DHQ: Digital Humanities Quarterly
2014
Volume 8 Number 3

Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment

Martin Mueller <martinmueller_at_northwestern_dot_edu>, Northwestern University

Abstract

This is the text of the Hilda Hulme Memorial Lecture given at the Institute for English Studies at the University of London in July 2013.

Introduction

[1]

1

Hilda Hulme was a philologist working in the border area between linguistics and literature at a time when work in this field produced some of the finest work around. Since then Linguistics and Literary Studies have parted company, with gains and losses on both sides. The digital turn offers the promise of productive rapprochements, and part of my talk will be about that promise.

2

To judge from her book *Explorations in Shakespeare's Language* Hilda Hulme was a gifted miniaturist, practicing, in her own words a "disciplined imagination" that combined "the sober ant-like industry of the professional scholar" with "the grasshopper swiftness of the crossword puzzle addict (as well as the relaxed alertness of the confidence trickster)." She admired earlier generations of scholars who (in Malone's words) "carefully examined" the "meanest books" because they "might happily throw light on some forgotten custom, or obsolete phraseology." But she also took those critics to task for not looking closely enough. Scholarly ant and grasshopper, she shuttled between text and context, Shakespeare and the language of his time, claiming that for about two hundred words she had discovered "new elements of meaning," while for three dozen words she argued that "the original text has more meaning than the emendations currently accepted." "On the negative side," she asserted that "I can confidently make two claims: I have invented no new linguistic terminology and have put forward no new Shakespearean emendations."

3

Explorations in Shakespeare's Language is a work of modesty and cunning, well reviewed at the time and well worth studying half a century later. The "context" for Hulme's microanalysis of Shakespearean textual cruxes was the Oxford English dictionary, her wide reading in local and regional records, both printed and in manuscript form, and encyclopedic works like Thomas Cooper's *Thesaurus*, the most ambitious Latin dictionary of the late 16th century. The evidence is often very sparse, as Hulme reminds us at every turn. The sparser the evidence, the more ingenious its parsing, with the strengths and weaknesses that come from this situation.

The allographic journey of texts, the query potential of the digital surrogate, and scalable reading

4

Let me turn from Hilda Hulme's work to my topic "Shakespeare His Contemporaries," which you note implies a similar shuttling between text and context. I will talk about the ways in which digital resources, tools, and methods have greatly increased the range, granularity, speed, and accuracy of such shuttling. I will tell you about how I came to the "digital turn," and I will use three mouthfuls of terms, the "allographic journey of texts," the "query potential of the digital surrogate," and "scalable reading." I will also say a little about DATA, my acronym for digitally assisted text analysis.

"Assisted" is an essential ingredient in my understanding of that term.

Nelson Goodman, in his *Languages of Art*, uses the opposition of "autographic" and "allographic" works to distinguish between physical artifacts that are uniquely embodied (autographic) and symbolic artifacts that can be embodied in different notational systems (allographic). There are "originals" of the Mona Lisa and of Michelangelo's David, but the autograph of Bach's *Art of Fugue* is not the "original" of that work, although it is the "original" of that manuscript considered as a physical artifact. 5

Texts are allographic objects par excellence. We like notations that we are used to, and we like to think that there is a close relationship between a work and its original system of notation. Early Modern scholars often are attached to "original" spellings, just as some music lovers prefer performances on "original" instruments. But what we hear in such a performance is something very different from what Bach's contemporaries would have heard. They would have heard the ordinary sound of the strings and winds of their day. We hear the difference from the ordinary instruments of our day. The difference may be illuminating in various ways, but it is a delusion to think that what we hear is closer to some putative original. Similarly, texts are not defined by the ways in which they are spelled or by the paper on which they are written. 6

Consider the texts of ancient Greek. The familiar notation we encounter in the pages of Teubner, Budé, or Oxford Classical has a lot to do with Porson's handwriting, but would have been unintelligible to Aeschylus or Plato. For several decades, betacode was the most common form of writing and reading Greek on a computer. Betacode used the symbols available on a standard English IBM keyboard of the fifties to represent accented Greek through a combination of Roman Capital letters, parentheses, forward, and back slashes – very ugly to look at and detested by many classicists as barbarian technology. But to judge from inscriptions on Greek vases from the sixth and fifth centuries, betacode stripped of its accents and breathing markers is arguably closer to what Homer's first readers would have seen. 7

