Contribution ID: **158** Type: **Presentation**

AI-Ready Data Workflows for Social Science and Humanities

Tuesday, October 14, 2025 12:03 PM (11 minutes)

Recent advancements in artificial intelligence (AI) and access to new types of data have led to increased applications of AI in computational social science and humanities (SSH). A wide range of cutting-edge examples shows the results of bringing AI and SSH together, from the latest computer vision AI models used to detect archaeological traces in satellite imagery or to identify mounds on historical maps [1], to recent language models used to analyze social network behaviors1 or to perform longitudinal studies on the entire scientific corpus between others [2]. Furthermore, multimodal AI solutions combine different data types and are rapidly gaining popularity in social science and humanities. For instance, they are being used to generate language models capable of understanding ancient regional languages, thereby helping to enhance our understanding of history [3].

Despite AI's undeniable benefits in these fields, there are plenty of challenges to solve. Due to its inherent complexity and the need for high computational power, social and humanities scientists usually face a huge entry barrier to integrating these methods into their research [4]. In addition, using AI methods entails a set of dangers that must be carefully considered through proper guardrails and validation methodologies [5] over the AI models and the data used to train them. Still, they are usually tied to specific use cases and continuously evolve in parallel with the evolution of AI technologies, which may be difficult for scientists to follow.

To address this challenge, we have been developing a set of computational workflows for social science and humanities at the Barcelona Supercomputing Center. Since every computational experiment can be described as a workflow, creation, execution, tracing, and validation techniques for workflows become essential to address the problems mentioned above. The proposed presentation aims to review the opportunities and challenges in the workflow design and gather insights from the data expert community.

A main goal of the workflows is to increase accessibility to the use of AI and computational power. First, the workflows aim to lower technical barriers by abstracting complexity as much as possible, letting scientists focus on the iterations between research questions, results, and refinements. Second, the workflows aim to optimize costs and allow for scaling-up experiments by optimally using shared public computational infrastructures (e.g., Exascale Supercomputers belonging to the EuroHPC network), making cutting-edge research more accessible and affordable to a broader community. The workflows community [6] tackles many of these challenges to lower the burden for end users while enabling the efficient and scalable execution of computational experiments, such as the convergence of AI and HPC workflows, the support of multi-facility workflows, exploiting heterogeneous HPC environments (GPUs, NVMs), and the achievement of FAIR computational workflows, among others.

However, workflows are not only built of technical components that abstract code complexity or optimize costs and performance. Workflows should also aim to provide state-of-the-art validation and guardrails techniques for integrating AI responsibly into scientific research. To this end, the goal is to facilitate the adoption of best practices throughout the validation of the AI models' outputs and the data curation and documentation. For instance, leveraging current AI-ready and machine-actionable metadata initiatives, such as Croissant [7] and DDI-CDI (project homepage https://zenodo.org/records/11236871), we can make the research data more interoperable and discoverable by design. Another key technique is what in the literature is known as eXplainable AI (XAI) [8], which encompasses the capture of relevant metadata during the execution of AI experiments that later helps to understand in detail both training and inference processes of AI algorithms. XAI techniques provide a way for users to learn not only how AI models have been trained, but also to better trust and understand the decisions taken by AI systems.

One of the pillars of science is to enable research reproducibility. Workflows, and in particular, tracking the provenance within a workflow, are key to enabling automatic computational reproducibility, as shown in [9].

By establishing repeatable methods through workflows, we can define detailed logs and provenance records of the experiments, making them computationally reproducible by design and facilitating posterior verification processes. By integrating community-driven specifications such as RO-Crate [10] approaches into workflows and high-performance computing environments [11, 12], we can make computational AI workflows reproducible on demand. This work can allow computational social and humanities scientists to verify prior AI-based studies with less complexity.

In conclusion, our work aims to address the following issues with AI-Ready Data workflows: (i) Lowering the entry costs and reducing technical barriers for social and humanities scientists; (ii) Aiding in the adoption of best practices during data curation, preparation, and the validation of AI outputs; and (iii) Improving research reproducibility and posterior audit processes by integrating the AI workflows with RO-Crate-like solutions. While our proposal represents an initial effort to accelerate research and innovation in computational social science and humanities, maturing these AI workflows requires domain-specific expertise and a large variety of use cases. Consequently, we advocate for creating open workflows as a community-driven effort to build better validation methodologies and practices that can be shared and utilized by a wide research community.

References:

[1] I.Berganzo-Besga, et.al, Scientific Reports,2023.
[2] A.Castro Torres et.al., ICCSI,2025.
[3] M.Coll-Ardanuy et.al., Digital Humanities Conference,2025.
[4] J.Calder, et.al., IEEE BITS,2022.
[5] C.A.Bail, PNAS,2024.
[6] R.F.Da Silva et.al., arXiv:2410.14943,2024.
[7] M.Akhtar et.al., NeurIPS,2024.
[8] R.Dwivedi et.al., ACM Computer Surveys,2023.
[9] R.Sirvent et.al., IEE/ACM WORKS,2022.
[10] S.Soiland-Reyes et.al., Data Science,2022.
[11] S.Leo et.al., PLoS One,2024.
[12] R.F.Da Silva et.al., Computer,2024.

Primary authors: GINER MIGUELEZ, Joan (Barcelona Supercomputing Center); Dr SIRVENT, Raül (Barcelona Supercomputing Center); Mr LERGA FELIP, Eudald (Barcelona Supercomputing Center); Dr BADIA, Rosa M. (Barcelona Supercomputing Center); Dr CROSAS, Mercè (Barcelona Supercomputing Center)

Presenter: GINER MIGUELEZ, Joan (Barcelona Supercomputing Center)

Session Classification: Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI

Track Classification: SciDataCon2025 Specific Themes: Rigorous, responsible and reproducible science in the era of FAIR data and AI