Contribution ID: **139** Type: **Presentation**

Building a data infrastructure for Social Science and Humanities: A double perspective on quality and community from Italy and France

Thursday 16 October 2025 12:17 (11 minutes)

The Social Sciences and Humanities (SSH) disciplinary sector encompasses research studies that share an epistemological commitment to the critical investigation of human experience, cultural expression, and social organization. SSH research contributes to the development of theoretical frameworks and methodological approaches that are essential for understanding complex societal transformations.

European institutions, including the European Commission, recognized the importance of SSH researchers and research projects by integrating it into broader research frameworks, such as Horizon Europe and NextGeneration EU, and supporting international cooperation through initiatives like the European Research Infrastructure Consortium (ERIC). Yet, despite this support, SSH research faces distinct challenges when compared to STEM disciplines:

1. sparsity and fragmentation of its data;
2. heterogeneity of sources - textual, audiovisual, or contextual - generated under diverse, often non-standard conditions;
3. isolated datasets within institutional silos, limiting their discoverability, accessibility and reuse.

Such variability underscores a systemic challenge in ensuring data quality across SSH research, which involves evaluating the accuracy, completeness, consistency, and documentation of data to enable its meaningful interpretation, reuse, and long-term preservation.

Moving towards data-intensive and AI empowered research, RIs will see a further shift of their role from data suppliers to primary users of machine enabled workflows. To ensure sustainability of workflows, given by their reproducibility and replicability, data quality becomes a fundamental aspect of good research and reliable results, in particular in the case of AI mediated processes and workflows.

DARIAH.it, as part of the H2IOSC project, and DARIAH-FR, particularly through the IR* Huma-Num, are two national research infrastructures that exist to serve SSH researchers and research projects in their respective countries. Both initiatives are inscribed in the ERIC DARIAH-EU, which seeks to structure the development and interaction of national research infrastructures in Europe. Though both Italy and France have different approaches, the common problem remains: how to improve quality of data, metadata, paradata and related tools based on an approach involving users?

Italy

In Italy, DARIAH.it is working to strengthen the national infrastructure, focusing on promoting standards for data, services and workflows, as well as developing platforms to support users in producing and managing FAIR data and resources, access a wide range of services and combine them into meaningful scientific workflows requiring the interaction of different & independent services, also leveraging on AI modules (i.e. AI-mediated DH).

Recently, due to the increasing number of attacks brought to different institutions (British Libraries) and resources (e.g.: Archive.org) dealing with cultural content - Italy started a process of transition towards Critical Infrastructures for SSH, bringing cybersecurity and resilience into DARIAH.it national infrastructure development plans.

Being a socio-technological environment, and aiming to bring measurable advancements for the research community, the upgrade of the technical infrastructures requires a strong investment on the human component: RIs can support research by providing quality data, tools and processes but researchers are encouraged to be the owner of their own algorithms and to know how they work to ensure they get the right answers to the right questions, hence transparence, explainability, standardization, alignment and training are the drivers to successfully complete this transition.

France

Huma-Num, the French national infrastructure for SSH has built a robust technical research infrastructure fed by community feedback. After its first decade, certain aspects need to be updated, most notably, the approach to data & metadata quality in its research data repository, NAKALA. Considering the large amount of existing data in NAKALA (over 1.4 million files), it was deemed unfeasible to manually curate all existing and future deposits..

To implement this quality plan, and in addition to purely technical controls, Huma-Num relies on communities to develop a network of curators, in articulation with the national ecosystem Recherche Data Gouv„ with their harmony assured by the development of a curation guide. Alongside this, Huma-Num has extensively developed its documentation, with a focus on data preparation, and has organized a series of webinar sessions for users. Finally, Huma-Num launched a global content analysis of the repository to gain a better understanding of current practices and develop quality indicators.
By combining technical developments with community feedback to improve quality, Huma-Num gradually evolves from a purely technical infrastructure to a knowledge infrastructure.

Conclusion

This paper will develop the two Italian and French approaches to building an infrastructure for SSH centered on quality and sustainability. The problems to be solved are quite similar despite differences in national organizations, but the crucial common point is the necessary involvement of users, without which actions are doomed to failure. These dual approaches are fruitful examples for others that are confronted with the thorny problem ensuring research data and metadata quality in SSH.

References

- Bellini, Emanuele and Emiliano Degl'Innocenti. 2024. Transitioning SSH European Research Infrastructures to Critical Infrastructure Through Resilience. IEEE International Conference on Cyber Security and Resilience (CSR): pp. 801-806, doi: 10.1109/CSR61664.2024.10679383.
- Edmond, Jennifer, ed. 2020. "Digital Technology and the Practices of Humanities Research."Open Book Publishers. https://doi.org/10.11647/obp.0192.
- Gray, Edward J., Nicolas Larrousse. Huma-Num IR*, 10 Years of Building a Research Infrastructure at the European level. Huma-Num, 10 Years of Building a Research Infrastructure at the European level, 2024. ⟨halshs-04573643⟩
- Lacagnina, Carlo, et al. «TOWARDS A DATA QUALITY FRAMEWORK FOR EOSC». Zenodo, 9 January 2023. https://doi.org/10.5281/zenodo.7515816.
- Spadi, Alessia, Emiliano Degl'Innocenti, and Carmen Di Meo. 2024. «DARIAH.It: Data Integration Strategies and Solutions for Digital Resources Management and Research in the Arts and Humanities». Mimesis Journal 13 (2):119-34. https://doi.org/10.13135/2389-6086/9920.

Primary authors: SPADI, Alessia (Consiglio Nazionale delle Ricerche (CNR), Italy); GRAY, Edward (IR* Huma-Num / DARIAH-EU); LARROUSSE, Nicolas (Huma-Num CNRS); Dr DEGL'INNOCENTI, Emiliano (CNR, DARIAH-IT)

Presenter: GRAY, Edward (IR* Huma-Num / DARIAH-EU)

Session Classification: Presentations Session 10: Infrastructures to Support Data-Intensive Research - Local to Global

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global