Opening up Classics and the Humanities: Computation, the Homer Multitext Project and Citizen Science

http://tinyurl.com/okg6uqc
Gregory Crane
University of Leipzig, Department of Computer Science
Tufts University, Department of Classics
September 2014

Abstract: Increasingly powerful computational methods are important for humanists not simply because they make it possible to ask new research questions but especially because computation makes it both possible -- and arguably essential -- to transform the relationship between humanities research and society, opening up a range of possibilities for student contributions and citizen science. To illustrate this point, this paper looks at the transformative work conducted by the Homer Multitext Project (http://homermultitext.blogspot.de/; http://www.homermultitext.org/).

My departure point for this paper was a set of questions during a workshop on Computational Humanities, at Schloss Dagstuhl in July 2014 that Manfred Thaller posed to me in response after a talk I delivered, and then as part of a podium discussion. I had argued that we needed to advance research projects that provided our students with an opportunity to make substantive contributions to research as a central part of their education. In so doing, I echoed Wilhelm von Humboldt, who argued that students in a university should always be engaged in advancing human understanding -- mastering set curriculum was for primary and secondary school. For Humboldt, the challenge of new questions and interests from students was one of the great intellectual advantages of working in a university rather than in a research institute. Manfred reminded me that if we focused too much upon serving the students, then we would do a disservice to research.

The challenge is to establish a productive tension between the established interests of the faculty and the fresh questions of the students. I conclude this paper by describing the Homer Multitext Project and how, in opening up new opportunities for student contributions and research, this project shifted my idea of how the field should move forward. Research should not solely serve students (then it runs the risk of being dumbed down) but it should not only serve specialist researchers (then it runs the risk of becoming detached scholasticism). In the ideal case, we identify research that challenges (and captivates) the most advanced researchers but that also engages a broader audience and provides multiple opportunities for contribution. In the study of Greek and Latin, I see this productive tension leading to re-organization (and, in the US at least, a revival) of very traditional philological questions with very new methods and in a much more decentralized and collaborative culture.

This paper thus argues that the most important consequence of computation for the Humanities lies not in the new research questions that are now appearing but in the fact that research with

increasingly large bodies of digitized source materials opens up opportunities, indeed the necessity, for a new, more open culture of intellectual production, one that is less hierarchical, more focused on collaborative inquiry, more dynamic, and, in my view, more effective as an environment for broad and deep intellectual development. Computation allows -- perhaps more accurately, challenges -- humanists to redefine their relationship to their students and to society as a whole.

In my view, the rise of computation -- and, for humanists, computation over large data sets -- makes possible this emerging shift in academic culture. Computation unleashes two complementary changes, one affecting the kinds of research that specialists pursue and the other, potentially of far greater impact, affecting the culture of intellectual production and the relationship between what specialists do and the intellectual life of society as a whole. The first is essentially incunabular in nature: just as Gutenberg and other early producers of incunabula (as early printed books are called) sought to produce more efficiently with print technology the same kinds of documents that circulated in manuscript form, the first generation of humanists have commonly measured the importance of digital technology by the impact that this technology had upon the specialist articles and books that they could produce for their professional colleagues.

The real significance of digital methods lies in their potential to make humanities research fundamentally more accessible, first across traditional disciplines (and thus energizing disiplinary research) but also, and more tellingly, across boundaries not only of space and time but also of language and culture. This in turn raises the need for research on how to make intellectually accessible primary sources that are increasingly available in digital form. Where professionalized scholarship was written by experts for experts, we need scholarship that enables as many people as possible to approach (and in some ways go beyond) the capacity of experts when they encounter a digitized source, in whatever medium and of whatever kind. Such work, of course, advances specialist research (e.g., the historian of science who can now make effective use of materials in Arabic as well as in Greek and Latin) but it also can also can make the human record increasingly increasingly accessible to a broader, indeed, to a global audience. As barriers shift, the culture of humanities research will shift as well -- whether by engaging the potential for new forms of dialogue or by trying to maintain boundaries between closed scholarly networks and a wider community.

First, consider the effects that computation has already begun to exert on the scale of research at both the macro- and micro-levels. The initial point is not whether or not most, or even many, humanities researchers have felt these effects. The first question is whether we can see some scholarly activity that does take advantage of the potential changes in scale.

