DARIAH Annual Event 2024

Workflows: Digital Methods for Reproducible Research Practices in the Arts and Humanities

Book of Abstracts

Programme Committee

Toma Tasovac (Chair)
Andrea Scharnhorst (Co-Chair)
Kim Ferguson (Co-Chair, BoA Editor)
Daniel Alves (NOVA FCSH)
Marita Everhardt
Adeline Joffres
Amelia McConville
Anne Grésillon
Arnaud Roi
Carmen Di Meo
Elena Gigliarelli
Françoise Gouzi
Georgios Artopoulos
Laure Barbot
Maria Ilvanidou
Natalia Ermolaev
Sally Chambers
Tanja Wissik
Tibor Kálmán
Tomasz Parkoła
Tomasz Umerle

Links:
DARIAH AE Website Archive
DARIAH AE 2024 Zenodo Community
DARIAH AE 2024: Workflows (Video)

DARIAH-EU
Celebrating the first decade

Table of Contents

DOI: 10.5281/zenodo.13654378

Conference Schedule

Date: Tuesday, 18/June/2024

8:30am - 9:00am	**Welcome coffee**
9:00am - 1:00pm	**SSH Open Marketplace Workshop** Session Chair: **Laure Barbot**, DARIAH-EU; laure.barbot@dariah.eu Session Chair: **Michael Kurzmeier**, ACDH-CH; michael.kurzmeier@oeaw.ac.at
10:00am - 11:00am	**JRC + WG Chairs** Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl Session Chair: **Agiatis Benardou**, DARIAH-EU; a.benardou@dcu.gr
11:00am - 11:30am	**Morning coffee break**
11:30am - 1:00pm	**Joint NCC/JRC Meeting** Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl Session Chair: **Tibor Kálmán**, GWDG; tibor.kalman@gwdg.de
11:30am - 1:00pm	**WG Meeting: Research Data Management** Session Chair: **Francesco Gelati**, Universität Hamburg; francesco.gelati@uni-hamburg.de Session Chair: **Françoise Gouzi**, DARIAH Open Science Officer; francoise.gouzi@dariah.eu
1:00pm - 2:00pm	**Lunch break**
2:00pm - 3:30pm	**NCC Meeting** Session Chair: **Nanette Rissler-Pipka**, Max Weber Foundation; rissler-pipka@maxweberstiftung.de Session Chair: **Edward Joseph Gray**, CNRS; edward.gray523@gmail.com
2:00pm - 3:30pm	**WG Meeting: #dariahTeach** Session Chair: **Susan Schreibman**, Maastricht University; susan.schreibman@gmail.com Session Chair: **Costas Papadopoulos**, Maastricht University; k.papadopoulos@maastrichtuniversity.nl
2:00pm - 5:30pm	**Workshop: Emerging job profiles for DH Graduates: Bridging gaps between Industry & Education** Session Chair: **Amelia Sanz**, Complutense University of Madrid; amsanz@ucm.es Session Chair: **María Goicoechea**, University Complutense of Madrid; mgoico@filol.ucm.es
2:00pm - 5:30pm	**SSH Open Marketplace ATRIUM workshop** Session Chair: **Laure Barbot**, DARIAH-EU; laure.barbot@dariah.eu Session Chair: **Michael Kurzmeier**, ACDH-CH; michael.kurzmeier@oeaw.ac.at
2:00pm - 5:30pm	**Theatralia WG Workshop: "Developing a Controlled Vocabulary for Performing Arts Data"** Session Chair: **Anamarija Žugić Borić**, Institute of Ethnology and Folklore Research; zugicboric@ief.hr Session Chair: **Cécile CHANTRAINE BRAILLON**, La Rochelle Université; cecile.chantraine_braillon@univ-lr.fr
3:30pm - 4:00pm	**Afternoon coffee break**
4:00pm - 5:30pm	**WG Meeting: Digital Numismatics** Session Chair: **David George Wigg-Wolf**, Deutsches Archäologisches Institut; david.wigg-wolf@dainst.de Session Chair: **Rahel C. Ackermann**, Swiss Inventory of Coin Finds; rahel.ackermann@fundmuenzen.ch
4:00pm - 5:30pm	**WG Meeting: Multilingual DH** Session Chair: **Maroussia Bednarkiewicz**, University of Tübingen; maroussia.bednarkiewicz@uni-tuebingen.de Session Chair: **Alíz Horváth**, Eötvös Loránd University; aliz.horvath06@gmail.com

Date: Wednesday, 19/June/2024

8:30am - 9:15am	**Welcome coffee**
9:30am - 11:00am	**Opening & Keynote: Meredith Martin, "Worked Up About Data"** Session Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl
11:00am - 11:30am	**Morning coffee break**
11:30am - 1:00pm	**Tools and Workflows for Collaboration between Researchers, Research Data Management Professionals and GLAMs** Session Chair: **Francesco Gelati**, Universität Hamburg; francesco.gelati@uni-hamburg.de Session Chair: **Françoise Gouzi**, DARIAH Open Science Officer; francoise.gouzi@dariah.eu
11:30am - 1:00pm	**AI for DH** Session Chair: **Natalia Ermolaev**, Princeton University; nataliae@princeton.edu
11:30am - 1:00pm	**WG Meeting: Bibliodata** Session Chair: **Vojtěch Malínek**, Institute of Czech Literature, Czech Academy of Sciences; malinek@ucl.cas.cz Session Chair: **Tomasz Umerle**, Instytut Badań Literackich Polskiej Akademii Nauk; tomasz.umerle@ibl.waw.pl
11:30am - 1:00pm	**WG Meeting: DH Course Registry** Session Chair: **Iulianna Van der Lek**, CLARIN; i.vanderlek@uu.nl Session Chair: **María Goicoechea**, University Complutense of Madrid; mgoico@filol.ucm.es
1:00pm - 2:00pm	**Lunch break**
2:00pm - 3:30pm	**Discussing SSH Open Marketplace Workflows: User Experiences, Editorial Policy and Future Development** Session Chair: **Mikko Tolonen**, University of Helsinki; mikko.tolonen@helsinki.fi Session Chair: **Vojtěch Malínek**, Institute of Czech Literature, Czech Academy of Sciences; malinek@ucl.cas.cz
2:00pm - 3:30pm	**Organizing Knowledge** Session Chair: **Agiatis Benardou**, DARIAH-EU; a.benardou@dcu.gr
2:00pm - 3:30pm	**WG Meeting: Ethics and Legality in the Digital Arts and Humanities (ELDAH)** Session Chair: **Koraljka Kuzman Šlogar**, Institute of Ethnology and Folklore Research; koraljkak@gmail.com Session Chair: **Walter Scholger**, CLARIAH-AT; walter.scholger@uni-graz.at
3:30pm - 4:00pm	**Afternoon coffee break**
4:00pm - 5:00pm	**DARIAH Portugal Showcase** Session Chair: **Daniel Alves**, NOVA FCSH; dra@fcsh.unl.pt
5:00pm - 5:30pm	**Group photo**
5:30pm - 6:30pm	**Poster & Demo Session**
6:30pm - 7:30pm	**Working Group Community Meeting** Session Chair: **Kim Ferguson**, DANS; kim.ferguson@dans.knaw.nl Session Chair: **Marita Everhardt**, DANS-KNAW; marita.everhardt@dans.knaw.nl

Date: Thursday, 20/June/2024

8:30am - 9:00am	**Welcome coffee**
9:00am - 11:00am	**Auditory Workflows** Session Chair: **Marita Everhardt**, DANS-KNAW; marita.everhardt@dans.knaw.nl
9:00am - 11:00am	**Ethical Workflows** Session Chair: **Andrew Janco**, University of Pennsylvania; apjanco@upenn.edu
11:00am - 11:30am	**Morning coffee break**
11:30am - 1:00pm	**The RSE Turn in Digital Humanities** Session Chair: **Natalia Ermolaev**, Princeton University; nataliae@princeton.edu
11:30am - 1:00pm	**Documenting Workflows** Session Chair: **Agiatis Benardou**, DARIAH-EU; a.benardou@dcu.gr
1:00pm - 2:00pm	**Lunch break**
2:00pm - 3:30pm	**DiMPO WG Revival Meeting** Session Chair: **Costis Dallas**, Vilnius University; konstantinos.dallas@kf.vu.lt Session Chair: **Maciej Maryl**, Institute of Literary Research of the Polish Academy of Sciences; maciej.maryl@ibl.waw.pl
2:00pm - 3:30pm	**Literary Collections** Session Chair: **Anne Baillot**, DARIAH; anne.baillot@dariah.eu
2:00pm - 3:30pm	**Visual Cultural Heritage** Session Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be
3:30pm - 4:00pm	**Afternoon coffee break**
4:00pm - 5:00pm	**DARIAH Themes 2022-2024 Showcase** Session Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org Featuring updates and reports from projects that were funded under the 2022-2024 DARIAH Theme funding. "Open Bibliodata Workflows" - Vojtěch Malínek "MobileGIS workflow in archaeological prospection: the case study of a rural site near the Roman city of Mustis (N Tunisia)" - Julia Chyla "White paper and podcast on workflows in digital projects" - Lucas Burkart "Harmonizing Workflows in HTR/OCR Publication Pipelines of Textual Heritage" - Mareike König "MotiveR: a workflow designed to detect significant lexico-grammatical patterns for automatic stylistic analysis" - Antoine de Sacy
5:00pm - 6:00pm	**Poster & Demo Session**
7:00pm - 10:00pm	**Social Dinner in Lisbon**

Date: Friday, 21/June/2024

8:30am - 9:00am	**Welcome coffee**
9:00am - 11:00am	**Workflows for Cultural Heritage** Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl
11:00am - 11:30am	**Morning coffee break**
11:30am - 12:30pm	**DARIAH: Reflection on 10 Years** Session Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org
12:30pm - 1:00pm	**Closing Remarks** Session Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org
1:00pm - 1:45pm	**Brown Bag Lunch**
2:00pm - 3:30pm	**Women Writers in History WG Workshop** Session Chair: **Amelia Sanz**, Complutense University of Madrid; amsanz@ucm.es Session Chair: **Isabel Maria Lousada**, NOVAFCSH; isabel.lousada@fcsh.unl.pt

Opening Keynote

Time: Wednesday, 19/June/2024: 9:30am - 11:00am
Session Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org
Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl

Meredith Martin, "Worked Up About Data"

“Worked Up About Data” explores the history of and controversies around humanities data to help understand why the humanities have not been considered data-driven, even though scholars in these fields have long pioneered studying culture as essentially data and today are relying increasingly on digital platforms in their research. Looking at both critical histories of humanities data and the subsequent controversies about the concept of data as somehow opposed to humanism, I argue that our distraction about what counts, or doesn’t count, or shouldn’t count, as humanistic labor has obscured our ability to adequately track, theorize, and formalize our methods.

Recording: https://www.youtube.com/watch?v=7ua8EdspuyY

Presentations

Panel Session: Tools and Workflows for Collaboration between Researchers, Research Data Management Professionals and GLAMs

Time: Wednesday, 19/June/2024: 11:30am - 1:00pm
Session Chair: **Francesco Gelati**, Universität Hamburg
Session Chair: **Françoise Gouzi**, DARIAH Open Science Officer

Tools and Workflows for Collaboration between Researchers, Research Data Management Professionals and GLAMs

https://doi.org/10.5281/zenodo.12819051

Francesco Gelati[1], Tom Gheldof[2], Bianca Gualandi[3], Tugce Karatas[4], Françoise Gouzi[5], Ines Vodopivec[6], Johan Van der Eycken[7]

[1]Universität Hamburg, Germany; [2]Katholieke Universiteit Leuven, Belgium; [3]Alma Mater Studiorum – Università di Bologna, Italy; [4]Université du Luxembourg, Luxembourg; [5]DARIAH-EU; [6]National and University Library of Slovenia; [7]Belgian State Archives;
francesco.gelati@uni-hamburg.de, tom.gheldof@kuleuven.be, bianca.gualandi4@unibo.it, tugce.karatas@uni.lu, francoise.gouzi@dariah.eu

DARIAH Research Data Management Working Group is a place where researchers, Research Data Management (RDM) professionals and GLAM personnel have come together since 2020. On the basis of this interdisciplinary, inter-professional and inter-institutional collaboration, we would like to reflect in the panel on:

1. How open-source, by-researchers-for-researchers tools and their workflows shape collaborations, and establish trust and transparency in interdisciplinary communities (Tugce Karatas);
2. How brand-new research data repositories can create trust from scratch among researchers also thanks to their workflows (Johan Van Der Eycken and Tom Gheldof);
3. How RDM professionals connect researchers working on cultural heritage and the institutions involved, harmonising the respective needs, policies and workflows (Bianca Gualandi);
4. How inter-institutional scientific collaboration can be documented by means of proprietary and open-source tools (Francesco Gelati).

1) Creating a plan for digital infrastructure for research data involves several crucial steps to ensure the smooth transition and effective management of data for research institutions. This paper investigates the challenges of designing and implementing a structured workflow that can transform research data into a machine-actionable format using digital curation strategies where researchers can streamline data exchange processes in the context of contemporary historical research at the Luxembourg Center for Contemporary and Digital History (C2DH). It aims to dissect the intricacies of integrating digital tools and workflows to facilitate computational analysis, reproducibility, transparency, publication, metadata standardisation and dissemination. Furthermore, it delves into the infrastructural dimension of developing a workflow to facilitate the requirements deriving from research data management, data protection policies, and FAIR principles using several bespoke tools including Data Steward Wizard in collaboration with Luxembourg's National Data Service, Learning Center and Research Fund.

2) There are numerous online data repositories available, so researchers often get lost. However, one repository is not the same as another. Are the rights of the depositor guaranteed? What type of data is the repository for? What about sensitive data? Are there quality requirements?, etc. The uncertainty creates some distrust among researchers, especially when it comes to new repositories. Using the concrete example of www.sodha.be (the Belgian Social Sciences and Humanities data archive), which was developed in the framework of the ESFRI's CESSDA and DARIAH, we will explain step by step what measures we have taken to gain the trust of researchers.The aim of this paper is to explain the difficulties we faced (and sometimes still face), how we were able to identify them and what solutions we came up with. Finally, we want to address the challenges of maintaining the reliability of the platform. We will discuss We will discuss the current challenges and the next steps (technically, legally, organisationally and communicatively) to maintain the trust of the researcher.

3) RDM professionals, such as data stewards, act as conduits between researchers on one side and universities and GLAM institutions on the other. A sort of human infrastructure, they provide strategic guidance, and connect researchers to support services and technical solutions available within their institutions and beyond. We present the experience of the University of Bologna and particularly the collaboration currently happening within the NRRP project CHANGES - Spoke 4 "Virtual technologies for museums and art collections". The project investigates the different value-creation processes behind digital cultural heritage (DCH) and (in)tangible heritage, and the additional meaning that DCH brings into the knowledge and narrative of museums and art collections. We discuss how high-level data management strategies have been defined and a common data management plan is being maintained, creating a common framework for the many partners and GLAM institutions involved, and the several planned pilots.

4) The paper would like to reflect on how inter-institutional collaboration has been documented at the Hamburg University Archives.

- Strictly formalised but interoperable: the proprietary tool "Pure" is a Research Information Management System (RIMS). Interoperability is granted, since information can be exported and exchanged in the open-source file-format standard CERIF, but data is not freely available.
- Formalised but little interoperable: administrative and operational information is entered into the records management system (RMS) too. Since the RMS offer limited file-format conversion options, it seems we have a data silo with little data reusability;

- Little formalised but FAIR: Universitätsarchiv Hamburg's GitLab instance also contains metadata about scientific collaborations, e.g. about the research project "Integrating Voices of the Community". Data is available there in a FAIR but little structured form.

The presenter will invite the audience to interact and to share tools they use for this goal.

Paper session: AI for DH

Time: Wednesday, 19/June/2024: 11:30am - 1:00pm
Session Chair: **Natalia Ermolaev**, Princeton University

Building a Fichero: New Tools, Old Documents, and Machine Learning Workflows with an Endangered Afro-Colombian Archive

Andrew Janco[1], Kelly López Roldán[2], Daniel Tubb[3]
[1]University of Pennsylvania, United States of America; [2]Independent Researcher, Colombia; [3]University of New Brunswick, Canada; apjanco@upenn.edu

This paper describes outcomes and challenges in human-scale document processing. We discuss a workflow that begins with document preservation, moves through text recognition, and ends with a catalogue that demonstrates capabilities of LLMs for research and archival work, while remaining attuned to the vision of research partners.

Until 2022, the Istmina Circuit Court archive, with documents from the 1870s to 1930s, was rotting, disorganized, and in garbage bags. Yet, this archive is a crucial source of Afro-Colombian history in an often-marginalized region of the Chocó in Colombia. In 2023, seven young people from Istmina and Quibdó worked with the Muntú Bantú Foundation, a community center focused on Afro-diasporic memory. With researchers from various universities, they were able to digitize the archive, which is available online at the British Library. While the project was successful, the digitization has enabled new workflows to catalogue and interpret the archive. This paper explores these workflows.

Throughout, we are interested in a key problem of equity in knowledge production: How can new tools be used to the benefit of local knowledge-producers? Our paper focuses on the work of cataloguing archival materials, a first step in enabling local researchers (and others) to make meaning. Project interns catalogued 330 Case Files and wrote a book of micro-history. Yet, the task of cataloguing 470 more cases remains daunting. We reflect on machine learning pipelines to extract text, catalogue the archive, and understand 61,000 images using Weasel, a digital workflow system.

To extract text, we built a workflow which fetches images hosted on the British Library; uses Kraken to segment text in each image; deploys Google Vision to extract handwritten and typewritten text; sends the image, the polygonal representations, and text to eScriptorium, where users can review it and correct the text. Here, we discuss both positive outcomes and ongoing challenges in recognizing text in typewritten versus handwritten documents, in segmenting images into regions, and working with these tools.

From there, to create a catalogue, we built workflow that downloads transcriptions from eScriptorium; uses spaCy named entity recognition to extract and link the names of people, places, dates, events, organizations; and employs open-source large-language models running locally via Ollama (mistral:instruct and mistral:8x7b) to generate summaries, timelines, catalogue entries, and other catalogue material. We run experiments with RAGatouille, ColBERT, LangGraph powered agent-based workflows to do further work on the archive. Finally, we export this material as a catalogue of extracted text, summaries, named entities, keywords, etc. into a fichero, a box of linked and tagged digital index cards in Markdown format which Obsidian and similar tools make accessible to non-technical users. Additionally, we use Nomic's Atlas to map the data and metadata. The collection can be published to the web using 11ty or other static generators.

Here, we discuss steps to create a machine-generated catalogue, challenges to choose the right approaches, the costs of online commercial AI models, and the importance of the right tool.

Automatic Clustering of Hebrew Manuscripts

Daria Vasyutinsky Shapira, Berat Kurar-Barakat, Mohammad Suliman, Sharva Gogawale, Nachum Dershowitz
Tel Aviv University, Israel; dariashap@tauex.tau.ac.il, berat@tauex.tau.ac.il

This paper presents the interdisciplinary research conducted at Tel Aviv University in the framework of the ERC Synergy project, MiDRASH. Our work combines scholarly domains of Hebrew paleography and deep machine learning. We aim to automatically cluster medieval Hebrew script types-modes beyond the limits of contemporary human paleography. Currently, we are working on Ashkenazi square script.

Successful algorithms will be applied to other medieval Hebrew script type-modes, allowing improved clustering of the least-studied script types, such as Byzantine and Yemenite, and deepening our understanding of the sub-clustering of Oriental, Sephardic, and Italian scripts. This, in turn, will lead to the discovery of new paleographic patterns, improved layout segmentation based on script types, and more.

Panel Session: Discussing SSH Open Marketplace Workflows: User Experiences, Editorial Policy and Future Development

Time: Wednesday, 19/June/2024: 2:00pm - 3:30pm
Session Chair: **Mikko Tolonen**, University of Helsinki
Session Chair: **Vojtěch Malínek**, Institute of Czech Literature, Czech Academy of Sciences

Discussing SSH Open Marketplace Workflows: User Experiences, Editorial Policy and Future Development

Edward Gray[1], **Sally Chambers**[2], **David Lindemann**[3], **Vojtěch Malínek**[4], **Róbert Péter**[5], **Nanette Rißler-Pipka**[6], **Mikko Tolonen**[7]
[1]DARIAH-EU / IR* Huma-Num (CNRS UAR 3598), France; [2]DARIAH-EU / Ghent Centre for Digital Humanities, Faculty of Arts and Philosophy, Ghent University, Belgium / KBR, Royal Library of Belgium; [3]UPV/EHU University of the Basque Country, Dept. of Linguistics and Basque Studies, Vitoria-Gasteiz, Spain; [4]Institute of Czech Literature, Czech Academy of Sciences, Czech Republic; [5]Institute of English and American Studies, University of Szeged, Hungary; [6]Max Weber Foundation – German Humanities Institutes Abroad, Bonn, Germany; [7]Department of Digital Humanities, University of Helsinki, Finland;
edward.gray@dariah.eu, sally.chambers@dariah.eu, david.lindemann@ehu.eus, malinek@ucl.cas.cz, robert.peter@ieas-szeged.hu, rissler-pipka@maxweberstiftung.de, mikko.tolonen@helsinki.fi

One of the primary goals of the European Open Science Cloud (EOSC) research infrastructure initiative is to provide integrated access for research data, tools, and services across all scientific domains in Europe. Although EOSC currently houses millions of research outputs, the issue of workflows – step-by-step instructions of how a research process is conducted, enabling reproducibility of research – is considerably more complex and represents only a fraction of this content. This challenge becomes even more pronounced in the field of Social Sciences and Humanities (SSH) due to its diverse and heterogeneous nature across various disciplines.

The SSH Open Marketplace, developed during the Horizon 2020 Project SSHOC, acts as a thematic entry door into the EOSC and is maintained by DARIAH, CLARIN, and CESSDA. The development work behind the SSH Open Marketplace was inspired by previous ventures such as the DiRT directory (Dombrowski 2014), TAPoR, and Standardization Survival Kit (Riondet & Romary, 2018), and indeed the initial data aggregation came in part from these sources (Gray et al., 2021). Aside from this initial data aggregation, the SSH Open Marketplace also developed a sustainability plan and governance scheme that integrated the lessons learned from these previous experiences in order to integrate community feedback and provide solid infrastructure and human resource support from the ERICs in the interest of maintaining platform viability (Dombrowski, 2021; Petitfils et al., 2021).

A cornerstone feature of the SSH Open Marketplace is its Workflows, which leverage the platform's inherent contextualisation to provide a systematic, step-by-step representation of a research scenario. Workflows offer an ideal means of sharing one's research resources and maximising the potential of the SSH Open Marketplace. By integrating tools and services with publications, datasets, and training materials, workflows offer a comprehensive view of a research endeavour, making it both reproducible and user-friendly.

However, workflows are inherently reliant on users, both for their creation, and those that will implement them to reproduce research methods. Consequently, within the Bibliographical Data Working Group of DARIAH, members organised a booksprint in October 2023 with the dual purpose of generating new workflows (Lindemann & Klaes, 2023; Moretti & Heibi, 2023; Péter, 2023; Umerle & Malínek, 2023; Tomczyńska & Korytkowski, 2023; Tolonen et al., 2023) and addressing fundamental questions surrounding them (Malínek et al., 2024). This included inquiries into the essence of constructing a SSH workflow, the potential criteria for their development, the expectations of the DH community regarding SSH workflows, and an evaluation of their strengths and weaknesses.

Panel Structure:

The aim of the proposed panel, moderated by Mikko Tolonen, is to analyze and discuss possibilities of use, maintenance and further development of SSH Open Marketplace as a key vehicle for collecting and presenting the SSH workflows. Our panellists will further develop and build upon results achieved within Bibliographical Data Working Group´s Open Bibliodata Workflows project as well as introduce their own experiences of SSH Open Marketplace editors and representatives.

The panel will be centred around three main parts (each approx. 20-25 minutes) and a final general discussion. In the first part, two bibliodata related workflows will be introduced as model case studies: "LODification of bibliographical data: Zotero to Wikibase migration with ZotWb" (Lindemann & Klaes, 2023) and "Multilingual analysis and visualization of bibliographic metadata and texts with AVOBMAT" (Péter, 2023). These workflows were selected with regard to their interdisciplinarity and can serve as practical model demonstrations of SSH workflows. First one showcases how to convert bibliographical data into a wikidata structure with the use of open software solutions. The second case study presents the workflow of the AVOBMAT multilingual research tool (Péter et al., 2020, 2022), which enables researchers to critically analyse, enrich bibliographic data and texts at scale with the help of NLP methods. The implemented analytical and visualisation tools provide close and distant reading of texts and bibliographic data with time-based components.

Following these case studies, the rest of the panel will be arranged in an interactive way in the form of a round table where these and other practical experiences with SSH Open Marketplace (e.g. Candela et al., 2023 & Chambers et al., 2023) as an environment for creation and storing of such workflows will be discussed and evaluated from the perspective of SSH Open Marketplace editors and curators (S. Chambers, E. Gray) as well as from the position of the external users. Subsequently, the panel will discuss the future development plans of the SSH Open Marketplace platform and its strategic goals. Each part will be followed by discussion with the audience and a general discussion as the last part of the panel.

