Towards interoperability: Introducing the Impresso data lab for the enrichment and analysis of historical media

Düring, Marten
marten.during@uni.lu
C2DH, Luxemburg
ORCID: 0000-0001-7411-771X

Guido, Daniele
daniele.guido@uni.lu
C2DH, Luxemburg
ORCID: 0000-0003-1601-4274

Kalyakin, Roman
roman.kalyakin@ext.uni.lu
C2DH, Luxemburg

Humanities research practices are commonly characterised by iterative workflows using a wide variety of different sources and modalities (text, image, audio, video, objects, ..) which stands in contrast to typically rather static and homogeneous cultural heritage collections. This challenge is increasingly addressed by computational DH research projects and infrastructures (“Twi-XL Project” 2024; “CLARIAH Media Suite” 2024; Kemman and Kleppe 2013) .

This half-day workshop invites DH scholars and librarians to explore the intersection between four fast-paced developments which have the potential to improve the interoperability between static infrastructures and dynamic research needs:

First, changing DH research practices are shifting towards computational processing. We define the subfield “computational humanities” as a distinct user group of computer-savy humanists who wish to analyse cultural (heritage) data at scale harnessing advanced methods from data science and machine learning. Within the “big tent” Digital Humanities community, the decidedly computational analysis of cultural heritage data has emerged as a vibrant subfield with dedicated journals and conference series (Karsdorp, Kestemont, and Riddell 2021; “Journal of Cultural Analytics,” n.d.; “Computational Humanities Research 2024,” n.d.) .

Second, NLP-based semantic enrichment of cultural heritage data collections begins to happen at scale. Several projects have successfully combined NLP and historical research interests to advance the study of historical newspaper collections using large scale semantic enrichments such as topic modelling, text reuse detection or named entity recognition, see e.g. (Langlais 2019; Doucet et al. 2020; Cordell and Smith 2024; Ahnert and Demertzi 2023; Impresso 2024) .

Third, the integration of LLMs in research workflows. In parallel, generative AI and especially Large Language Models are beginning to find their place in (Digital) Humanities research practices (see e.g. (Pichler and Reiter 2024; Karjus 2023)) and sharing platforms such as Hugging Face (“Hugging Face – The AI Community Building the Future.” 2024) are transforming the way researchers engage with cultural heritage data.

Fourth, the development of “data labs” by cultural heritage institutions to enable access to machine-readable data. This paradigm shift aligns with the The “Always already computational. Collections as data" paradigm coined by a research project of the same name that has since 2016 received a strong resonance among libraries, archives and other GLAM institutions worldwide (“Always Already Computational” 2024) . Many of them strive to offer access to their data using dedicated platforms. (Candela et al. 2022) . The scope and capabilities of such platforms vary considerably and are shaped by copyright concerns: The spectrum ranges from closed computing infrastructures such as the HathiTrust Research Centre (“HTRC Analytics” 2024) or commercial platforms like Constellate (“Constellate” 2024) to publicly accessible public domain data such as those provided by the National Library of Scotland’s Data Foundry (Ames 2021) .

Taken together, these developments constitute an opportunity to overcome some of the main limitations researchers face when dealing with diverse data and large scale collections: Pre-defined corpora such as those provided by infrastructures do not often match the needs of researchers; NLP-based enrichment of large collections is resource-intensive and complex; copyright-concerns restrict access; and exploratory and analytical tools developed for platforms typically focus on generic tasks rather than question-specific ones.

Using the Impresso data lab (public release in October 2024) as a starting point, we will discuss how the strengths of research infrastructures (e.g. access to data, large-scale processing) can be combined with researcher needs (e.g. specific datasets and objectives) to increase interoperability.

The workshop is organised by the interdisciplinary research project Impresso Media Monitoring of the Past — Beyond Borders which leverages an unprecedented corpus of newspaper and radio archives and uses machine learning to pursue a paradigm shift in the processing, semantic enrichment, representation, exploration and study of historical media across modalities, time, languages, and national borders (Impresso 2024) .

The Impresso data lab offers access to a growing Western European newspaper and radio corpus. It has two primary purposes: First, to complement the inherently limited analytical capabilities of the Impresso web app by enabling flexible computational analysis via API (see also (Kemman and Claeyssens 2022)), a dedicated Python library and an environment of interactive Jupyter notebooks. Second, to respond to user needs to freely link and analyse external research data to Impresso using a variety of semantic enrichments such as named entities or topics.

To this end it offers access to a dedicated API that also enables document annotation services. This with the goal to establish a transparent and versatile framework for data-driven comparative analysis of internal and research-specific external documents. Overall, the data lab provides the following services:

- API accessibility: Opening the corpus, enrichments, and tools for programmatic exploitation.
- Dynamic research workflows: Experimenting with modular, dynamic, and personalised research workflows to bridge the gap between digitised collections and data-driven historical research.
- Annotation service: Enabling the enrichment of user-provided documents by project-based NLP components, supporting e.g. topic models, named entity recognition, keyphrase extraction, and vectorized representations.
- Enrichment import: Allowing users e.g. to import external enrichments of project documents and empowering researchers to work with self-generated topics in the interface.

The user-oriented API provides convenience modules for programmatic exploitation of data and enables researchers to process the Impresso corpus and enrichments along with other documents and using (external) libraries relevant to their research needs. In other words, while the Impresso web app offers powerful search, filter and curatorial capabilities, the data lab offers space for question-specific operations with the data which go beyond the capabilities of a generic user interface.

Agenda

The workshop strives to balance instructive segments with free exploration to foster active debate among participants. We will begin with an introduction to the Impresso project, an overview of the current landscape and the results of a survey of data labs in the cultural heritage sphere.