One change on the macro scale is well known: we have millions of documents available online for at least limited analysis and we have begun to use, for example, data and text mining tools to visualize patterns that are not visible to the reader working word by word and text by text -- the term “distant reading” became popular in discussion of, for, and against the use of

technology in the humanities, but this provocative phrase (challenging, as it does, the idea of “close reading”) simply designates a special case of scientific visualization. We can now detect patterns in thousands of novels or hundreds of newspapers published over decades and, in so doing, we can pose new questions about genre, style, culture, and, more generally, the circulation and evolution of ideas. We can’t, of course, carefully scrutinize all the sources that our text mining systems can detect but we can learn the principles of sampling data, look at some passages carefully and then assess how our confidence that we have detected the basic patterns in the collection as a whole. If we then feel that we know less than we did when we only relied upon intensive reading of small corpora, we may in fact only better understand the limitations of what we think we know.

Such macroscopic work is at a very early stage -- we have scarcely begun to exploit the millions of books already available in digital form and most work focuses on relatively small clusters with thousands, rather than millions, of documents. And many of those collections that humanists exploit are locked away behind the paywalls of commercial sites and, even when researchers can download a collection, they have only a limited ability to publish the data which they extract and upon which their conclusions depend. But enough can already be done to see what is possible now and what can be done as we organize and expand the data that is open for unfettered research.

The macro scale leads us, however, to the challenges of the micro-scale. We can identify patterns not only within, but across, languages -- topic modelling can identify clusters of words associated with particular themes and ideas (e.g., tracking the appearance of associated words such as “nasa satellite launch soyuz rocket” can allow us to track the frequency of space exploration in very large collections) but we can track these across languages automatically (e.g., finding similar clusters in German, Russian, Chinese, Arabic etc.). But this capacity poses its own challenges: we may be able to view the (apparent) rise and fall of topics such as space travel (or the Pelagian heresy or a shift from religious to secular language in scientific publications or a range of other trends) but there are limits to the number of languages that human researchers can study, much less master and, as noted above, we cannot draw substantive conclusions from a scientific visualization without checking the data.

There are various ways to define where big data begins. We could choose that point where we have so much data that statistical methods elicit general patterns. We could choose the point where calculations require so much computation or so much data that we cannot use desktop machine. Or we could choose that point where we simply can’t analyze all the data to our own satisfaction ourselves. This final point can occur when, for example, we have several thousand novels in our native language -- but it also occurs when we have five short collections of poetry in languages that we have not studied. Or, I should say, that particular language problem represented, for all practical purposes, an absolute dead end in print culture when all we had was a page of text and print grammars, lexica, translations and other static reading aids.

We now have at our disposal qualitatively new reading environments that already allow readers to work with more linguistic sources than was feasible in print. These reading environments are built from networks of links that associate words and phrases in source texts to modern language translations, dictionaries, morphological, syntactic and other categories of linguistic analysis, maps, timelines, encyclopedic resources, and virtually any category of background information. While such reading environments are still rapidly evolving and the richly annotated corpora upon which they depend are growing, there is already, for example, no reason that students in a university level survey on Greek literature, with no training in Classical Greek, cannot immediately begin to work with the Greek text. In my own teaching, we have had students (1) compare multiple translations directly against the Greek original, (2) explore the meaning of a Greek cultural term by tracing it through the Greek and seeing how it is translated, and (3) intensively study a short passage of poetry, learning the function of every word. Such word by word reading does not replace -- or even diminish the value of -- mastery: no one who has slogged word-by-word through a few lines of Homer will confuse what they can do with fluency. Our preliminary hypothesis is that direct exposure to the Greek will intrigue readers and increase the number of those who do choose to study Greek and other historical languages -- and we are beginning to see evidence that supports that hypothesis.

The most important project that I have seen in Classics -- to my mind, the most important project that I have seen anywhere that we might count as Digital Humanities -- is the Homer Multitext Project. The HMP addresses the challenges of both the macro and the micro level, but, in so doing, it upends the culture of scholarly production, indeed it does something that many Classicists (myself included) would have thought impossible, and, in so doing, it fundamentally changed my view (and runs directly against the prevailing views among many colleagues) of where scholarship should go and how we should organize the study of Greek and Latin. At a lecture at one elite US university, a friendly colleague commented, with shock, that the work that I was advocating “so… so … philological,” by which she meant that I was advocating the study of the original sources in terms of their material production and their linguistic and cultural content. I understood her surprise -- certainly within an American context.