Paper Session: Organizing Knowledge

Time: Wednesday, 19/June/2024: 2:00pm - 3:30pm
Session Chair: **Agiatis Benardou**, DARIAH-EU

A Digital Workflow for Transforming Unstructured Text from Humanities Publications into a Scholarly Knowledge Graph

Vayianos Pertsas, Sofia Nasopoulou, Marianna Tsigkou
Athens university of economics and business, Greece; vpertsas@aueb.gr

The significant surge in research publications across disciplines poses a growing challenge for experts in maintaining a comprehensive understanding of their respective fields [1]. Especially in multidisciplinary fields like Digital Humanities, it becomes notably harder to retrieve and interconnect knowledge, particularly for cases of unstructured text from OCRed papers. This situation calls for new "strategic-reading" methods that transform the essence of knowledge encoded in textual form into structured formats like Knowledge Graphs (KG), thus changing the way researchers engage with literature [2]. In this paper we present a digital workflow for creating such a KG from Humanities' publications through the identification, extraction and interrelation of entities deriving from Scholarly Ontology [3], specifically designed for documenting scholarly work. A specialization of SO for DH is known as NeMO [4].

We focus on four types of entities: 1) *Activity,* noun phrases denoting research processes like archeological excavations, surveys, experiments etc.; 2) *Finding,* denoting the results of the above activities; 3) *Reference,* spans representing the individual entries within the article's bibliography; 4) *CitationPointer,* textual elements within text that refer to an entry in the article's reference list. Figures 1,2&3 show indicative examples.

Initially the text is segmented into sentences using spaCy[1]. For the entity extraction task we employ four RoBERTa-base Transformers from Hugging-Face library which we fine-tune using two manually annotated datasets: one with sentences from articles' main text containing annotations of Activities, Findings and CitationPointers and another with entire pages containing annotations of References. Both were randomly sampled from 25,681 OCRed articles of Humanities disciplines such as Archeology, Paleontology, Anthropology, etc., years 2000-2021, retrieved from JSTOR repository. Datasets' characteristics are presented in Tables 1&2.

Evaluation results measured in Precision, Recall and F1 are presented in Table 3. Error analysis showed that performance depends on the complexity of the textual spans: *Activities* impose higher lexico-syntactic variation compared to *Findings,* while References and *CitationPointes* follow very specific and easily recognizable patterns.

The extracted entities are then interrelated via *resultsIn(Activity,Finding) and indicates(CitationPointer,Reference) relations, using inference rules based on the collocation of the entities in the same sentence and regular expressions respectively.* All interrelated entities are associated with their source sentence, publication's metadata and further transformed into URIs, yielding the KG as RDF triples. The entire workflow is shown in Figure 4.

Constructing KGs from academic publications is an area of active research. Works like [5,6,7,8] focus on integrating bibliographic content, whereas research in [9,10,11] applies off-the-shelf NER models to extract entities, like materials and tasks, from scientific papers. While these efforts deal with already structured articles and standard named entities, our workflow enables the extraction of semantically complex (and of variable length) concepts like research activities and findings and deals with unstructured -OCRed- text, while focusing on the domain of Humanities research. The produced KG allows for answering complex queries such as: "find all the references of the articles containing activities that deal with 3D Reconstruction" or "retrieve the findings of the articles that have as reference a specific DOI".

[1] https://spacy.io/

Unveiling the Veil: reflecting on the omission of workflows in Digital Humanities research projects and their implications for reproducibility

https://doi.org/10.5281/zenodo.12687124
João Oliveira[1], Constança Almeida[1], Filipa Magalhães[2]
[1]NOVA University of Lisbon, School of Social Sciences and Humanities; [2]Centre for the Study of the Sociology and Aesthetics of Music (CESEM), NOVA University of Lisbon, School of Social Sciences and Humanities; oliveirajoaopedro79@gmail.com

To what extent are workflows documented, accessible, and comprehensible? If achieving reproducibility is a structural objective, why are the definition, clarity, and explanatory fluidity of Digital Humanities project workflows not readily available or occasionally, inaccessible to the interested scientific community? How significantly will this issue impact the replicability of research project results? When analysing these issues and comparing it with research projects in the so-called 'exact sciences', it appears that the majority of research projects in Digital Humanities omit their workflows. This research delves into whether the current state of Digital Humanities projects conforms to the global objectives of Open Science, Open Data and FAIR principles. It examines whether there is any misalignment compromising reproducibility and replicability of the results.

Weiland (2018) defines workflows as a deliberate or rational organization of any purposeful activity, typically specifying its steps in the form of a process directed at a particular result. Therefore, to address our concerns about the omission of workflows in Digital Humanities research projects, we sought to discuss this definition through an in-depth qualitative analysis of some case studies.

To achieve this, we looked at three interdisciplinary projects ranging from contemporary performing arts to digital environmental history: "ERC Project: From Stage to Data, the Digital Turn of Contemporary Performing Arts Historiography (STAGE)"; "In Search of the Drowned in the Words: Testimonies and Testimonial Fragments of the Holocaust" and "The Making of a Forest: landscape change at the Argentine-Brazilian border, 1953-2017".

Preliminary findings suggest that while digital humanities projects often articulate overarching principles of Open Science, FAIR principles, and general methodology the overall granularity of workflows required for enhancing reproducibility is often lacking.

Decisions made at critical junctures of the research process, nuances in information manipulation, and the specific configurations of digital tools employed need to be enhanced. (Liu et al., 2017) This raises concerns regarding the reproducibility and replicability of the knowledge generated within the Digital Humanities domain as a direct consequence of the lack of workflow evidence.

We consider the need to deepen the disparity in the advancement of Open Science and similar principles amidst the underdeveloped state of workflows in Digital Humanities research projects. By implementing clearer workflows, the replicability of results would become evident and could be more readily integrated into various ongoing projects. The clarification of individual research decisions, information handling, and exploration sources throughout the project lifecycle lacks a comprehensive definition, posing challenges for e-science scholars in method development based on previously used methods in research projects (Antonijevic & Cahoy, 2018).

In this paper, we propose to reflect on the importance of making information about workflows accessible and comprehensible through documentation, giving concrete examples from the case studies analysed.

Guarding accessibility - AI supported ontology engineering in the context of the MuseIT repository design

https://doi.org/10.5281/zenodo.12533424

Vyacheslav Tykhonov[1], Nasrine Olson[2], Moa Johansson[3], Kim Ferguson[1], Nitisha Jain[4], Lloyd May[5], Andrea Scharnhorst[1]
[1]Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science, Netherlands, The; [2]University of Boras, Sweden; [3]ShareMusic, Sweden; [4]Kings College London; [5]Dartmouth College, Stanford University;
kim.ferguson@dans.knaw.nl

The MuseIT project explores new technologies that facilitate and widen access to cultural assets and enables the co-creation of them. It builds upon the UN Universal Declaration of Human Rights and the Convention on the Rights of Persons with Disabilities, both proclaim human rights for all people, including people with disabilities. One of the key outcomes of MuseIT is to set up a content repository capable of indexing, storing and retrieving multisensory, multi-layered digital representation of cultural assets respecting FAIR principles and long term preservation. Dataverse is the envisioned platform to be used. This paper zooms into one important step: namely the creation of specific knowledge organisation systems suitable for the later archiving process. Based on ontology engineering we look for new ways to describe accessibility facets of artefacts in the MuseIT repository. We use the term accessibility here in the context of disability, following the Wikipedia definition that "accessibility is the design of products, devices, services, vehicles, or environments so as to be usable by people with disabilities".

If it comes to categorising disability we depart from the International Classification of Functioning, Disability and Health (ITC) released by the World Health Organisation. As in its name, this classification provides a medical view on disability. Additionally, we also use Wikipedia.EN (or a collection of web resources (MuseIT disability collection) stored as part of the Now.Museum (Vion-Dury et al. 2023) on a dataverse platform). In short, material which represents a more cultural view on a phenomena (e.g., Salah et al. 2012).

In this paper, we present a new innovative knowledge engineering solution (EverythingData) through which 'proto-ontologies' are created by combining Large Language Models with structured data in the form of Knowledge Graphs (Pan et al 2024). It is schematically depicted in the figure below, and in essence relies on a set of APIs. Those enable workflow as such: (a) start with a term (group of terms) for instance 'disability' from the ITC (b) identify a related wikipedia (eng) page (or other texts) (c) feed these texts into a LLM instance and receive a group of terms and their relations which in essence represent a proto-ontology (d) iterate this process and (e) evaluate the variants of a proto-ontology by a group of experts.

Deeply concerned with issues of equality, democratisation and social inclusion, MuseIT researchers are aware of the power of knowledge organisation when it comes to documenting content and organising access. Knowledge organisation systems (KOS) in all their forms (classifications, thesauri, ontologies etc.) are a double-edged sword. They make things visible but they also represent a certain normative stance (Bowker, Star 2000). With experts we will organise the human evaluation of those proto-ontologies to select those set of terms and relations which best suit a rich description of cultural assets with respect of their consumption by people with impediments, while at the same time also to give credit to the variety of producers of cultural assets, acknowledging specific conditions under which such a creation took place.

Paper Session: Auditory Workflows

Time: Thursday, 20/June/2024: 9:00am - 11:00am
Session Chair: **Marita Everhardt**, DANS-KNAW

Challenges and Opportunities in Intermedial Auditory Workflows: The Sound-Text Archive of Muslim Women of West Africa (STAMWWA)

Estrella Samba-Campos
Nodo CLARIAH-CM, Spain; essamba@ucm.es

The Sound-Text Archive of Muslim Women of West Africa (STAMWWA) serves as a repository of "voices," conveying the auditory narratives of Muslim women across West Africa and its diaspora. Originating from a teaching module and a questionnaire sent through WhatsApp, this archive amalgamates audible accounts addressing contemporary issues surrounding faith and practice, while simultaneously promoting the use of African languages in responses.

My presentation first presents the workflow conceived and implemented during the assembly of the archive, aiming to progressively discuss higher levels of granularity concerning contemporary ethical practices in sound repositories. Central to this discussion are inquiries regarding the nature of a sound archive and the implications posed by the ethical, reproducible, and sustainable characteristics of a workflow. As the archive organically expands, there arises a necessity to confront the underlying questions about accessibility, digital humanities, and open data science within the context of West Africa.

I draw on diverse theories such as the problem of knowledge as visualized pages following Walter Ong's foundational theory (Ong, 1982), the concept of archival amnesty or those underrepresented in and by the textual archive as argued by Tonia Sutherland (Sutherland, 2017), or the statement made by Lawrence Kramer, who asserts that "the kingdom of text has been steadily shrinking since the Enlightenment" (Kramer, 2019).

By integrating these perspectives into the workflow process of the STAMWWA, my paper seeks to discuss the inherited power of epistemologies born textual as opposed to emerging, intermedial epistemologies through multimedia. With this in mind, established concepts within disciplines such as Digital Humanities and its pedagogy are examined constructively, aiming to understand their evolution in new cultural landscapes and trans-academic frontiers.

Bibliography

Dussel, E., «Transmodernidad e interculturalidad», Astrágalo: Cultura de la Arquitectura y la Ciudad, n° 21, 2016 (pp. 31–54).

Kramer, Lawrence. 2019. The Hum of the World: A Philosophy of Listening. California: University of California Press.

Jensen, K. (2016). «Intermediality» eds. Jensen K., Craig R., The International Encyclopedia of Communication Theory and Philosophy, Wiley-Blackwell, Chichester.

https://onlinelibrary.wiley.com/doi/10.1002/9781118766804.wbiect170

Ong, Walter J. [1982], 2002. Orality and Literacy: The Technologizing of the Word. Routledge.

Samba-Campos, E. (2017-). Archivo Sonoro-Textual de Mujeres Musulmanas de África-Occidental. http://dynamicreadings.com/investigacion/archivo-sonoro

Sutherland, Tonia (2017). Archival amnesty: In search of Black American transitional and restorative justice. Journal of Critical Library and Information Studies, 2, 1–23. https://doi.org/10.24242/jclis.v1i2.42

AI-Driven Workflows for Unlocking Switzerland's Collective Memory: Distant listening of the RTS Archive

https://doi.org/10.5281/zenodo.12191436
Giacomo Alliata, André Rattinger, Kirell Benzi, Sarah Kenderdine
EPFL, Switzerland; giacomo.alliata@epfl.ch

In the era of digital transformation, the vast digitization of audiovisual archives poses significant challenges for researchers in the arts and humanities (Fossati et al., 2012; Jaillant, 2022). This contribution presents a pioneering approach to address these challenges, utilizing cutting-edge AI technologies to design reproducible workflows for the efficient processing, analysis, and access to large audiovisual archives (Colavizza et al., 2021). Focusing on the case of the Radio Télévision Suisse (RTS) archive and its 200,000 hours of footage, essentially serving as Switzerland's collective memory, we demonstrate a comprehensive pipeline that integrates multiple algorithms and frameworks to facilitate the extraction of meaningful content.

The digitization of extensive audiovisual archives creates an overwhelming volume of data, making manual processing impractical. Our approach leverages the power of AI to operationalize these archives, unlocking new possibilities for researchers and the broader public to engage with the rich cultural and historical content within. The core components of our pipeline include speech-to-text conversion, speaker diarization, and named entities recognition (NER). These algorithms work synergistically to transform raw audiovisual content into structured and annotated data.

The speech-to-text module plays a pivotal role in making spoken content accessible, allowing for the operationalization of speech within the archive. Thanks to the Whisper algorithm (Radford et al., 2023), we open audiovisual archives to the lens of NLP and "distant listening" enquiries (Clement, 2012, 2016). The speaker diarization component further enhances this process by segmenting the content into distinct "paragraphs", facilitating efficient navigation and analysis by extracting individual units of content. Incorporating NER adds a layer of richness to the data, connecting individuals, locations and events recorded in the archive to broader knowledge repositories. In particular, we have integrated the work previously described in Alliata et al. (2023) to connect the rich metadata extracted to open databases such as WikiData, expanding the full pipeline to the benefits of large knowledge bases (Rudnik et al., 2019).

The potential applications of our AI-driven workflows are diverse and impactful. One notable application is the creation of a dynamic "Map of Switzerland" visualizing the geographical distribution of voices recorded in the RTS archive. This serves as a novel exploration tool and contributes to a nuanced understanding of linguistic variations and cultural diversity over time, as well as geographical representations in the archive itself. Additionally, our workflows enable the implementation of Retrieval-Augmented Generation (RAG) pipelines (Lewis et al., 2020) to augment searchability within the archive leveraging Large Language Models and embedding models, enhancing the discoverability of specific topics, individuals, or historical events through more human-like queries.

This presentation emphasizes the technical, methodological and conceptual aspects of our AI-driven workflow, developed to operationalize large audiovisual archives through the dimension of spoken text. Such a "distant listening" approach can enrich these collections with a wealth of semantic information that opens up novel avenues of access and analysis. Furthermore, our case study on the RTS archive offers a compelling example of potential applications and thus contributes to the advancement of reproducible research practices in the arts and humanities.

Recognising all forms of knowledge: What can we learn from creative practice research to benefit all and enable reproducible research?

https://doi.org/10.5281/zenodo.12187604

Jenny Evans[1], **Adam Vials Moore**[2]

[1]University of Westminster, United Kingdom; [2]Jisc, United Kingdom; J.Evans2@westminster.ac.uk, adam.vialsmoore@jisc.ac.uk

Have you ever witnessed the dynamic, iterative process of a musician experimenting with sound, only to see this represented as a static sound file and no easy way to represent the narrative that evidences its research element? This disconnect between creative practice and traditional research frameworks hinders discoverability, transparency of access and re-use. We propose a solution that recognises the "hidden side" of research, empowering creative disciplines like music, dance, and visual arts to be fully recognised within a Findable, Accessible, Interoperable and Re-usable (FAIR) Research landscape.

Research reward and recognition systems, infrastructure, and policies often assume single, static, text-based outputs, with non-text outputs an add on, making it challenging for creative practice research to be captured let alone discovered. Open access is seen as simply making the entirety of an output 'open', neglecting the uniqueness of practice disciplines where the connections between the elements matter as much as the individual output. This mismatch stifles the richness of creative research, hindering its potential for interdisciplinarity, collaboration, and societal impact.

We present key findings of the Practice Research Voices (PR Voices) and the complementary Sustaining Practice Assets for Research, Knowledge, Learning and Engagement (SPARKLE) projects, funded by the UK's Arts and Humanities Council in 2022.

One of the outcomes of this work was the development of a workflow framework capturing both process (practice) and product in diverse formats (e.g., audio, video, movement notation). It respects the unique challenges of non-linear, iterative research, empowering researchers to share their "hidden side" data alongside traditional outputs. By fostering deeper understanding of practice-based research processes, we aim to enhance transparency, reproducibility, and collaboration within FAIR research principles.

We explore how emerging digital methods and standards like the framework can enable capture and sharing of research data in A&H, addressing:

- **Workflow integration:** Seamlessly integrating the framework within existing repository platforms.
- **Non-linearity challenges:** Capturing the iterative and embodied aspects of practice research.
- **Skills development:** Building researcher capacity for platform adoption.
- **Interdisciplinarity potential:** Facilitating collaboration across diverse research disciplines.

Our analysis demonstrates the framework's positive impact on research transparency, reproducibility, and collaborative potential.

Conclusion:

This work-in-progress presentation invites feedback and discussion. We aim to raise awareness, encourage adoption of the framework, and initiate collaborations to further develop and refine this approach. By embracing the multifaceted nature of creative practice research, we can unlock its full potential within the FAIR Research landscape, enriching the scholarly ecosystem for all, and fully recognise all forms of knowledge.

For clarity, we use the definition of practice research articulated by Bulley & Şahin (2021): "*An umbrella term that describes all manners of research where practice is the significant method of research conveyed in a research output. This includes numerous discipline-specific formulations of practice research, which have distinct and unique balances of practice, research narrative and complementary methods within their projects*."

Paper Session: Ethical Workflows

Time: Thursday, 20/June/2024: 9:00am - 11:00am
Session Chair: **Andrew Janco**, University of Pennsylvania

Closing the loop: integrating students and the community in the Creolistic research workflow

Carlos Silva[1,2,4], Luís Trigo[1,3,4], Vera Moitinho de Almeida[3,4]
[1]CLUP - Centre of Linguistics of the University of Porto; [2]DEPER - Department of Portuguese and Romance Studies; [3]CODA - Centre for Digital Culture and Innovation; [4]FLUP - Faculty of Arts and Humanities, University of Porto, Portugal;
cssilva@letras.up.pt

CreoPhonPt is an interdisciplinary and collaborative research project on phonological, lexical and sociolinguistic information of Portuguese Creoles [1][2]; and based on the integration of Open Science, Citizen Science, and FAIR [3] and CARE [4] principles, into the flow of scientific production in the humanities and social sciences fields [5]. To foster students and citizen engagement, we integrated this project in the classroom environment.

The main goals of the Phonology seminar of the Master in Linguistics, at FLUP, are that students (I) acquire knowledge inherent to phonological theory and (II) can apply phonological theory to the description of natural languages. To this end, Portuguese-based creoles and African languages were selected for teaching how to build a linguistic-analysis workflow from bottom-up, while enhancing and taking advantage of CreoPhonPt's potentialities.

Across two years, we applied project-based learning [6] and cooperative learning [7] methods to encourage creative and critical thinking in the classroom and to foster collaboration and social responsibility outside the classroom. Seminars were split into ten blocks, each one comprising theoretical exposure and practical work. All stages of the workflow (Fig.1) were monitored by the teaching team. A set of freeware and opensource online tools, also easily available outside the classroom, were used. Students archived the data in GitHub's CreoPhonPt open access and version-controlled repository.

Year 1: focused on language data and metadata from existing bibliography. Each student was given a set of raw data, which had to be structured, manually/automatically processed (incl. data wrangling and reconciling) and archived.
Year 2: we took students a step back and focused on phonological data and metadata from field work. Students were divided in groups of four and made responsible for collecting, structuring, manually/automatically processing and archiving phonological/audio data from fellow colleagues Guinean-Bissau creole speakers, who assessed them throughout the workflow. Two interdisciplinary open seminars were organised, bringing together researchers from sociology, anthropology, education science, culture and communication, as well as technology.

RESULTS:

1) Learning: students acquired a full set of skills (incl. soft/hard, theoretical, technical, research, management) that can be useful beyond the scope of this Master course.
2) Community: Afro-Portuguese communities, which have these languages as their cultural and social asset, were actively involved in field data collection, authority/responsibility, archiving, (re)use and dissemination.
3) Research: the open access CreoPhonPt repository was largely enriched with further socio-historical, cultural and linguistic data/metadata for preservation, dissemination and further investigations.

The lack of previous studies and the fact that these languages are endangered was per se a motivation for all the participants. Likewise, the co-creation of scientific data and its collective benefit were very stimulating, as evidenced by the pedagogical surveys' results. Recently, the CreoPhonPt incursion in the classroom was awarded an honourable mention "Innovative Pedagogical Practice", by UPorto. More recently, students presented their work at a public event [8], the collected corpus was submitted to an international workshop [9], while the collective written essays were submitted to a special issue of *elingUP* (online open access journal of UPorto's Linguistics students).

INEL workflows for creating digital corpora of minority languages: Lessons learned

Alexandre Arkhipov, Elena Lazarenko, Aleksandr Riaposov
Universität Hamburg, Germany; aleksandr.riaposov@uni-hamburg.de

Since 2016, the INEL project has been working on creating richly annotated XML-based corpora of various minority languages of Northern Eurasia. Four corpora have been published by 2023 – Kamas, Selkup, Dolgan, and Evenki – with several more currently under development.

As we are dealing with severely endangered, low-resource languages, the source materials tend to be of diverse form and origin, including data from fieldwork archives containing audio recordings and handwritten texts. The texts that end up in the corpus encompass various genres such as folklore, personal narratives, and conversations on everyday topics. As a starting point, the available data need to be digitized and/or converted to a format suitable for import into SIL FLEx, a linguistic analysis software. Digitization may involve transcribing audio with a native speaker consultant (ELAN), manual typing text from manuscripts, Handwritten Text Recognition or OCR (Transkribus, ABBYY FineReader), transcoding into a unified Latin-based transcription in Unicode, etc. The next step in our workflow is linguistic annotation, which is divided in two stages: first, interlinear morpheme glossing in SIL FLEx; second, providing additional layers of annotation (e.g., syntactical functions, information structure) and corpus metadata via the EXMARaLDA package. At the latter stage, corpus data start being version-controlled in Git and undergo continuous curation through Corpus Services, an in-house developed Java package that helps maintain data consistency by pinpointing errors in the files and automatically correcting them if possible. The finalized corpora are provided in several XML-

based formats – EXMARaLDA Basic- and Segmented Transcription Data, ISO/TEI, and ELAN Annotation Format. ISO/TEI also serves as input format for Tsakorpus, the platform providing an online search interface for the published corpora.

In our presentation we will showcase some aspects of the evolving workflows, raising attention to the following considerations:

- We aim to progressively automate as many parts of the workflow as we feasibly can without compromising the data quality. In particular that concerns tedious manual tasks where a machine, given proper instructions, performs exceptionally well.
- The tools we use shape our workflow. Each workflow component brings its own set of limitations and requirements that will influence other stages of the workflow, sometimes in non-obvious and subtle ways. An example of that are (non-identical) algorithms of text segmentation in FLEx and EXMARaLDA, which affect the transcription conventions used throughout the workflow.

As a recent optimization step, starting from the desire to streamline the configuration of the Tsakorpus interface (which comes very late in our workflow), we arrived at the need to extend and standardize the EXMARaLDA annotation panel so that the annotations at specified tiers, including affix glosses, could be checked against a closed set of values during the automated data curation process. However, the values and associated descriptions should be prefilled at the analysis stage in FLEx. They are then stored in a hierarchical structure that follows the way they are to be displayed later in the search interface.

Combining and Merging Workflows: A Case Study from the Time-Layered Cultural Map of Australia (TLCMap)

Paul Arthur[1], Isabel Smith[1], Bill Pascoe[2], Jane Lydon[3], Hugh Craig[4]

[1]Edith Cowan University, Australia; [2]University of Melbourne, Australia; [3]University of Western Australia, Australia; [4]University of Newcastle, Australia; paul.arthur@ecu.edu.au

This paper reports on workflows utilising the Time-Layered Cultural Map of Australia (TLCMap) national research infrastructure (https://www.tlcmap.org/). TLCMap is a set of online tools that allows humanities researchers and the public to compile, analyse, and visualise humanities data using spatio-temporal coordinates. This resource has been developed to help create digital maps from cultural, textual, and historical data, layered with datasets registered on the platform. TLCMap is not a singular map, but rather a range of software tools, or ecosystem.

The presentation focuses on a particular case study utilising TLCMap – the Western Australian Legacies of British Slavery (WALBS) Australian Research Council Discovery Project (https://australian-legacies-slavery.org/). This project has been uncovering Australia's colonial ties to Atlantic slavery by tracing the movement of people, wealth, and culture from slave-owning Britain to Australia in the nineteenth century. The research has been primarily biographical, identifying hundreds of individuals including British slaveowners, beneficiaries of slavery, and more recently, enslaved people.

Using TLCMap's Gazetteer of Historical Australian Places, the journeys of six individuals have been plotted out on a 3D map. Each journey is visualised by a series of geographical points that users can click on to view information including the dates the individual was there, a brief biographical summary of what occurred, or the significance of the place, and links to relevant images or other resources. There are multiple visualisations, including a mode featuring a moveable timeline, and another that links to each point of an individual's journey.

Though the WALBS research began by uncovering slaveowners and colonists who benefited from slavery, attention has recently shifted to tracing the lives of the enslaved. This has raised challenges around the bias of data. Details of colonists and slaveowners record and glorify their own acts, while entire lives of the enslaved are written out of the records. Saidiya Hartman (2008) describes how the only glimpses we have of the enslaved are via their brief, distorted, and violent encounters with power. (Projects such as the SlaveVoyages database https://www.slavevoyages.org/, which includes data on some 92,000 Africans forced to make the voyage from Africa to and within the Americas, offer notable and large-scale counters to the silences.) To accommodate gaps in our data, we have moderated our templates, including 'notes' fields that allow us to comment on the way a particular date or location has been obtained. Discussion includes how estimates have been devised, as well as the limitations of the records available. This allows critical reflections on the biases of the archives, and less rigidity in the requirements for data. We have also incorporated other materials, beyond geographical and temporal data, that can be visualised and narrated to tell the story of a person's life. In particular we have placed emphasis on biographical text, first-person quotations, and images, to add more richness and colour to the narrative despite limited archival records.

