the Greek root of "archive" is the word that means 'government'. Archives also strive for completeness in the records they keep (which is one reason why intentionality is an issue). If an archive, then, is (ideally) a complete collection of authentic and authoritative records of great importance, then it also might supercede the edition, or take it to a new level: if you can accurately represent all states of a work, rather than choosing among them according to some principle like authorial intention, then perhaps you could escape the choices and compromises which print, for so long, imposed on the telling of history (through the lives of the famous) or on critical editions (focused on best texts, or last versions).

This moment in humanities computing, the moment in which we first passed from tools to texts as primary objects of attention in the actual scene of computing, and the moment in which we first realized the computer's ability to represent qualities of the human record beyond one and zero, beyond number and letter, beyond black and white, quite oddly coincided with the heyday of a theoretical turn in the humanities that was largely suspicious of representation. Looking back on that time, from the year 2000, one critic wrote that

> in a number of competing and conflicting narratives, the turn towards modernity is diagnosed as a fall into representation—and this diagnosis has been adopted both by nostalgic anti-modernists who see representational epistemology and ethics as a break with the lived immediacy of the Greek polis, and by postmodernists who define representationalism as the persistent vestige of a Western metaphysics unable to truly confront the death of God. But the nosology of representationalism also has a strange habit of passing over into a celebration. While the concept of representation is held responsible for generating certain illusions of reason, reason can only be disabused of these illusions by the notion that the presence promised by representation is nothing other than representation itself. Thus representation functions as an inherently contradictory notion: at once the sign of reason's failure as well as being the goal of philosophical maturity. [3]

Paradoxically, though, editing—already out of fashion in an age of critical theory—did actually acquire some new power in this moment, even though representation itself was suspect, on theoretical grounds. It did so not because of the increased representational capacity of the digital medium, but because that capacity allowed editors of a theoretical bent to interrogate representation, and editorial theory and practice, in new ways. Perhaps these were just illusions of reason, but perhaps in some cases they amounted to philosphical maturity.

[new slide: editions]The Rossetti Archive became one of the best known of these theoretical editing projects, and not least because its editor (Jerome McGann) was eager to tackle the unknown and push the capabilities of the medium—and his colleagues—to the limit. The Rossetti Archive had, and has, the grand ambition not only to bring together all the states of all the work (text or picture) produced by Dante Gabriel Rossetti, but also to relate those items to one another and to document each of them with bibliographic and textual and historical commentary. The subject for this project was chosen because of his appropriateness to a multimedia treatment, but I doubt that Jerry would disagree with the assertion that in this long experiment far more has been learned about editing than about Rossetti. All to the good: it is much more important, at the moment, to reimagine the practice of editing as a form of attention paid to primary source material, in this new medium, than it is to reimagine Rossetti (who would have been a good candidate for Kermode, having been in, and out, and then back in vogue, during the last century or so).

"Archive" editing projects like the Rossetti Archive do not only represent individual texts: they represent textual genealogies, the relationship between texts and works in other media, and often the intertextuality of works by different creators, across time, place, and media. We'll return to this point, but relationships **are** the point in this sort of project, and it is ultimately not the object in isolation that poses the greatest challenge, but the object in context, in collections, and in relation. Editorial work, at its most fundamental level, consists of specifying those relationships. And that specification opens the way for a tipping back in the direction of tools—but more on that later.

[new slide: comparison] With storage costs dropping from about $2000/gigabyte in 1993 to about $20/gigabyte in 2000 to about $1/gigabyte today, image-intensive projects like the Rossetti Archive and The Blake Archive were bound to appear. And while it has been less revelatory in terms of editorial theory and method, the Blake Archive has demonstrated that editing in this medium can really change the way we understand and value significant figures in the humanities. In Blake's case, this has meant a restoration of his work as an artist, and his practice as a printer, to equal importance with his (more abstractable) prowess as a poet. The Blake Archive has, I think, forever changed the way students and scholars alike will encounter Blake. The editors have accomplished this by foregrounding the artifact as an object of attention, subordinating commentary and apparatus, and enabling some operations, such as comparison, that become more important when the artifact is the object of attention.

