Reproducibility and explainability in digital humanities

Ries, Thorsten; van Dalen-Oskam, K.H.; Offert, Fabian

published in
International Journal of Digital Humanities
2023

document version
Publisher's PDF, also known as Version of record

International Journal of Digital Humanities (2023) 5:247–251
https://doi.org/10.1007/s42803-023-00078-7

INTRODUCTION

Reproducibility and explainability in digital humanities

Thorsten Ries[1] · **Karina van Dalen-Oskam**[2,3] · **Fabian Offert**[4]

Published online: 4 December 2023

Keywords Reproducibility · Explainability · Digital humanities · Methodology · XAI

This special issue *Reproducibility and Explainability in Digital Humanities* of the *International Journal of Digital Humanities* (IJDH) has been planned in the wake of the controversy about Nan Z. Da's article "The Computational Case against Computational Literary Studies" in *Critical Inquiry*, 2019 (Da, 2019a, b). While many contributions to this discussion were keen on proving that Da's criticism was either methodologically flawed or biased, that it focused on a small spectrum of studies, and disregarded the exploratory purpose of some methods, many also implicitly or explicitly conceded that reproducibility and explainability were fundamental issues of several digital methods that deserved more methodologically rigorous attention. Furthermore, it was repeatedly stated that DH is currently missing a theoretical framework to interpret and critically evaluate "data" and the results of digital methods (Algee-Hewitt et al., 2019; Arnold and Buell, 2019). Following this rationale, the present special issue of IJDH provides a forum to find answers to the question "How can DH design research methodologies, procedures, and tools that provide reproducibility and explainability in order to safeguard methodological rigor, verifyability, and trans-

✉ Thorsten Ries
thorsten.ries@austin.utexas.edu

1 Department of Germanic Studies, University of Texas at Austin, Austin, Texas, USA

2 Huygens Institute KNAW, Department of Computational Literary Studies, Amsterdam, The Netherlands

3 Department of Dutch Studies, University of Amsterdam, Amsterdam, The Netherlands

4 Department of Germanic & Slavic Studies, University of California, Santa Barbara, Santa Barbara, California, USA

parency?" The articles have been drafted, peer-reviewed, and published during the advent of Large Language Models (LLM) such as GPT-3, 3.5, 4, Llama-2, Bard, and Alpaca, and others, in the years 2022 and 2023, which fundamentally change the game on multiple technological and conceptual levels, but also stress the timeliness of this special issue. Some of the contributions, for instance Chun and Elkins (2023), already reflect this recent and rapid development. As the hype of LLM development and the ensuing advances of other types of AI models will likely lead to a more widespread adoption of AI in DH applications (e.g. via LangChain and similar combined AI stacks and approaches), the core issues that this special issue and its online collection *Reproducibility and Explainability in Digital Humanities* address will become even more prevalent within DH. At present, we observe consequential shifts in the conceptual role of the term "model" in DH, e.g. as critical modeling of subjectivity (Underwood, 2020), or as subject of humanities criticism (Bode, 2020).

While around the time and in the wake of the 2019 debate the first studies emerge involving critical reproduction of results of previous research in the context of tools criticism (e.g. van Es et al., 2018; van Es, 2023; Herrmann et al., 2023), we see that also in the sciences that standards, parameters, methods, and even the subject of meaningful reproducibility, especially AI-based research, are constantly subject of domain-specific research and debate (Heil et al., 2021). This is exactly the methodological question that Christof Schöch is addressing for the humanities in his contribution on cycles of "repetitive research" and the role of the dimensions "question", "method", "data" in this special issue (Schöch, 2023). Toby Burrows critically explores the role of verifiability of sources, and of the impact of several digital methods on reproducibility in historical research (Burrows, 2023). Joseph Rudman argues for a high forensic standard for studies to be "repeatable, reproducible, and accurate", and explores which elements are needed to meet this standard (Rudman, 2023). The contributions by Sarah Middle, Nabeel Siddiqui, Samuel Huskey, Jyothi Justin and Nirmala Menon survey approaches to advance reproducibility by promoting standards for verifyable research procedures, methods, and datasets, and explore the role of open, shareable, reproducible workflows in international research settings (Middle, 2023; Siddiqui, 2023; Huskey, 2023; Justin and Menon, 2023).