References

Hartman, S. (2008). Venus in Two Acts. *Small Axe*, 12(2), 1-14.

TLCMap. https://www.tlcmap.org/.

Western Australian Legacies of British Slavery. https://australian-legacies-slavery.org/.

Inclusive Workflows to Address Vicarious and Secondary Trauma in Humanities Research: Insights and Suggestions from a UK/IE Community Interest Group

Kristen Michelle Schuster[1], Vicky Garnett[2]

[1]University of Southampton, United Kingdom; [2]Trinity College Dublin; k.m.schuster@soton.ac.uk, Vicky.Garnett@tcd.ie

At some point during our research careers, we will grapple with an institutional review board or research ethics application. Questions about data management, informed consent and recruitment all require some degree of planning and preparation and are inevitable within the overall workflow of a project. Most of the questions we prepare responses for are designed to facilitate reflective and ethical practices for the treatment of research subjects.

The value of considering how research affects subjects is undeniable. However, at what point in the workflow of a project do we stop to consider the effect data and associated methodologies will have on us as researchers?

Our paper will outline resources researchers and practitioners can draw on to scope and implement practices that raise awareness about vicarious and secondary trauma in research. We will situate this outline within a discussion of the work conducted by the Protecting the invEstigator in Traumatic Research Areas (PeTRA), a UK/Ireland Digital Humanities Association Community Interest Group (CIG). PeTRA has invested in building networks between researchers and practitioners interested in developing documentation and resources to promote research practices that address both participant and researcher experiences.

We will review the existing resources PeTRA is collecting around the issue of vicarious trauma in (digital) humanities research. Following this review, we will discuss findings from a pilot study designed to gauge the manner in which ethics is taught as part of a research methodology. Initial results revealed that some training occurs formally as part of degree programmes, and that risks to the researcher are not covered as often as issues of consent, GDPR and risks to the participant. These findings have informed the development of a wider project within PeTRA to reflect on both the differences in approaches to ethics in digital humanities scholarship and practice across the UK and Ireland. We will conclude by reflecting on strategies we can adopt to integrate trauma informed practice into research workflows, networking and infrastructure building in the digital humanities.

Panel Session: The RSE Turn in Digital Humanities

Time: Thursday, 20/June/2024: 11:30am - 1:00pm
Session Chair: **Natalia Ermolaev**, Princeton University

The RSE Turn in Digital Humanities

Natalia Ermolaev[1], **Mary Naydan**[1], **Rebecca Koeser**[1], **Jeri Wieringa**[1], **Arianna Ciula**[2], **Pamela Mellen**[2], **Miguel Vieira**[2], **Alexander Czmiel**[3], **Maciej Maryl**[4], **Laure Thompson**[1]

[1]Princeton University, United States of America; [2]Kings Digital Lab; [3]Berlin-Brandenburg Academy of Sciences and Humanities; [4]Institute of Literary Research of the Polish Academy of Sciences; nataliae@princeton.edu, mnaydan@princeton.edu, rebecca.s.koeser@princeton.edu, jeri.wieringa@princeton.edu, arianna.ciula@kcl.ac.uk, pamela.mellen@kcl.ac.uk, jose.m.vieira@kcl.ac.uk, czmiel@bbaw.de, Maciej.Maryl@ibl.waw.pl, laurejt@princeton.edu

The term 'Research Software Engineering' (RSE) has been in use for just over a decade to describe the professional practice of creating and sustaining high-quality, reusable software for research applications. RSE teams at academic and other research institutions are expanding rapidly, and are considered integral for the success of today's cutting-edge computational research.

Though much of the RSE community is oriented toward the STEM fields, there has been a growing adoption of the RSE framework – its roles, terminology, workflows, processes – in the arts and humanities domain. Beginning with Kings Digital Lab (KDL) in 2016, the use of 'RSE' to describe technical profiles and work has been steadily increasing among digital humanities individuals, centers and research groups. This panel brings together scholars and professionals from a sampling of DH initiatives that have adopted the RSE framework for their staffing and practices to discuss its various impacts, benefits, and challenges for the digital humanities.

Each group on the panel will focus on one relevant workflow related to their RSE activities as a way of illustrating broader methodological, technical, infrastructural or conceptual themes. Topics covered may include: how do roles, workflows, processes, collaborations, outputs change when our technology-based research is cast as 'Research Software Engineering'? How do the principles and practices of the larger RSE profession align with, support or enhance software development for arts and humanities research? How are other DH development approaches or research cultures distinctive or different? What changes has the adoption of RSEs meant for the various infrastructures (technical, institutional, social, financial) of the DH group or center? How has this change enabled (or prevented) new partnerships and collaborations, both internal and external? What new possibilities can the RSE framework afford for the future of computational humanities research and the DH community?

Panel participants will speak from various types of research teams and institutions, and from perspectives that range from earliest adopters to relative newcomers to the RSE framework. For audience context, we will begin with a general overview of the origins and current state of the global RSE community. We will provide a list of readings and resources to assist DH groups and individuals interested in implementing the RSE framework at their own institutions.

Paper Session: Documenting Workflows

Time: Thursday, 20/June/2024: 11:30am - 1:00pm
Session Chair: **Agiatis Benardou**, DARIAH-EU

Reproducible methods in the Arts and Humanities through workflows: the case of the SSH Open Marketplace

Laure Barbot[1], Elena Moro Battaner[2], Stefan Buddenbohm[3], Maja Dolinar[4], Edward Gray[1], Cristina Grisot[5], Klaus Illmayer[6], Alexander König[7], Michael Kurzmeier[6], Barbara McGillivray[8], Clara Parente Boavida[9], Christian Schuster[10]
[1]DARIAH; [2]Universidad Rey Juan Carlos; [3]Göttingen State and University Library; [4]ADP; [5]University of Zurich & DaSCH; [6]OEAW; [7]CLARIN; [8]King's College London; [9]Iscte-Instituto Universitario de Lisboa; [10]Babeș-Bolyai University Cluj-Napoca;
christian.schuster@ubbcluj.ro

The Social Sciences and Humanities Open Marketplace - marketplace.sshopencloud.eu/ -, one of the flagship services of the "Social Sciences and Humanities Open Cluster" (SSHOC)[1], is a discovery platform for new and contextualised resources from the Social Sciences and Humanities (SSH). Its main aim is to support researchers in discovering, accessing, and comparing digital tools and methods for their research. With its five content types - Tools & Services, Datasets, Training Materials, Publications, and Workflows - the SSH Open Marketplace covers a large range of research practices.

Workflows are defined as "Sequences of steps that one can perform on research data during their lifecycle. Workflows can be achieved by using diverse tools, resources and methods, and the useful resources are connected to each step" (SSH Open Marketplace (2023)). The primary value of workflows in the SSH Open Marketplace lies in their ability to be reused and applied to different research contexts or projects. This is enabled by the basic structure of workflows, which can be adapted to a range of real research use cases, while capturing tools, methods and processes that can be reused beyond individual research projects. By doing so, SSH Open Marketplace workflows support reproducible endeavours highlighting which nascent methods are in use in a given research community or how already agreed standards are applied.

At the time of writing this abstract, the SSH Open Marketplace counts 48 workflows,[2] most of which were created during in-person events. Based on this experience, face-to-face workshops have been found to be one of the most effective ways of popularising workflows and guiding researchers to upload their methodology in a new and not necessarily familiar format. While some workflows are "service-oriented", highlighting what can be done thanks to tool chains, like for example the *LODification of bibliographical data: Zotero to Wikibase migration with ZotWb*[3], other workflows focus on the functioning of a given service, see *ArkeoGIS how to share a dataset on the platform.*[4] Contrasting this contextualisation of services, some workflows are more oriented towards standards in practice, as it is the case for *Create a dictionary in TEI*[5] or *Collaborative Digital Edition of a Musical Corpus.*[6]

Promoting workflows, not only as an innovative way to document a research project, but to open up research methodologies and shed some light on processes rather than research outputs only, is crucial. Indeed, workflows can play a critical role in harmonising research methods and contribute to foster reproducibility of these methods across projects. Our paper also aims to highlight the challenges faced by the collective workflow collection as developed in the SSH Open Marketplace: generalising from practices is a complex epistemological task, and maintaining a coherent and up-to-date collection of SSH workflows and promoting them to ensure their reusability requires considerable resources. This work is coordinated by the SSH Open Marketplace Editorial Board and we would like to hear from the DARIAH Annual Event audience what could be improved so that the existing and future Marketplace workflows can best benefit the Arts and Humanities research communities.

Using GitHub for digital editions. From Transkribus to static websites

https://doi.org/10.5281/zenodo.11609624
Laura Untner, Peter Andorfer
Austrian Academy of Sciences, Austria; laura.untner@oeaw.ac.at

Arthur Schnitzler (1862–1931) is one of few Austrian authors today who is read and received in-ternationally. Shortly after his death on October 21, 1931, Clara Katharina Pollaczek gathered her memories of her former partner. The resulting typescript, which was typed by Schnitzler's secretary Frieda Pollak and is now kept in the Vienna City Library (ZPH 242), comprises 990 pages and mainly contains letters, diary entries and commentary notes. Entitled Arthur Schnitzler und ich, the memoir was first published in the form of a digital edition in 2023 (Müller/Untner/Mangel/An-dorfer 2023; see Fig. 1–2).

Because the digital edition was intended as a work in progress from the very beginning, it was published soon after an OCR was executed in Transkribus (using the »Text Titan I« model, cf. Transkribus 2023) and some other basic steps (initial collation, editorial commentary, table of con-tents) were completed. After that, we invited everyone to partake in improving the transcription – which is what happened. The project turned into a type of citizen science project, so the workflow had to meet this requirement. Especially the ongoing corrections made by volunteers in Transkri-bus were to be integrated into the digital edition with as little effort as possible. A workflow via GitHub that allows us to go from Transkribus to a static website in just two clicks made it possible to do justice to these dynamics.

To export the transcripts from Transkribus, we developed a GitHub action that accesses the Tran- skribus API using a Python package (Andorfer/Schlögl/Haak). Then, the exported files are trans- formed into valid XML/TEI documents using an adapted version of the XSL stylesheet developed for the default TEI export in Transkribus (Kampkaspar/Boenig/Stadler/Grallert). This stylesheet converts the METS and PAGE files, including the tags inserted in Transkribus (e. g. paragraph markings), into TEI files. Then, a separate XML document already containing metadata is created for each page via an automated comparison with the table of contents. To finally rebuild the web- site, another GitHub action developed for the DSE Static Cookiecutter (Austrian Center for Digital Humanities and Cultural Heritage) is used. After just a few minutes, the workflow is completed, and the data is updated.

Although the workflow was originally developed for a rather simple project, it is now also being tested for the digital scholarly edition Arthur Schnitzler: Briefwechsel mit Autorinnen und Autoren (Müller/Susen/Untner 2018–[2024]) and various other

projects at the Austrian Centre for Digital Humanities and Cultural Heritage (e. g. those on the origins of the Austrian Federal Constitution (FWF P I 5679) and on Hanslick's critiques (FWF P 35379)). Basically, the procedure remains the same, with the exception that in some cases more XSL transformations are required, e. g. to export a whole series of customized tags in Transkribus like greetings and farewells in letters and to create individual documents not for each page but for each letter.

[References & Figures in PDF]

The Polifonia Research Ecosystem: an Executable Data Management Plan

https://doi.org/10.5281/zenodo.12515624

Enrico Daga[2], Andrea Scharnhorst[1], Raphael Fournier- S'niehotta[3], Marco Gurrieri[4], Jacopo de Berardinis[5], James McDermott[6], Marilena Daquino[7], Jason Carvalho[2], Marco Ratta[2], Christophe Guillotel-Nothmann[4]

[1]Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science, Netherlands, The; [2]The Open University; [3]CNAM, Université Pierre et Marie Curie; [4]CNRS; [5]King's College London; [6]National University of Ireland Galway; [7]University of Bologna; andrea.scharnhorst@dans.knaw.nl

This paper introduces an innovative approach to documenting research components (including research data). The Research Ecosystem approach (Daga et al. 2023) gives guidelines to determine which components are relevant in the light of certain research questions, how to annotate them as semantic, machine-readable artifacts, and how to validate, control and preserve them. Developed in the context of Polifonia, a semantic-web-based research project to improve access to musical cultural heritage, the framework is re-usable in all projects which rely on collaborative, open software development. We zoom into specific challenges which emerge when dealing with digital objects from the cultural heritage domain - in our case music. In particular, we showcase how to achieve machine-readable expressions of license information, and enrichment of metadata supported by Large Language Models.

Ultimately, the Research Ecosystem fosters the management of Data in their context, namely together with Tools and Reports. However, it goes beyond pure documentation. In the case of Polifonia, archetypical users with their specific information needs (described as Personas and Stories) constitute one important component type. Components are connected via annotations. This way, the digital-humanities-born specific methodological approaches, the data used, the tools built and the documentation produced, all form a network which shows the interdependencies of research questions, methods, and data. In this component-based structure, traditional project management elements such as Work packages and Tasks also appear as part of the annotation scheme.

The beauty and efficiency of the approach lies in the consequent use of existing platforms such as GitHub, which already contain elements to build a machine-based Ecosystem. One result is a machine-executable Research Data Management plan, which takes standardization in RDM to a next level.

The challenge in re-using the Ecosystem approach is to find answers to classic research management questions such as what are the right components, which annotations are needed to express the interdependencies, and how both components and their links need to change over time. To support re-use of concept, Polifonia defined workflows for building and evolving an Ecosystem next to workflows which form part of the Ecosystem. This way, documentation and management are intrinsically interwoven with classical research management questions: how to find the right questions, how to find the right methods, and how to organize collaboration in an interdisciplinary setting (Guillotel-Nothmann et al., 2022).

Our Research Ecosystem approach builds on other approaches to formal description of research assets (such as FAIR Digital Objects, or RO-CRATE), but its essence is to address a middle level: above the elements of research but below the large-scale research objectives. We will show how the Polifonia Ecosystem evolved, and how the Ecosystem improves efficiency when tracing the quality and FAIRness of the research output and processes.

Paper Session: Literary Collections

Time: Thursday, 20/June/2024: 2:00pm - 3:30pm
Session Chair: **Anne Baillot**, DARIAH

Working around Walled Gardens: The Princeton Prosody Archive as Workflow

https://doi.org/10.5281/zenodo.12575879
Meredith Martin, Rebecca Sutton Koeser, Mary Katherine Naydan
Princeton University, United States of America; mm4@princeton.edu, rebecca.s.koeser@princeton.edu, mnaydan@princeton.edu

The Princeton Prosody Archive (PPA) is an open-source, full-text searchable database of 6,000+ English-language digitized works about the study of poetry, versification and pronunciation. But the PPA is also — technically and conceptually — a workflow that brings together materials from both HathiTrust Digital Library and Gale/Cengage's Eighteenth Century Collections Online into one searchable interface (figure 1). This workflow is necessary because today's digital research landscape consists of silos or "walled gardens" of scholarly materials controlled by proprietary vendors and digital libraries with layers of copyright restrictions. Designed to keep scholars inside them, these "walled gardens" limit scholars' ability to work across collections. What if scholars could analyze materials from multiple collections at once? What discoveries could scholars make if they weren't limited by a particular vendor's metadata, OCR quality, indexing practices, or built-in Natural Language Processing tools? [1]

The PPA has asked and answered these questions over a sixteen-year process of negotiating with vendors, collecting and curating materials, building the public-facing web application, and making the data usable for computational analysis. Our short paper will discuss how we gained access to additional levels of data by securing Memoranda of Understandings from HathiTrust and Gale/Cengage and worked within their data infrastructures to pull in bibliographic metadata, full-text OCR, and page images from their APIs and servers. Although our work improved their infrastructures and data, our workflows did not always align: for instance, we waited over a year in Gale's development queue for minor yet necessary enhancements to their API to access basic metadata. Similarly, we spent years correcting HathiTrust's inaccurate metadata for hundreds of works but could not feed those corrections back into HathiTrust because of HathiTrust's understandably limited ability to develop individual workflows with each partner research library.

The roadblocks we encountered while trying to integrate our workflows starkly illustrate how Big Tech and for-profit companies circumscribe how scholars conduct academic research, even on public-domain materials. For instance, despite HathiTrust being a not-for-profit collaborative of academic and research libraries that own the physical copies of the digitized works, we needed special permission from *Google* to display page image thumbnails because Google digitized 95% of HathiTrust works in the mid-2000s [2]. We had to prove that our archive of highly specialized texts about prosody — miniscule compared to HathiTrust's 18+ million volume collection (see figure 2) — wouldn't compete with Google Books! This example is one of many that exposes the fragility of the current research ecosystem, in which scholars encounter often invisible limitations around who can access these materials, how they can be used, and even which materials get included in the first place [3].

The PPA imagines an ideal scholarly information landscape in which the curation of data is driven by research questions, rather than by profit or by who happens to have digitized what. Rather than developing new resources from scratch, we imagine sharing workflows, ideas, and code with other scholars to remove barriers and work across — rather than just within — our research library's subscriptions.

Consolidating the heterogeneous landscape of literary corpora

https://doi.org/10.5281/zenodo.12807802
Ingo Börner[2], Vera Maria Charvat[1], Matej Ďurčo[1], Adrián Ghajari[3], Lukas Plank[1], Salvador Ros[3]
[1]Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), Austrian Academy of Sciences (OEAW), Austria; [2]University of Potsdam, Germany; [3]Spanish National Distance Education University (UNED), Spain; ingo.boerner@uni-potsdam.de, veramaria.charvat@oeaw.ac.at, matej.durco@oeaw.ac.at, sros@scc.uned.es

The Computational Literary Studies Infrastructure (CLS Infra) project aims to consolidate the heterogeneous landscape of data and tools in the CLS domain, with a specific focus on literary corpora. This contribution describes a data integration task against the background of developing a bespoke data model.

As part of the project (D6.1, Ďurčo et al., 2022), a metadata inventory of literary corpora was compiled to provide an overview of existing datasets, including their formats and mode of access. Based on the information derived from this initial collection of metadata and the Metamodel for Corpus Metadata (MKM; Odebrecht, 2018) as a conceptual starting point, the CLSCor ontology has been developed within the project as a unifying conceptual model to describe various aspects of literary texts. It introduces the basic entities: Corpus, Corpus Document and Feature, Feature being a generic mechanism to capture any specific characteristics of manifestations of literary works or collections thereof. These can be structural or semantic phenomena, like paragraphs, distinct speakers in a drama or verses in a poem. To guarantee its interoperability, the CLSCor ontology is based on CIDOC CRM and its extensions CRMdig and LRMoo. The CLSCor ontology will be accompanied by a set of controlled vocabularies (modeled in SKOS) that are currently in development.

In order to validate the applicability of the proposed ontology, data from three distinct datasets covering main literary genres were processed: novels (ELTeC), drama (DraCor), and poetry (POSTDATA). Not only do these sample datasets represent the different genres, but also three different underlying data structures and solutions for providing the data, making them ideal as proof of concept. The general workflow comprised mapping the respective source models to the CLSCor data model, transforming the data into RDF using custom scripts and merging the converted datasets into one consolidated knowledge graph. Given the distinct ways these three datasets were offered, a custom solution was implemented for each:

ELTeC data, exposed as TEI files in a github repository, was obtained by accessing the Github API.

Relevant information was extracted using XPath expressions and transformed into CLSCor-compliant RDF via a Python script.

Data from DraCor – TEI encoded corpora of plays – was integrated by transforming the JSON data returned by the DraCor API to RDF with a custom Python script.

In the case POSTDATA, the legacy tool "Horace", originally used in the project's data generation workflow, has been adapted to generate RDF data conforming to the CLSCor ontology.

After the milestone of merging three distinct sources into one graph, the graph is being evaluated for coherence, consistency and usability for exploration via a discovery application. The outcomes of this evaluation will inform further iterations and the integration of further sources.

Next to a rich catalogue of literary corpora, another envisioned outcome is robust conversion workflows for a variety of data sources and formats, accompanied by tutorials and promoted in training events to foster their reuse.

From NEWW to SHEWROTE: Developing a workflow to retroactively document a research dataset lifecycle and to ensure future data sustainability

Alicia Montoya[1], **Amelia Sanz**[2]

[1]Radbout University, The Netherlands; [2]Complutense University of Madrid, Spain; alicia.montoya@ru.nl, amsanz@ucm.es

Between 1997 and 2023, a network of 100+ scholars across Europe worked collaboratively, in a series of funded and unfunded projects, to produce a corpus of highly granular data on the reception of women authors from the Middle Ages to circa 1940. Embracing almost 7,000 women writers and 30,000 individual receptions of their works or persons, the data thus produced was successively stored in different databases: a first one developed at Utrecht University (1997-2014), and a second one, NEWW, built and hosted by the KNAW-Huygens Institute until 2023. At present, our WWIH DARIAH WG is developing a completely new, third iteration of the database, SHEWROTE, in partnership with the Humanities Lab of Radboud University (The Netherlands).

The redevelopment of the database has entailed a complex process of a) documenting the history of the database's previous iterations, particularly (implicit) decisions made in data harvesting and enrichment, b) rethinking the database's basic concepts and research questions, and formulating requirements in line with the research focuses of a new generation of scholars, and c) designing a new datamodel that is responsive to current needs, and adaptable to future needs, especially in data interconnectivity and sustainability. This conceptual rethinking preceded the work, which has now begun, of retooling and crosswalking data from the old (now archived) version of the database to the new one.

This paper presents some of the challenges faced so far in redeveloping the database. These have included, first, the move from a non-relational to a relational database structure, and designing a new datamodel that introduces several additional fields. Thus, the addition of the FRBR fields of Work and Manifestation will allow us to link data to existing identifiers, thereby prepping the data for data-linking with databases such as Cambridge's *Orlando* database, in the future. By introducing more granular place data, we anticipate mapping functionality and visualizations, to reflect new research focuses on the physical movement of texts and their authors in colonial and other settings. Further, by including alternative names, alternativ.e genders (for pseudonyms) and also dates during which alternative names (pseudonyms, married names) were used, we are making the existing Person data more granular, linking it to VIAF identifiers and creating new CERL identifiers when necessary, thus contributing to other collaborative datasets and to the broader (bibliographic) research community. Throughout this process, we strive to make decisions taken in the past explicit, documenting these systematically, and thereby retroactively making the tacit knowledge of generations of scholars, working in very different technical and institutional settings, compliant with modern FAIR data principles. Finally, we discuss the key issue of credits according to COARA Guiding Principles: how to document the contribution of the dozens of scholars, students and developers who contributed their labor and expertise to the database , as record creators and editors, and in other capacities, and how to acknowledge different levels and kinds of involvement in a manner that does justice to the complexities of digital, collaborative workflows and the research data lifecycle.

Paper Session: Visual Cultural Heritage

Time: Thursday, 20/June/2024: 2:00pm - 3:30pm
Session Chair: **Sally Chambers**, DARIAH-EU

Workflows for Digital Scholarship in Three Dimensions

Susan Schreibman, Costas Papadopoulos, Kelly Gillikin-Schouer, Alicia Walsh
Maastricht University, Netherlands, The; susan.schreibman@gmail.com

Reproducible workflows in the humanities have traditionally been embedded in research publications in the forms of footnotes, endnotes, citations, and bibliographies. This scaffolding allows researchers to trace the arguments in a publication to its antecedent roots. Reproducible workflows have taken on a new urgency with the advent of big data: making not just the underlying dataset available, but the algorithms used to parse the data so that the findings may be analysed, confirmed, or refuted.

We would argue that scholarship in 3D sits somewhere between these two modalities. Many of the workflows in 3D scholarship, particularly if that scholarship involves reconstruction of physical spaces (which may or may not currently exist or exist in a different form from the digitally reconstructed state), remain invisible to those outside of the team that created the model. The workflows in 3D reconstruction resembles traditional humanities scholarship in which a plethora of decisions are taken in the creation of an argument resulting in the 3D model. Here, it is virtually impossible to take a big data approach to make visible the workflow: even if the entire dataset (eg the model files) is deposited in a repository, it will not reveal the myriad of decisions taken in the model's construction. Perhaps the nearest equivalent for documenting the modelling workflow might be the lab notebook. But even with the notebook, researchers would also need access to all the paradata the modeller/researcher used in the decision-making process, in effect, creating an archive that does not benefit from the archivist's expertise in ordering and cataloguing.

Nevertheless, there is no doubt that scholarship in 3D lacks models so that it can become, in itself, a scholarly argument, as opposed to the ways in which 3D models are currently engaged with: as 'twirly things' available on a platform such as SketchFab (with limited annotation available) or in surrogates including videos and 2D images in articles.

PURE3D (funded by the Dutch PDI-SSH (Platform Digitale Infrastructuur–Social Sciences and Humanities) is filling this gap, creating both an infrastructure and a workflow from which 3D scholarship can be published, argued, and interrogated. The infrastructure is modelled on traditional text-based editions with one caveat: the text in a PURE3D edition is the 3D model, surrounded by multimodal annotation (text, images, video, structured/unstructured data). The editions also document the creation process (paradata), providing users with direct access to modelling/interpretative decisions, and source material (which can be embedded in the edition or hyperlinked online). The goal of the PURE3D platform is not to provide researchers with toolkit to create a reproducible model, but provide researchers with the underlying dataset (which is deposited in a trusted digital repository), the final published model/dataset, along with the decision-making process and source material. A peer review process provides an additional layer of transparency so that the edition becomes a knowledge site, making visible decisions, assumptions, and levels of certainty, along with the paradata, on which modelling decisions have been made.