When I was a graduate student in the early 1980s, I never took a course on palaeography and I never had much interest in manuscripts or textual criticism. I understood, of course, the importance of consulting the textual notes of one edition and of checking different editions against each other. Occasionally, a variant would be important and I learned how to highlight that importance, but I was interested -- I might say, the field was interested -- in interpreting, rather than establishing, our texts. Classical Philology was out. Literary criticism was in -- and anyone who wanted a job and a career had to internalize that. As I review a 2010 ranked list of US Phd Programs in Classics, I don’t in the most highly rated programs see faculty younger than myself who have made a career in producing editions or in conducting traditional philological work (I would be happy to be reminded of any whom I have missed). Undergraduate classics programs have, almost entirely, given up asking their students to work through traditional reading lists -- even the Harvard Classics Department, drawing upon a very talented

pool of students, did away with its reading list in 2009. As far as I know, Yale is the only US undergraduate Classics program that still maintains a reading list for Greek and Latin.

The larger trend in the United States away from engagement with Greek and Latin per se and towards topics that can be discussed (even if not properly studied) in English became dramatically clear when our professional association narrowly voted to change its name. The American Philological Association has become the Society for Classical Studies. Even senior members of the field argued that they could not explain what philology was and that we needed a new name for our field. I don't find it difficult to explain philology (basically, it defines anything you can learn about the human past from the textual record) but the new term reflects the new business model upon which Departments of Classics depend: we teach large classes where we work with sources in English translation, we offer Classical Civilization Programs that require little or no knowledge of the original languages and we use the numbers in these large courses and more general programs to keep our language classes and even majors in Greek and Latin alive.

Against this context, I was surprised when Gregory Nagy, my former thesis advisor, set out in the early 2000s to have the 10th century Venetus A manuscript of Homer Digitized -- first by scanning a print facsimile of the manuscript and then, after extensive negotiations and at great expense, by arranging to have the original manuscript digitized at very high resolution and in multiple wavelengths of light. In this focus upon a crucial manuscript, my colleague seemed to be running against the tide of scholarship. I was far more surprised, however, when I saw how the Homer Multitext Project was editing the digitized manuscripts.

Undergraduates at the College of the Holy Cross, Furman University, the University of Houston, Brandeis University, and elsewhere are collaboratively editing the Venetus A as well as other manuscripts of Homer. These manuscripts are, of course, written in a script that is very different from our modern printed editions and contains a wide range of abbreviations and symbols -- the study of such manuscripts is called palaeography (the course I did not take as a graduate student) and has traditionally been reserved for advanced graduate students. In addition, the Venetus A and subsequent manuscripts that the HMT has digitized include elaborate commentaries, with multiple categories of annotation, all written in a very different Greek than the text of the Homeric epics, studded with technical terms about grammar and literary analysis. It is difficult enough for students to work with the Greek text of the Homeric epics. It is clearly impossible for introductory students to contribute to editions of these manuscripts -- they have their hands full learning grammar, vocabulary, and background information about different genres, as they work their way through the canon of Greek and/or Latin. Except, of course, for the fact that Classics Departments in the US have given up on reading lists and students in the US don't work their way systematically through the canon any more.