The *Iliad* may or may not have existed in anything like its current form before it was written down, but since then it has been on an allographic journey with many steps. The young Plato would have read it in the Old Attic alphabet, which is like Latin in not distinguishing between short and long vowels. The *metagrammatismos* of the fourth century distinguished between long and short forms of "e" and "o." Accents and breathing marks would have been unknown to Aristotle. Venetus A, the most important of the *Iliad* manuscripts stands on our side of the transitions from scroll to codex, scriptura continua to space between words, and the division of letters into upper and lower case. Villoison's 18th-century edition of Venetus A is on our side of the print revolution. Not many years from now readers of the *Iliad* in Greek are likely to encounter it first in a digital format on some mobile device. 8

Each of the steps in this allographic journey has its own "affordances" and gives to a text a particular query potential. The transition from scroll to codex is particularly interesting. A codex is a random access device, greatly speeding up forms of non-contiguous reading. It encourages the creation of metadata in the form of tables of content, indexes, and the like. From a conceptual perspective, the promise of such metadata is clearly envisaged in the Biblical concordance invented by 13th-century monks, but the realization of its full potential is the work of 16th-century publishers and scholars, to whom we owe the conversion of canonical texts into aggregates of small citable chunks that serve as the "hyperlinks" of a print-based structure of external look-up tools. The chapter-and-versification of the Bible by Estienne is the most radical example. The same publisher segmented Plato's texts into chunks of approximately 100 words each, identified by the "Stephanus numbers" that have for four centuries served as the universal identifiers of Platonic passages. 9

So much for the query potential of the printed text. What about the query potential of the digital surrogate? We use the term "digital surrogate" to describe a digital version of a text that originated in the world of print or manuscripts. The digital surrogate differs from a "born-digital" document. But is a digital *Iliad* any more of a surrogate than Venetus A or Villoison's edition of it? Or does it simply represent another step in the allographic journey of texts, a big step to be sure, but not a step that is usefully described as a descent from the more to the less real? 10

I have been a Homer scholar for much of my life, and my interest in digitally assisted text analysis has its origin in my 11

trying to get a better handle on the problem of repetitive or formulaic language in the *Iliad* and *Odyssey*. My take on Homer was deeply shaped by Karl Reinhardt's essay of 1937 about the judgment of Paris. While conceding that the poems are rooted in an oral tradition, I never believed that they were themselves "oral" (whatever that means), and I was always attracted to some form of the hypothesis that their distinctive features were the result of a productive collision between two language technologies, orality and writing. I also thought that repetitions came in quite different forms and for a while tried to keep track of different kinds and their occurrences on the 3x5 index cards of an earlier world. Then I discovered the power of relational databases. If you were a business man and wanted to know which of your customers owned properties in both Colorado and Connecticut, that was the kind of question "Oracle" excelled at answering with great speed and accuracy. The philological cousin of that question is very hard to answer in a print-based environment. If you come across a repetition, you can look up its other occurrences. 19th-century commentaries like Ameis-Hentze are very good at keeping track of occurrences and letting you move from one to the other. But the printed commentary offers little help in finding the locations of all repeated strings that occur in *Iliad* 1 and *Odyssey* 8 but nowhere else. By contrast, the relational oracle (whether Oracle, SQLServer, MySQL, SQlite) offers a lot of fast and accurate help, if (and that is not a small "if") you give it the data in a form it can handle.

12 With the help of two programmers I built a list of all strings of two or more words that occurred at least twice in Early Greek epic, all 36,000 of them with their 192,000 occurrences. Each item in that list consisted of a string, its length, and its citation. From these three data points, as well as counts and summaries derived from them, I was able to construct a simple taxonomy of repetitions, which I classified in terms of how they answered to the question "Who is speaking?"

13 There is a small but important body of "explicit" repetitions in which a Homeric character orders another character to deliver a message to a third person, and to do so verbatim. These passages are important because they prove that the concept of deliberate and literal repetition was familiar to the poets and their audience. At the other end of the spectrum are the many strings of words without which it is hard to say anything in any language, things like "I will not," "of the," and the like. You might say that the language is the "speaker" of such phrases. Between these end points of the spectrum there is a large body of phrases that are a genre specific version of what the linguist John Sinclair called the "idiom principle." Here we are in the world of "rosy-fingered dawn," "swift-footed Achilles" and the "wine-dark sea." They are the "formulae" that Homeric epic is famous for. They are rarely longer than three words, although there are some whole-line formulae, notably in lines that introduce or follow a speech. Genre speaks in those phrases.

