Towards a Conceptualisation of Replication in the Digital Humanities

Verhaar, P.

Towards a Conceptualisation of Replication in the Digital Humanities

Peter Verhaar

Peter Verhaar works as an Assistant Professor at the Leiden University Centre for Arts and Society. He teaches courses for the MA *Book and Digital Media Studies* and the BA minor *Boek, Boekhandel en Uitgeverij*. His research is mostly in the field of Computational Literary Studies. Next to his position as a lecturer and researcher, he also works at Leiden University's *Centre for Digital Scholarship*, where he offers advice on various aspects of digital scholarship, including open science, data science and research data management.

During the last few decades, researchers, funding agencies, academic publishers, and governments have increasingly exerted themselves to make the final and the intermediate results of academic research publicly available, with as few legal, practical or financial restrictions as possible (OECD 15; European Commission, *Open Innovation* 52). Nowadays, such efforts to enhance the transparency and openness of scholarly workflows are often subsumed under the wide umbrella of the Open Science movement. Advocates of Open Science mostly aim to ensure that the data, the publications, and the software that are created by researchers become accessible under terms that foster reuse and collaboration (Vicente-Saez et al. 433). This endeavour to transition to open research is frequently motivated by the conviction that such transparency can help to safeguard academic integrity and accountability. It is often assumed that when researchers share their data and methods alongside their publications, this enables peers to scrutinise the findings and to verify the conclusions. Eventually, this possibility to replicate the findings of research projects should instil a higher degree of trust in science and in scholarship, both within society at large and among the academic community itself (European Commission, *Open Innovation* 15).

Such attempts to replicate the results of earlier research projects have often proven problematic, however (Ioannidis 696). In a meta-analysis of 100 studies in the field of psychology, it was found that only 39 of the results could be corroborated (Open Science Collaboration 6; Baker, "First Results from Psychology's Largest Reproducibility Test" 1). Similarly, in a survey conducted for *Nature*, it was found that 70% of all researchers who have tried to reproduce the experiments of their colleagues failed to arrive at identical results (Baker, "1,500 Scientists Lift the Lid on Reproducibility" 452). Reproducibility issues such as these came to the surface initially within disciplines such as psychology, biomedicine (Bustin 36), and pharmaceutical sciences, but, at the moment, it is often argued that the problem is more widespread (Munafò et al.). It has been claimed that, at present, numerous academic disciplines are in the throes of a rebarbative "replication crisis" (Baker, "1,500 Scientists Lift the Lid on Reproducibility" 453). Evidently, such failures to confirm research results can have

deeply detrimental effects on the esteem of scientific work. Academic research often builds on the results achieved by predecessors, but when such cumulative work is based on spurious findings, this will invariably lead to a waste of time and financial resources.

At the same time, it can be observed that the debate regarding the reproducibility of research has predominantly taken place within fields employing empirical and quantitative methodologies, such as the life sciences, the natural sciences, and the social sciences. So far, the concept of reproducibility has been of lesser relevance within the humanities, a field in which findings and analyses are often based on interpretation and on qualitative methods (Britt Holbrook et al.). Humanities scholars are generally interested in the meaning or the value of cultural or historical artefacts, and the interpretation of such artefacts often depends strongly on the background knowledge, the intentions, and the theoretical perspective of individual scholars. Even when the steps that have been followed within a study have been meticulously documented, it may still be the case that different researchers, faced with the same data and applying the same analytic methods, will arrive at divergent interpretations. In many cases, peers also lack access to the sources that have been consulted in a study, especially when such objects are located at distant sites or in remote heritage institutions. Scholars often place a degree of trust in the claims of their peers, accepting the findings that are published without examining the primary sources these claims are based on themselves (O'Sullivan; Bode 84ff).

The prominence of hermeneutics and idiosyncrasy notwithstanding, there are distinctly several humanistic subfields whose research protocols and outcomes may reasonably form the object of independent replication. Studies in the field of linguistics, for instance, typically draw on strict statistical methods, applied to carefully delineated corpora of language samples. By the same token, and more generally, the call of the Open Science movement for more transparency can presumably be answered readily by humanities researchers who make use of computational methods. In the research field that is commonly referred to as the *digital humanities*, scholars typically make use of digital data and

of algorithmic methods to address humanistic questions (Fitzpatrick 12). The application of digital tools and algorithms invariably requires a standardisation and an objectivisation of scholarly practices. This article aims to develop a conceptualisation of the term *replication* that is appropriate for the form of research that takes place in the digital humanities and intends to clarify the conditions that contribute to the replicability and the reproducibility of this specific area of research.