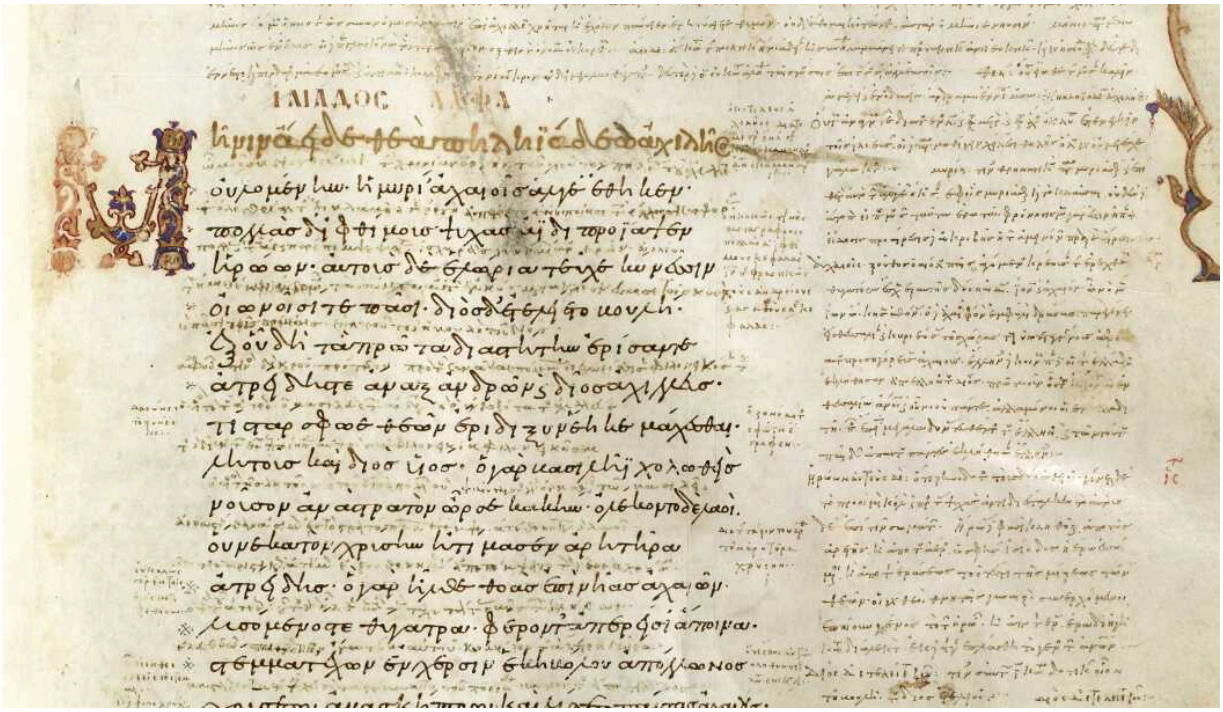

Figure 1: A page image from the Venetus A Manuscript of Homer with various classes of commentary, including notes in between the lines, in the margins, and in a stream of annotations surrounding the text.

But, of course, students at Holy Cross, Furman and elsewhere have shown themselves to be quite capable of doing important work -- as many faculty in the sciences have long known. The Homer Multitext Project has brought a laboratory culture into the study of Greek and Latin, one where students begin immediately making useful contributions and where they earn increasing responsibility as they develop and demonstrate their skills and commitment. First year students of Greek may not be able to read Homer, much less Byzantine commentaries, but they know alphabet well enough to see which lines of Homer appear on a particular manuscript page and to draw boxes around the different classes of ancient commentary, effectively classifying the structure of each page (figure 2).

Figure 2: A page of the Venetus A where different scholia have been identified and classified.