14 Finally, there is a body of phrases that vary in length from two words to twenty lines and occur twice or at most four times. There are about 14,000 of them with a total of 34,000 occurrences. These are the cases for which a model of predominantly oral production has no good explanation. When you have a problem it is helpful to know how much of it there is and how it is distributed. For instance, even without looking at the text of phrases shared between books, it is noteworthy that there are a lot of phrases shared between the first and last books of the *Iliad* and a lot of phrases shared between those books and the *Odyssey*.

15 The size of the corpus of Greek epic is only about a third of the Shakespeare corpus. Many of what you might call "problem phrases" may well be idioms that just happen not to occur a lot in what is quite a small corpus. With other phrases — and especially longer phrases — it is more plausible to argue that the poet speaks in the sense that the author of one phrase unconsciously remembers a phrase from another context or deliberately alludes to it. I have called such phrases "interdependent repetitions." In the absence of external evidence it is typically difficult to determine whether A copied B or B copied A. But the evidence for a model copy relationship is often very strong, and if you accept the cumulative evidence of such phrases, you must postulate the existence of a fairly stable text, because without it you cannot have context-specific allusions.

16 Such was my entrée into the world of DATA or digitally assisted text analysis. Having access to the Homeric poems in the latest stage of their allographic journey from the pre-literate or recently literate to the digital, I made use of the query potential of the digital surrogate to re-examine aspects of the Homeric Question, for the Homeric Question is fundamentally the question of how to account for repetitions in the *Iliad* and *Odyssey*. This work was done in the nineties and I was aware of the irony of applying the quite recent business technology of relational databases to a set of literary phenomena that probably had something to do with "Homer," whoever he or they were, applying the recent

business technology of writing, adapted from Phoenician traders, to the genre of orally produced epic narrative and transforming it in the process.

17 My work on Homer is available on a website called the Chicago Homer. For modern readers with or without Greek, the digital text of that site with its associated list of repetitions has some affordances that cannot be matched in print. Whether the *Iliad* was literate, oral, or a hybrid of both, its reception for at least the first five hundred years of its life was predominantly "aural." Greeks knew the *Iliad* as Italians know *Rigoletto*. Repetitions, then, are something the expert listener hears. In the Chicago Homer the novice reader of the *Iliad* can "see" repetitions because they are embedded as links that can be displayed or hidden at the reader's discretion. Repeated phrases can also be filtered by length or count. Thus link technology can be used to help readers "navigate the neural pathways of bardic memory" as I once called it.

18 Let me draw some general conclusions from my work with Homer before turning to Shakespeare. My conclusions were encouraging in some ways but disappointing in others. Disappointing because what the database told me about the distribution and density of Homeric repetitions closely mirrored the observations of generations of 19th-century scholars. Why bother with a tool that only tells you what is already known? But you can be encouraged by that very fact. An IBM manager once observed that computers are dumb but fast, while humans are smart but slow. Considered as a database, the quarter million words of Greek epic are very small fry. Quite a few 19th-century scholars knew all or much of Homer by heart. So you can think of the collective memory of those scholars as a very smart, if slow, computer that performed well when it worked on a problem within its range and was given enough time. If an algorithmically produced and organized inventory of repetitions "merely" replicates the findings of that "human computer," you may turn things around and say that this agreement confirms the utility of the machine and increases your trust in its performance when it scales up to tasks that are clearly beyond the scope of human memory.[2]

19 Many people think that the main point of digitizing a text is to deliver readable stuff to the screen of some digital device. More books for more readers is certainly a noble goal. If you look at the world in 1983 and 2013 and compare the amount of readable texts freely available then and freely available now to anybody who is eager to read them and has access to the Internet, the progress is remarkable. But beyond the question of extending there is the question of enhancing access and exploring more fully the query potential of the digital surrogate, which I should probably stop calling a surrogate. There is a lot more that you can do with a digital text than read it, but what you can do with it depends on what you do to it, and what you do to it is a matter of making the text or a whole set of texts in a corpus "machine actionable." To return to Hilda Hulme's lovely phrase, a text that is "machine actionable" in the right way can do a great deal to make "the sober ant-like industry of the professional scholar" more productive and its inevitable tedium more bearable.