Machine learning and artificial intelligence models are "black boxes", and the representational problem of this opacity has been discussed in multiple recent studies in the area of DH (Fazi, 2021; Offert, 2023). Research on the problem of "explainability" or "interpretability", especially of machine learning based methods and AI models (Explainable AI, XAI), is currently a very active research area in Computer Science. Many of the current approaches to trace the decisions of predictions of AI's are implemented in packages such as SHAP, FastSHAP and XGBoost (Lundberg, 2023; Covert and Lee, 2021; Jethani et al., 2022; Lin et al., 2023; Covert et al., 2022). On the humanities side, this development motivated a series of publications that turn to redefined explainability concepts that share little ground with the aforementioned Computer Science research, but seek to establish a humanities research area of "critial AI" (Goodlad, 2023; Bode and Goodlad, 2023) and "visual epistemology" (Drucker, 2020) in its own right (Berry, 2023; Floridi and Chiriatti, 2020; Floridi, 2023). The contributions in this special issue take the computational XAI route in combination with humanities approaches to open the "black boxes" and explore the "grey boxes". In

the present special issue, James Dobson gives an in-depth introduction to XAI methods and techniques from Computer Science research for "Reading and Interpreting Black Box Deep Neural Networks", arguing for a "research program informed by tool criticism in which the use of computational tools is conceived of as a metainterpretive act" (Huskey, 2023). Jon Chun and Katherine Elkins propose an advanced XAI approach and workflow for LLM-based research (GPT-4) using diachronic text sentiment analysis and narrative generation (Chun and Elkins, 2023). El-Hajj, Eberle, Merklein et al. contribute an equally advanced visual XAI approach for historical research on primary visual sources (history of the sciences) (El-Hajj et al., 2023). Pandiani, Lazzari, van Erp, and Presutti are using computer vision and visual XAI technology to interpret the ARTstract dataset (Pandiani et al., 2023).

This special issue *Reproducibility and Explainability in Digital Humanities* is published in a two-stage process: first, in print form, and second, as online collection on the website of the journal, where a few more articles will be added.

The editors of this special issue would like to thank all authors for the valuable contributions we were privileged to edit and publish. We also thank all peer reviewers for their important work.

References

Algee-Hewitt, M. A., Bode, K., Brouillette, S., Finn, E., Klein, L., Long, H., et al. (2019). Computational literary studies: a Critical Inquiry online forum. Available from: https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/

Arnold, T., & Buell, D. (2019). More responses to "the computational case against computational literary studies". Available from: https://critinq.wordpress.com/2019/04/12/more-responses-to-the-computational-case-against-computational-literary-studies/

Berry, D. (2023). The Explainability Turn. *Digital Humanities Quarterly, 17*(2)

Bode, K. (2020). Why you can't model away bias. *Modern Language Quarterly,81*(1), 95–124. https://doi.org/10.1215/00267929-7933102. https://read.dukeupress.edu/modern-language-quarterly/article-pdf/81/1/95/1567705/95bode.pdf

Bode, K., & Goodlad, L. M. E. (2023). Data worlds: an introduction. *Critical AI, 1*(1-2). https://doi.org/10.1215/2834703X-10734026

Burrows, T. (2023). Reproducibility, verifiability, and computational historical research. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00068-9

Chun, J., & Elkins, K. (2023). eXplainable AI with GPT4 for story analysis and generation: A novel framework for diachronic sentiment analysis. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00069-8

Covert, I., & Lee, S. I. (2021). Improving kernelshap: Practical shapley value estimation using linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* vol. 130 (pp. 3457–3465). PMLR

Covert, I., Lundberg, S., & Lee, S. I. (2022). Feature removal is a unifying principle for model explanation methods. arXiv. arXiv:2011.03623. [cs.LG]