As students enhance their skills, they can take on more challenging tasks, such as creating a diplomatic edition of the textual data in the manuscript. Figure 3 illustrates the elaborately encoded TEI XML data itself.

```
            <editor> Annalisa Quinn</editor>
            <editor> Christine Roughan</editor>
            <editor> Emily Schurr</editor>
            <editor> Neel Smith</editor>
            <editor> Melanie Steinhardt</editor>
            <editor> Megan Whitacre</editor>
         </titleStmt>
         <publicationStmt>
            <p> The Homer Multitext. Casey Dué and Mary Ebbott, editors; Christopher Blackwell and
               Neel Smith, project architects.</p>
         </publicationStmt>
         <sourceDesc>
            <p> Text edited directly from photography of the Venetus A manuscript, Marciana 454.</p>
         </sourceDesc>
      </fileDesc>
   </teiHeader>
   <text>
      <body>
         <div n="1">
            <div type="scholion" n="25">
               <div type="lemma"/>
               <div type="comment">
                  <p> διὰ τί <supplied reason="lost"> εὐθὺς</supplied>
                     <seg type="word"> ἀ <supplied reason="lost"> πὸ</supplied></seg> τῶν <seg
                        type="word"> τε <supplied reason="lost"> λευταίων</supplied></seg> τοῦ
                        <supplied reason="lost"> πολέμου ἤπξατο; καί φαμεν ὅτι ἅπας μὲν ὁ χρόνος ὁ
                        πρὸ τοῦ δεκάτου ἔτους οὐκ ἔσχεν οὕτω συνεχεῖς τὰς μάχας, διὰ τὸ καὶ τοὺς
                        Τρῶας αὐτοὺς</supplied> φόβῳ <supplied reason="lost"> τοῦ</supplied>
                     <persName n="pers1">
                        <seg type="word">
                           <supplied reason="lost"> Ἀχιλλέ</supplied> ως</seg></persName> ἐντὸς <seg
                        type="word"> κατα <supplied reason="lost"> κε</supplied> κλεῖσθαι</seg>
                     <seg type="word">
                        <supplied reason="lost"> τ <unclear> οῦ</unclear></supplied></seg> τείχους·
                     τὸ δε δέκατον πλείονας ἔσχε <unclear> ηναι</unclear> πράξεις καὶ τοὺς <seg
                        type="word"> πολέ <supplied reason="lost"> μους</supplied></seg> ἰσσπάλου
                     τοῦ <persName n="pers1"> Ἀχιλλέυ</persName> ὀργιζομένου· ὁ δὲ ποιητὴς <seg
                        type="word"> οικο <supplied> νομικῶς</supplied></seg>
                     <supplied reason="lost"> κἀν</supplied> τούτῳ ἤρξατο μὲν ἀπὸ τῶν τελευταίων·
                     δια δε τῶν σποράδην αὐτῷ λεχθεντων περιέλαβε καὶ τὰ προ τούτου πραχθέντα‡</p>
               </div>
            </div>
         </div>
```

Figure 3: Part of a TEI XML diplomatic edition for one of the categories of scholia in the Venetus A, showing part of the credits as well as the edited text. A description of the published data can be found here.

The Homer Multitext Project inspired a range of research projects on other manuscripts (such as the Archimedes Palimpsest and on inscriptions. The resulting Homer Multitext Blog has produced a stream of publications. Some of these publications present new interpretations of the manuscripts and documents under analysis, others announce the publication of data, and still others describe the activities of the Homer Multitext Project and, equally important, information about how the community of young researchers has come together and defined itself. The Holy Cross work has also produced peer-reviewed publications produced by undergraduate researchers (such as Christine Roughan, Digital Editions and Digital Diagrams, a very important analysis of how to represents diagrams in digitized manuscripts on mathematics)

but peer-reviewed publications are only possible because a new culture of intellectual production has taken place. Scholars may hunt through the blog for analyses of manuscripts or for publications of data but the most serious students of Greek -- indeed, of any historical language -- should pay particular attention to posts such as "How to build a community of scholars through pancakes," "Collaboration and new tools are keys to success for Holy Cross research teams," and "Undergraduate interest in manuscripts."

I have spent this much time on the Homer Multitext Project because I think that it illustrates a number of critical points.

First, Greek is a hard language. Manuscripts of Homer really are quite complicated (in 2010, Google singled out the Venetus A Homer Manuscript as a model for organizing complex content). If students can do useful work on the Homer Multitext Project, then they can do quite a bit in many other subjects, where primary sources are not as difficult.

Second, the vast majority of work on the Homer Multitext Project, on the Archimedes Palimpsest and on other projects at Holy Cross, Furman and elsewhere is voluntary: students receive neither academic credit nor payment during the year but work for intrinsic rather than extrinsic motivations. Some students have been able to win summer research support but only after they have show their commitment to, and enthusiasm for, their work during the academic year and there has not been any guarantee that summer support would be available. The five years of Homer Multitext Blog posts document a sustained interest in making primary sources such as manuscripts useful. When these students have had an opportunity to produce something new and tangibly useful, they have responded -- and probably done so with more passion and learned more than when they are ploughing through settled curricula and simply absorbing new information and skills.

Third, the Homer Multitext Projects provides one example of citizen science. Where crowdsourcing tries to get members of the public to contribute to research, but contributors remain members of a largely faceless crowd and the focus is upon what the crowd produces, not on what members of that crowd learn. Each member of the Homer Multitext Project receives named credit for his or her work, and all activities either increase their skills directly or provide added motivation to increase their skills. We don't need crowds. We need in our republic of letters a broad community of citizens, each of whom has responsibilities as well as rights. A citizen has a voice and can contribute to a larger conversation.

