CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The humanities have a 'reproducibility' problem
Authors	O'Sullivan, James
Publication date	2019-07-09
Original Citation	O'Sullivan, J. (2019) 'The humanities have a reproducibility problem', Talking Humanities, School of Advanced Study, University of London, 09 July. Available at: https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/
Type of publication	Contribution to newspaper/magazine
Link to publisher's version	https://talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/, http://jamesosullivan.org/
Rights	© 2019. Curated by the School of Advanced Study
Download date	2026-11-16 14:08:58
Item downloaded from	https://hdl.handle.net/10468/8208



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

talkinghumanities.blogs.sas.ac.uk

The humanities have a ‘reproducibility’ problem - Talking Humanities

8-10 minutes

We’ve all heard about the digital revolution in the arts and humanities: digital humanities some call it, a major part of which is using computers to conduct data-driven analyses of complex materials like literature. ‘There are a lot of drawbacks (and benefits) to this “new” discipline, or set of disciplines, or passing fancy, however you conceive it,’ explains Dr James O’Sullivan, digital arts and humanities lecturer at University College Cork, as he examines one such shortcoming – reproducibility.

Digital humanities has strengthened pre-existing syntheses between the sciences and critical explorations of the human condition, but it has also transferred challenges from the former to the latter. Chief among these is that of reproducibility, and the essential requirement that any experiment claiming to be scientific can be faithfully and independently replicated.

The history of the humanities is one of privilege, and there are countless pre-digital examples of studies based on content inaccessible to the wider scholarly community, so ‘transferred’ isn’t quite the right word. ‘Exacerbated’ is

perhaps more appropriate, in that humanities scholars are increasingly expected to accept the findings of their peers without access to the data from which discoveries are drawn. Access to data is just part of the problem. The relative obscurity of computer-assisted techniques has also contributed to the rise of our discipline's reproducibility problem.

Computational methods are central to a range of disciplinary processes, the digital means by which we produce new knowledge and meaning of significance to humanities scholarship. While process can itself be an act of interpretation, this act is always in the service of the product, the new insights, be that into the literary or otherwise, offered by contemporary scholarship's many esoteric approaches. Herein lies part of the value of the digital humanities: the way we approach research allows for new questions and the revival of existing debates.

But if the methodological foundations of the *digital* humanities are to continue to mature, then we must continue to be critical of all those limitations which become pronounced when we engage in computer-assisted criticism.

Katherine Bode's paper in *Modern Language Quarterly* highlights several instances where researchers have not shared, or in some cases sourced, their data. Many of us are guilty of such transgressions: some of my computational work relies on literary datasets that, while not necessarily restricted, are difficult for peers to replicate. Many of the works used in some of the most high-rpofile examples of macro-analyses remain under copyright, prohibiting researchers from sharing the texts (though we could be better

at sharing other elements). This restriction precludes our peers from validating findings, and offering further iterations of our work. Should scholars who create datasets hold power over digital artefacts of cultural significance? How can we validate new insights offered by scholars in our field? Should we sacrifice access in the name of exploration, or do we need to, at least strive, for balance between the two?

I am not attempting to detract from the value of new ways of reading, but rather, warning against an overreliance on principles that are, to some of us, as much about magic as they are mathematics. We need to dispel the mysticism embedded in digital humanities. Scholars with technical proficiencies have a responsibility to explain their methods clearly, while the less technical need to increase their familiarity with new practices. It is frustrating that there are still journals that will not consider articles for peer-review because ‘the method has not been fully explained’. It has, follow the citation trail.

The potential of digital humanities will never be realised if practitioners must continue to publish within self-serving siloes. When I use a computer to analyse Joyce, I want Joyceans to read the results – the method is boring, what is of interest is the interpretation that I have offered, and the most appropriate readership is not to be found in DH journals. Yet, that is where I will inevitably have to publish, because there remains a disconnect between what I have done with the machine, and the conclusions I have drawn from such calculations. DH scholars need to describe their work for non-DH audiences, making it reproducible and transparent. Where such work has been done editors and readers need to

recognise and appreciate the effort.

The reproducibility problem is particularly problematic in research contexts where the subject matter is as culturally and socially sensitive as it is intriguing. For illustrative purposes, I draw reference to a study I completed with a colleague at Pennsylvania State University, Sean G. Weidman. It builds on work by David L Hoover, who produced a comparative 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 female markers like ‘children’ and ‘mirrors’ contrasting with male markers such as ‘beer’ and ‘lust’. Our study attempts to check if Hoover’s results would reproduce using a larger dataset, drawn from across distinct literary epochs, namely, Victorian, modernist and contemporary. We produced 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 debates on gender theory or the potentiality of a distinct form of écriture feminine.

Our results are not important to this article, what is of significance here is that this is a sensitive study, the findings of which are based on a corpus we largely cannot share, using a method with which many of the subject’s most engaged scholars may not be able to interact. That, in a nutshell, is the problem. Should we be using, promoting and teaching methods researchers and their peers do not fully understand? Is ‘interdisciplinarity’ merely an excuse for the application of methods which are ‘black boxes’ to many

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 humanities scholars our duty is to, at least, be aware of their presence.

This article has been partly reproduced from panel discussions and conference proceedings on the issue of ethics in the digital humanities: **Access, ownership, protection: the ethics of digital scholarship**: Katherine Mary Faull, Diane Jakacki, James O'Sullivan, Amy Earhart, and Micki Kaufman; **Digital Humanities, Kraków (July 2016)**: Diane Jakacki, Laura C Mandell, Paige Morgan, James O'Sullivan and Katie Rawson; **Digital scholarship in action: research**: presided by Patricia Hswe. *MLA Annual Convention*, Austin (January 2016);
The ethics of data curation: the quandary of access vs. protection: Diane Jakacki, Katherine Faull, Dot Porter and James O'Sullivan. *Keystone Digital Humanities*, Philadelphia (July 2015).

Dr James O'Sullivan (@jamescosullivan) is a lecturer in digital arts and humanities at University College Cork (National University of Ireland). He has previously held faculty positions at the University of Sheffield and Pennsylvania State University. His work has been published in a variety of interdisciplinary journals,

including *Digital Scholarship in the Humanities*, *Digital Humanities Quarterly*, *Leonardo*, and *Hyperrhiz: New Media Cultures*. His writing has also appeared in such venues as *The Irish Times* and *The Conversation*. He and Shawna Ross are the editors of *Reading Modernism with Machines* (Palgrave Macmillan 2016). See josullivan.org for further information.