20 The fully machine actionable text is a data structure in which the text as a sequence of words is supplemented by an inventory of its parts that are classified in various ways. These supplementary data, often called metadata, can be retrieved by those classifications and in a "just in time" manner. In the Chicago Homer, a phrase in a particular place can lead you instantly to all its other occurrences. The procedure is a very basic philologic look-up: you go from a place in the text to a list that tells you more about the phenomenon in that place, whether it is a word, a phrase, or a grammatical condition. But in a well-designed machine actionable environment the time cost of look-ups drops by orders of magnitude, and this drop changes the calculus of what is quite literally "worthwhile."

Shakespeare His Contemporaries

21 I have left Homer and for the past several years have focused on the question of how to promote corpus-wide inquiries into Early Modern Drama by leveraging the query potential of textual data in digital form. There are some 800 plays or play-like texts printed before 1660 — the world of "Shakespeare His Contemporaries" generously construed. They amount to at least 15 million words, 60 times the size of Homer, 20 times the size of Shakespeare, and well beyond the scope of individual memory. Bibliographical and other records about them have been exhaustively studied by generations of scholars. Most of the text themselves are available in the digital transcriptions of the EEBO-TCP project. Students of Early Modern Drama live in a data-richer world than students of Homer. They have access to the TCP

transcriptions as well as to digital facsimiles of the printed originals, which are rarely excellent, sometimes very bad, but often good enough. The quality of the digital transcriptions is largely a variable of the facsimile the transcribers had in front of them.[3]

22 Compared with a hundred or fifty years ago, there are a lot more literary scholars today whose career depends on publishing something that will get them promoted. The MLA bibliography for 2012 contains ~64,000 entries compared with some 15,000 entries in 1962. We also live in an age that has been suspicious of the highly canonical, celebrates the margin, and likes to decenter or open up things. You would think, then, that both in absolute and relative terms, non-Shakespearean plays would receive a lot more attention than they have in the past. But this is not the case. Instead we have more scholars chasing the same three dozen Shakespearean and at best two dozen non-Shakespearean plays. Consider *Titus Andronicus* and *Lust's Dominion*, both of them plays about a lascivious queen and an ambitious Moor. There are 2915 references to Shakespeare's play in JSTOR, compared with 133 for *Lust's Dominion*. And *Lust's Dominion* does relatively well because it is about a hot topic. If you compile the bibliography for a minor Elizabethan play, you will often find that it is very short and that the more substantive treatments are quite old. The highly canonical is alive and well or, to put it a little more cynically, scholars like to attach themselves to celebrities, celebrities are "positional goods," and you can't have more than a few dozen at a time. You can see why professional scholars stay away from minor plays, unless they explicitly deal with hot topics. A play may interest them, but how will an entry about it look on a c.v.?

23 But what if you made the object of inquiry not this or that play, but the corpus as a whole and you presented the corpus in such a way that it fully leverages the query potential of a digital corpus? It is a corpus of ~15 million words. It includes just about all dramatic texts that have survived, and from what we know about lost plays, it is a substantial portion of what was ever printed. And what if you set out to recruit a new group of young explorers, who either need not worry about their c.v. or whose c.v. would be improved by respectable work on minor Early modern plays? How about undergraduates? Almost forty years ago I served on a search committee for a position in the Italian department while at the same serving as second reader for an honors thesis written by a very gifted undergraduate in that department, who went on to law school. Her written work towered above that of the job candidates, and ever since then I have believed that bright and ambitious undergraduates can make substantial scholarly contributions.

24 How do you leverage the query potential of a digital corpus of plays? Some years ago I was part of the MONK Project, which took some steps towards creating a "digital environment designed to help humanities scholars discover and analyze patterns in the texts they study." MONK was an acronym for "metadata offer new knowledge." Metadata are data about data. If you know about library catalogs or email headers you know about metadata.

25 The diabolical side of metadata has been much in the news lately. Here is a quotation from the *New York Times* (June 9):

> American laws and American policy view the content of communications as the most private and the most valuable, but that is backwards today, said Marc Rotenberg, the executive director of the Electronic Privacy Information Center http://epic.org/, a Washington group. The information associated with communications today is often more significant than the communications itself, and the people who do the data mining know that.