Fourth, the Homer Multitext Project illustrates how work on the micro level responds to the challenges posed by the macro level. We have thousands of manuscripts already available under a as high resolution scans but the transcription and analysis of manuscripts is only one of many tasks that we need to perform as we work with the growing digitized collections from the human record. Those of us who are advanced researchers or library professionals need to engage our students and fellow-citizens as colleagues and collaborators.

Fifth, preliminary reports suggest that the ability to contribute to projects such as the Homer Multitext Project brings new students to the study of Greek. Students want to contribute and to learn at the same time, and they supposedly learn better when learning allows them to make a contribution of some sort. While this general thread needs formal study, my own experiences suggest that we will find that the ability to contribute has a substantial positive effect upon learning.

Sixth and perhaps most-importantly, the Homer Multitext Project illustrates a principle that appears in other discussions of citizen science project. Citizen scientists may begin by contributing to a research project but they quickly begin to provide their own input about what matters and to influence the goals of research. In the case of the Homer Multitext Project, this principle is, at least from my perspective, very much in play.

Here I return to the question that Manfred posed to me -- what should be the role of student research in choosing among the various directions that professorial research can take. The contributions that the HMT students have made and the extended, disciplined enthusiasm that they have shown over a period of years (even before the HMT Blog began in 2010) has transformed my own view of how to advance the study of Greek and Latin. I find myself somewhat bemused to be focusing on tasks that are both radically new, insofar as they are only possible because of computation, and radically traditional, insofar as they restart, or perhaps more properly, reinvent activities, such as editing, that have long traditions. The most important output for the study of Greek and Latin is not the research output but the appearance of a research culture that flattens hierarchies, reduces barriers to entry, energizes students, and provides a new strategy for re-integrating the study of Greek and Latin within Classical Studies.

I see now a dialog between students and faculty. In the short run, faculty have almost absolute control because we establish and maintain our programs. But over time, we are entirely dependent upon our students because if the students stop coming, our jobs and our fields will vanish. If we focus solely upon the questions that we inherited, we may produce contributions that are of immense scholarship but those contributions will become mere information rather than knowledge if our fields of inquiry die out and the results of our work no longer find their way, directly or indirectly, into the minds of living human beings. Put another way, does my field as a Classical Philologist consist of the particular questions I learned to ask as a graduate student or does that field more generally consist of a sustained engagement with Greco-Roman culture in general and with Ancient Greek and Latin in particular? So now I find myself advocating the study of manuscripts and (in other contexts) the intensive analysis of Greek and Latin language because these are topics that demand great intellectual rigor and that can now, because of the computational environment in which we live, (re-)engage our students. Perhaps we will see a generation of such newly philological work as we rebuild the infrastructure for the study of Greek and Latin from ground up and then experience, predictably, a new shift away from digital philology. What matters to me as a student of Greek and Latin is that these languages and the cultures associated with them remain vigorously alive in the intellectual life of humanity.

This shift from mastering a set curriculum to the integration of learning and research is, on the one hand, a radical shift away from the top-down but very demanding program of study that I encountered as an undergraduate in the 1970s. There we were challenged to master -- in large measure, on our own -- extensive reading lists of Greek and Latin and to produce essays and a senior thesis, but we also knew -- and were reminded if we forgot -- that we had no significant voice and were not partners in the advancement of knowledge. If such programs are no longer viable, this change does not necessarily reflect a decline in the number of students who come to college after having studied. Latin has, reportedly, rebounded a bit since the 1970s when I entered college. I think that the motivated students who might have majored in Classics a generation ago go into fields where they can participate in research and make a difference. This goes beyond the relentless political campaigning for so-called STEM disciplines (Science, Technology, Engineering and Mathematics -- in Germany, the corresponding acronym is MINT). Students want to learn and they want to make a difference, and they learn more when they do both at once.

This is not a new idea. Two hundred years ago, Wilhelm von Humboldt, who helped frame the structure for the modern university system, argued that the purpose of a university was to advance human understanding (here I use a non-positivist translation of the German Wissenschaft, which might be more properly rendered “science” or “knowledge”). What is new is that computation has created a new space within which we conduct intellectual activity, one where there are vast new fields of activity, where students can now learn as they contribute, and where Humboldt’s vision becomes a realisable goal.