Bridging the gap between visual cultural heritage collections and digital scholarship in DARIAH-FI

Inés Matres
University of Helsinki, Finland; ines.matres@helsinki.fi

Despite advances in the field of computer vision that have inspired some to call for a visual turn in digital humanities [1], [2], there are multiple challenges that prevent researchers in humanities and cultural studies to benefit from these advances or uptake digital workflows. Humanities scholars are often sensitive regarding the risk of bias in any automated way of interpretating images, and critical discrepancies have been found between the most advanced AI image labeling services [3]. Before digital workflows can be made accessible to research communities interested in extensive historical and contemporary imagery, research infrastructures (RIs) must be developed to enhance the state of data that is often unstructured, challenging to discover, and lacks proper contextualization.

In this presentation, we focus on research workflows familiar to humanists interested in visual cultural heritage materials, highlighting the insights from qualitative interviews with six visual researchers and findings of recent studies on their data practices. Many researchers still resort to hybrid approaches, which involve visiting archives, digitizing resources themselves, or acquiring materials from both open and private sources [4]. As a result, researchers examining historical or contemporary visual artefacts are facing similar challenges. A growing trend is the accumulation of mid- to large-sized visual corpora, alongside documents, interviews, or archival materials that are susceptible to be transformed into well-contextualized research datasets. Here, we emphasize the terms "susceptible" and "datasets" because, despite the existence of computer-assisted software for qualitative data analysis or for annotating visual collections, digital means to conduct research and generate data are not systematically adopted by the community [5]. Furthermore, recent research in Sweden and Finland have shown that sharing data is difficult for visual researchers due to GDPR, third-party ownership, copyright and the inadequacy of data sharing platforms for their often heterogeneous research materials [6], [7].

To address the challenges encountered by visual research communities, DARIAH-FI, currently in its developmental stage, will prioritize issues of technological and legal nature over the next two years. DARIAH-FI builds on years of experience in text-driven computational humanities in Finland, currently consolidating as a network of researchers in diverse fields of digital SSH from six universities, in collaboration with major hubs for computer science and digital cultural heritage aggregators (see www.dariah.fi). The RIs vision is to now include visual research communities and institutions giving access to Finnish photographic and other visual cultural heritage. The goal is to begin closing the workflow gap that exists among various scholarly practices in the humanities and cultural research. This will involve the participation of researchers whether they are familiar with digital humanities or not. The objective is to: 1) create AI models that can generate usable metadata for researchers, 2) connect visual objects to archival and textual sources essential for their interpretation, and 3) alleviating the visual culture research

community, whose workflows are constrained either by cumbersome accumulation of research materials, or barriers to publishing, sharing and archiving data.

Improving workflows in digital art history: the usefulness of patrimonial images segmentation

Léa Maronet[1], Truc Alice[2]

[1]École Pratique des Hautes Études, France - Université de Montréal, Canada - Centre National de la Recherche Scientifique, France; [2]Université Rennes 2, France - Université de Montréal, Canada; lea.maronet@huma-num.fr

Since Johanna Drucker questioned the existence of "digital art history" (Drucker, 2013), the last decade has seen a growing number of art historical projects make use of computer vision methods. However, digital art history remains a fragmented field: until now, these projects prefer to develop their own technical and methodological solutions. This project-based logic prevents the reproducibility of workflows and, thus, requires a significant financial and time investment, large quantities of data and technical skills that are not accessible to all research units, let alone all researchers (Romein et al., 2020, 310).

This logic ultimately hampers the standardization of digital practices, particularly in terms of data recording and algorithm training. Particularly, segmentation of patrimonial images is not subject to any standard or harmonization. It's not a question of trying to standardize the description of image content - an attempt that falls within the realm of ontology - but proposing a way to harmonize the recording of objects of interest within these images. If initiatives are being produced in the field of literary studies (Chagué, Clérice & Romary, 2021), in digital art history, the question remains little addressed (Bardiot, 2021). Yet having corpora of segmented images, whose objects have been identified by their coordinates and recorded in a hierarchical and standardized way, would make it possible to create ground truths that would serve as basis for training new algorithms and be profitable for every workflow in art history.

As workflows in digital art history are generally based on the technique of transfer learning, which reduce the amount of data required to train algorithms and improve the results obtained, this question is crucial in the discipline. However, the algorithms available are for the most part trained on "natural" images, which means they are hardly relevant to the specificity of patriomonial images. The absence of harmonization in the solutions proposed project after project culminates in a bitter realization: the tools available swiftly evolve, rendering them obsolete in the short term. Therefore, it appears vital today, in the age of Open Science, to pool our segmented images so they can serve as ground truths for future research, minimizing model training while enhancing the results obtained.

Segmentation always proposes an interpretation of the images. As such, how can we harmonize data segmentation, while remaining relevant to the epistemological requirements of art history? Asking this question implies putting into tension issues of interoperability and reproducibility, on the one hand, and epistemological reasoning in relation to the tools developed, on the other hand. These questions should not be addressed after tools development but upstream and should accompany the entire workflow (Stutzmann, 2010, 247-278). To address these tensions, the specific needs of digital art history about questions of image segmentation have to be identified, especially in the case of small and limited corpora, and existing standards in other fields will be discussed. The aim is twofold: to produce ready-to-use models and workflows adapted to patrimonial data, and to share data enabling these models to be produced.

Paper Session: Workflows for Cultural Heritage

Time: Friday, 21/June/2024: 9:00am - 11:00am
Session Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science

Bidirectional Workflow for Planning Data Stewards' Educational Activities at the National Level: An Example of Good Practice from Slovenia

Ines Vodopivec
National and university library of Slovenia, Slovenia; ines.vodopivec@nuk.uni-lj.si

Due to the fast and extensive transformation of the scientific digital landscape, which is largely oriented towards Open Science and constantly growing in complexity and size, scientific management is pushing for the development of resources and information access in the remodelling of associated working profiles. Open Science and Research Data Management (RDM) require new skills, not only in the technical field and in educational institutions (universities, faculties), or research institutions, but also in heritage institutions, where data management specialists have been trained for decades to work with data systems, databases, data strategies, and data storage.

The transition of the new profiles in the RDM field to data stewards and data managers is most obvious in libraries. Libraries, besides being heritage institutions, are also core providers of RDM activities and are therefore very active in the area of Open Science. National and research libraries have traditionally played an important role in knowledge exchange by collecting and preserving (digital) knowledge, exchanging metadata, and providing end users support.

In Slovenia, libraries have been recognized as the most relevant institutions for long-term sustainable data stewardship service delivery, as well as environments for the development of new skills and working profiles. The education system for library professionals in Slovenia has traditionally been provided through university programs for formal education, and by the National and University Library for early-stage and ongoing professional development in the field. Additionally, a new Open Science project started last year, which will provide educational activities in collaboration with the Data Stewardship, Curricula, and Career Paths EOSC Task Force, enabling the employment of new working profiles.

However, a statistical analysis was conducted, revealing that only five percent of professionals younger than 35 years are currently employed in all university and scientific libraries in Slovenia. With anticipated progress in the next ten to fifteen years, relying solely on such small percent of younger generations for planning the development and advancement of the profession is not feasible. For this reason, a special national educational model was developed in Slovenia to enhance the knowledge of all generations working in libraries.

This paper addresses the question of how to establish sustainable, long-term research support through Open Science educational activity workflows. It presents a newly developed national model for integrating Open Science educational programs aimed at teaching new data stewards and data managers in Slovenian libraries, as well as enhancing the knowledge of senior librarians to upgrade the profession. The workflow of this national model is built on three main stakeholders: the policy makers as the financers, libraries as the providers, and universities/researchers as the end-users. This workflow represents bidirectional planning, anticipating both top-down and bottom-up approaches, and offering knowledge to the wider professional field (not limited to library types).

Digital Revival of the Slavonic Manuscript Collection of the Plovdiv National Library

https://doi.org/10.5281/zenodo.11234633
Maxim Krasimirov Goynov[1], Ivan Georgiev Kratchanov[1,2], Detelin Mihaylov Luchev[1], Desislava Ivanova Paneva-Marinova[1], Radoslav Dimov Pavlov[1]
[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria; [2]National Library "Ivan Vazov", Plovdiv, Bulgaria; ivankra@gmail.com, dml@math.bas.bg

One of the largest collections of manuscripts in Bulgaria is housed in the National Library "Ivan Vazov" (NLIV) in the city of Plovdiv. The library's collection consists of 362 exemplars, most of which are Slavonic, but also many Greek, Ottoman Turkish, and Persian manuscripts, dating from the 11th to the 19th centuries, written on both parchment and paper. The Slavonic manuscript collection contains highly valuable medieval works that are well-known in academic circles and that have undergone extensive paleographic analysis and description. These include the Kyustendil Palimpsest from the late 12th century, the Apostle of Slepcha (fragment) from the second half of the 12th century, the Kichevski Triod from the second half of the 12th century, and the Trebnik of Daniil of Etropole from 1592, among others.

NLIV is one of the institutions in Bulgaria responsible for preserving the manuscript heritage of the nation, as well as for promoting and providing access to it. In fulfillment of these expectations, since 2016 NLIV digitized and made the Slavonic manuscript collection available online at the federated digital repository of the National Academic Library and Information System (NALIS), which includes content from several universities and libraries. Because of the detailed paleographic description of the manuscripts, the metadata standard of choice was MARC 21, which offers an appropriately high degree of granularity.

In October 2021, NLIV launched a new online digital library (digital.libplovdiv.com), based on the CultIS web-based software platform for intelligent digital management and presentation of large datasets and knowledge in the fields of culture, humanities and social sciences

(cultis.math.bas.bg). CultIS is developed and maintained by the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, within the CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH.

CultIS incorporates some of the most current and advanced software technologies, giving a wide range of functionalities for content manipulation, descriptive cataloging, indexing, intelligent data searching, curation, and collection development, aiming to meet the functional needs of different users.

Metadata structures in CultIS are managed using a dynamic models building service. It provides the ability to maintain complex structures (including arrays and recursive relations), and to extend them anytime. Because of the CultIS platform's better performance, advanced features and instrumentarium, a decision was made to migrate the Slavonic manuscript collection from NALIS (using dSpace v6.2) to the new NLIV digital library.

The process was automated by data extraction tools, developed for loading and transforming data from the old system. Data loaders for processing the METS and MARC 21 standards used in dSpace were implemented, and metadata was mapped to the manuscript descriptive model defined in the new NLIV digital library. Using the module for the management of the metadata model, the descriptive model could be continuously modified by adding new categories and fields as well as by changing the condition of the fields as access levels, and reviewing the changes made to the system by the users.

Culture Data Aggregator - data processing and enrichment workflow

Marcin Heliński, Aleksandra Nowak, Tomasz Parkoła

Poznan Supercomputing and Networking Center, Poland; helin@man.poznan.pl

For over 20 years, Poznan Supercomputing and Networking Center has been supporting the work of scientists and researchers of cultural heritage and digital humanities. Between 2021 and 2023 PSNC was a leader of the DARIAH-PL project, carried out by the members of the Polish DARIAH-PL network. As a result, the digital humanities infrastructure called Dariah.lab was created. The infrastructure includes hardware and software tools using state-of-the-art technological solutions and integrated digital resources from various fields of humanities and arts sciences. The network of distributed research laboratories constituting Dariah.lab was designed to meet digital humanities researchers' needs.

Fundamental activity in digital humanities and cultural heritage researchers work is the use of digital libraries and repositories. Metadata of digital objects collected in such services may differ significantly between instances run by different institutions due to the nature of the provided digital resources and institutional policies. To simplify and improve access to the information from those various sources PSNC decided to work on the data aggregation components. The intention is to provide a single access point (called Culture Data Aggregator) to the variety of digital resources that can be reused in multiple scenarios, including aspects related to science, business or communities. The portal allows users to discover digital objects both using their original metadata as well as metadata enriched using, among others, semantic tags. However, before the data is visible in the portal it has to go through the processing path, starting from aggregation, through mapping, unifying, enrichment, or indexing.

This data processing workflow is supported by DACE (Data Aggregation and proCessing Engine) that has emerged from more than a decade of PSNC's activities on data aggregation pipelines. DACE consists of multiple scalable microservices that communicate with each other via the Apache Kafka messages carrying information necessary to process data in each of the workflow steps. Microservices involved in data processing have their own specific tasks. These include aggregation of digital objects metadata both via the OAI-PMH protocol and a selection of other well-known or very specific APIs (e.g. Z39.50, Mediawiki API, Wordpress API), XSLT transformation of records from the source schema to the internal schema used by the DACE system, unification and cleaning of records metadata, enriching objects with metadata from the LOD dictionary and indexing objects metadata in the Solr search engine (for further use in the discovery portal).

The mentioned enrichment step uses the instance of WikiBase service prepared specifically for Culture Data Aggregator. This service contains entities retrieved from the catalog of master records run by the National Library of Poland. While processing the certain digital object metadata the enrichment microservice detects values present in a configurable set of schema fields and tries to associate them with entities from the LOD dictionary.

In our paper we would like to dive into the details of the Culture Data Aggregator architecture and elaborate on the data processing workflow showcasing how it can be used to help digital humanities in their research work.

Validating a reproducible workflow for publishing Collections as Data: the case of Europeana

Sally Chambers[1], Gustavo Candela[2,1], Alba Irollo[3], Vicky Dritsou[4,1], Agiatis Benardou[1,4], Nuno Freire[3], Vicky Garnett[1], Toma Tasovac[1]

[1]DARIAH-EU, Paris, France.; [2]Department of Software and Computing systems, University of Alicante, Spain; [3]Europeana Foundation, The Hague, The Netherlands.; [4]Digital Curation Unit, R.C. "Athena", Institute for the Management of Information Systems, Maroussi, Greece.; sally.chambers@ugent.be, gcandela@ua.es, v.dritsou@dcu.gr, nuno.freire@europeana.eu

Cultural Heritage institutions have been exploring new ways to make their digital collections available for arts and humanities research. Examples include the Data Foundry at the National Library of Scotland, which publishes data openly and in reusable formats (Ames, 2021); the open data platforms facilitating data-level access to digitised and born-digital collections at BnL and KBR, and the BnF Data Lab and KB Lab as experimental labs to actively engage and support researchers in the use of cultural heritage data. Many of these initiatives are inspired by the international 'Collections as Data' movement (Padilla et al., 2019) which has most recently resulted in the Vancouver Statement on Collection as Data (Padilla et al., 2023).

This paper aims to build on the legacy of the DARIAH Annual Event 2023 on cultural heritage data as humanities research data. Within the context of the International GLAM (Galleries, Archives, Libraries and Museums) Labs Community, a Collections as Data Checklist (Candela et al., 2023) was published as an easy-to-apply method to encourage small and medium-sized cultural heritage institutions to publish their digital collections as 'Collections as Data'. The deployment of a common European data space for cultural heritage has opened up new possibilities towards a decentralised approach for publishing cultural heritage data, with the Europeana Foundation and DARIAH leading the engagement with academic and research communities. To meet their needs, especially in relation to computational methods, the checklist became the basis for the development of a Workflow to publish Collections as Data, published in the Social Science and Humanities (SSH) Open Marketplace.

The workflow is designed to make cultural heritage institutions aware of the essential steps to publish datasets suitable for reuse. It offers a range of resources for each step and aims to cater to the needs of those interested in reusing datasets for research, teaching and learning purposes at the academic level. It also describes a number of potential use cases from different

domains. This paper will focus on Europeana.eu and new possible scenarios in the European data space for cultural heritage. While researchers are used to considering access to the Europeana platform at the item level or via the APIs, the Collections as Data approach led to new reflections on the need for the publication of datasets with characteristics, such as those described in the workflow, more suitable for reuse in research. Furthermore, it paved the way for new features enabling researchers to curate and publish datasets on the platform.

Complementary to this is the ongoing development of a DARIAH Campus curriculum focussing on the reuse of digital cultural heritage in Higher Education and Research, which will include courses on the Europeana APIs and the workflow. This paper also aims to investigate which other courses to include in the curriculum, so that such training materials can be used to encourage uptake and use of Collections as Data within the DARIAH community and beyond.

Posters & Demos

VELD: Versioned Executable Logic and Data - Making digital workflows reproducible in a reliable way

Stefan Resch, Matej Durco
ACDH-CH OEAW, Austria; stefan.resch@oeaw.ac.at

Making digital workflows reproducible in a reliable way is challenging due to the inherent complexity of software environments, since it involves installing language compilers and libraries, configuring them, managing correct paths of data and code dependencies, as well as external resources such as data APIs or repositories. While there are various packaging and building tools to help resolving dependencies (e.g. pip, maven, npm, etc.), these cover only a part of the overall setup. Also, once a stable configuration has been found, oftentimes little effort is spent to ensure it is usable on different systems and over time.

The EU project Computational Literary Studies Infrastructure[1] aims at consolidating the heterogeneous landscape of data (literary corpora) and tools (like NLP or stylometry) and making them more easily available for literary studies. This goal implies the need for composable and reproducible workflows. In answer to this need we conceived VELD: Versioned Executable Logic and Data. VELD is not a tool itself, but rather a design pattern relying on two widely used and proven industry technologies: docker and git.

Docker[2] is a virtualisation technology that encapsulates arbitrarily complex software setups in an isolated environment, avoiding conflicts with the host OS and other docker environments. This enables high reproducibility over varying systems and over time, as well as high portability between local and remote (server infrastructure) execution, easing integration and maintenance costs.

Git[3] is a de facto standard for collaborative version control. In the context of VELD, it is used for structuring reusable components (data, code) into separate git repositories that are then grouped per workflow or project using git submodules[4].

VELD combines the execution isolation of docker with the state management of git by defining three objects:

- Data veld: a git repository that contains purely data, to be used as input or output or both.
- Executable veld: a git repository that contains source code and docker context to encapsulate the code.
- Chain veld: a git repository that wires together the above git repositories containing data velds or executable velds by linking them as git submodules, thereby referencing their most recent state and the entire history. A chain is thus the reproducible persistence of a workflow.

The key to the composability of VELD is a metadata schema describing the expected file formats, interfaces, mount points, documentation, based on a bespoke model for capturing characteristics of corpora and tools being developed in the CLS INFRA project. The metadata is expressed in yaml[5] format and included in the docker compose file[6] within the git repository capturing the structure and mechanics of individual veld components and governing their execution.

The VELD design was already used successfully for real-world tasks, while simple examples[7] with documentation have been implemented as well for a quick introduction. As a next step and a goal of the CLS INFRA project, an initial set of pre-defined and adaptable workflows based on selected NLP tools will be made available to the community, both for local and remote usage through dedicated infrastructure.

Echoes of Discord: Unveiling Violent Whispers in 1920s Slovenian Newspapers

Marko Milosev[1], Álvaro Pérez[2], Salvador Ros[2]
[1]Central European University; [2]UNED, Spain; alvaro.perez@linhd.uned.es, sros@scc.uned.es

Digital humanities, with its interdisciplinary approach, holds the potential to uncover obscured aspects of historical narratives. This paper presents a novel approach to analyzing historical media content for calls to violence using advanced natural language processing techniques. The main source for analysis is the bulletin *ORJUNA*, published by a radical right extremist organization ORJUNA (Organization of Yugoslav Nationalists) during the 1920s. This organization was notorious during the interwar period in the Kingdom of Yugoslavia for perpetrating acts of political violence against minorities and political opponents. By employing Large Language Models (LLMs), we systematically label incitement towards violent action, and any covert messaging that may lead to violence within this media and correlate these instances with actual events of violence. This study underscores the importance of understanding the role of media in inciting or reflecting societal violence, and disseminating extremist views, providing insights into the dynamics between media narratives and real-world actions.

Our implementation showcases the utilization of LLMs within a conversational pipeline. The benefits of this methodology are twofold: it reduces the need for extensive labeled datasets, which are often scarce in historical media analysis, especially in low-resource languages such as Slovenian, and it offers an alternative to the significantly more resource-intensive approach of fine-tuning our own model.

We employ a combined methodology of "few-shot learning" (Brown et al., 2020) and "chain of thought" (Wei et al., 2022) reasoning to give context to our model with minimal examples while encouraging it to articulate the reasoning process behind its classifications. This approach not only enhances the model's accuracy in identifying nuanced expressions of violence but also offers transparency in its decision-making process.

The "few shot" learning technique allows LLM to glean insights from a limited set of examples, enabling it to generalize its understanding of radical language patterns in 1920s Slovenian newspapers. This approach enhances the language model's adaptability to historical contexts with scarce labeled data, making it a valuable tool for digital humanities research. On the other hand, the "chain of thought" technique unfolds as a structured narrative, guiding the LLM through a meticulous process of recognizing and labeling linguistic nuances associated with radicalization and violence. The combination of these two techniques not only enriches LLM's interpretative abilities but also contributes to the broader understanding of the socio-political landscape of 1920s Slovenia. The significance of this research lies in its potential to shed light on historical events that may

have been overshadowed or misunderstood. By uncovering patterns of extremist radicalization within the context of 1920s Slovenian newspapers, it contributes to a deeper understanding of the socio-political landscape of the time.

In conclusion, this research contributes to the evolving landscape of digital humanities, showcasing the practical implementation of advanced language models in historical research, casting light on historical events obscured by the shadows of time. The methodologies introduced in this paper offer a framework for leveraging Mistral (Jiang et al., 2023) LLM in similar studies across different historical periods and regions, thereby expanding the scope and impact of digital humanities research.

Fostering reproducible research by linking data and publications

Nicolas Larrousse[1], Olivier Baude[1], Marie Pellen[2], Dominique Roux[3]
[1]Huma-Num, CNRS, France; [2]OpenEdition Center, CNRS, France; [3]METOPES, CNRS & Université de Caen, France;
Nicolas.Larrousse@huma-num.fr

In order to support research projects in SSH (Social Sciences and Humanities) in building Open Science, France has developed complementary infrastructures, Huma-Num, OpenEdition and Metopes:

- Huma-Num[i] provides a set of platforms, tools and support for the processing, dissemination and preservation of digital research;
- OpenEdition[ii] provides a set of four scholarly communication platforms and Metopes[iii] provides a set of tools and methods, built according to single source publishing model, enabling the creation of natively structured editorialized context.

The aim of the COMMONS[iv] project is to create links between data and publications by building bridges between platforms provided by the infrastructures in a seamless way[v]. But what is exactly a link between data and publication and how to define it?

There is no single answer to this question:

- A simple and not normalized citation in a form of text reference in a publication which is not easy to process and produces not so reliable results;
- A more reliable way is to express the link between resources on the different platforms, a minima in one metadata, in the respective information systems of the different platforms. The information is then made available for reading and processing, but remains relatively "hidden" in the information systems;
- A step forward is to publish the information linking data and publications in a standardized way, for instance using SCHOLIX[vi]. Information thus becomes consumable by machines and can feed aggregators like ScholeXplorer[vii];
- Another way of representing the link between data and publication is to show the data embedded in a publication, even allowing to manipulate this representation in order to obtain different views of it. In that case, more specific metadata is required to feed properly visualization tools.

The COMMONS project aims to implement these different levels of linkage between Data and Publications. This can then serve as the basis for other "sub-products" such as Data Papers. This workflow of multiple publications is more likely to occur at the end of the research process, but should be considered at an early stage of the project to be efficient. COMMONS is going to develop tools to facilitate this multiple publication on the platforms provided by the infrastructures. By facilitating access to data used to write articles this will improve the reproducibility of SSH research, which is currently rather limited especially compared with other disciplines. On the other hand, being able to access the publications associated with a dataset will increase trust in the dataset and will foster the motivation to reuse it.

To ensure that this system is stable over time and therefore fulfills its purpose, a number of issues will have to be addressed during the COMMONS project, in particular to maintain the consistency of information: for example, what to do when a new version of the dataset is submitted.

However, one of the most important challenges is to motivate and train users to use the tools developed by the project.

[i] https://huma-num.fr - https://documentation.huma-num.fr/humanum-en
[ii] https://www.openedition.org
[iii] https://www.metopes.fr
[iv] COnsortium of Mutualised Means for OpeN data & Services for SSH
[v] https://shs.hal.science/halshs-03881307
[vi] http://www.scholix.org/
[vii] https://scholexplorer.openaire.eu

Weasel: A Tool for Creating Reproducible Research Workflows

Nick Budak[1], Andrew Janco[2], David Lassner[3]
[1]Stanford University, United States of America; [2]University of Pennsylvania, United States of America; [3]Independent Scholar, Germany; budak@stanford.edu, apjanco@upenn.edu, david@kapitel-software.de

Weasel is an open-source tool designed to facilitate the management and sharing of end-to-end workflows. Initially developed for machine learning tasks, weasel's capabilities are well-suited to the needs of art and humanities researchers who need reproducible pipelines to support open research practices. Because it encourages depositing and sharing scripts that prepare data along with research, Weasel supports data-driven humanities researchers beginning at an early stage and throughout the iteration process.

This demonstration shares two case studies that demonstrate the versatility of weasel. The first case study explores how Weasel manages data and code used to add a new phonology-predicting component to an NLP pipeline. The second case study

showcases how Weasel transforms a large collection of historical sources into research data. Through these examples, we discuss the pros and cons of using Weasel. One highlight compared to other workflow management tools is that Weasel has minimal setup complexity as it does not require containerization.

Furthermore, we argue that Weasel addresses a critical need in arts and humanities research by providing a tool that enforces reproducible workflows and adheres to open research principles. This paper delves into the challenges researchers face in consistently following open research values and practices and emphasizes the importance of a tool that enforces open practices through its structure.