Such clarification of the central concepts is necessary, as the terminology used in this context continues to be convoluted (Plesser 2). The ambiguity stems in large part from the fact that different academic disciplines also have dissimilar notions of the rationale and the exact nature of reproducibility. Differences with respect to methodology, types of research data, epistemic values, and analytic tools may all affect the views on how studies can or should be replicated. Terms such as replication, reproduction, and repeatability have sometimes been defined in distinct and sometimes contradictory ways.[1] Despite such disagreements, it can be stated, in very broad terms, that reproduction and replication both describe a relationship between two studies. The first of these is an original study, and the second study aims to reuse or recreate specific elements of this earlier work.[2] The objective of such a revisitation is usually to determine whether the new study can achieve the same results as the earlier study. Reproductions and replications are considered to be successful if the results of the two studies are sufficiently congruent. Peels also emphasises that the results of the new study should not be determined or affected in any way by the results of the earlier study, and that it ought to reach its results independently, reusing only the data, the protocols, or the tools of the earlier study (3). While the general principles underlying reproduction and replication may be clear, researchers planning to reuse

1 Plesser notes that Jon Claerbout and the Association for Computing Machinery have proposed contradictory definitions of the terms 'replicability' and 'reproducibility' (2). This article largely follows the definition given by Claerbout.

2 The KNAW defines a replication study as "a study that is an independent repetition of an earlier, published study, using similar methods and conducted under similar circumstances" (18).

or to evaluate parts of an earlier study will still need to translate these general principles into concrete operational steps, closely attuned to the methodological and epistemological traditions of particular academic fields. As part of such an operationalisation, a decision needs to be taken on whether the study aims to reproduce or to replicate the earlier work. The term *reproducibility* refers, more specifically, to the ability to repeat all the stages of the workflow that were followed in a study precisely, based on data and methods documented or made available by the original author.[3] This specific conceptualisation of *reproducible research* was developed by computer scientist Jon Claerbout. He encouraged his colleagues to ensure that the readers of a textual publication could duplicate the full analytic process that was implemented for the study, starting from raw data and eventually leading to the exact same results (Claerbout and Karrenbach 602; Donoho et al. 4). Reproduction implies the ability to redo a full experiment by following a set of instructions, and it may, for this reason, be viewed as a mechanical activity.[4] To underscore the idea that procedures such as these concentrate on the re-enactment of a method, Goodman et al. refer to reproducibility as *methods replication*.[5]

3 The question whether two separate datasets can be considered identical may of course give rise to a more principled debate about the nature of identity and of equality in this context. In the case of a collection of digital data, we can assume that the data values can be copied from one location to another without any loss of information or quality. If no digital data are available, however, researchers aiming to reproduce a study may attempt to recreate the original results, by following the exact same protocol that was followed in the earlier study. Researchers aiming to reproduce results obtained in a laboratory setting, for instance, may try to regenerate these results by repeating the experiment under identical circumstances, as far as possible, but the data that result from the repeated experiment may not be fully identical. Evidently, for observations of unique historical events or one-off natural phenomena, such a faithful recreation of a data set would be impossible. To forgo difficulties such as these, the current discussion will limit itself to the case of digital data used in computational research.

4 Facilitating this stringent form of reproducibility can obviously pose some challenges. Research software may have a dependency on specific code libraries, for instance, and these libraries may also be available in different versions.

5 Goodman et al. define the term as "the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results" (2).

Importantly, reproducibility does not imply that we can verify whether a certain finding is correct or truthful. If there are mistakes in the reasoning underlying the code, these will clearly be reproduced as well. Computational reproducibility is concerned first and foremost with creating transparency. Because of such openness, peers can ascertain whether the data and the methods are indeed as they are represented in a publication. To be able to corroborate academic results, it is usually necessary to replicate the study. In contrast to reproduction, which essentially implies an exact repetition of an experiment based on the same data and the same code, replication refers to a re-implementation of the experiment. It demands a critical revisitation of the study, which attempts to determine whether the findings of this earlier study were correct (Plesser 1). A replication may potentially be based on data and code made available by the original author, but, if this is the case, the data and the code must also be subjected to a close and careful examination. Alternatively, scholars carrying out a replication can reconstruct or re-engineer a method, based on the documentation of the original methodology (Rougier et al. 5). In addition, they may choose to recreate or recollect the dataset discussed in the earlier work. Replicability demands transparency, which may result from the original author's attempts to make the research reproducible.[6] This type of replication, in which all the steps of the original study are revisited critically, can be pertinent if there is a suspicion of academic misconduct or fraud. Through a duplication of the study, peers can evaluate whether the data and its analysis genuinely lead to the findings as represented in a publication. Framed more positively, this strict form of replication can also be motivated by a need to corroborate or confirm earlier findings, and to ensure that follow-up research can reliably build on robust results. Next to such stringent forms of replication, in which researchers attempt to approximate the original methodology faithfully,

6 Goodman et al. describe replication as the process of "obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible" and suggest that replication can also be referred to as "results reproducibility" (3).

a number of other laxer forms of replication can be distinguished. For example, researchers may apply the method that was developed within an earlier study to a new data set, to evaluate the generalisability of this approach, or attempt to answer the same question using the same data, but with a different method. This enables them to gauge the robustness of specific findings.

