The Presence of Digital Process
James O'Sullivan, http://josullivan.org

Computational methods are an essential part of the Digital Humanities, in that they are central to a range of disciplinary processes. By "process", I refer to the digital means by which we produce new knowledge and meaning of significance to Humanities scholarship. While process—the application of the computer-assisted methods we develop, manipulate, and adopt—can be an act of interpretation in itself, I would argue, and I am sure that few would disagree, that this act is always in the service of the product, the new insights, be that into the literary or otherwise, offered by our fields' many esoteric approaches. Herein lies part of the value of the Digital Humanities: the way we approach research allows for new questions to be asked and existing debates to be revived. While it is now comprised of a great many, and often dissonant, scholarly, and indeed creative, activities, our community first emerged out of a fascination with the potential for the computer to be utilised as an instrument for scholarly enquiry. The very essence of the Humanities is criticism, and so if the methodological foundations of the *Digital* Humanities are to continue to mature, then we must continue to be critical of this essential element—repeatedly, we must ask of our machines, *how* and *why*.

This session is about digital research as a means of discovering the "unimaginable", a means of constructing and supporting "new paradigms" for scholarship. Digital research can deliver all of these promises, and more, but the practical application—digital research in action—must be implemented in an appropriate and thoughtful manner. In this respect, there are two issues which I would like to address—method and data—this arena's most radicalised, but equally problematic, components.

Some of my work is concerned with computational analytics, which, typically, relies on literary datasets that, while not necessarily restricted, are difficult for peers to replicate. Literary datasets are particularly susceptible to computational approaches—text is as malleable as it gets for a machine—and the new findings that such techniques reveal have the potential to add considerable value to our core pursuits. However, in research contexts where the subject matter is as culturally and socially sensitive as it is intriguing, scholars are presented with an ethical dilemma as far as data is concerned.[i] Many of the works used in macro-analyses are often still under copyright, and so researchers are prohibited from sharing the texts. This restriction precludes our peers from doing two important things: validating our findings, and offering further iterations of our work. Considering the effort that is required in digitising certain datasets, our discipline is fast becoming one where much of the work that claims to be empirically valid cannot in fact be validated. Much of our research is conducted on datasets which take the researchers years to acquire and prepare. If datasets are not shared—and oftentimes they cannot be for legal reasons beyond the control of the researcher—then one must turn to replication, which, requiring much time and institutional support, is often unfeasible. As a result, there is no realistic mechanism by which we, as scholars working within the Digital Humanities, can query the validity of many of our interpretations. Should

scholars who create datasets hold power over digital artefacts of cultural significance? How can we validate the new insights being offered by scholars in our field? Should we, as scholars, sacrifice access in the name of exploration, or do we need to at least strive for balance between the two? The same can be said of method, wherein researchers both develop and apply techniques which are not generally understood by the wider community. I am not attempting to detract from the value of new methodologies, but simply warning that our field is one in which much is being based upon approaches that are not entirely robust. We use and teach tools based on algorithms we do not understand—how many of us fully comprehend each of the phases required in the production of the wonderful visualisation upon which we have based such bold claims? In essence, many, if not all of us, are guilty of claiming the mantle of scientific without knowing much about the science.

For the purposes of illustration, I will draw reference to a study completed by a collaborator at Penn State, a graduate student in the College of the Liberal Arts, Sean G. Weidman. Soon after my arrival in State College, Sean came to me with the proposal for a study: he wanted to repeat David Hoover's experiment in which he produced a set of gendered wordlists. In the study, Hoover presents a list of the one hundred most distinctive words in the works of twenty-six poets, equally split between male and female authors. Hoover remarks that some aspects of his findings are "almost stereotypical", with "[f]emale markers like *children* and *mirrors* and male markers like *beer* and *lust*".[ii] Sean was curious to see if Hoover's results would be replicated using a larger dataset, drawn from across a number of distinct literary epochs, namely, Victorian, modernist, and contemporary. Furthermore, Sean intends to produce the paper within a more qualitative context, acknowledging, for example, the extent to which the scope of such research does, or does not, immediately contribute to the debates of gender theory or the potentiality of a distinct form of écriture feminine. We gathered corpora for 54 authors, for a combined total of 212 novels and collections of short stories. Craig's Zeta was the primary methodology, as was the case in Hoover's initial study. Our Zeta analysis was conducted using a text slice length of 2,000, text slice overlap of 1,000, and an occurrence and filter threshold of 2 and 0.1 respectively. At this point, I want you all to be honest—how many of you completely understood the description of this method? Our results are not important at this particular juncture—what is of significance here is that this is a sensitive study, the findings of which are based on a corpus we cannot share, using a method with which many of the subject's most engaged scholars may not be able to interact. And that, in a nutshell, is the problem. As scholars, we have a responsibility to ensure that the application of digital methods does not damage the Humanities—we cannot continue to use computation as an excuse for claims which, while "technically" accurate, are contextually misrepresented, or through some nuance that a machine cannot detect, utterly misinterpreted. The nature of experimentation and calculation are such that these issues will always be a part of our field, but as humanists, our duty is to at least be aware of their presence.

---

[i] Yes, I use "data" in the singular!

[ii] Hoover, David L. 'Textual Analysis'. *Literary Studies in the Digital Age*. Ed. Kenneth M. Price and Ray Siemens. Modern Language Association of America, 2013. Web. 3 Nov. 2014.