New NLP and spaCy projects

The lack of existing models for historical and low-resource languages presents a significant barrier to research. The "New Languages for NLP" project provides instructional materials and workflows to enable researchers to create the necessary training data and models. For this project, spaCy projects (and now weasel) provide a structured and reproducible workflow for training language models in currently unsupported languages.

Colombian Court Documents

This case study details a project that uses weasel to transform scans of archival documents from Colombia into structured research data. The documents are nearly all handwritten and have significant mold and water damage. They are court records from a local courthouse in Istmina, Colombia that were digitized through the British Library's Endangered Archives Program (EAP1477). Each section of the workflow focuses on a specific task related to transforming the raw materials into structured data.

Making a Weasel projects repository targeted at DH community

Weasel GitHub repository offers various project examples that can serve as a starting point for many data science or machine learning projects (template projects).

In the NewNLP institute [1] that heavily relied on spacy projects (the predecessor of weasel), we developed template projects that were specifically tailored toward a Digital Humanities user base. For example, we developed a project for creating the relevant assets for a spacy model for entirely new languages (new in the sense of languages that were not supported by spacy).

To that end, we are preparing a new GitHub organization and repository where we provide our weasel projects as templates for everyone to use. This place shall serve as a hub for sharing DH-specific weasel projects from the community in the future.

Domain-Specific Digital Scholarly Editing as an integration of traditional and digital methods: the case study of GreekSchools

https://doi.org/10.5281/zenodo.12077552

Angelo Mario Del Grosso[1], **Federico Boschetti**[1,3], **Simone Zenzaro**[1], **Graziano Ranocchia**[2]

[1]CNR-ILC, Italy; [2]Università di Pisa; [3]VeDPH Università Ca' Foscari Venezia; angelomario.delgrosso@cnr.it, federico.boschetti@ilc.cnr.it, simone.zenzaro@ilc.cnr.it, graziano.ranocchia@unipi.it

The goal of scholarly editing is to reconstruct and publish texts by exploring philological phenomena and documenting them in critical apparatuses or editorial notes. Textual scholarship involves complex, multi-stage processes, and the digital landscape requires even additional effort from textual scholars, as it necessitates formal protocols and standard representations of workflows and resources. This formalization step is often perceived as burdensome by traditional scholars and may be error-prone and time-consuming. An intermediary approach for DSE is based on DSLs. It is aimed to bridge the gap between traditional methodologies applied to printed editions and the recommended digital techniques endorsed by DHers. The latter refers to the established practice of producing editions by XML-encoding in compliance with TEI guidelines.

The proposed DSL-Based DSE aims to preserve scholars' longstanding ecdotic knowledge while leveraging computational capabilities compliant to the open science digital agenda, based on the FAIR principles. Our approach follows two principles: 1) enabling scholars to work in a familiar text-focused environment, and 2) ensuring machine actionability and interoperability. The GreekSchools ERC-885222 project offers the perfect setting for developing the aforementioned workflow. Specifically, the aim of the project is to publish scholarly editions of a portion of the Herculaneum papyri - a collection of carbonized scrolls with unedited Greek texts.

The potential of computational advancements to enhance the workflow of papyrologists has been demonstrated by recent AI initiatives, such as the Ithaca project and the Vesuvio Challenge, and by platforms like the papyri.info. Moreover, the GreekSchools aligns with the goals of the H2IOSC consortium targeting the following RIs: CLARIN for the computational linguistic technologies; DARIAH for the digital humanities technologies; E-RIHS for the advanced imaging techniques; and OPERAS for the open scholarly communication technologies. In this framework, we propose CoPhiEditor, a web-based digital scholarly environment designed to enhance the GreekSchools workflow. Compared to similar initiatives such as Proteus, Textual Communities, and LEAF-writer, CoPhiEditor implements the DSL-based DSE approach.

The primary goal is to facilitate a collaborative and cooperative workflow, allowing scholars to work on the same text simultaneously or on different parts of the text asynchronously. Building on this workflow and principles, CoPhiEditor leverages formal DSLs, language technologies, and IIIF framework to provide automated support, enhancing the quality of the editing process and the editorial work. CoPhiEditor reduces the need for manual checks to ensure adherence to editorial conventions, such as encoding lacunae, unclear characters, and apographs' readings. It also verifies cross-textual constraints, such as the consistency between diplomatic and literary transcriptions, and the correctness of the apparatus entries. Moreover, it maximizes time spent on textual phenomena, such as evaluating hypotheses for reconstructing the text with readings and conjectures.

In addition, CoPhiEditor maintains the adherence to standard formats like EpiDoc. This guarantees interoperability of the edition across various software, enabling processing, such as search capabilities, information extraction, and the creation of training-sets for machine learning. CoPhiEditor is an open-source component, adaptable to different philological projects. Indeed, the tool extends to initiatives like the DiScEPT project, demonstrating its potential to meet new requirements.

The ATLAS of Italian Digital Humanities: a knowledge graph of digital scholarly research on Italian Cultural Heritage

https://doi.org/10.5281/zenodo.11569280

Alessia Bardi[1], Marilena Daquino[2], Riccardo Del Gratta[1], Angelo Maria Del Grosso[1], Roberto Rosselli Del Turco[3]

[1]CNR, Italy; [2]University of Bologna; [3]University of Torino; angelomario.delgrosso@cnr.it

ATLAS is a project funded by the Next Generation EU program of the European Commission for 24 months (October 2023 - October 2025) that aims to improve the FAIRness and exploitation of Digital Humanities (DH) projects and scholarly data about Italian cultural heritage (https://dh-atlas.github.io).

DH research outputs are often not easy to discover, and risk obsolescence if not well documented and based on shared guidelines and standards. Moreover, projects are often self-referential, meaning that they may not follow metadata standards and best practices. In addition, users' experience is limited in exploration, since there is a lack of interlinking across projects with a clear content overlap and explanations of such overlaps - including contradictory statements or disagreement.

The goal of ATLAS is to identify shared metadata standards, protocols, reusable workflows, good practices, guidelines and evaluation frameworks in the Digital Humanities. To tackle the aforementioned problems in real-world scenarios and ensure the representativeness of identified guidelines, a pool of selected sources will be integrated in a knowledge graph, and reengineered with state-of-the-art Semantic Web technologies and Natural Language Processing methods. Pilot projects will help us to define guidelines and create a golden set of reference projects in the Digital Humanities. The aim is to collect sources that are published according to shareable criteria, that can be easily mined to extract research topics, inter and intra-textual relations, as well as bibliographic, literary, and thematic data. This will allow us to define quality criteria for recommending best practices to future projects.

Moreover, the data extracted will be reconciled with international authority records (e.g. VIAF) and open data sources (e.g. Wikidata) to facilitate their reuse and the development of mashup applications. Finally, such enhanced data will be preserved and leveraged in a dedicated platform to support exploration and discovery of the landscape of DH projects, and will provide suggestions on tools and resources to scholars that are planning new projects.

In summary, the ATLAS project will contribute to the Italian DH research community with four main results:

- A whitebook including results of the analysis of the state of the art and good practices for FAIR scholarly data
- A knowledge graph on DH projects and scholarly data on Italian Cultural Heritage, accessible online via the ATLAS web application and preserved in CLARIN [1].
- The pilots evaluation, highlighting differences and strategies to cope with mapping knowledge, data manipulation, access and persistence of different types of digital artefacts.
- A search portal dedicated to scholarly literature and data relevant to the pilots and beyond, built on top of the OpenAIRE CONNECT Gateway on Digital Humanities and Cultural Heritage [2,3].

References

[1] CLARIN IT. https://www.clarin-it.it/it

[2] OpenAIRE CONNECT Gateway on Digital Humanities and Cultural Heritage. https://dh-ch.openaire.eu

[3] Baglioni, M. et al. "The OpenAIRE Research Community Dashboard: On Blending Scientific Workflows and Scientific Publishing." Digital Libraries for Open Knowledge. Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-30760-8_5.

Documenting sustainable workflows for a multilingual publishing project and the case of the Programming Historian

https://doi.org/10.5281/zenodo.12167753

Anisa Hawes[1], Charlotte Chevrie[2], Joana Vieira Paulino[3], Anna-Maria Sichani[4], Eric Brasil[5]

[1]ProgHist Ltd, United Kingdom; [2]ProgHist Ltd, United Kingdom; [3]Programming Historian em português; Universidade Nova de Lisboa, Portugal; [4]ProgHist Ltd, United Kingdom; School of Advanced Study, University of London, United Kingdom; [5]Programming Historian em português; Universidade da Integração Internacional da Lusofonia Afro-brasileira, Brazil (IHLM/UNILAB); admin@programminghistorian.org, publishing.assistant@programminghistorian.org, jpaulino@fcsh.unl.pt, annamaria.sichani@sas.ac.uk

Programming Historian is the title shared by four online journals published by ProgHist: *Programming Historian in English* (launched in 2012), *en español* (launched 2016), *en français* (launched 2019), *em português* (launched 2021).

Our publications empower the development of digital research skills in the humanities, through article-length lessons on digital techniques and computational methods.

The lessons are created by a global community of authors, editors, peerreviewers and translators who have collaborated to shape more than 240 lessons across the four editions.

Within the past three years, ProgHist has professionalised the publishing services that subsist, support and sustain our journals. The transition from volunteer initiative to professional publisher has involved re-thinking and reconstructing our workflows to improve editorial efficiency and enhance contributors' experience, while achieving consistently high quality, and ensuring that the learning resources we publish are accessible and sustainable.

GitHub Pages is the hub of our editorial and pre-publication workflows. GitHub hosts our website architecture, while also providing an infrastructure for editorial exchange and peer review discussion.

For DARIAH's Annual Event 2024, we propose a poster that showcases the workflows we have designed, developed and documented within this platform. In particular we want to share our inventive use of GitHub Project Boards and labels to track lessons from Proposal, through Open Peer Review and Revisions, towards Publication in a form that we think empowers contributors in their roles and supports cross-journal collaboration. Our poster will also provide insight into how we are creating Mermaid diagrams to document our workflows, defining activities, responsibilities and timeframes at each phase.

DARIAH Media Hub - optimized data delivery service for digital humanities

Marcin Heliński, Aleksandra Nowak
Poznan Supercomputing and Networking Center, Poland; helin@man.poznan.pl

For over 20 years, Poznan Supercomputing and Networking Center has been supporting the work of scientists and researchers of cultural heritage and digital humanities. This cooperation led to creating many systems widely used for digitization, archiving, processing and sharing digital resources on the Internet. PSNC was also a leader of the DARIAH-PL project (2021-2023), carried out by the members of the Polish DARIAH-PL network. As a result, the digital humanities infrastructure called Dariah.lab was created. The infrastructure includes hardware and software tools using state-of-the-art technological solutions and integrated digital resources from various fields of humanities and arts sciences. The network of distributed research laboratories constituting Dariah.lab was designed to meet digital humanities researchers' needs. It consists of Source Laboratory, Automatic Enrichment Laboratory, Supervised Semantic Discovery Laboratory, Intelligent Analysis and Interpretation Laboratory, and Advanced Visualization Laboratory. The purpose of building the Dariah.Lab research infrastructure was to expand the scope of research in the humanities and arts in Poland, both in the purely scientific context, as well as in the area of economic applications.

In recent years, PSNC has also intensively focused on the active use of IIIF technology which is an international standard for delivering high-quality images and audio / visual materials to different end-user environments. The variety of formats of digital objects in digital libraries and repositories has been a challenge for software developers for many years. IIIF which stands for International Image Interoperability Framework is a very efficient way to solve this problem by providing interoperability and a consistent way of accessing the content of digital resources.

The international success of IIIF technology encouraged us to use it in Dariah Media Hub (DMH), an optimized data delivery service created within the DARIAH-PL project. This system allows users to upload and host their files in multiple formats. It supports various image formats (PNG, JPEG, TIFF) and PDF documents. Files of this type are internally converted to the IIIF compliant image format (pyramidal TIFF). Moreover, users are able to store audio, video and 3D materials within the service. All the mentioned file types are accessible directly or via a IIIF manifest file available for use in any IIIF-compatible viewer (e.g. Mirador or Universal Viewer).

If we look more broadly at the elements of the Dariah.lab infrastructure, the Dariah Media Hub service is a step in a larger workflow that involves services producing images, audio / visual or 3D materials that are considered as input data to DMH, and services that use IIIF-compliant images converted in DMH and IIIF manifests for further research work. One of the services that use the results of DMH is the Virtual Transcription Laboratory (VTL). It allows users to prepare various types of text layers (transcriptions, transliterations, translations or annotations) based on scans.

In our poster, we would like to present the data processing workflow both in terms of the entire Dariah.lab infrastructure and within the optimized data delivery service. We will also demonstrate the use of DMH to deliver IIIF-compatible resources.

Proto4DigEd: Prototypical Workflows for Digital Editions

https://doi.org/10.5281/zenodo.11632615
Ursula Bähler[1], Elias Zimmermann[1], Yann Stricker[2], Reto Baumgartner[1], Rita Gautschy[3]
[1]UZH, Switzerland; [2]UB/UZH, Switzerland; [3]DaSCH, University of Basel, Switzerland; rita.gautschy@dasch.swiss

Goal

The main goal of the Open Research Data (ORD) project Proto4DigEd funded by swissuniversities for a duration of 18 months (07.2023–12.2024) is to test and further develop prototypical workflows in the conception phase of digital edition projects. Although fitted into the Swiss research landscape, its outputs are tailored to international requirements in the spirit of DARIAH-EU, offering solutions to its communities: Proto4DigEd answers to a growing need for standardisation concerning digital edition projects (Pierazzo 2019: 2014). For this aim, the project uses a small subset of the planned binational digital letter edition PARES (Gaston Paris: des archives aux reseaux, Switzerland and France).

Standardised Design

Proto4DigEd is concerned with all stages of the data life cycle from the conception phase until the publishing and archiving of the data in a repository. The project is not intended to develop new technical solutions, but focuses on a combination of easily available, long-time proofed and thus sustainable tools and on seamless transitions between them. For the automated transcription of manuscripts (HTR) the tool *Transkribus* offers general text recognition models as well as the training of specific models. *TEI Publisher* is a publishing toolbox developed in Switzerland and Germany that is supported by a growing community and allows standardised markup of texts in TEI-XML and their online publication. The *DaSCH Service Platform (DSP)* allows long-term archiving and long-term visualisation of edition data.

International Scope: Multilingual Editions

Multilingualism is an important concern in our workflow since Switzerland is a country with more than one official language and strong connections to French, German and Italian communities. The multilingual international correspondence of Gaston Paris is therefore a testcase for not only Swiss, but any pan-European project. Proto4DigEd develops versatile workflows for the use and training of multilingual HTR-Models and for the linking of metadata to different national and international authority databases with different language sets.

Community Driven Output

The completely open-source output consists of a model edition linked to a handbook of our workflows. The latter is based on MkDocs and continuously enriched and will be made available on the GitLab instance of the University of Zurich, available from spring 2024 (an URN will be available by the time of the conference). The Center Digital Editions & Edition Analytics, ZDE UZH, helps in sharing the workflow and offers discussion and training to any interested project; for this we envision to put the workflow – beside other modes of communication – on the SSH Open Marketplace. We thus focus on the needs of national and international communities, as the project aims for their needs in reproducible, long-time available and sustainable workflows.

References

DaSCH Service Platform (DSP): https://docs.dasch.swiss/DSP-API/01-introduction/what-is-knora/

Pierazzo, Elena. 2019. "What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter". International Journal of Digital Humanities 1 (1): 209–220. https://doi.org/10.1007/s42803-019-00019-3 (accessed 10.01.2024)

TEI-Publisher: https://teipublisher.com/index.html

Transkribus: https://readcoop.eu/de/transkribus/

ZDE UZH: https://www.zde.uzh.ch

Literary Methods for All: CLS INFRA

Sarah Hoover[1], Julie M. Birkholz[2,14], Ingo Börner[3], Joanna Byszuk[4], Sally Chambers[2,12,14], Vera Maria Charvat[5], Silvie Cinková[6], Anna Dijkstra[7], Julia Dudar[8], Matej Ďurčo[5,12], Maciej Eder[4], Jennifer Edmond[9], Evgeniia Fileva[8], Frank Fischer[10], Vicky Garnett[11,12], Françoise Gouzi[12], Serge Heiden[13], Michal Křen[6], Els Lefever[2], Michał Mrugalski[15], Ciara L. Murphy[16], Carolin Odebrecht[15], Eliza Papaki[12], Marco Raciti[12], Emily Ridge[17], Salvador Ros[18], Christof Schöch[8], Artjoms Šeļa[4], Justin Tonra[17], Erzsébet Tóth-Czifra[19], Peer Trilcke[3], Karina van Dalen-Oskam[7], Vera Yakupova[11]

[1]CLS INFRA, University of Galway; [2]Ghent University; [3]University of Potsdam; [4]Instytut Języka Polskiego Polskiej Akademii Nauk; [5]ACDH-CH; [6]Charles University; [7]Huygens Institute for History and Culture of the Netherlands; [8]University of Trier; [9]University of Dublin Trinity College; [10]Freie Universität Berlin; [11]Trinity College Dublin; [12]DARIAH-EU; [13]École normale supérieure de Lyon; [14]Royal Library of Belgium; [15]Humboldt-Universität zu Berlin; [16]Technological University Dublin; [17]University of Galway; [18]UNED; [19]Coalition for Advancing Research Assessment (CoARA); sarah.hoover@universityofgalway.ie

EU-Funded Computational Literary Studies Infrastructure (CLS INFRA) is helping to build the shared and sustainable infrastructure needed to reinvent approaches to Europe's multilingual literary heritage. Working within the FAIR and CARE principles (Wilkinson et al. 2016; 2019; Carroll et al. 2020) we are standardizing, aligning resources, and widening access for all. The aim is to build on recently-compiled high-quality literary corpora, such as DraCor and ELTeC (Fischer et al. 2019; Schöch et al. 2021; Odebrecht, Burnard, and Schöch 2020), and tools e.g. TXM, stylo, multilingual NLP pipelines (Heiden 2010; Eder, Rybicki, and Kestemont 2016), demonstrating how they can be used by anyone engaged with textual analysis. We will present four recent achievements of the CLS INFRA project, focusing on the tools and workflows that make CLS methods accessible to all.

The Survey of Methods, (https://zenodo.org/records/7892112) presented on an interactive grid (https://clsinfra.io/resources/d3-2-methods) provides an accessible introduction to practices, methods and issues that are prominent within the current landscape in CLS (Schöch, Dudar, Fileva, *eds.* 2023). The Survey Grid provides targeted information in a focused way that is suitable for a summary overview of CLS subfields like authorship attribution, genre analysis, literary history, gender analysis, and canonicity. Each of the areas contains smaller methodological steps: corpus building, pre-processing and annotation, data analysis, and evaluation. All surveys were based on a large collection of publications from the last ten years.

Non-academic Applications: By profiling potential users from diverse fields such as policy, consultancy, journalism, publishing, medicine/psychology (bibliotherapy), and artistic practice, the project envisions CLS as a versatile tool. User scenarios, drawn from interviews with non-literary research users, outline tasks in fields ranging from history research to narrative-driven futures institutes. The research identifies four models to help non-academics take advantage of the opportunities offered by CLS tools as a key driver for the sustainability and impact of the infrastructure.

Inventory of Existing Data: Here we compile a comprehensive overview of the landscape of literary corpora and sources currently available, describing our methodological approach and analysing the various challenges encountered in the effort to collect information about these resources and consolidate them into a structured form. Based on an initial inventory of 86 corpora or corpus sets, we exemplify their wide variety with respect to structure, context and purpose, and consequently the differing modes of provisioning. We also propose a technological path towards making this information searchable via a central discovery catalogue by discussing principal design decisions regarding the data model and the technology stack needed for such a task.

Data for the People: Reproducibility of research is at the core of open science. The relationship between measurable reproducibility and shared insights takes center stage in philosophy of science and methodology of literary studies. This report addresses the technical, institutional, legal/regulatory and ethical/social issues that help and hinder sharing data and tools. Based on case studies of researchers who overcame hindrances or took advantage of affordances, it provides template policy instruments as a set of recommendations for researchers and institutions, particularly libraries.

SIstory 5.0: The Flow of the Past, Upgraded at Last

https://doi.org/10.5281/zenodo.11617427

Katja Meden[1,3], Vid Klopčič[2], Ana Cvek[1], Mihael Ojsteršek[1], Matevž Pesek[1,2], Mojca Šorn[1], Andrej Pančur[1]

[1]Institute of Contemporary History, Slovenia; [2]Faculty of Computer and Information Science, University of Ljubljana, Slovenia; [3]Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia; katja.meden@inz.si, mihael.ojstersek@inz.si

The portal History of Slovenia – SIstory [1] is a platform for publications and sources on historical and cultural heritage, which has been in operation since 2008 and is an integral part of the Research Infrastructure for Slovenian History at the Institute of Contemporary History within the national research infrastructure DARIAH (DARIAH-SI). The portal is primarily aimed at

researchers in the field of history and other related disciplines (especially the humanities and social sciences) to support their research work. While the portal mainly refers to publications and textual sources, several extensive databases have also been compiled and published as part of the portal (e.g. the list of victims of World War I and II and censuses).

The portal has undergone several changes over the course of its existence, from minor updates (e.g. moving from unstructured metadata to a custom metadata application profile) to major updates of the entire portal. However, the most recent update (presented in this poster, the beta version is available at [2]) served as a complete redesign from the ground up, prompting us to re-evaluate the fundamentals of the system, taking into account the legacy issues and solutions from previous versions of the portal, in order to improve the functionality of the system on the one hand and provide the familiar user experience on the other. The development of the new version took place through a series of discussions and feedback from various stakeholders involved in the process of acquisition and classification of the materials, librarians, data owners and, most importantly, users of the portal. Based on their feedback, we have focused on the key features that will improve the usability and overall user experience of the portal.

This poster will therefore present the infrastructural view of the SIstory portal, focusing on the workflows that are an integral part of the portal and its development. It will highlight not only the content that the portal reflects but also the life cycle of the data within the portal (which includes several steps from the initial acquisition and valorization of the documents, digitization, (meta)data processing to the first publication of the documents on the SIstory portal).

In addition, we would like to emphasize the above-mentioned upgrade and the general design workflow related to the reconstruction of such a system, which at the time of writing contains more than 55.000 publications and other sources (excluding the additional databases), as well as smaller, more technical workflows within the upgrade (e.g. narrowing the metadata application profile or providing a metadata crosswalk to improve interoperability) to illustrate the importance of addressing the challenges encountered during the redesign process and emphasizing the need for a balance between improved system functionality and maintaining user familiarity. By providing insight into the intricacies of this major upgrade, we hope to provide a valuable perspective for future efforts to improve digital platforms for historical research and ultimately promote the accessibility and utility of historical data within the academic community and beyond.

Leiden University's Faculty of Humanities: A preliminary study on publishing reproducible workflows

https://doi.org/10.5281/zenodo.11656520

Femmy Admiraal, Myrte Vos, Alma Strakova
Leiden University, The Netherlands; f.admiraal@library.leidenuniv.nl

Over the past decade or so, we have seen a tremendous increase in attention for data management and data publishing at Leiden University, both from an institutional as well as from a researchers' perspective. In 2016, a first version of the Data Management Regulations was published[1], and in 2021 it was revised, providing the necessary policy framework.[2] In the course of 2022 and 2023, data stewards were appointed in each faculty to support the implementation of the data management policy, and moreover to support the researchers in making their data FAIR. The uptake by researchers is evident from the increase in published data sets per year, as shown in Figure 1.[3]

(see attached pdf)

Figure 1: Total number of datasets published by researchers affiliated to Leiden University between 1983 and mid-2022 (Raga Raga, 2022).

Leiden University has a longstanding track record in Humanities research, and the faculty covers a broad range of domains with its seven institutes. While some of these institutes remain faithful to traditional methodologies, such as the Institute for History, others have fully embraced digital methods, such as the Leiden University Centre for Digital Humanities. And yet others conceptualize research data and the corresponding workflows entirely different, such as the Centre for Arts in Society.

With this preliminary study, we aim to uncover the variation in the documentation of workflows of Leiden's Humanities scholar. We will take a closer look at the datasets published after the implementation of the first Data Management Regulations in 2016 until 2023. Within that subset, we will investigate which type of documentation is included describing the workflows, if any at all. Based on those findings, we will look for patterns across research disciplines, across institutes, as well as across repositories. The poster will present the results of the study, as well as some lessons learned for improving the support for publishing reproducible workflows, especially in those domains in which it is not so common yet to describe workflows in detail.

References

Raga Raga, Núria. 2022. *Dataset publication by Leiden University researchers: research on the impact of FAIR data*. Digital Scholarship @ Leiden: https://www.digitalscholarshipleiden.nl/articles/dataset-publication-by-leiden-university-researchers-research-on-the-impact-of-fair-data.

Schoots, Fieke and Femmy Admiraal. 2023. *Updated Data Management Regulations for Leiden University*. Digital Scholarship @ Leiden: https://www.digitalscholarshipleiden.nl/articles/updated-data-management-regulations-for-leiden-university.

[1] See Schoots & Admiraal (2023) for a reflection on the policy process throughout the years: https://www.digitalscholarshipleiden.nl/articles/updated-data-management-regulations-for-leiden-university.

[2] The current Leiden University Data Management Regulations are available at https://www.library.universiteitleiden.nl/binaries/content/assets/ul2ub/research--publish/research-data-management-regulations-leiden-university.pdf.

[3] Note that this study was based on datasets published until mid-2022, which accounts for the ostensible decrease in the last year, while in fact the data for 2022 are incomplete.