Following the conviction that, for the digital humanities, transparency is not a goal in itself, but rather a means to cultivate reuse and collaboration, it may be argued that replication is ultimately more relevant than mere reproduction. Replication actually implies a critical examination of the data and of the method, as well as an attempt to corroborate the findings of earlier research. To repeat and scrutinise the work that was performed in an earlier research project, first, it is necessary to acquire the raw data, and, potentially, the processed data. As is the case in virtually all disciplines, researchers in the humanities are increasingly stimulated to publish their dataset in agreement with the FAIR data management principles (Wilkinson et al.) and openly, if this is permitted within the constraints of intellectual copyright laws and data protection regulations. Within the digital humanities, primary data may consist of a wide range of resources, including images, videos, sound recordings, social media posts, or machine-readable text. Evidently, it is not possible in all cases to facilitate open access to all the materials that have been used. Research may be based on resources that are still protected under copyright law, or on resources that have been shared with specific teams of researchers following bespoke agreements with publishers of content providers. When data has been generated from dynamic online resources, such as social media platforms, researchers can also decide to develop and share code for the reconstruction of these datasets. This solution was implemented in the code developed for the research project described in Bourrier and Thelwall's article "The Social Lives of Books", published in the *Journal of Cultural Analytics*. The article examines readers of Victorian literature by comparing data from the Open Syllabus Project and the MLA International Bibliography with data extracted from Goodreads (Bourrier and Thelwall). The article contains a link to the GitHub repository

containing software that can download ratings of books directly from the Goodreads website. The dynamic nature of the Goodreads data creates a challenge for reproducibility, however. Since many new reviews and ratings have been added since the publication of the article, it can be difficult to confirm whether the results that are reported in the publication are accurate. Such differences are less problematic, nonetheless, when the aim of the replication is mostly to evaluate the validity of the method that is discussed.

Secondly, in addition to acquiring or reassembling the original data, scholars who seek to replicate a study need to criticise the method that was applied to analyse these data. In many cases, such criticism can take place through a reconstruction of the method, based on the documentation of the methodology offered in a publication. When the researchers responsible for the original study have published their research software in open source, the necessary criticism can obviously focus on the code. Fortunately, many scholars have begun to share the code they have developed openly using software repositories such as GitHub or Zenodo, under a licence that stimulates reuse. Such activities are propelled in part by the Open Science movement, which often sees open software as one of its central pillars (Lowndes et al.). In recent years, many developers of research software have also begun to make use of notebooks which combine code and human-readable documentation. Such notebooks are often based on RMarkDown or iPython, and enable programmers to develop code in the spirit of literate programming, a concept developed by Donald Knuth (Knuth 1ff). The central concept underlying literate programming is that programmers should intersperse the computer code with messages in a natural language explaining what they want the computer to do. Subsequently, these messages enable peers to develop an understanding of the logic that is implemented in the software. Such documentation can clearly be very helpful during replications, although it needs to be added that, while the development and the use of open source code continue to increase (Open Source Initiative), it can still be difficult occasionally to match code repositories to publications. At present, there is no firm tradition of citing software in scholarly articles yet.

Growing numbers of journals have adopted a data availability policy, implying that papers can be submitted exclusively if the reviewers and the readers can access the data, but very few journals have put in a place an analogous software availability policy.

Replication work may thirdly concentrate on the conclusions that are drawn from the results. Even if it is found that the data and the method indeed generate the results that are reported, scholars may still disagree, fully or partially, with the concluding statements inferred from these results. Goodman et al. refer to this aspect of replication as "inferential reproducibility".[7] Scholars engaged in replication studies should consider whether the methods that were applied were suitable for the research question and whether the central claims that are made in the publication follow logically from the outcomes of the experiment. In fact, such attentiveness to conclusions is essential, as replication studies generally aim to determine whether particular methodologies can genuinely advance the production of credible and reliable knowledge.