Brief demos will showcase Impresso's capabilities. This includes an overview of the Impresso web app and its main search and filter operations, a first data lab demo to show how users can adapt and create Jupyter notebook templates to model, curate and visualise spatial and relational data and thereby generate their own, personalised views on the Impresso corpus. A second demo will present opportunities to enrich external sources using annotation services (named entity recognition, press agency detection, topic model inferences) and to link them back to the Impresso corpus.

During hands-on sessions, participants will have the opportunity to test the web app and notebook templates for their real-world utility.

It is in the nature of things that neither data nor methods are perfect. Demo and hands-on sessions are intended as starting points for exchanges among participants about links between research infrastructures and the needs of the researcher community. This with the goal to identify and document strengths, shortcomings, opportunities and aspirations for future work against the background of real-world needs, pragmatics and restrictions.

The workshop is structured as follows:

20": Welcome, introduction round and setup.

30": Where we stand today: Impresso2 and results from a survey of cultural heritage data labs

20": Demo I: From web app to data lab - Extracting and exploring spatial and network data in interactive notebooks

45": Hands-on experimentation with notebook templates

20": Exchange among participants

15": Break

15": Demo II: Enriching external sources via data lab - Named entity recognition, topic modelling and press agency detection

50": Hands-on experimentation with notebook templates

25": Exchange among participants and summary of discussion

We invite all members of the DH community. Coding experience is an asset but not a requirement to participate. The number of participants is limited to 25.

Convenors

Marten Düring holds a PhD in Contemporary History from the University of Mainz and is part of the C2DH's Digital History Unit. Marten's research focuses on contemporary history, digital history and more specifically network analysis in the historical disciplines. Within Impresso, Marten contributes to project management, coordinates interface development, digital history research, and project dissemination and contributes a historical case study on discourses surrounding nuclear power technologies in historical media.

Daniele Guido is a designer and full-stack developer specialising in data visualisation, network visualisation and digital methods. He designs and develops experimental web applications and tools to improve information retrieval in the digital humanities. Within Impresso, Daniele is responsible for the design and development of the Impresso web app and the Impresso data lab.

Roman Kalyakin is a full-stack engineer. He holds an M.Sc in Electrical Engineering and Audiovisual technologies from Saint Petersburg State University of Aerospace and Instrumentation. Roman has extensive experience in building and managing software products for startups

and large enterprise companies. Roman has a strong background in machine learning, having worked on projects involving ML pipelines, graph databases, and comprehensive visualization tools. Within Impresso, Roman is responsible for the design and development of the Impresso API, Impresso data lab, Impresso web app and DevOps.

Bibliographie

Ahnert, Ruth, and Léllé Demertzi. 2023. "Living with Machines Final Report." The Alan Turing Institute. https://doi.org/10.23636/PSQ5-6A91.

"Always Already Computational." 2024. Always Already Computational - Collections as Data. July 23, 2024. https://collectionsasdata.github.io/.

Ames, Sarah. 2021. "Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry." *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31 (1): 1–13. https://doi.org/10.18352/lq.10371.

Candela, Gustavo, María Dolores Sáez, MPilar Escobar Esteban, and Manuel Marco-Such. 2022. "Reusing Digital Collections from GLAM Institutions." *Journal of Information Science* 48 (2): 251–67. https://doi.org/10.1177/0165551520950246.

"CLARIAH Media Suite." 2024. July 23, 2024. https://mediasuite.clariah.nl/.

"Computational Humanities Research 2024." n.d. Accessed April 18, 2024. https://2024.computational-humanities-research.org/.

"Constellate." 2024. July 23, 2024. https://constellate.org.

Cordell, Ryan, and David Smith. 2024. "Oceanic Exchanges." *Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914* (blog). July 23, 2024. http://oceanicexchanges.github.io/.

Doucet, Antoine, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, et al. 2020. "NewsEye: A Digital Investigator for Historical Newspapers." In . https://hal.science/hal-03029072.

"HTRC Analytics." 2024. July 23, 2024. https://analytics.hathitrust.org/.

"Hugging Face – The AI Community Building the Future." 2024. July 19, 2024. https://huggingface.co/.

Impresso. 2024. "Media Monitoring of the Past." Impresso. impresso. July 23, 2024. https://impresso-project.ch/.

"Journal of Cultural Analytics." n.d. Accessed April 18, 2024. https://culturalanalytics.org/.

Karjus, Andres. 2023. "Machine-Assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence." arXiv.Org. September 24, 2023. https://arxiv.org/abs/2309.14379v1.

Karsdorp, Folgert, Mike Kestemont, and Allen Riddell. 2021. *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Kemman, Max, and Steven Claeyssens. 2022. "User Demand for Supporting Advanced Analysis of Historical Text Collections." May 30. https://doi.org/10.5281/zenodo.6595769.

Kemman, Max, and Martijn Kleppe. 2013. "PoliMedia." In *Research and Advanced Technology for Digital Libraries*, edited by Trond Aalberg, Christos Papatheodorou, Milena Dobreva, Giannis Tsakonas, and Charles J. Farrugia, 401–4. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-40501-3_46.

Langlais, Pierre-Carl. 2019. "Distant reading the French news with the Numapresse project: toward a contextual approach of text mining." *Numapresse* (blog). February 7, 2019. http://www.numapresse.org/2019/02/07/distant-reading-the-french-news-with-the-numapresse-project-toward-a-contextual-approach-of-text-mining/.

Pichler, Axel, and Nils Reiter. 2024. "»LLMs for Everything?«Potentiale Und Probleme Der Anwendung von In-Context-Learning Für Die Computational Literary Studies." In *Book of Abstracts of Dhd*. Passau, Germany.

"Twi-XL Project." 2024. July 23, 2024. https://twi-xl.humanities.uva.nl/.