The Networking of Female Translators in Spain (1868-1936): A Study of Their Representation in the Mnemosyne Digital Library

https://doi.org/10.5281/zenodo.11221937
Emilio José Ocampos Palomar, Dolores Romero López
Universidad Complutense de Madrid, Spain; dromerol@ucm.es

From the research group La Otra Edad de Plata (LOEP) we present a workflow on one of the collections of the *Mnemosine* project. The *Mnemosine* project is a Digital Library of Rare and Forgotten Literary Texts (1868-1936) whose aim is to select, categorise and make visible in digital format literary texts belonging to a forgotten repertoire in order to allow the historical revision of the period; and the collection in question is entitled "Women Translators in Spain (1868-1936)", which aims to be a field of international experimentation for the creation of interoperable semantic networks through which a wide group of scholars can generate innovative research and theoretical reading models for literary texts.

Studies on Spanish literature in the late nineteenth century and the first third of the twentieth century are evolving from research on canonical writers towards the study of "rare and forgotten" authors, themes and genres during what is now called *The Other Silver Age*. And in this approach to the margins of the canon, women's writing and translation are fundamental: for long before the professionalisation of translation, some Spanish women saw in the task of translating a work that provided them with a livelihood and some public recognition for their collaboration with literature, bequeathing us magnificent translations of classics and moderns. The collection presented here rescues their names and in some cases their vital data; it also provides access to some of the resources located, giving an account of the works produced by women translators of *The Silver Age*: of them and of the authors who served as their source, of the originals and their translations, of the connections of these women in the world of culture.

This collection also fills a gap in critical studies on gender and translation. Thus, by locating and systematising women translators and their works, it is possible, on the one hand, to show how the translations of these women gave rise to valuable advances towards the awakening of a different female identity committed to her social and cultural environment, and, on the other hand, to reflect on how, throughout *The Silver Age*, women dignified their work as translators by developing professionally.

In short, if we make intelligent use of the potential of new technologies (macroanalysis, stylometry, topic modelling, etc.) for philological purposes, and, at the same time, we start from a specific data model such as *Mnemosine*, which provides the relational search of linked data for a more selective and better targeted retrieval of information, it is possible to demonstrate the richness of the translations made by women in Spain and their commitment to modern women; it is possible to reconstruct, digitally through the *Mnemosine Digital Library*, a forgotten literary history: *The Other Silver Age*.

Development of an educational program and its role in communicating individual DH research workflows at different universities and institutes

Masao OI
National Institutes for the Humanities, Japan; oi-masao519@g.ecc.u-tokyo.ac.jp

Digital Humanities (DH) is attracting increasing attention in Japan. In recent years, new courses on DH have been offered at universities around Japan, and related positions have been created. The government has also allocated a new budget for the construction of a digital research infrastructure for the arts and humanities. On the other hand, a shortage of human resources for DH has become an issue. Education is important for human resource development. In Japan, however, institutions that can promote DH education on an organizational basis are rare. As a result, it is difficult to conduct DH education in a systematic framework because there is no accumulated curriculum as an organization.

Therefore, the National Institutes for the Humanities (NIHU) has developed the "NIHU DH Lecture Program," that maintains and communicates the workflow of DH human resource development among different organizations and can be referenced and utilized at any university or research institute. The basis of this project is the production and release of video content, which serves to inspire and motivate students and young researchers who feel difficulties and mental barriers in DH, as well as senior researchers willing to newly start DH research, to become interested in and want to practice DH research in an enjoyable way. It also aims to encourage the sublimation and sharing of knowledge that is reproducible and versatile from practices that have been tacitly practiced by individual researchers until now.

To this end, the "NIHU DH Lecture Program" has two layers:

The first layer is a series of video dialogues that acts as an entry function. Researchers of various generations, from veterans to young researchers, are invited to discuss the appeal and excitement of DH research, sharing their own research topics and passions. The program is unique in that the announcer also participates in the discussion, eliciting basic questions from beginning researchers and the general public, and attempting to "translate" the content into easy-to-understand terms. By using professional listeners as intermediaries to elicit narratives, the series has succeeded in lowering the hurdles in DH research and education. In addition, Dr. Toma Tasovac, director of DARIAH, was a special guest in this series, and important discussions connecting Japan and DARIAH began to flow.

The second layer is a practice on technologies and methods to develop individual specific research series. These can be used as on-demand learning materials for faculty members who have little experience in organizing educational programs, for example, when they are newly in charge of DH research.

In addition, "NIHU DH Lecture Program" is also working on storytelling using digital earth, podcast contents that link researchers' narratives with resources, and the production of Linked Open Data of these contents. By translating these contents and providing them back to DARIAH, it aims to promote DH research and DH education efforts in Japan and Asia on an international scale, and to contribute to the development of new global co-creative research.

Optimizing Data Collection Methodology and Workflows in Research on Open Access

Gabriela Ewa Manista, Maciej Maryl, Magdalena Wnuk
IBL PAN, Poland; gabriela.manista@ibl.waw.pl

PALOMERA (Policy Alignment of Open Access Monographs in the European Research Area) project, supported by funding from Horizon Europe, aims at exploring Open Access (OA) policies pertaining to academic books in European Research Area and to facilitate access to this documentation through a tailored Knowledge Base designed in DSpace. The initial phase of the project focused on gathering various forms of data, encompassing both qualitative and quantitative.

During only one year of the project, the research team collected a diverse array of materials on policies, including legal documents, grey literature, research articles, reports, statistics, and outputs from other projects, along with transcripts from interviews conducted over the course of the project. The methodology of data collection was drawn from the identification of the resources from 39 countries in 29 languages (later machine translated to English and corrected by researchers), through collection and pre-processing in the Zotero Library, to pre-coding using the MaxQDA program. At the same time, the PALOMERA team set the survey on the LimeSurvey, and conducted 42 interviews – individual and group. Ultimately, around 900 documents with 1500 excerpts were collected, 454 surveys were completed, ~4 million items concerning bibliometric data aggregated, and more than 40 in-depth interviews transcribed. Throughout the process, data were collected so that later the whole set was ready for swift analysis. This comprehensive gathering process was facilitated by the application of various methods and digital research tools, which effectively structured activities, directed them towards specific objectives, and enabled the tracking of progress throughout the project, **making the whole workflow clear to the researchers involved in the operations even when tasks overlapped** (see: Fig.1). Those research and data collection practices were later translated into a presentation in the online Knowledge Base.

Fig.1. PALOMERA Data Collection Workflow

The poster aims to showcase the efficient data collection process, spanning from the development of the methodology to the preparation of data for analysis within an international research setting and publishing the outcomes in an online repository. It underscores the significance of employing established project management methodologies that are instrumental in research endeavours. This methodology holds potential for replication in similar (policy) landscape analyses and future collaborations within the European scholarly community.

Mapping Thonet

https://doi.org/10.5281/zenodo.11397095
Sarah Kindermann
Donau University Krems, Germany; fraukindermann@gmail.com

The project "Mapping Thonet" is the visual processing of the image research for my master's thesis "The Public Image of Thonet, Photographs 1850-1900". Mapping Thonet is an open research project on photographs of Thonet furniture from the period 1850-1900. For the research, national and regional archives as well as museum collections were searched worldwide and selected photographic image contents of the research were territorially located with geodatas. The thematic background of the work lies in the context of Christian expansion and the associated colonialism of the 19th century and the introduction of a commodity as a representative of a religious affiliation and its visual implementation in the collective memory. The image research is divided into 2 main categories, which concern on the one hand representations of people with a piece of Thonet furniture and on the other hand interior and exterior spaces on which a piece of Thonet furniture is located.

For the visualization and processing of the data, the collected results of the research were processed with the data visualisation programme Github Collection Builder[1], with the aim of recording image authors, persons depicted (if not anonymous), place and time (if available) as well as the data providers (image source owners). The preliminary results are available at the following link: https://fraukindermann.github.io/Mapping-Thonet/

The demo metadata that can be downloaded as csv are available down under the following link: https://github.com/Fraukindermann/Mapping-Thonet/blob/main/_data/demo-metadata.csv?plain=1

[1] *You can find a tutorial under the website of Programming Historian. Programming historian is a open source training website, that give historians and related scientists the possibility to learn the basics of programming and shows up tools to work with: https://programminghistorian.org/es/lecciones/exhibicion-con-collection-builder*

Corpus annotation and dictionary linking using Wikibase

https://doi.org/10.5281/zenodo.12078616
David Lindemann
UPV/EHU University of the Basque Country, Spain; david.lindemann@ehu.eus

This poster presents a data model and two first use cases for the representation of contents of text corpus data on Wikibase instances,[1] including morphosyntactic, semantic and philological annotations as well as links to dictionary entries. Wikibase[2] (cf. Diefenbach et al. 2021), an extension of MediaWiki, is the software that underlies Wikidata (Vrandečić & Krötzsch 2014).

The use case for which the model has been proposed is documents that belong to the Basque Historical Corpus. That corpus today exists in several versions stored in separated and incompatible data siloes (based on relational databases) and made available through different online user front ends. A second use case is an experiment for linking a Serbian literature corpus in NIF format to a Serbian dictionary in Ontolex-Lemon.[8]

Heavily inspired by the latest trends in the field of Linguistic Linked Open Data,[9] we model a corpus token as node in a knowledge graph, and link it (1) to the respective paragraph (Basque) or token (Serbian) in the source document ; (2) to a lexeme node, which is annotated with the standard lemma; (3) to a lexical form associated to that lexeme, which is annotated with the

grammatical features describing the form; (4) to a lexical sense associated to that lexeme, which is annotated with a sense gloss; (5) to an ontology concept representing the word sense; and (6), to a text chain containing philological annotations. Furthermore, we represent token spans as separate nodes; these are linked to the contained tokens, and to annotations that apply to the whole span. We implement and populate the model on our own Wikibase instance hosted on Wikibase Cloud. Core classes and properties used on a Wikibase by default for describing lexemes deploy Ontolex-Lemon (McCrae et al. 2017), the W3C-recommended model for lexical data, so that the created datasets are compatible with the Linguistic Linked Open Data Cloud. We define properties that describe corpus tokens as equivalent to NIF, a standard for corpus annotation (Hellmann et al. 2013).

We are currently populating the proposed model with tokens from a 1737 Basque manuscript, the transcription of which has been carried out on Wikisource,[10] and inserting annotations of the above described types including philological annotations by Lakarra (1985), as well as direct links to the corresponding paragraph in the manuscript transcription on the Wikisource platform.

The poster will contain the data model as visual graph, textual summaries, and sample SPARQL results showing token annotations, and linked dictionary entries and referenced Wikidata entities.

1 Accessible at https://monumenta.wikibase.cloud (Basque data), and https://serbian.wikibase.cloud (Serbian data).

2 See https://wikiba.se, and https://en.wikipedia.org/wiki/Wikibase.

3 See https://www.wikidata.org (Vrandečić & Krötzsch 2014).

4 Accessible at https://klasikoak.armiarma.eus/.

5 Accessible at https://www.ehu.eus/etc/ch/.

6 Accessible at http://bim.ixa.eus/ (Estarrona et al. 2022).

7 For example, at https://www.ehu.eus/monumenta/lazarraga/ (Bilbao eta al. 2011).

8 The source datasets are described in Stanković et al. (2023).

9 See literature discussed in Stanković et al. (2023), and Pedonese et al. (2023).

10 See Lindemann & Alonso (2021), accessible at https://eu.wikisource.org/wiki/Azkoitiko_Sermoia.

Wielding the Magic of Words: Unleashing Text-to-SPARQL Spells for Seamless Semantic Querying

ADRIAN Ghajari, SALVADOR ROS, ALVARO PEREZ
Spanish National Distance Education University,UNED, Spain; sros@scc.uned.es

In the realm of the digital humanities research process, one critical workflow is related to obtaining data from different sources. Scholars often grapple with the challenge of navigating databases or knowledge graphs to extract valuable insights for their research. This process necessitates a fundamental understanding of query languages, as SQL the most popular language for relational databases **(Popescu, Etzioni, and Kautz 2003**) or, SPARQL, the standard language for querying knowledge graphs based on ontologies such as DBpedia, Wikidata or Wikipedia, (**Pourreza and Rafiei 2023**).

However, the complexity of creating suitable queries poses a significant obstacle for these workflows. The situation is exacerbated when SPARQL is used.

To overcome this situation the use of Generative AI can play an essential role since these models can be used as interfaces which can interpret queries expressed in natural language and respond with the information required, thereby empowering scholars with a more intuitive and user-friendly means of accessing and analyzing data within digital humanities triple stores.

The methods for achieving this goal are based on prompt engineering **(Liu et al., 2023)**, and there is no formal way to build them. It is a heuristic process that remembers ancient spells.

For this purpose, we have built a set of modern spells called prompts. We have selected eight prompt combinations based on few-shots, chain of thoughts and Retrieval Augmented Generation, RAG techniques. These spells have been launched over three different Large Language Models, LLMS, and their performance has been evaluated on a dataset of 1,000 SparQL natural language query-consult pairs extracted from the LC-QUAD test Split dataset, which is based on the DBpedia knowledge graph. Notably, the most proficient model, GPT-3.5 Turbo LM from OpenAI, achieved an accuracy of 48% with the few-shot prompt. These preliminary findings indicate that achieving parity with human-level performance remains a formidable challenge. However, these findings yield valuable insights into the development of effective prompts. Therefore, this work not only underscores the complexities inherent in developing prompts that facilitate nuanced interactions with LLMs but also sets a foundation for future explorations aimed at bridging the gap towards human-equivalent performance in natural language query processing.

Acknowledgements

This research has been carried out in the Grant CLSINFRA reference 101004984 framework funded by H2020I NFRAIA-2020-1.

References

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55 (9): 195:1-195:35. https://doi.org/10.1145/3560815.

Popescu, Ana-Maria, Oren Etzioni, and Henry Kautz. 2003. "Towards a Theory of Natural Language Interfaces to Databases." In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 149–57. IUI '03. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/604045.604070.

Pourreza, Mohammadreza, and Davood Rafiei. 2023. "DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction." https://doi.org/10.48550/arXiv.2304.11015.

PyMotifs: a workflow designed to detect significant lexico-grammatical patterns for automatic stylistic analysis.

Antoine Silvestre de Sacy[1], Bruno Spilak[2]
[1]Université Paris III - Sorbonne-Nouvelle, France; [2]NodenAI; antoinedesacy@gmail.com

PyMotifs is a workflow situated at the frontier of stylometric and stylistic questions and is intended to be used by both disciplines: as a complement to stylometry, whose orientations are essentially quantitative, seeking to answer questions such as: what are the subjects addressed by a given text? By whom was it written? When was it written? But also as a complement to stylistics, whose orientations are essentially qualitative, focusing on particular and specific facts of language: is this form significant for this author or this text? How can we interpret this form as a style of this author, text, genre or corpus?

Contrary to many approaches in digital humanities and in big data which stick to distant reading, but also in contrast to stylistics which sticks to singular facts of language, PyMotifs proposes a mixed method, relying on statistical methods but nevertheless enabling a systematic return to the texts to interpret the results. PyMotifs thus allows for both inductive approaches (data-driven and bottom-up approaches) and deductive approaches (top-down, qualitative stylistics). Even when introducing features of the Scikit-Learn machine learning framework, PyMotifs provides a return to text that puts the interpretability of results at the heart of its system, thanks in particular to algorithms using rolling stylometry.

Preserving Humanities Research Data: Data Depositing in the TextGrid Repository

https://doi.org/10.5281/zenodo.11279675

Stefan Buddenbohm[1], José Calvo Tello[1], George Dogaru[2], Ralf Klammer[3], Alex Steckel[1], Lukas Weimer[1], Stefan E. Funk[1]
[1]Göttingen State and University Library, Germany; [2]Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen; [3]Center for Information Services and High Performance Computing TU Dresden; buddenbohm@sub.uni-goettingen.de, steckel@sub.uni-goettingen.de

If preservation is use, what are the implications for a humanities research data infrastructure?

Preserving research data long-term, making it accessible and reusable for the scientific community is a fundamental concern for research infrastructures. This boils down to an alignment of expectations between data depositor and data recipient. For example, the research data infrastructure has requirements relating to format, metadata, responsibilities and licenses.

When research data has been successfully deposited/published in a repository, the next pitfall looms: the often unappealing presentation or constrained findability of data. Applying the famous words of John Cotton Dana: If "Preservation is use", the research data infrastructure has to emphasize potential and re-usability of data.

The paper introduces the data depositing workflow of the TextGrid Repository (TGRep).

The TextGrid Repository

TGRep is a pioneer of the Digital Humanities in the German-speaking area. Today, TGRep is part of the Text+ portfolio, the NFDI consortium for language and text-based research data in Germany. Each Text+ data center offers a workflow for incorporating research data that complies to its scope, making it available for reuse.

ELTeC

ELTeC in TGRep, the European Literary Text Collection, serves as an example for the identification, consultation, ingest, transformation, enrichment, publication and integration in the portfolio of Text+, spelling re-usability and interoperability. ELTeC is a state-of-the-art, open access multilingual collection of corpora containing novels from several European traditions developed for several reasons, among them the development of tools and methods in Computational Literary Studies. Currently, ELTeC contains more than 2000 full-text novels in XML-TEI in 21 languages. They are distributed via multiple platforms (such as GitHub and Zenodo). 1365 full-texts in 15 languages are also published in the TGRep.

The Data Depositing Workflow in TGRep

As TGRep is of great relevance for Text+ and its community, so is the task of minimizing the effort spent when publishing data there. The solution implemented in Text+ consists of a workflow that automates the creation of the technical files required when importing into the repository, while allowing for as much manual intervention as needed.

To use the system, the user interacts with a web-based user interface running inside a Jupyter notebook. After specifying the location of the TEI files to be imported, the data is analyzed in an automated step, which finds and extracts metadata common to all files and makes this available for verification and manual improvement, if necessary. In a subsequent manual step, the user can check and edit the extracted metadata, but also change how the metadata is identified (in which case the previous step can be executed again). In the last step, the technical TGRep metadata files are generated.

The new workflow not only improves the data import process, but also serves as a blueprint for further easy-to-build applications that combine libraries and notebooks and rely on the versatile Jupyterlab environment, which can be deployed both locally and in the cloud.

An exploration of an evaluation framework for digital storytelling outcomes in the AI age

Yaming Fu[1], Simon Mahony[2]
[1]School of Cultural Heritage and Information Management, Shanghai University; [2]Department of Information Studies, University College London; s.mahony@ucl.ac.uk

Digital and Semantic Web related technologies facilitate the possibility of automatically generated digital storytelling products. This is closely related to the construction of data and technical infrastructure with a human-machine collaborative digital

storytelling workflow. This presentation introduces our design for a workflow and evaluation framework (Figure 1) based on our digital storytelling practice since 2021.

The knowledge databases which record narrative elements, described in unified resource description frameworks, provide a crucial basis for creating human-machine collaborative digital stories by automatically identifying links and storylines. They also make it possible to create story-themes or plots that are machine-identifiable or -editable. A crucial factor in the workflow is the design of algorithms and logic rules, drawing on traditional narrative grammar, story organization, and narrative structure. Digital storytelling outcomes need to be accessed and experienced by people, so the aim is to present them by adding interaction methods (touch, voice, visual, augmented reality) to construct scenes.

To be a robust scholarly method for research, the reliability, validity, and evaluation of digital storytelling methodology and its outcomes need to be transparent, standardized, and reproducible to ensure that they meet academic rigour and scientific method (Figure 2). This is essential for effectively evaluating the credibility, advantages, and effectiveness of digital storytelling outcomes.

Practitioners and researchers need to be able to evaluate whether the technology used in generating digital storytelling outcomes is standardized, unified, and reusable, and whether the data is managed according to FAIR principles. They need to be able to verify that the ethical evaluation of the technology is appropriate so that the results can be trusted. It is also necessary to evaluate coverage, reliability, verification, and the credibility of the materials themselves. Trust is a significant issue, particularly if the providers do not make it clear that these digital stories are produced algorithmically. Trust is highlighted in the European Commission's White Paper on AI. A lack of trust 'is a main factor holding back a broader uptake of AI' and prompts developing a clear regulatory framework (European Commission, 2020: 9). Trust is often reduced regarding confidence in AI generated content because of a lack of verifiability (Samek and Müller, 2019). Hence the need for what has become known as *explainable AI*; 'in part motivated by the need to maintain trust between the human user and AI' (Jacovi et al, 2021: 624). Whether this trust can be achieved seems uncertain, particularly as the risks of the use of AI in different scenarios are as yet unclear and not fully understood. When a human-machine collaborative digital storytelling product is presented or put into use, it is important to evaluate user experience; how the audience feels about the outcome, whether they are satisfied with it, as well as the accessibility, and ethical evaluation of the interactive effect.

A framework such as we suggest here will go some way towards building the necessary trust for digital storytelling outcomes using AI where human and machine gradually build a collaborative workflow.

IT-inclusive workflow in data steward education

Bence Vida, Eszter Szlávich, Zsuzsa Sidó
Eötvös Loránd University ; Faculty of Humanities, Hungary; vida.bence@btk.elte.hu, szlavich.eszter@btk.elte.hu, sido.zsuzsa@btk.elte.hu

One of today's most pressing issues in the organization of science is the decapitalization of scientific data, ensuring the transparent and free flow of scientific information shared within and beyond academia. Guiding principles for this are the principles of Open Science, FAIR use and Open Access that offer accessible and usable solutions for the scientific community.

Data Stewards are professionals putting these principles into practice, managing large volumes of data (digitalized and born-digital) and keeping track of technological innovations and legislative frameworks.

In autumn 2023, the Department of Digital Humanities at Eötvös Loránd University launched a Data Stewardship Postgraduate Specialist Training Course. The first cohort includes primarily researchers, research infrastructure managers, university and research librarians. Among the participants of the training are also staff members of public institutions or companies in charge of data management and data security, staff members of cultural institutions, public and private collections (GLAM and memory institutions included), and collection managers.

In our paper we would like to present the workflow developed in designing the training course in relation to Open Science workflows. The Train the Trainer Workshop organized by OpenAIRE in 2023 supported us in developing this workflow, and we were able to use the methodology learnt there. The uniqueness of our training lies in extending the traditional understanding of the data stewardship field that we were introduced to at the OpenAire workshop.

The basic workflow includes the development of e-learning materials and project assignments, the recruitment of an interdisciplinary team of teachers and the creation of an inclusive and collaborative environment for the participants. The Data Stewardship Training Workflow emphasizes ensuring the security and authenticity of data, creating the conditions for FAIR research data management, the designation and use of processed data for secondary uses, and the operation of appropriate data management systems and repositories.

Our unique addition is that we consider IT knowledge and skills as important as knowledge about research data management and data governance in general. The design places emphasis on incorporating basic and intermediate level of IT skills into our training. Therefore, participants will gain systematic knowledge of data management models and workflows, as well as data management technologies. The courses have also been designed to consider the fact that the storage, movement, planning and enforcement of research data according to appropriate standards are only one part of the data management workflow. Basic knowledge of the conditions under which research data is generated and stored, automated management of data sets, the structure and operation of repositories, technical knowledge of the digital data processing workflow and the query languages, and programming basics required for database management are essential.

Finally, we would like to emphasize the need to involve IT professionals in the research data cycle, and for our data stewards to be able to involve them in project-based teamwork. This collaborative approach can be conveyed in an environment which promotes interdisciplinarity and teamwork. Our hope is that the workflow we have described/designed will be able to comply to this.

Training design as a workflow: producing adaptable and reusable learning pathways for Arts and Humanities

https://doi.org/10.5281/zenodo.12078139
Fotis Mystakopoulos, Karla Avanço
OPERAS; fotis.mystakopoulos@operas-eu.org, karla.avanco@openedition.org

This paper explores the application of the workflow concept to training design, which translates into treating the training preparation as a workflow as well as organising learning modules based on the research lifecycle. The paper also shows how to tailor the training program to the specific needs of Arts and Humanities.

OPERAS is a European Research Infrastructure dedicated to open scholarly communication in the Social Sciences and Humanities (SSH). The strategic objectives of the infrastructure include having a knowledgeable and empowered competence-based community. Training plays an increasingly significant role in building such a community, highly qualified to act and lead in open science ecosystems as it promotes a continuous learning culture, fosters collaboration, and builds collective expertise in a dynamic environment.

OPERAS has been carrying out different actions in line with its objectives, among which the contribution to the Horizon Europe project Skills4EOSC by leading a Work Package focused on thematic training. In particular, OPERAS is developing and implementing a pilot program dedicated to instructing a cohort of trainers in Open Science and Research Data Management (RDM) for the Social Sciences and Humanities (SSH).

The training design followed a workflow that started with the definition of competencies and skills for researchers (Catalogue of Skills), followed by the instructional design phase anchored on FAIR principles to make the whole training not only findable and accessible but also interoperable and reusable (FAIR-By-Desgn Methodology). In this framework, each step of the design undergoes FAIR compliance checking. As co-design was also part of the process, other SSH research infrastructures (DARIAH, CESSDA, and CLARIN) reviewed the final learning pathway.

Research and data lifecycle various phases gave the structure for the training program itself. Emphasis is placed on exploring the workflows involved in managing research data and establishing connections between the RDM process and a broader publication strategy that aligns with Open Science practices. Particularly, in managing research data during the different phases, the authors attempt to understand and collate information on the level of granularity required to properly document research materials according to FAIR principles, always having reproducibility in mind.

This paper addresses several key points, including the significance of developing training materials that adhere to the FAIR principles from conception to materialisation. Additionally, it discusses the creation of a tailored learning pathway designed specifically for delivering training sessions to SSH researchers and adaptable to the different disciplines. Crucially, the authors also discuss the feedback from pilot participants on how the material could be improved, enhancing adaptability and reproducibility. Whereas the SSH are often categorised together in broader disciplinary taxonomies, this paper presents the nuances and peculiarities that emerged from the pilot program, concentrating primarily on the humanities aspect.