Da, N. Z. (2019). The digital humanities debacle. *The Chronicle of Higher Education*

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry, 45*(3), 601–639. https://doi.org/10.1086/702594

Drucker, J. (2020). Visualization and interpretation: Humanistic approaches to display. *MIT Press*. Available from: https://doi.org/10.7551/mitpress/12523.001.0001

El-Hajj, H., Eberle, O., Merklein, A., Siebold, A., Shlomi, N., Büttner, J., et al. (2023). Explainability and transparency in the realm of digital humanities: toward a historian XAI. *International Journal of*

Digital Humanities. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00070-1

Fazi, M. B. (2021). Beyond human: deep learning, explainability and representation. *Theory, Culture & Society, 38*(7–8), 55–77. https://doi.org/10.1177/0263276420966386

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology,36*. https://doi.org/10.1007/s13347-023-00621-y

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds & Machines, 30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Goodlad, L. M. E. (2023). Editor's introduction: humanities in the loop. *Critical AI, 1*(1-2). https://doi.org/10.1215/2834703X-10734016

Heil, B. J., Hoffmann, M. M., Markowetz, F., Lee, S. -I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 1122–1144. https://doi.org/10.1038/s41592-021-01256-7

Herrmann, B., Bories, A. S., Frontini, F., Jacquot, C., Pielström, S., Rebora, S., et al. (2023). Tool criticism in practice. On methods, tools and aims of computational literary studies. *Digital Humanities Quarterly, 17*(2)

Huskey, S. (2023). Committing to reproducibility and explainability: using git as a research journal. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities

Huskey, S. (2023). On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities

Jethani, N., Sudarshan, M., Covert, I. C., Lee, S. I., & Ranganath, R. (2022). FastSHAP: real-time shapley value estimation. In *International Conference on Learning Representations*. Available from: https://openreview.net/forum?id=Zq2G_VTV53T

Justin, J., & Menon, N. (2023). Reproducibility of Indian DH projects: a case study. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00071-0

Lin, C., Covert, I., & Lee, S. I. (2023). On the robustness of removal-based feature attributions. arXiv. arXiv:2306.07462. [cs.LG]

Lundberg, S. (2023). SHAP documentation. Available from: https://shap.readthedocs.io/

Middle, S. (2023). A documentation checklist for (Linked) humanities data. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00072-z

Offert, F. (2023). Can we read neural networks? Epistemic implications of two historical computer science papers. *American Literature,95*(2), 423–428. https://doi.org/10.1215/00029831-10575218. https://arxiv.org/abs/https://read.dukeupress.edu/american-literature/article-pdf/95/2/423/1891570/423offert.pdf

Pandiani, D. S. M., Lazzari, N., van Erp, M., & Presutti, V. (2023). Hypericons for interpretability: decoding abstract concepts in visual data. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities

Rudman, J. (2023). Reproducibility and non-traditional authorship attribution: Invitatio ad Arma. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00067-w

Schöch, C. (2023). Repetitive research: a conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00073-y

Siddiqui, N. (2023). Minimal research compendiums: an approach to advance statistical validity and reproducibility in digital humanities research. *International Journal of Digital Humanities*. Special issue: Reproducibility and Explainability in Digital Humanities. https://doi.org/10.1007/s42803-023-00074-x

Underwood, T. (2020). Machine learning and human perspective. *PMLA, 135*(1), 92–109. https://doi.org/10.1632/pmla.2020.135.1.92

van Es, K. (2023). Unpacking tool criticism as practice, in practice. *Digital Humanities Quarterly, 17*(2)

van Es, K., Wieringa, M., & Schäfer, M. T. (2018). Tool criticism: from digital methods to digital methodology. In *Proceedings of the 2nd International Conference on Web Studies* (pp. 24–27). WS.2 2018.

New York, NY, USA: Association for Computing Machinery. Available from: https://doi.org/10.1145/3240431.3240436