[new slide: investigation] Kevin Kiernan's Electronic Beowulf provides an excellent example of the fact that tools for working with artifacts, digitized as images, have gradually developed, over the last decade. Logically enough, those tools are image-based, exploiting the qualities of the "carrier medium" rather than the statistical qualities of textual data. Note, though, that the text is still in this picture, of course, as the object of investigation, and as annotations and other forms of apparatus—but access to the text is by way of an image.

[new slide: mapping] Salem Witch Trials: A Documentary Archive and Transcription Project is another "archive" project, one which began by wishing to collect, edit, and cross-index the documentary record of the Salem Witch trials, held in a number of different (real) archives. One of the things that such projects do (and the Whitman archive has been doing it too) is to reconstruct and virtually unite a documentary record that may be spread across many real archives. In the case of the Salem Witch Trials, the virtual archive contains not only legal records, but also maps, paintings, religious texts, and a partial record of the culture's reprocessing of the event, in various forms of popular culture. Interestingly, mapping turns out to be very important to all sorts of "text-oriented" projects, and tools developed for use with maps, namely Geographic Information Systems, turn out to have unexpected and profound application in these textual projects. That's because one large subset of the humanities is about events that unfold in a particular space (real or imagined), over some period of time, and mapping the events is an excellent way of understanding them.

[new slide: modeling] Perhaps the extreme or limit case of the representational impulse in humanities computing is the modeling of buildings and sites of historical interest. These have been produced in projects like the Amiens Cathedral project at Columbia, or the Roman Forum (and many others) modeled at the Cultural Virtual Reality Lab at UCLA, led by the person—Bernie Frischer—who has just been appointed director of the Insitute for Advanced Technology in the Humanities at the University of Virginia, or in this project, The Crystal Palace, produced by Chris Jessee and Will Rourk at Virginia, working from facsimiles of the original architectural drawings. Modeling projects such as these take a step beyond digital imaging of an existing artifact—usually they are reconstructions, in part or in whole, and rather than showing us what is there, they strive to represent what is no longer there. What are the benefits of this activity? As in all modeling projects, these make explicit what you think you know about the object of your attention—an exercise that often ends up showing you what you **don't** know, what your forgot, or what you simply got wrong. What are the risks? A spurious sense of certainty is probably the biggest risk, owing to the uniform polish and apparent completeness of the models.

[new slide: spaghetti] Earlier, in speaking about the Rossetti Archive, I said that "Archive" editing projects represent textual genealogies, the relationship between texts and works in other media, and often the intertextuality of works by different creators, and in fact those relationships are the point in this sort of project, This is a picture of those relationships in a small part of the Rossetti Archive—in a very direct sense you're looking at a picture of the work of scholarly editing. It's also a picture that, for me at least, represents the turning point in my own thinking, the point at which it started to become clear that you could work with these artifactual representations in some more abstract way. Apparently it struck Jerry in the same way: in an essay called Textonics, he wrote:

> The interface we have built for The Rossetti Archive is dismayingly inadequate to the Archive's dataset of materials. At present the Archive organizes approximately 9,000 distinct files, about half of which are SGML/XML files. When the Archive reaches its scheduled completion date some four years from now, it will have about twice that many files. Here is a directed graph of a tiny subset of the Archive's current set of analytic relations. We call this "Rossetti Spaghetti," and I show it to give you a graphic sense of the scale and complexity of this grain of Rossettian sand on the shore of the internet's opening ocean. One can indeed, even here, see an infinite world dawning in that grain of sand.

At about the same time that we cooked up Rossetti Spaghetti, the gods arranged for Geoffrey Rockwell to come and spend a year at IATH, on sabbatical. During that year, Geoff made an enormous contribution to what was going on at Virginia, in curriculum planning, in thinking about research, and in teaching us all something about the importance of play. He also brought with him an abiding interest (going back at least to his work with John Bradley on TACT) in tool-building.

[new slide: taporware] TAPoRware is one manifestation of that interest, not only on Geoff's part, of course, but on the part of a whole community. The fact that TAPoR exists is evidence that humanities computing is returning to a focus on tools, and as ever, the tools will bear the imprint of the technological moment in which they are created. What that means, in the present moment, is that these tools will be

- networked and distributed, with design influenced deeply by the Web and by Web Services
- based on open standards and therefore (somewhat) interoperable
- modular and (if properly designed) relatively easy to extend, modify, and personalize.