The central objective of reanalyses, as discussed earlier, is to establish whether a new study can obtain the same results as an older study, but currently, there is no discipline-wide consensus concerning the precise nature of this sameness. When the numbers that are generated in a replication study are not fully identical to the numbers that are presented in the earlier paper, this evidently poses the question to which degree discrepancies are still acceptable. When should we conclude that replication is unsuccessful? It has been suggested that results agree sufficiently if the "replication shows a statistically significant effect ($p < .05$) with the same direction as the original study" or when "the original effect size is within the 95% confidence interval of the effect size estimate from the replication" (Open Science Collaboration 77). Peels argues that two independent findings can be considered commensurate when they "have the same direction", when they have a "similar effect size", or when they display

7 Goodman et al. define the term as "the making of knowledge claims of similar strength from a study replication or reanalysis" (3).

a "similar p value, confidence interval, or Bayes factor" (4). It may be argued, for example, that, when one study discovers a positive correlation between two variables, a second study can be seen to confirm these earlier results if it likewise establishes a positive correlation between these variables. Ultimately, it is probably undesirable to define fixed and generic margins for such deviations since the importance of such variations will often strongly depend on the goals of individual research projects. In the end, it will be the responsibility of individual researchers engaged in replication work to determine whether the results that are obtained in the new study still support the main claims that are made in the study that is revisited.

In sum, it can be stated that replication in the digital humanities entails a critical evaluation of the results of an earlier study, based on a close examination or a re-implementation of its scholarly workflow, including the data and the analytic software. Such replication studies can be expedited considerably via the open availability of all the scholarly resources that were generated in the original study, in step with the central tenets of the Open Science movement. When the eventual results of the new study agree with the results of the study under scrutiny, this is a good indication of the reliability of these outcomes. It simultaneously suggests that the original study's methods can be reused productively in follow-up research. The conceptualisation of replication that is given above is admittedly rather lenient, and, given this leeway, it may prove difficult to create exact numbers about the number of studies whose results can genuinely be corroborated. If it is accepted that the aim of replication is mostly to prompt methodological discussions, the question of whether the digital humanities face a replication crisis immediately becomes less relevant. A recent EU report about reproducibility and replication also argued that, rather than framing the high number of irreproducible studies as a crisis, it is more beneficial to view replicability as an ideal to be pursued and to concentrate more closely on its benefits (European Commission, *Reproducibility of Scientific Results in the EU:* 8–9). Replicability implies a high level of transparency, which should enable peers to evaluate the correctness and relevance of findings. It can lead to a better understand-

ing of the methodology of academic fields, and, as such, it can buttress collaboration and interdisciplinarity. Taken together, these various consequences eventually help to strengthen the reliability and integrity of academic research.

Considering that the methodology of the digital humanities is, to a large extent, still in development, all attempts to replicate published findings can be highly beneficent for the advancement of the field. Next to evaluating the technical accuracy of findings, replication studies additionally ought to stimulate debates on the overall relevance and scholarly value of specific methodologies. Such tests of the suitability of these methods may stimulate other researchers to develop these analytic processes further. If a method has been shown to yield useful results within a specific context, colleagues can attempt to test the generalisability of the approach by applying it to other data and other domains. Replication studies accordingly form a vital prerequisite for the maturation of the field's methodology. For researchers in training and early career researchers, replication work may also entail education. By redoing some of the field's seminal studies, and using the numbers mentioned in published articles as a benchmark, scholars can effectively familiarise themselves with the practicalities of specific research methods and begin to learn about the history of their discipline.

The Swiss painter Paul Klee famously stated that art does not aim to reproduce what we can see, but, instead, to make us see.[8] Instead of solely imitating an existing reality, a painting should bolster a specific way of looking at the world. It ought to expose what is essential. Along similar lines, a scholarly replication can be viewed as a creative and critical interaction with an earlier work of scholarship, and the aim of such an exercise is not necessarily to replicate every single detail of the research design with the highest level of exactitude. The goal is first and foremost to revisit the essence of the study and to evaluate whether all the activities followed were adequate and effectual. Digital humanities research

8 "Die Kunst gibt nicht das Sichtbare wieder, sondern macht sichtbar" (Klee 28).

ultimately seeks to discover those computational methods that enable us to explain languages, cultures, and societies, and, if executed well, such revisitations of past scholarly activities make us see and understand those methods that work.

Works Cited

Baker, Monya. "1,500 Scientists Lift the Lid on Reproducibility." *Nature*, vol. 533, 2016, doi:10.1038/533452a.

–. "First Results from Psychology's Largest Reproducibility Test." *Nature*, 2015, doi:10.1038/nature.2015.17433.