26 The metadata mind turns from the content of an individual text to a network of often very abstract relationships between properties of many texts, and these properties may not be limited to what is traditionally considered to be "in" the text.

27 Drama is a particularly productive target for some metadata-oriented inquiries, because a quite explicit and rigid system of metadata is part of the genre itself. Plays are divided into speeches, scenes, and acts. They have stage directions and speaker labels. The metadata status of these things is nicely highlighted by the use of Latin: *Actus primus, dramatis personae, exeunt omnes, moritur*. The dramatis personae or cast list often include descriptions like "Gonzalo, a humorous old lord," which is an informal classification by sex, age, social status, and temperament. A well-structured cast list (not all of them are) is a little prosopography. Can you construct from individual cast lists a machine actionable prosopography of the genre? The power of such a thing would go considerably beyond Thomas Berger's very helpful

Index of Characters in Early Modern English Drama, Printed Plays, 1500-1660. You would have something like a census of ~10,000 dramatic characters who speak more than a sentence or two and their interactions over a period of 150 years.[4]

28 You learn a lot about a large corpus if each of its analyzable items is consistently described in terms of very few and quite simple data points. The power of business and political intelligence derives from the patient pursuit of connecting a few dots across millions or billions of items. I do not want to claim that in applying such techniques to a corpus of plays we will "pluck out the heart of their mystery." For that I am too much of a student of Karl Reinhardt and his wonderful philological credo of 1946, which I now like to read as a prophetic warning about the opportunities and limits of the digital turn:

> Philology becomes more questionable, the less it can let go of itself. This does not mean that it loses its confidence within its own domain. Nor does it mean that it excludes itself from the expansion of its borders that developments in the humanities have opened to it. But it is part of philological awareness that one deals with phenomena that transcend it. How can one even try to approach the heart of a poem with philological interpretation? And yet, philological interpretation can protect you from errors of the heart. This is not a matter of the *ars nesciendi* that Gottfried Hermann insisted on in his day. There it was a matter of things you could not know because of accidents, the circumstances of transmissions, or of inadequate acuity. Here it is a matter of something that is necessarily beyond reach but our awareness of it effects our way of reaching towards it. It is not a matter of the ultimate *Ignoramus* but of a methodological modesty that is aware of something that must be left unsaid and that with all perceptiveness or intuition you cannot and should not trespass on.

29 But Reinhardt's methodological modesty should not keep us from learning the many things that can be learned from what Michael Oakeshott in another context has called mapping operations. Because its metadata are so explicit and consistent across many texts, drama is a genre that lends itself to digital mapping and offers quite precise directions about how to do it.

30 The TCP texts of Early Modern drama are encoded in TEI-XML. Every speech by a character is captured in a container or "element" that has a speaker label. If each speaker label maps consistently to the appropriate character in the cast list you can create an abstract model of the text in which for a while you ignore what is said but focus instead on who speaks at what length in the presence of what other characters. You see the play as a network of communicative acts among named individuals. If your cast lists contain classifications of characters by sex, age, and social status, a play becomes a network of communications among characters of a certain kind. If you have data of this kind for all or most Early Modern plays, you can use those abstract models as the basis for analysis by genre, period, or author. A lot of labour is buried in this succession of "ifs," but the pay-off can be considerable. Over the course of the coming year I hope to turn at least some of these "ifs" into readily available data.

31 Franco Moretti has called such procedures "distant reading," a term that poses a polemical challenge to the time-hallowed practice of "close reading." I prefer the term "scalable reading." Digital tools and methods certainly let you zoom out, but they also let you zoom in, and their most distinctive power resides precisely in the ease with which you can change your perspective from a bird's eye view to close-up analysis. Often it is a detail seen from afar that motivates a closer look.[5]

32 "Scalable reading" of this kind is at the heart of new forms of intertextual analysis opened up by digital tools that let you discover shared verbal patterns across large data sets. These tools are critical to genomic research and to what you might call forms of forensic philology where you focus on plagiarism detection or authorship attribution. But their true power for Literary Studies lies in their potential for enhancing our understanding of the shared verbal patterns that make up a genre or tradition. In this regard Homer's formulae are only a very special and marked case of a universal rule: to understand a genre is to understand its ways of repeating and varying.