By extending the FAIR principles beyond research data to encompass the whole workflow of training design and locating the training within the research and data lifecycles, we aim to support practices in training development across different disciplinary communities, improving their quality and fostering reusability.

Translation Loops: Shaping and Reshaping TaDiRAH

https://doi.org/10.5281/zenodo.12161125
Luise Borek[1], Canan Hastik[2]
[1]Technical University of Darmstadt, Germany; [2]Interessengemeinschaft für Semantische Datenverarbeitung e.V.;
luise.borek@tu-darmstadt.de

In this contribution we would like to propose a workflow that offers low-threshold access for the multilingual DH community to jointly maintain and further develop the taxonomy of digital research activities in the Humanities (TaDiRAH). This builds on outcomes of the workshop "Multilingual taxonomy initiative – TaDiRAH as community of practice" at DH2023 conference organised by active members and users presenting the current status of the taxonomy and its translations to collect requirements for the development of a workflow.

The aim is to design a workflow for the collaborative further development of the taxonomy that brings all language versions up to a standardised level. This workflow leaves room for, and inevitably requires, the existing terms and scope notes to be discussed and renegotiated between the cross-language and cross-disciplinary user communities. Interesting aspects of digital research activities will also be highlighted and the question of whether there will also be continuous structural and/or content-related adjustments.

The extensive revision in the course of the SKOSification of the taxonomy has led to varying states of translations which are now no longer congruent with the revised English version (version 2.0 issued 2020.09.29). However, one of the strengths of the taxonomy is its multilingualism as it makes research activities that do not operate in English more visible. However, it also shows that there are language areas in which terms relating to digital research activities do not yet exist and will now be introduced with TaDiRAH. The existence of an internationally harmonised vocabulary supports disciplinary activities in the research landscape. In order to achieve this, the active participation of researchers from the respective specialist communities is necessary.

The Vocabs service as the primary home (https://vocabs.dariah.eu/tadirah/en/) provides the latest version of TaDiRAH. For further developments, templates for translation are available in a platform-independent exchange format. The processing is organised individually by the respective working groups according to their own needs and preferences. In order to merge the different language versions into the Simple Knowledge Organization System (SKOS) in one way, we are initiating collaboration with CESSDA Vocabulary Service (https://vocabularies.cessda.eu/) where authorised users can create, manage and translate vocabularies. Once a harmonised version is achieved with an export to SKOS a new version in Vocabs is published by national moderators and supervised by the TaDiRAH board.

The need for workflows for collaborative and cross-disciplinary development, publication as linked open data, maintenance of taxonomies, and collated vocabularies is great and has not yet been satisfactorily resolved. As a rule, Git-based workflows have become established in projects. However, in less tech-savvy communities, this usually leads to an exclusion of motivated

contributors. And, the work around TaDiRAH requires more administrative and coordination effort for harmonisation and standardisation. A browser-based tool is essential for working with terms, descriptions, and their translations. CESSDA fulfills all the requirements for the TaDiRAH user community leaving the language communities room to organise themselves according to their own needs and preferences.

From Manuscript to Metadata: Advancing Digitizing Workflows in Serbian Literary Research

https://doi.org/10.5281/zenodo.11390926
Larisa Kostić, Jelena Lalatović
Institute for Literature and Art, Serbia; zlatnalala@gmail.com

The Institute for Literature and Art has undergone a pivotal transformation in digitizing Serbian literary research workflows, transitioning from manuscript-centric methodologies to metadata-driven paradigms. Digitization brought about twofold discussion: the one primarily concerned with technological issues and the one dealing with thematic challenges that digital visibility enforces (Bode, 2018). Since the 1970s, the Institute for Literature and Art researchers have been nurturing approaches that nowadays can be labeled as avant-garde and non-canonical. Research practice says that exploring minority cultures and women's contributions to the cultural capital as literary translators, editors, and authors constitutes a tradition of significant longevity. Digital repositories, which could be centered around this criterion, enable and facilitate a comprehensive integration of previous and emerging trends and cultural and literary studies shifts (Longley & Bode, 2014).

A comparative examination of two so-far repositories within the Institute's purview unveils distinct efficacy differentials and their consequential impact on scholarly visibility and citation metrics.

The initial repository, founded on the Islandora platform in 2019, grappled with inherent deficiencies detrimental to scientific documentation depositing. At the time, the absence of experienced practice in adhering to metadata standards and a permissive approach to non-Latin scriptural entries within the Islandora platform compromised the repository's discoverability, particularly on scholarly indexing behemoths like Google Scholar.

The transition to the EPrints platform in 2021 heralded a seismic advancement in the Institute's scholarly landscape, unequivocally corroborating the necessity of tailored platforms for managing scientific documentation. Replete with meticulously crafted features, including citation tracking, metadata standardization, peer review workflows, and download statistics, EPrints emerged as the quintessential repository solution. The salubrious effects of this transition are manifest in the palpable amplification of scholarly visibility and citation metrics (Simpson & Hey, 2005). Works hosted on EPrints, adorned with comprehensible metadata descriptors, manifestly accentuate their presence on Google Scholar, EBSCO, and Scopus, warranting prominent indexing and augmented citation accrual. This workflow transition heightened visibility confers manifold advantages upon authors, increasing their scholarly influence while elevating the institution's academic prestige within the scholarly community.

These developments within the Institute for Literature and Art portend a seminal evolution in scholarly workflows within the Serbian literary domain. Could we ponder whether adopting EPrints as the repository of choice highlights the transformative power of digital archives platforms in enhancing scholarly visibility and dissemination effectiveness? Moreover, could this catalyze an epistemic renewal within Serbian literary research?

Digitizing seminal monographs, articles, and collections of papers that belong to the corpus of minority voices in the cultural capital plays a vital role in epistemic revitalization. This vision and a project proposal consist of seven platforms, each belonging to specific departments (as research units) at the Institute for Literature and Art. Digitized texts should be followed by annotations that govern and accompany advanced searches with limited keywords. We want to explore how searches based upon narrow keywords may contribute to further identification and expansion of the target groups beyond academia (curation specialists, human rights activists, feminist groups), thus contributing to the popularization of research results.

The numismatics digital workflow at archaeological excavations

Wojciech Ostrowski[1,3], Barbara Zając[2], Łukasz Wilk[3], Grzegorz Sochacki[1], Paulina Zachar[3], Jakub Modrzewski[3], Jarosław Bodzek[1]
[1]Jagiellonian University, Poland; [2]National Museum in Krakow, Poland; [3]Warsaw University of Technology, Poland;
w.d.ostrowski@gmail.com

Digital numismatics is a well-established ecosystem full of standards and good practices, from high-quality digitalization procedures meant to be used for coins through data gathering and sharing using Linked Open Data (LOD) and the Semantic Web approaches up to methods and technologies used for dissemination of graphical data - like the International Image Interoperability Framework (IIIF) combined with image web annotation tools.

Recently, at the Archaeological Institute at Jagiellonian University (Cracow, Poland), we tried to develop our internal standard for numismatic works during archaeological excavation led by our field teams. Together with the Warsaw University of Technology and the National Museum in Krakow (The Numismatic Room), we started with laboratory experiments on efficient digitalization on a large scale using the RTI (Reflectance Transformation Imaging) technique. The object of our study was the hoard from Nietulisko Małe, Ostrowiec County (Poland), containing approximately 3,000 ancient Roman coins.

We came up with an RTI dome combined with a cloud processing service and automatic publication of the final results in a web-based RTI viewer connected to a database built on international standards. During our work, we also tried to facilitate semantic annotation of a coin using the Segment Anything Model - a general approach image segmentation model developed by Meta- and make our (new and old) data compatible with the international LOD standard of nomisma.org.

Next, we tried to confront our experiences with the workflow used by numismatists in the field and found hardware and software solutions for issues with our "laboratory" approach. We found that applying RTI as a documentation method may solve problems with image quality and provide reproducibility of the proposed workflow. On the other hand, one of the most significant issues

which we needed to address during this process was how to establish such a workflow (including hardware) that was easy to use and robust enough to be reusable by many different specialists in different field conditions.

Finally, we had to find how to provide easy-to-use access to the proposed workflow results (with Web-based RTI viewers) and manage and share these data according to FAIR principles.

Literature review matrix as a tool for collaboration and reproducible research synthesis

https://doi.org/10.5281/zenodo.12518783

Anna Matysek, **Jacek Tomaszczyk**

University of Silesia in Katowice, Poland; anna.matysek@us.edu.pl, jacek.tomaszczyk@us.edu.pl

The application of literature review matrices as a collaborative and reproducible research synthesis tool holds significant promise across various disciplines. These matrices play a crucial role in the organization and synthesis of information from diverse sources, allowing researchers to discern patterns, identify gaps, and observe trends in existing literature (Kemmerer et al., 2021). In addition to enhancing the systematic approach to evidence synthesis, literature review matrices contribute to the transparency and reproducibility of research endeavors (Ghezzi-Kopel et al., 2021; Spillias et al., 2023).

Moreover, these matrices foster collaboration among researchers by facilitating joint analysis and interpretation of data, resulting in more comprehensive and reliable research outcomes (Murniarti et al., 2018). They prove instrumental in defining and summarizing effective confounding control strategies in systematic reviews of observational studies, thereby elevating the validity of the drawn conclusions. Furthermore, the incorporation of literature review matrices into the systematic review process aids in planning, identifying, evaluating, and synthesizing evidence, thereby bolstering the rigor, transparency, and replicability of the overall review process (Murniarti et al., 2018; Sjöö & Hellström, 2019).

In our paper, we propose a comprehensive workflow for creating literature matrices, utilizing a combination of conventional and artificial intelligence (AI) tools. This workflow encompasses various search services, a reference manager, a word processor, and illustrative AI tools to streamline the process of extracting and comparing information from research papers. By providing a structured framework for searching, organizing and analyzing the literature, a literature review matrix can facilitate the efficient screening and selection of relevant papers, as well as the extraction and synthesis of data.

References:

Ghezzi-Kopel, K., Ault, J., Chimwaza, G., Diekmann, F., Eldermire, E., Gathoni, N., … & Porciello, J. (2021). Making the case for librarian expertise to support evidence synthesis for the sustainable development goals. Research Synthesis Methods, 13(1), 77-87. https://doi.org/10.1002/jrsm.1528

Kemmerer, A., Vladescu, J., Carrow, J., Sidener, T., & Deshais, M. (2021). A systematic review of the matrix training literature. Behavioral Interventions, 36(2), 473-495. https://doi.org/10.1002/bin.1780

Murniarti, E., Nainggolan, B., Panjaitan, H., Pandiangan, L. E. A., Widyani, I. D. A., & Dakhi, S. (2018). Writing Matrix and Assessing Literature Review: A Methodological Element of a Scientific Project. *Journal of Asian Development*, *4*(2), 133. https://doi.org/10.5296/jad.v4i2.13895

Sjöö, K., & Hellström, T. (2019). University–industry collaboration: A literature review and synthesis. *Industry and Higher Education*, *33*(4), 275–285. https://doi.org/10.1177/0950422219829697

Spillias, S., Tuohy, P., Andreotta, M., Annand-Jones, R., Boschetti, F., Cvitanovic, C., Duggan, J., Fulton, E., Karcher, D., Paris, C., Shellock, R., & Trebilco, R. (2023). *Human-AI Collaboration to Identify Literature for Evidence Synthesis* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-3099291/v1

The CLaDA-BG-Research System: Data Management, Knowledge Organization and Services

https://doi.org/10.5281/zenodo.11057595

Kiril Simov, **Petya Osenova**

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria; kivs@bultreebank.org, petya@bultreebank.org

Our main goal is to construct an ecosystem for supporting research within social sciences and humanities (SSH). Thus, CLaDA-BG[1] needs to provide information management of a huge variety of research objects including different kinds of texts.

The top unification of these data is the metadata, but only scarce common information can be represented in this way. On the other hand, the tools for creation (digitization), representation, generalisation, search, etc. are specific and thus have to be customised accordingly or developed specially for these tasks. As one of the steps we consider the identification of information of interest and its simultaneous observation within context.

We follow the statement „Language Technology is a core Big Data technology. Growth in the volume and variety of data is mostly due to the accumulation of unstructured text data; in fact, up to 80% of all data is unstructured text data." (Strategy report on research infrastructures. ESFRI report, p. 108)[2]. Thus, we focus mainly on textual data.

Our system can store data in other media like audio, video, images, but such research objects need to be linked to structural or textual descriptions. We also construct a layer of knowledge – interpretation of data with respect to their creation, content and interconnection. The knowledge is implemented in the form of a Bulgarian-centric Knowledge Graph that represents factual information from data integrated with the conceptual structure of Bulgarian WordNet and other language resources.

The main software components of the ecosystem are:

Instance Register provides a parameterized database which allows the creation of different registries for various kinds of objects. Each register is defined as a schema which presents the conceptualization of the objects. The researchers working on a given

type of objects could specify the properties of the objects and their values. The values could be of different types like numbers, coordinates, textual elements (ranging from a single word to long documents), identifiers of other objects (usually represented in different registers). Thus, the researchers could construct structural descriptions, link them with each other, connect them to the supporting documents, or index the documents with appropriate objects.

Document Referent supports the work with documents. The users can edit documents, describe them with different metadata, annotate them with objects (e.g. Named Entities) from the knowledge graph, with senses from WordNet, with terms from various terminological lexicons. The *Document Referent* component supports the users in their research. They could create their own document database necessary for their research, index it in an appropriate way and create their own research in the form of articles, books. The result can be loaded in the specified components of the system.

Document Annotation and *Lexicon Creation* components support the extension of the linguistic layer of the knowledge represented within the system.

Currently, the system is used for the modelling of cultural monuments and artefacts in the capital of Sofia. The focus is also on biographies of people related to these monuments or artefacts; on the history of Street Names/Name; historical events related to them.

Digitization in context: understanding institutional practices through grant applications

https://doi.org/10.5281/zenodo.11859561

T. Leo Cao

University of Texas at Austin, Leiden University; leo.t.cao@gmail.com

Utilizing computational methods based on a reproducible workflow designed for arts and humanities researchers with little to no coding background, this paper examines the digitization of museum collections as a response to the devastating impacts of the COVID-19 pandemic, during which most cultural institutions were forced to shut down their physical locations, with some closures becoming permanent. Even those that managed to survive struggled to engage with their audiences online, unless they had already cultivated an active online presence and established a robust digital infrastructure to support their operations remotely. Scholars and industry stakeholders have researched the ways cultural institutions strove, or failed, to make the transition and the institutional factors that influenced these outcomes.

Theoretically, this paper seeks to make a contribution to the ongoing scholarly and professional discourse around the digital transformation among cultural institutions by highlighting the role played by cultural policy on a national level. Digitization as a cultural practice depends in large part on the operational capacities of the cultural institutions, which in turn depend on the resources they can utilize and the priority they assign to other activities that necessarily compete for the same resources. Cultural policy, particularly funding opportunities through competitive grants aimed at cultivating digital capacities, therefore plays a crucial role in mediating the dynamic of capacity building and resource deployment, ultimately impacting the implementation of digitization projects in cultural institutions.

Methodologically, this paper utilizes topic modeling to analyze data extracted from the awarded grants database provided by the United States Institute of Museum and Library Services (IMLS) to examine how cultural institutions in the U.S. articulate digitization work in their federal grant applications. This paper first outlines existing research related to digitization as a cultural practice and grantmaking as a policy instrument. Then, it explains how the method of topic modeling is used to analyze the data collected from the IMLS database. Following the results of the analysis, it discusses the main argument of the paper that digitization should be understood, rather than as a standalone technical process, within the broader institutional context in which other activities may compete for operational priority and limited resources.

This paper seeks to contribute to the conference by presenting a reproducible and adaptable workflow, with the help of generative AI tools, for researchers in arts and humanities who may not have the methodological capacities to design and implement similar digital projects from the ground up. This paper therefore addresses one of the thematical questions of the conference: To what extent is the increasing use of artificial intelligence affecting our research workflows? This paper advocates the use of generative AI tools for research and training purposes in arts and humanities research, providing a template of how researchers with little to no coding background can still utilize computational methods to address various research questions in their own fields.

Reproducible Workflows for Innovative Scholarly Outputs

Maciej Maryl, Magdalena Wnuk, Tomasz Umerle

Institute of Literary Research of the Polish Academy of Sciences, Poland; maciej.maryl@ibl.waw.pl, tomasz.umerle@ibl.waw.pl

This paper presents the current research of OPERAS Innovation Lab dedicated to innovative scholarly outputs and their evaluation. The results are presented in the form of guidelines and actionable workflows for researchers to follow.

OPERAS, a European research infrastructure for scholarly communication in Social Sciences and Humanities, has established an Innovation Lab to support the innovative potential of scholarly communication. The Lab provides up-to-date knowledge on innovative communication practices in research as well as targeted guidance for scholars seeking to disseminate their outputs innovatively. The Innovation Lab observatory collects and stores information on innovative outputs as well as on current research in that field. It provides case studies and research papers from different countries with additional insights and conclusions that could be useful for other members of the SSH community interested in starting and maintaining digital initiatives.

This paper will outline the iterative process of research and expert consultations carried out within the lab around three types of innovative outputs:

1. scholarly toolkit (SHAPE-ID toolkit);
2. interdisciplinary online journal (Journal for Digital History and DARIAH Overlay Journal);

3. scholarly code for open science on the example of a recommender system for open-access books based on text and data mining (Snijder 2021).

The case study process is based on the principle of an open collaboration whereby different users will be able to engage with the process. Thus the workflow itself will be open for contributions from the wider community through feedback and consultation events during dedicated workshops. The methodology used qualitative methods (interview) and OPERAS Service Design and Transition Package (SDTP), a standardised template to describe and support the development of new services.

A case study workflow is firstly discussed with the scholars responsible for the project during an in-depth interview focused on identifying their needs regarding the process and the challenges they are encountering, such as issues of intellectual and technological sustainability or evaluation of non-traditional outputs. Then, in an iterative process, solutions are prototyped, involving various stakeholders, like publishers or e-infrastructures (Ell and Hughes 2013), to forge best practices and provide practical advice. Finally, an important part of the process is to prototype the evaluation of the innovative output. It is discussed firstly with the creators of the output and then with the panel of experts. The task is to develop the scholarly quality assurance mechanism which can be embedded in the output (e.g. open peer review) or used in research assessment.

Each case study will be discussed as a reproducible workflow for researchers willing to engage with similar formats and face related challenges. Specifically, it will detail issues around: (1) setup and planning; (2) consecutive steps leading to establishing an innovative project; (3) evaluation guidelines; (4) sustainability considerations. In the conclusions, we will discuss how we aim to use those workflows in the everyday operations of the lab in supporting the creation of innovative services for OPERAS.

"Atlas of Holocaust Literature" - a case study of an workflow in interdisciplinary innovative project combining Holocaust studies and digital cartography of literary studies

Bartłomiej Szleszyński, Konrad Niciński
Institute od Literary Research, Polish Academy of Sciences, Poland; bartlomiej.szleszynski@ibl.waw.pl, konrad.nicinski@ibl.waw.pl

The aim of the paper will be to present a workflow of the Atlas of Holocaust Literature - project co-managed by the Department of Digital Editions and Monographs and Holocaust Literature Research Group of IBL PAN - as an extensive case study. This workflow is noteworthy for project wide scope, diverse objectives and high interdisciplinarity.

The project aims to present testimonies of the Holocaust from the Warsaw and Lodz ghettos in the form of interactive maps combined and linked with topographical passages from the source texts.

It therefore requires the joint work of representatives of various disciplines (literary scholars, historians, digital humanists, cartographers, graphic designers), as well as research teams dealing with two very different urban centers.

The premise of the project is to “translate” the testimonies of the Holocaust and the literary texts on Holocaust into topographical categories. Maps developed during the project are not meant to be historical maps, but rather "maps of experience" of individual authors - creating them requires a close combination of "close reading" and "deep mapping" methods, which lead to the construction of a data network rather associated with "distant reading."

The scale of the project is shown by the numbers - pilot version of the project, created in 2019 and available online, is based on 17 testimonies (exclusively from the Warsaw Ghetto) and contains more than 1,000 topographical units and 3000 passages; eventually the project is to include 12 000 passages and approximately 3,000 space units from Warsaw and Łódź - both in Polish and English.

The project team is divided into four sections

· analyzing selected texts, isolating topographical passages and marking them with established categories;

· entering the extracted fragments into the appropriate places of the digital structure and linking them with other elements of the content

· creating static and interactive maps on the basis of the extracted and described fragments

· ensuring the flow of knowledge and data between the various sections, setting the general scholarly directions the project

The project is also managed by two closely cooperating managers, one in charge of the Holocaust studies aspect of the work and the other in charge of digital implementation, which ensures the project's high quality implementation on both sides.

During the presentation we want to discuss the workflows of each section, connections between each section and general structure of workflow in the project in the context of other digital humanities projects in the field of digital cartography and Holocaust studies, as well as theoretical studies covering close reading and deep mapping in digital humanities.

Scalable data science workflow for the Finnish national bibliography

Julia Matveeva[1], Kati Launis[2], Osma Suominen[3], Leo Lahti[1]
[1]University of Turku, Finland; [2]University of Eastern Finland; [3]National Library of Finland; yulia.matveeva@utu.fi

Statistical analyses of bibliographic metadata catalogs can provide quantitative insights into large-scale trends and turning points in publishing patterns, enriching and even challenging the prevailing views on the history of knowledge production. The use of bibliographic catalogs has become a well established tool in literature history and helped to renew research methodology. However, the efficient utilization of large-scale data collections as research material depends critically on our ability to critically evaluate data representativeness, completeness, quality and trustworthiness. Our earlier work has demonstrated how remarkable fractions of the bibliographic metadata curation and analysis process can be automated through dedicated bibliographic data science workflows.

Here, we report the development of an open and scalable data science workflow supporting the research use of Finnish National Bibliography, Fennica. This expands our previous work to cover all one million records in Fennica from 1488 to the present day

and provides refined bibliographic metadata for the entire Fennica catalog, covering a variety of product types, including monographs, newspapers, and audiovisual material. The refinement strategies are contingent on specific research use cases, however, and devising generalizable solutions is frequently complicated by the nuanced nature of questions posed in the application field. The scalability of the solutions varies by data type, and the refinement process must strike a balance between accuracy and scale. Our reproducible workflows emphasize transparency, consistency, and provenance as key elements of this process; we show how standardized refinement procedures and automated generation of versatile statistical summaries of the refined data can be used to monitor the curation process while supporting in-depth statistical analyses and modeling of publishing patterns over time and geography.

The growing research use of bibliographies reflects the broader expansion in the use of large-scale metadata, which has boosted the development of data science techniques for automated data management, refinement, integration, and computational analysis. We share recommended good practices and conceptual solutions supporting the development of research-oriented bibliographic data science workflows, and use this to demonstrate how automatically refined national bibliographies can help to build a data-rich view of the history of Finland's publishing landscape.

Research workflows in the digital age: A qualitative workflow study in the field of terminology science

Tanja Wissik
Austrian Academy of Sciences, Austria; tanja.wissik@oeaw.ac.at

Changes regarding information and language processing and especially artificial intelligence have led to a series of changes in the research landscape, and terminology science is no exception. As Roche et al. (2019) stated, these technological changes have for example influences on the design of terminological resources, the way data and knowledge are represented and the way users access data (Roche et al. 2019). So, the technological change influences the final product, which is also visible by a number of new or updated digital terminological resources, published in recent years (Roche et al. 2019). However, little is known, how the terminological practices and the workflows, that lead to these new terminological resources, are affected by technological change and emerging paradigms like open data.

This and related questions were explored in a project on terminology workflows carried out in 2023. The project researched changes in terminology practice induced by factors such as rapid technological change, the growing significance of infrastructures, emerging paradigms like openness (e.g., open data, open source), the increasing importance of data sharing and new forms of cooperation with qualitative methods, namely semi-structured expert interviews. During the project, 15 interviews were conducted with individuals involved in terminology practice (e.g. terminologists, terminology managers, tool managers, members of standardizing committees) in different institutions (e.g. university departments, public administrations, language institutes, international organizations). Afterwards the interview data was transcribed, anonymized and analyzed.

This paper reports on the findings based on the analysis of data collected in the before mentioned project on terminology workflows. The aim of the study was to explore how collaborative terminology workflows, where stakeholders from different disciplines (terminologists, domain experts and IT specialists) work together, are affected by technological change. In this paper, the following questions will be addressed:

(1) What are the (new) challenges in the collaboration between the different stakeholders?

(2) Which steps are automated or supported by tools?

(3) How different workflow steps are documented and at which granularity?

(4) How human decisions in the workflow are recorded and documented? What are the difficulties in modelling these decisions?

The findings of this study can be used to compare it with the results from other studies in DH for example with the results of the study "Project Endings" on long-term curation and preservation of DH projects (Comeau 2023). The findings can also be used as a starting point for the discussion, which role research infrastructures can play in the context of workflows.

References

Roche, Christophe / Alcina, Amparo / Costa, Rute (2019). Terminological Resources in the Digital Age. *Terminology* 25(2), 139–145.

Comeau, Emily (2023). The Project Endings Interviews: A Summary of Methodological Foundations. *digital humanities quarterly* 17(1).