[new slide: hyperpo] Hyperpo demonstrates many of these qualities: I'm not sure how modular it is, but it certainly works on open standards and on the assumption that the network, and network resources, are available. The actual functionality of the tool is, in some ways, quite like what TACT offered, but it's networked, and it works with texts on the Web.

[new slide: stagegraph] The next step in developing a new paradigm for tools in humanities computing is visualization. This, all by itself, could in some cases make the investigation of statistical properties of humanities content more interesting to humanities scholars. StageGraph which is Steve Ramsay's project to graph certain features of dramatic texts, is an interesting example of this. This graph shows the paths of Antony and Cleopatra through the play that bears their name—where they intersect, where they don't, what locations they both pass through, etc.. This kind of birds-eye-view of a whole text, or a whole collection of texts, makes immediate sense to scholars, provided that the text, and the features mapped, are both of interest. For my purposes, though, an even better example of the new iteration of tools would be Stefan Sinclair's Hypergraph (which I don't have a slide for, having just seen it today for the first time). What's significant about Hypergraph, or things like it (the visual thesaurus, star-trees, etc.) is interactive visualization: the output is the interface.

[new slide: Grokker] Grokker is an excellent example of visualization as interface. It's effectively a search engine or a meta-search engine, and it clusters results based on their content, and presents them to you in an interactive visualization, so you can zoom in and out, pick and view items, rearrange groups and save your new "map" of the topic. The fact that this is out there, and works (and quickly, at that) makes it clear that this is technically possible: what's lacking is smart text, and tools that leverage that smartness, and visualizations that are suited to what we undeerstand as the significant features and dimensions of value in the objects of our attention.

[new slide: D2K] This notion that we might bring high-level and interactive visualization to bear on the properties of humanities artifacts (now not just texts in alphanumeric form, but perhaps also images, sound, models, etc..) brings me to D2K, and a project just funded by Mellon to build web-based text-mining and visualization tools for use with humanities digital libraries. That's a terrible thorny mouthful, and I'm sure it would send the board of Critical Inquiry scrambling for their bumbershoots, but I don't know a more direct or concise way of saying it. I'm not sure that "patacriticism" is less thorny, for example. Maybe "the figure in the carpet"? Finding patterns is the point—finding them, exploring them, explaining them. D2K stands for Data 2 Knowledge, and it's the name of a data-mining environment built by Michael Welge and his Automated Learning Group at NCSA. D2K is a general-purpose data-mining architecture, within which you can build specific data-mining applications, using a visual programming interface. You can use existing modules or create new ones, arrange those modules into what's called an itinerary, and then run data through that itinerary. Outputs can be used to create visualizations or fed to some other process. It's worth noting, in passing, that the itinerary answers quite precisely to one of the goals articulated in the Bradley and Rockwell essay from 1992, namely

> the need in computer-based text analysis to not only have environments that support the flexible linking of tools to meet immediate research demands, but to have the computer support the accurate preservation of these links, so that they can be fully and accurately evaluated and reused by others.[2]

[new slide: T2K] T2K D2K is all done in Java, and it actually works—and it wasn't designed with humanities content or humanities computing in mind. There is a text-mining adaptation of D2K, called T2K, but it knows very little about XML or structured text (in fact, so far, text-mining based on D2K has focused on unstructured text). My colleague at UIUC, Stephen Downie, has been using D2K for musical data, classifying music based on its properties, as part of a project on music-information retrieval. In the Mellon project, we'll be building a data store and D2K modules that leverage XML markup, designing visualizations for the output of D2k, and trying to use those visualizations as interfaces, to enable an iterative process of data-exploration.

[new slide: Shakespeare] Over the summer, as part of preparing the grant proposal, Steve Ramsay and a graduate student at the graduate school of library and information science, Bei Yu, used the insights from StageGraph as a basis for an experiment with D2K, in this case to see if structural characteristics of Shakespeare's plays, expressed in markup, could be used to classify those plays as tragedy, comedy, romance, or history. What you see before you is the result: it's a static visualization of the distribution of properties across texts. What's **un**interesting about this is that the software got it mostly right: it grouped the plays largely as they are traditionally grouped by scholars. What's **interesting** about the results is the one place where the software appears to have gotten it wrong: it puts Othello smack in the middle of the comedies. This struck Bei Yu as a disappointing failure; it struck Steve Ramsay (and me) as the most interesting and potentially valuable outcome of the experiment. As Jerry suggested this morning, it's not certainty but uncertainty, and surprise, that's a sign of value, of discovery, in the humanities.