In an earlier experiment I extracted a database of repeated phrases (trigrams or longer) from 320 Early Modern plays, using the same criteria as in the construction of the database of Homeric repetition. My findings are quite provisional, but I am encouraged to apply this procedure to a larger corpus of 600 plays or more as soon as they have been cleaned up sufficiently to remove unwanted "noise" from the data. With 320 plays you have 51,040 pairwise combinations of works. You count and measure the repeated n-grams shared between any such combination, give each of them a weighted value and define the degree of sharing as the sum of that value. With 600 plays you have almost 200,000 pairwise combinations, but the numbers are manageable: they are inputs for statistical summaries, and the larger numbers give you more confidence about the status of outliers that reward closer analysis. 33

My admittedly crude procedure arranges some 50,000 “shared rep values” on a continuum that ranges from hardly anything to quite a lot. My confidence in the procedure is strengthened by the fact that it accurately measures things you would expect. The overall degree of repetivity in Early Modern drama is much less than in Homer, but the distribution is quite similar if you accept a play as an equivalent of a Homeric book. The median “shared rep value” for repetitions within a play exceeds the median value between plays by a factor of 20, from which unsurprising fact you draw the useful conclusion that phrasal repetition between plays is actually quite rare.[6] 34

It is not surprising that plays by the same author share on average twice as many n-grams as plays by different authors. It is of some interest to note that the factor of authorship (~2) is much larger than the factors of genre or closeness in time (~1.2). Measurements of this kind hardly pluck the heart of mystery out of the genre, but they map out the territory in useful ways and give you a sense of threshold values. If you come across a pair of plays considered to be by different authors and you observe that their shared rep value lies two standard deviations above the average for same-author plays it is worth asking whether something is going on here. If one of these phrases strikes you as particularly “salient” the argument from salience is not answered by the objection that one swallow does not make a summer. 35

I see a bright future for Literary Studies in a world where well-curated corpora become objects of inquiry for scholars who have no qualms about putting simple but effective statistical routines in the service of the traditional virtues of philological tact and acumen. Last week I began a two-month project in which five undergraduates and I will work on collaborative curation of the EMD corpus and lay the groundwork both for a frequency-based lexicon of words and phrases that will support close intertextual analysis. I will apply to the 15 million words of Early Modern Drama the same techniques I applied to the 250,000 words of Early Greek epic. If things go well the resulting database will be in place by the end of this year. And I hope to take some steps towards a census of *dramatis personae* by age, sex, and social status. 36

This database will let the reader of a particular play, whether novice or expert, navigate its allusive pathways in the web of intertextual parallels, and get some initial measure of its closeness or distance from other plays. The 19th and early 20th century scholars who created much of the documentary infrastructure for the study of Early Modern texts had a deeply nuanced understanding of their materials, acquired over decades of continuing contact with them. Today's young scholars do not have this, and it would be foolish to promise that my database will give them similar mastery in days or weeks. But it will cut down the time cost of acquiring an initial sense of the lay of the land. And if they catch on to the spirit of corpus-wide inquiry and keep in mind Karl Reinhardt's warning about the need for methodological modesty, they may sometimes see details and patterns that Chambers, Greg, or Bentley could not see. The giants of those days relied on the Oxford English Dictionary and the texts they could touch with their hands. Today's readers, whether amateurs or professionals, cannot only look at the entire corpus of Early Modern Drama, but for any philological question arising from those texts they can also go to the corpus of EEBO texts before 1700, which now includes 45,000 titles and will grow to about 70,000 by 2015. Simple queries to that corpus can be answered in seconds. 37

If the texts of Early Modern Drama are made available to bright undergraduates in machine actionable forms that fully leverage the query potential of such texts for corpus-wide inquiry, and if several generations of such bright undergraduates do honours theses or other independent study projects that are based on corpus-wide inquiry, I am confident that they will find many interesting things to say about the language, forms, and themes as well as the historical and social contexts of Early Modern Drama. But their digital tools and resources will make no difference to the fact that their work will require the “disciplined imagination” that combines “the sober ant-like industry of the professional 38

scholar" with "the grasshopper swiftness of the crossword puzzle addict." You can say of their work what Greg Crane said in his recent manifesto about the Open Philology Project: "This is not a digital philology or digital humanities project. [It] is about philology."