Bode, Katherine. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly*, vol. 78, no. 1, 2017, pp. 77–106, doi:10.1215/00267929-3699787.

Bourrier, Karen and Thelwall, Mike. "The Social Lives of Books: Reading Victorian Literature on Goodreads." *Journal of Cultural Analytics*, vol. 5, no. 1, 2020, doi:10.22148/001c.12049.

Britt Holbrook, J., et al. "The Humanities Do Not Need a Replication Drive." *CWTS*, 2019, www.cwts.nl:443/blog?article=n-r2v2a4&title=the-humanities-do-not-need-a-replication-drive.

Bustin, Stephen. "The Reproducibility of Biomedical Research: Sleepers Awake!" *Biomol Detect Quantif*, 2015, doi: 10.1016/j.bdq.2015.01.002.

Claerbout, Jon F. and Karrenbach, Martin. "Electronic Documents Give Reproducible Research a New Meaning." *SEG Technical Program Expanded Abstracts*, 1992, doi:10.1190/1.1822162.

Donoho D. L., et al. "15 Years of Reproducible Research in Computational Harmonic Analysis." *Comput. Sci. Eng.*, vol. 11, no. 1, 2009, pp. 8–18, doi:10.1109/MCSE.2009.15.

European Commission. *Open Innovation, Open Science, Open to the World: A Vision for Europe*. 2016.

–. *Reproducibility of Scientific Results in the EU: Scoping Report*. Publications Office, 2020.

Fitzpatrick, Kathleen. "The Humanities, Done Digitally." *Debates in the Digital Humanities*, University of Minnesota Press, 2012.

Goodman, Steven N., et al. "What Does Research Reproducibility Mean?" *Science Translational Medicine*, vol. 8, no. 341, 2016.

Ioannidis, John P. A. "Why Most Published Research Findings Are False." *PLOS Medicine*, vol. 2, no. 8, Aug. 2005, p. e124. *PLOS Journals*, doi:10.1371/journal.pmed.0020124.

Klee, Paul. *Schöpfersiche Konfession*. Eric Reiss, 1920.

KNAW. *Replication Studies: Improving Reproducibility in the Empirical Sciences*. 2018.

Knuth, Donald. *Literate Programming*. CSLI, 1992.

Lowndes, Julia S. Stewart, et al. "Our Path to Better Science in Less Time Using Open Data Science Tools." *Nature Ecology & Evolution*, vol. 1, no. 6, 6, May 2017, pp. 1–7. *www.nature.com*, doi:10.1038/s41559-017-0160.

Munafò, Marcus R., et al. "A Manifesto for Reproducible Science." *Nature Human Behaviour*, vol. 1, no. 21, 2017, doi:10.1038/s41562-016-0021.

OECD. *Making Open Science a Reality*. OECD Science, Technology and Industry Policy Papers, Organisation for Economic Co-operation and Development, 205AD. *OECD*, www.oecd-ilibrary.org/content/workingpaper/5jrs2f963zs1-en.

Open Science Collaboration. *Estimating the Reproducibility of Psychological Science*. no. 6251, 2015, doi:10.1126/science.aac4716.

Open Source Initiative. *Ten Takeaways from the 2022 State of Open Source Survey*. 2022, opensource.org/ten-takeaways-from-the-2022-state-of-open-source-survey.

O'Sullivan, James. "The Humanities Have a 'Reproducibility' Problem." *Talking Humanities*, 9 July 2019, talkinghumanities.blogs.sas.ac.uk/2019/07/09/the-humanities-have-a-reproducibility-problem/.

Peels, Rik. "Replicability and Replication in the Humanities." *Research Integrity and Peer Review*, vol. 4, 2019, p. 2. *PubMed*, doi:10.1186/s41073-018-0060-4.

Plesser, Hans E. "Reproducibility vs. Replicability: A Brief History of a Confused Terminology." *Frontiers in Neuroinformatics*, vol. 11, Jan. 2018, p. 76. *PubMed Central*, doi:10.3389/fninf.2017.00076.

Rougier, Nicolas P., et al. "Sustainable Computational Science: The ReScience Initiative." *PeerJ Computer Science*, vol. 3, Dec. 2017, doi:10.7717/peerj-cs.142.

Vicente-Saez, Ruben and Martinez-Fuentes, Clara. *Open Science Now: A Systematic Literature Review for an Integrated Definition*. doi:10.1016/j.jbusres.2017.12.043.

Wilkinson, Mark D., et al. *The FAIR Guiding Principles for Scientific Data Management and Stewardship*. 2016, www.nature.com/articles/sdata201618.