There's one other thing I want to mention, in the context of tools (and what else can we call them?) that aim at visualization of the statistical qualities of texts in the aggregate. That point arises in part from a realization that came up early in the Horribly Named ACLS Commission on Cyberinfrastructure for Humanities and Social Sciences, namely that **the** primary resource constraint for the humanities, at least in an age of the digital representation of artifacts, is the current copyright regime.

According to "How Much Information? 2003" (by Peter Lyman & Hal Varian), our culture produces

- More than 300TB of print material each year (books, journals, newspapers, magazines)
- More than 25TB of Movies each year
- About 375,000TB of Photography each year
- About 987TB of Radio each year
- About 8,000TB of TV each year
- More than 58TB of Audio CDs each year

...and that doesn't include software (say, video games), or materials originally published on the Web. And almost all of that material, including the books, is born digital.

With all that digital cultural material swirling around us, why don't the humanities desperately cry out for computational methods, for big iron, for parallel processing? Because most of that material is inaccessible to scholarship based on representation of the artifact: Each of these books, journals, movies, TV shows, etc. will be copyrighted for the life of the author plus 70 years—at least. How will the record of the digital present be copied, even if only to be preserved? How will the history of the future be written? How will the culture of the future sample the culture of the past? If recent changes in the law indicate a trend, then copyright will extend effectively forever. On the other hand, if we had access to all this stuff (and not just from 2003, but, say, from the second half of the 20th century), then we'd need all the computational power and tools and methods we could get our hands on.

The reason I bring all this up is that tools which deal with texts in the aggregate, and produce as output visualizations of the properties of texts in that aggregation, provide a way around the copyright constraint. If we don't have to republish work in order to do digital humanities, perhaps we can get at a greater portion of that record, and perhaps computational methods will then come to seem useful, even necessary, and maybe tools won't be a dirty word.

[new slide: Ivanhoe] Ivanhoe But I agree that, at present, humanists don't like to think of themselves as using tools. I'm not sure they like to think of themselves as playing, either—I fear that many humanities scholars actually do believe that they are discovering things, and their discoveries should be taken seriously. I don't know whether the allopoetic will give way to the autopoetic, in our understanding of the goals and objects of humanities scholarship: if it does, then Ivanhoe will point the way to a generation of tools called GAMES. Game is a broad category, a category that could and does include simulation, reenactment, exploration, role-playing, detection and discovery, murder and mayhem, nurture, nonsense, and even sometimes sheer abstraction. Perhaps instead of saying to our colleagues "Let's use this web-based text-mining and visualization tool" we will one day say "would you like to play a game of text-exploration?"

In that ludic future, perhaps our tools will become our texts. As I look at the Ivanhoe interface, I think I hear it crying out for exegesis. What view of text and reading does it represent? What is the secret sin of the login window? What does it allow us to do, and what does it forbid? Where do its internal constraints, its natural laws, originate, and what does that origin say about the bias and the predispositions of its creator?

Enough: We've spent a generation furiously building digital libraries, and I'm sure that we'll now be building tools to use in those libraries, equally furiously, for at least another generation, and I look forward to it. I'm sure that the text won't go away while we do our tool-building—but I'm also certain that our tools will put us into new relationships with our texts. All we can really ask, in the end, is that those relationships be fruitful.

Endnotes

[1]: Tasman, Paul. 1958. *Indexing the Dead Sea Scrolls by Electronic Data Processing.* New York: IBM Corporation, 12; cited in Formatting The Word of God, An Exhibition at Bridwell Library Perkins School of Theology Southern Methodist University, October 1998 through January 1999. Ed. Valerie R. Hotchkiss and Charles C. Ryrie. Bridwell Library: Dallas, Texas, 1998.

[2]: "Towards new Research Tools in Computer-Assisted Text Analysis," John Bradley and Geoffrey Rockwell. Presented at The Canadian Learned Societies Conference, June, 1992.

[3]: Claire Colebrook, "Questioning Representation," *SubStance* 29.2 (2000). 49.