Postscript

39 The five Northwestern undergraduates referred to above – with one exception freshmen or sophomores at the time – were Nayoon Ahn, Hannah Bredar, Madeline Burg, Nicole Sheriko, and Melina Yeh. Over the course of eight weeks in the summer of 2013 they demonstrated remarkable persistence and ingenuity in working their way through 493 non-Shakespearean plays written between 1576 and 1642. The Text Creation Partnership (TCP) source texts contained 38,880 incompletely transcribed words, of which they corrected 32,300, using AnnoLex, aWeb-based collaborative curation tool designed by Craig Berry. The remaining 6,500 known errors include many difficult passages, but they add up to relatively "light work" for the "many hands" that can and should have a go at them, anywhere and anytime from http://annolex.at.northwestern.edu.

40 Since giving this talk, I have made progress with two other aspects of turning the texts into a machine-actionable corpus. Using Berger's *Index of Early Modern English Drama* I have mapped over 90% of some 300,000 speaker labels to unique corpus-wide ID's that can serve as the values for the "who attributes" of <sp> elements in the TEI encoding of the source texts. Phil Burns' MorphAdorner was an essential tool for that task, and it was equally important for mapping the original spellings of the texts to standardized forms. This mapping creates a virtual corpus in which the plays of Shakespeare and his contemporaries wear the same orthographical uniform. Orthographical or typographical "accidentals" provide crucial evidence for some inquiries, but for other inquiries they stand in the way and make different texts look more different than they are, especially for novices. The flexibility of the digital surrogate lets you filter out or focus on evidence, depending on your purpose.

41 Curation and exploration are two sides of the coin of working with textual data in digital form. Right now, Shakespeare His Contemporaries is mainly a data project in a predominantly curatorial phase. With a little luck by the time the quatrocentenary of Shakespeare's death rolls around in 2016, a user-friendly version of it will up and running in an open access environment that will cut down on the tedium of some older forms of exploration and enable new forms of exploration that previously were impracticable.

Notes

[1] The following is the text of the Hilda Hulme Memorial Lecture, which I gave on July 2, 2013 at the University of London's Institute of English Studies.

[2] Within the domain of Homeric scholarship, there is a narrower point. Proving that the other guy is almost certainly wrong is an important aspect of scholarly progress. And the machine-generated data made it very clear that the more radical claims about the traditional and formulaic quality of Homeric verse rest on a wilful neglect of critical evidence.

[3] It is well within the parameters of current technology to design an environment for collaborative and user-driven curation of these texts so that readers of this or that text can suggest corrections as they go along – a digital version of the printer's plea in the errata section of Harding's *Sicily and Naples*:

> Reader. Before thou proceed'ſt farther, mend with thy pen theſe few eſcapes of the preſſe: The delight & pleaſure I dare promiſe thee to finde in the whole, will largely make amends for thy paines in correcting ſome two or three ſyllables.

This has been a topic dear to my heart, but I will not say any more about it here except to repeat it is entirely possible to build environments with appropriate editorial supervision that allow for user-driven and incremental improvement of texts over time. Educated and interested readers with no special scholarly training will be able to spot and correct many errors in any environment that allows for easy alignment of the page image with the transcribed text.

[4] Imagine for a moment a crowdsourcing project in which readers were encouraged to write "tweets" of 140 characters each about characters and typical scenes or conflicts. You might end up with a useful "folksonomy" of "dramemes" and their distribution over time. With some skilful editing, these tweets could become a data structure in which some aspects of each play are catalogued in a reasonably consistent fashion.

[5] Shakespeare used the word “fate” 62 times in 27 plays. This seems to be a thoroughly unremarkable distribution of a word that is neither common nor rare. But if you look at the distribution of the word in the context of 320 plays (including Shakespeare's) you discover that “fate” is comparatively speaking the most underused noun in Shakespeare. A very simple statistical test shows that with regard to “fate,” a term central to the rhetoric of tragedy, Shakespeare is a negative outlier. The statistical oddity prompts a closer look. But it will take a very close look to figure out whether anything of interest is going on there.

[6] Just as in Homer, repetition rapidly declines with distance. Compared with a model of random distribution, a very high percentage of n-grams are separated by just a few hundred words.