The Digital Dictionary of Ancient Greek

Daniel Riaño Rufilanchas
CSIC, Spain; daniel.rianno@cchs.csic.es

Most ancient Greek dictionaries, including some of the most notable examples of European lexicography from the 19th and 20th centuries, were originally conceived as printed works. In these, space-saving considerations were critical, and they were created at a time when digital repositories of Ancient Greek texts did not exist. To make effective use of these dictionaries, users are presumed to possess a strong command of Ancient Greek, sometimes requiring advanced knowledge to navigate information related to dialectal vocabulary or terms derived from glossaries, papyri, or inscriptions.
The Digital Dictionary of Ancient Greek ("Diccionario Digital de Griego Antiguo", DDGA) derives lexical entries, parts of speech, morphology, prosody, citation sources, etymology, word formation, etc. from traditional lexicons. By consolidating information from all available dictionaries, DDGA fills in gaps and, more importantly, generates new data, such as declension types, conjugation paradigms, derivational morphology, etymology, etc. This information can be presented to users in a clear way or returned as a response to enable interoperability with other applications through an API.
DDGA provides users with comprehensive and precise information. It covers the usage of each form and lemma in nearly all extant Greek texts that predate the 6th century CE. These texts include literature, papyri, and a portion of already digitized inscriptions. DDGA employs a two-way strategy to lemmatize and analyze morphologically the texts: a) a list of over 2,000,000

forms to lemmatize (Madrid Wordlist of ANcient Greek) and b) UDPipe (a pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files https://lindat.mff.cuni.cz/services/udpipe/)
By using both procedures (UDPipe, in particular, to disambiguate forms) with our digitized Ancient Greek texts, we achieve a lemmatization accuracy rate of nearly 95%. Users can inquire about usage frequency by period, literary genre, author, region, and more, and receive a reference list for each appearance, and where this is possible, a link to the digital edition of Perseus texts or to the edition of the papyrus in https://papyri.info/. Users have the capability to search not only for individual lemmas but also for all words with specific endings, conjugations, those related to a particular root, those featuring a certain suffix, etc. DDGA generates the complete regular paradigm for each lemma, in addition to dialectal and irregular forms extracted from lexica. DDGA enhances information about composition and etymology, aspects that traditional lexicons often provide in an incomplete and unsystematic manner.
For terms well-documented in papyri and inscriptions, users obtain the actual forms found in historical documents, which often deviate significantly from the canonical form. DDGA supplements traditional dictionaries with several tens of thousands of proper names.
The ability to quantify observable grammatical phenomena in the largest Ancient Greek lexicon may distinguish this resource among the many tools dedicated to Indo-European language grammar. We anticipate wide usage of the paradigm modules for lexical forms among Greek students worldwide. This will enable them to enhance their foundational knowledge and compensate for deficiencies in instructional materials.

Odeuropa Workflow Use Case: Olfactory Digital Data from the Digital Library of Slovenia

Ines Vodopivec[1], Inna Novalija[2], Dunja Mladenić[2], Pasquale Lisena[3], Raphael Troncy[3], Inger Leemans[4]
[1]National and university library of Slovenia, Slovenia; [2]Jožef Stefan Institute, Slovenia; [3]EURECOM, France; [4]Royal Netherlands Academy of Arts & Sciences Royal Netherlands Academy of Arts and Sciences (KNAW), Vrije Universiteit Amsterdam, Netherlands; ines.vodopivec@nuk.uni-lj.si

The Odeuropa project integrated expertise in sensory mining, knowledge representation, computational linguistics, (art) history, and heritage science. Digital data were extracted from thousands of images and historical texts in six languages, all available in the public domain through the Smell Explorer or the Encyclopaedia of Smell History and Heritage.

Digital textual collections integrated into the workflow are represented in the Knowledge Graph, including eighteen GLAM institutions. Among others, these institutions include the British Library, Digitale Bibliotheek voor de Nederlandse Letteren, Deutsches Text Archiv, Digital Library of Slovenia, Europeana, Gallica, and WikiSource. Cooperating GLAMs collaborated with researchers from the UK, Netherlands, Germany, Italy, France, and Slovenia. In this paper, the results of inter- and transdisciplinary work are presented.

A variety of smell experiences, described by smell words, smell sources, qualities associated with smells, smell perceivers, etc., have been extracted from the available digital data. Books (monographs and dissertations) and periodicals (historical, scientific, general newspapers, and journals) have been the primary resources for Odeuropa smell data analysis. Additionally, manuscripts (mediaeval codices and literary manuscripts); images (photographs, postcards, posters); music (musical scores and audio recordings); and maps (maps and atlases) were used.

The Smell Experiences extraction workflow applied in the Odeuropa project encompassed six main steps: (1) Development of annotated benchmarks, (2) Analysis of benchmark statistics, (3) Development of a text processing system, (4) Extraction of Smell experiences, (5) Linking with Odeuropa semantic vocabularies, and (6) Exploration, visualization, storytelling. The workflow enabled the integration of a wide variety of resources for digital humanities research, from textual materials to visual representations.

Under step one, ten domains of interest were defined, such as Household & Recipes or Perfumes & Fashion. The project aimed to include 10 documents for each category in the benchmark, totalling 100 documents. Next, annotation setting was performed for each language using the INCEpTION tool. Quality control of the annotated datasets, addressing issues such as missing annotations, smell words with double annotations, unlinked frame elements, and relation errors, was then conducted by the researchers.

Analysing benchmark statistics was the second step. From the statistics of the Slovenian annotated benchmark, it became apparent that smell sources and qualities are frequently found together with smell words in historical texts.

Under step three, the annotated benchmark was utilized to develop models for extracting smell frame elements from the text corpora. One of the main outcomes was the development of state-of-the-art smell extraction models for included languages. Sloberta served as the basis for developing Slovenian one.

Subsequently, Odeuropa partners utilized the developed models to extract the actual smell data for the European Olfactory Knowledge Graph in step four. This was followed by linking of data with Odeuropa semantic vocabularies and embedding them into Odeuropa tools in step five. The Slovenian dataset was digitally processed, smell experiences were obtained and integrated into the Odeuropa Smell Explorer tool.

In the last stage, Odeuropa results were used for exploration, visualization, and storytelling. The number of tools and resources can be reused in the context of digital cultural heritage.

Emese Kún: Database for a Heritage - A Case Study Presenting Budapest City Archives' Thematic Websites

Emese Kún
Budapest City Archives, Hungary; kune@bparchiv.hu

In the last decade, Budapest City Archives has developed thematic websites sharing the documents of different provenance relating to the life and œuvre of several architects with a broad audience online. However, the elaboration of well-structured online databases confronted us with numerous theoretical and methodological challenges.

Our institution articulated the need to update an existing interface in 2019. We had to migrate data and a modular system from one software to another, and to make the concept applicable in case of further data augmentation. After finishing and publishing the first „digital archives" collection, we immediately started working on making the workflow and the toolset more effective for a future project. We found that, as a matter of fact, we lack the vocabulary to find the common denominator between our ideas and the technical possibilities. Moreover, the technical team was not familiar with the expectations and considerations of an institution working with archival material – artistic products, respectively architectural plans and/or buildings as artworks. Firstly, we had to define a flexible data structure in which manifold types of documents and artistic products can be presented. The process of arranging, describing, digitizing and publishing a group of archival documents whilst accumulating the relevant metadata is not as evident as it seems to be.

Several questions we managed to answer, but a few remained, open to international discussion. Most importantly, how can we form a data structure that provides a firm basis for a multi-levelled thematic framework, advancing analytical work and visual representation, when it does not correspond to the physical structure and arrangement of the material? How can we find a background system that enables the fluent migration of data from one software to another without any loss? It is unavoidable to reconcile the datasets of the existing Archival Information System with the one of the planned online platforms, even though they do not always overlap. The multi-layered platforms we endeavoured to create highly depend on the exact definition of requirements, in context of datasets, searchability and the reuse of data. The datasheets and datasets of our current websites allow us now e.g. the sorting of documents by geocoding and georeferencing (based on address). How can we make these processes more efficient and automatized? Which concepts and functions do we have to sacrifice in order to make the website operational?

We attempt to outline the ideal circumstances for a collaboration between an archival institution and software developers. At the same time, we would like to introduce the concept and workflow of representing the heritage of creative individuals, which contain a significant amount of metadata. Therefore, the (virtual or physical) private collections in the care of a public institution can serve as an actual case study in merging cultural heritage protection with data science.

List of our currently operating thematic websites:

https://kos.bparchiv.hu/en/homepage

https://lajta.bparchiv.hu/en/homepage

https://ybl.bparchiv.hu/en/node/4133

https://habitation.archivportal.hu/en/node/1020

Computational analyses of cultural production

Jeba Akewak, Pyry Kantanen, Julia Matveeva, Leo Lahti
University of Turku, Finland; yulia.matveeva@utu.fi

Quantitative analyses of cultural production increasingly rely on statistical integration of large and heterogeneous metadata collections. Data streams retrieved from statistical authorities, surveys, data observatories, and other sources are now routinely integrated into comprehensive research workflows. Ensuring the robustness and reliability of such procedures has become increasingly important, given the inherent complexity of the data and algorithm combinations in data-intensive mixed methods research. Furthermore, the tensions between openness and protection are further complicating the research and challenging transparency. A healthy development of this expanding field will depend critically on the community-driven model that supports continuous and transparent evaluation, improvement, and transfer of general-purpose methodology among the active data analysis practitioners. Whereas well-established traditions for collaborative development of modular algorithmic tools have emerged in quantitative research fields, the adoption of similar practices in computational humanities is lagging behind.

We will discuss these challenges by showcasing our recent work that focuses on data-driven analyses of cultural production of literature and music. This builds on open digital resources retrieved from European and national statistical authorities and memory organizations. Our work emphasizes the need to complement data management infrastructures with dedicated statistical analysis strategies that are adapted to the application domain; this can support robust conclusions from scarce and incomplete observations and help to distinguish between competing interpretations. We conclude by highlighting the collaborative development of an open source ecosystem that has facilitated the standardization and transparent evaluation of the alternative analysis strategies as part of the research workflow.

Building Bridges: Collaborative Approaches to Archiving and Accessing Urban History in Budapest

Ágnes Telek[1], András Sipos[1], Barbara Szij[2], Bíborka Anna Grócz[2], Viktória Sugár[3], András Horkai[3]
[1]Budapest City Archives, Hungary; [2]Hungarian Contemporary Architecture Centre, Budapest100, Hungary; [3]Óbuda University, Ybl Miklós Faculty of Architecture and Civil Engineering, Hungary; teleka@bparchiv.hu, barbara.szij@kek.org.hu, biborka.grocz@kek.org.hu

This paper explores the collaborative efforts of potential contributions between experts and perspectives on the subject of architecture and urban history. The Budapest City Archives is one of the most important and pioneering keepers of archival documents and drawings of the buildings of the Hungarian capital city. As such, we engage with various returning researchers and users, each with diverse project outcomes. The archival documents and the datasets they require are so heterogeneous it is a great challenge to have the proper way to present them to a wide variety of users.

The Budapest City Archives has been building databases for decades now, our most compound database is the Budapest Time Machine where we integrate more than 30 layers of georeferenced historical maps and four different types of archival documents. This way of accessibility initiates a new generation of research in the digital humanities and offers new opportunities to our stakeholders, who are not only using our online surfaces but creating new documents, (partly) based on ours, which are then returned to the archives for publishing.

One of our principal stakeholders is the 'Budapest100' project (coordinated by a Budapest-based NGO, the Hungarian Contemporary Architecture Centre) which is a pioneer in citizen science in Hungary. This public event is an annual celebration for the inhabitants and locals to uncover the past of their built environment. They work together with volunteer researchers to reconstruct the stories and histories of different buildings in Budapest. The other fundamental partner is the Ybl Miklós Faculty of Architecture and Civil Engineering, Óbuda University, which teaches the students 3D modelling by using our collection of building plans.

The archiving process of digital-born documents in both cases requires technical synchronicity. The metadata the different actors are using are similar, but the formatting of the data fields is different. Furthermore, they not only involve our datasets but use other sources as well, and those metadata which come from a different structure have to be integrated as well. The standardization of these fields is currently done manually. In the future, the automation of the data compilation would be necessary to expedite the archiving process.

We are putting the experiences of research workflows with this prime contribution into best practice guides within the framework of the **"City Memories" project.** This initiative brings together archives from three cities: Budapest, Copenhagen and Stockholm, dedicated to enhancing the interpretation and presentation of architectural drawings within the cultural heritage sector. The project integrates professionals from different fields, e.g. urban history, technology, and architecture. By fostering interdisciplinary collaboration, the project aims to address sector-specific priorities, reinforce the capacity of cultural heritage professionals, and engage with cross-cutting issues such as gender and environment.

Virtual Gallery: A Nexus of Artistic Practice and Research

https://doi.org/10.5281/zenodo.11193471
Jaroslaw Solecki[1,2], Justyna Gorzkowicz[2,3]
[1]Blue Point Art, United Kingdom; [2]Polish University Abroad, United Kingdom; [3]Research Center on the Legacy of Polish Migration (ZPPnO), United Kingdom; jaroslaw.solecki@puno.ac.uk, justyna.gorzkowicz@puno.ac.uk

This poster demonstrates the convergence of art, research methods and technological innovation within the digital humanities, specifically through the roles of artist-researchers and art anthropologists at the virtual Blue Point Art Gallery London. Created in response to the COVID-19 pandemic, this small but dynamic platform uses interactive 3D graphics accessible through web browsers to create a multifaceted space for creative expression, documentation and education. The gallery encourages both verbal and non-verbal interactions between artists, researchers and the public, with exhibitions delivered via WebGL and accessible directly on internet-connected devices without the need for additional installations.

An exemplary project is 'The Dystopia of Imitation', supported by Arts Council England. It features a transition from virtual installation to augmented reality (AR), starting with an animated 3D sculpture inspired by Johannes Vermeer's 'The Milkmaid'. This project illustrates the gallery's workflow by combining traditional sculptural techniques with digital methods, culminating in a generative art object in AR, signifying the merging of the physical and virtual realms. Such innovations demonstrate the potential of virtual galleries and AR to widen access to art, foster global collaboration and democratise the art process.

To further emphasise our interdisciplinary approach, the Gallery is also involved in educational collaborations. One notable project explores the reception of Stanisław Vincenz's work through a virtual exhibition of previously unknown photographs of the Hutsul region taken by Lidio Cipriani in 1933. These photographs, in the possession of Jacopo Moggi-Cecchi (University of Florence), have been digitised for this exhibition, which will be enhanced with additional multimedia elements and an accompanying bilingual book.

The Gallery also hosted a number of other exhibitions, including an exhibition of sketches by the Polish poet and artist Cyprian Norwid (along with an international conference and the development of a monograph). These projects underline the Gallery's commitment to non-commercial activities and its role as a nexus of artistic practice and research. They exemplify our commitment to integrating digital innovation into scholarly and artistic endeavours.

Blue Point Art Gallery is characterised by its fundamental research orientation and its focus on the digital value of design tasks. Our efforts extend beyond virtual exhibitions to include textual studies published as e-books and archived, for example, in the ZENODO repository.

In the context of gallery workflows, our approach integrates reproducible digital methods to standardise project management across artistic and academic collaborations. By using digital tools such as standardised digital archiving protocols, we can ensure consistency and high fidelity in the reproduction of results across different projects. For example, the transformation of The Milkmaid sculpture into AR can serve as a template for subsequent digital art transformations, using scripted automation to replicate and adapt the workflow for future exhibitions. This systematic application increases the scalability and applicability of our methods, enabling broader, more collaborative projects that maintain rigorous academic and artistic standards.

These initiatives highlight the Gallery's unique position within the digital humanities, combining artistic exploration with rigorous academic research to foster a deeper understanding and appreciation of art across global and digital landscapes.

Tracking fake news: Automatic construction of a dynamical knowledge graph

https://doi.org/10.5281/zenodo.12515239
Vyacheslav Tykhonov[1], Yves Rozenholc[2], Julliette Vion-Dury[3], Andrea Scharnhorst[1]
[1]Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science, Netherlands, The; [2]University Paris; [3]Université Sorbonne Paris Nord; yves.rozenholc@u-paris.fr, juliette.vion-dury@univ-paris13.fr

Web content represents both current activities and the memory of our internet based societies. The Now Museum (Dury et al. 2023) represents an approach to combine keyword driven scrapping of web-content with a research data sharing and archiving platform (Dataverse) delivering collections about any topic of interest. Hidden in such collections are knowledge graphs which "interlink entities (concepts) while also encoding the semantics or relationships" among them. Recently, in the context of the Now-Museum's infrastructural architecture, a new workflow has been established which extracts such knowledge graphs by a combination of the use of Large Language Models with querying structured data in an iterative way (Tykhonov 2024).

As a result, we obtain a knowledge graph whose edges link two keywords (nodes) related to the original concept, in which types of relationships are expressed in SKOS or OWL. But, we also see now new types of relationships between concepts not formalized or not recognized by human experts before.

We applied this technique to the subcollection on COVID-19 in the Now.museum. Concerning the necessary computer processing, it is important to point out that the original construction of such a thematic knowledge graph is resource consuming. However, once available the finally obtained knowledge graph can be updated almost instantaneously. We propose to use this latter property to derive a tool which is able to detect the apparition of a new concept by comparing an existing knowledge graph built with documents collected before time T to its evolution built from documents collected at times T+1, T+2, …. As a consequence, we receive a dynamic knowledge graph and can track its evolution. Our hypothesis is that 'fake news' appearing as part of the (evolving) source collection comes with a different context and thus changes the topology of the resulting knowledge graph.

Since January 2020, our team has captured the public parts of European newspapers in their original languages, with their publication date. We have put our focus on those containing either the term "Sars-Cov" or "Covid". This is the basis for a dynamic knowledge graph (recorded daily) of the Covid-19 pandemic. It enables us to visualize the evolutive topology or geometry of the graph. This will be especially striking around the periods of emerging of so newly emerging and debated terms as "Pangolin", "Masks", "vaccines", "Messenger RNA", …The underlying questions are: 1) How does the typology of the graph change when new terms emerge? 2) Is there a specific change of geometry for 'fake news'?.

The domain of application of such a tool would be both research and citizenship debates. We combine this workflow presentation with a digital humanities reflection. More specifically, we propose a philosophical and anthropological approach of the "fake news" phenomenon as well as an analysis of the psychological means through which they work out their way to the public misinformation. All of these elements underlie and underpin the adequation of the Now-Museum's scientifical, technological and ethical approach to such a complex collective issue.

Preservation as valorization: Strategic approaches to sustaining digital humanities workflows

https://doi.org/10.5281/zenodo.11533943

Elena Battaner Moro, Juan Alonso López-Iniesta

Universidad Rey Juan Carlos, Spain; elena.battaner@urjc.es

This paper presents a comprehensive strategy for preserving digital humanities workflows in the face of rapidly advancing digital technologies. Recognizing the multifaceted nature of workflows, which encompass datasets, analysis software, custom scripts, documentation, and metadata, we propose a structured approach to ensure their long-term viability, functionality, and accessibility.

Beginning by identifying some preservation challenges posed by the diverse components of digital humanities workflows, this paper emphasizes the need for a multifaceted approach to address the issues of digital obsolescence and evolving technological environments. The core of our strategy involves the integration of PID/Digital Object Identifiers (DOIs) and the use of Zenodo, an open-access repository, for the long-term storage and accessibility of these workflows. Assigning DOIs enhances the traceability and citability of workflows in scholarly contexts, while Zenodo facilitates their archiving with GitHub integration, particularly beneficial for workflows involving software development.

A significant focus of the paper is the development and implementation of a comprehensive metadata schema. This schema, e.g. incorporating elements of the TADIRAH taxonomy (see mainly https://vocabs.dariah.eu/en/), ensures complete coverage of workflow components, addressing both their descriptive and technical aspects. This approach not only aids in the standardization of metadata but also enhances interoperability within the digital humanities research landscape.

The paper describes the technical process of converting different workflow elements into a singular, cohesive digital object. This conversion involves advanced digital packaging strategies, employing container technologies for software and scripts, and bundling data with associated metadata. The creation of a detailed manifest file within the packaged workflow is highlighted, ensuring that the workflow is navigable and understandable as a unified entity.

In conclusion, this paper argues for the necessity of these preservation strategies to safeguard the long-term viability and accessibility of digital humanities workflows. By implementing these measures, researchers and institutions can contribute to the sustainability and advancement of digital humanities scholarship, ensuring that valuable digital methodologies remain accessible and usable in the face of ongoing technological change.

Furthermore, the recognition of digital humanities workflows as legitimate and citable research products may represent a paradigm shift in scholarly communication and evaluation. By employing structured preservation strategies, these workflows could thus transcend their traditional role and gain significance as autonomous contributions to knowledge. The ability to assign digital object identifiers (DOIs) to workflows, as discussed in this article, not only facilitates their citation and scholarly recognition, but also conforms to new trends in research metrics that value various forms of scholarly output.

Some references:

Borek, Luise; Hastik, Canan; Khramova, Vera; Illmayer, Klaus; Geiger, Jonathan D. (2021): Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR. In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th—10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 321—332. DOI: doi.org/10.5283/epub.44951

DOI Foundation (2023). DOI Handbook. DOI: doi.org/10.1000/182

Zenodo. www.zenodo.org/ Accessed 13 Dec. 2023

The study of mortars as a clue to discuss the controversial restoration of the reliefs on the façade of the 18th-century Saint Francis of Assisi Church in Ouro Preto, Brazil

Yacy Ara Froner[1], Luiz Antônio Cruz Souza[1], Willi de Barros Gonçalves[1], Ana Carolina Neves Miranda[2], Tiago de Castro Hardy[1], Izabelle Lohany[1], Lívia Freire[3], Hugo Marlon da Silva Nascimento[4], Alessandra Rosado[1], Antônio Gilberto Costa[1]

[1]Federal University of Minas Gerais (UFMG); [2]National Historic and Artistic Heritage Institute (IPHAN); [3]Federal Institute of Minas Gerais - Campus Ouro Preto (IFMG-OP); [4]Federal University of Itajubá, Itabira Campus (UNIFEI);
luiz.ac.souza@gmail.com

The Saint Francis of Assisi Church, in Ouro Preto, is one of the most important architectural monuments in Brazil. The building was entrusted to the Portuguese master mason, Domingos Moreira de Oliveira (c.1711-1794), with the contract dated 1766. The sculptural work was entrusted to the Portuguese craftsman José Antônio de Brito (?-1793). Upon completion of the main chapel, administrators began constructing the façade, featuring the entrance door designed by Antônio Francisco Lisboa (c.1738-1814), more commonly known as *Aleijadinho,* widely studied by art historians such as Germain Bazin (1901-1980) and Robert Smith (1912-1975). This church was one of the first architectural assets to be protected by national legislation, through its inscription in the Book of Fine Arts of the National Historic and Artistic Heritage Institute (IPHAN), in 1938. In 1980, it was included in the UNESCO World Heritage List.

Considering its historical and artistic importance, it was chosen as the research subject for the "Damage Map by Heritage Building Information Modeling (HBIM)" project, funded by the Brazilian National Council for Scientific and Technological Development. The project aims to characterize the deterioration patterns, propose damage assessment methodologies, and develop conservation recommendations for stone decorative building reliefs. To assist the project, the initial stage of the work consisted of archival research on the history of conservation-restoration and analytical methodologies already carried out at the monument. The first record of interventions is from 1822/1823, according to a payment receipt issued "to the bricklayer João Fernandes Coutinho for fixing the stonework on the Main Door by the Third Order of Saint Francis" (Box0259). Between 1935 and 1937, the National Monuments Inspectorate (IMN), led by Gustavo Barroso (1888-1959), carried out a second restoration of the church. The façade underwent cleaning, and two ornaments from the entrance door were removed, sparking significant controversy over the authenticity of these pieces. Members of the Historical Institute of Ouro Preto contended that the ornaments were original elements from the period of the entrance door's construction and were made of soapstone. Conversely, the IMN argued that these pieces were crafted from mortar and were removed because they were considered "dishonest and modern work" introduced in the 19th century, according to documentation from the IPHAN Central Archive (Box0259).

Was the restoration of the IMN appropriate or not? In search of answers, we sought to map all traces of mortar in the façade's reliefs and collect samples for their characterization. The field surveys incorporated close-range photogrammetry, as well as photography and aerial photogrammetry, using an Unmanned Aerial Vehicle (UAV). Subsequently, the KRITA software was used to generate layers and highlight the mortar elements for deciding the sample collection areas. Mortar analytical studies were carried out using diverse techniques: x-ray diffraction (XRD) to verify the phases present; probable trace calculations were performed by acid attack of the pulverized sample; Electron Microscopy images were generated to observe the particle morphology of the mortar constituents; scanning electron microscopy coupled with energy dispersive spectroscopy (SEM-EDS) was used to identify and map the elements present in the samples.

Telling stories from the past through digital collections: the PROMETHEUS project

Dimitar Ilkov Iliev[1], Elena Dzukeska[2]

[1]St. Kliment Ohridski University of Sofia, Bulgaria; [2]Ss. Cyril and Methodius University of Skopje, North Macedonia;
dimitar.illiev@gmail.com, elena@fzf.ukim.edu.mk

For the last 20 years, the encoding and the online publication of historical inscriptions on stone (known as epigraphic monuments) has been among the most fruitful and fastest-growing branches of Digital Humanities. The EpiDoc community, with its guidelines, schemas, and other tools, has proved essential for the creation of many digital collections of inscriptions from the Greco-Roman world and is currently surpassing the boundaries of both Classical Antiquity and epigraphy. One of the main challenges before the discipline today is connected with the greater interchangeability between separate collections by applying the FAIR principles as shown by initiatives such as epigraphy.info and the *FAIR Epigraphy* Project. Other steps further have been made by the EAGLE Europeana Project that addressed the issues of controlled vocabularies across all collections, as well as the public outreach of the digitized inscriptions for which it created its Storytelling platform.

Ancient inscriptions in Greek and Latin form a significant part of the heritage of South-Eastern Europe, spreading across national borders and connecting the region to the wider cultural and political context of the Mediterranean. Formerly accessible mainly to experts, epigraphic monuments are now made available to the general public by the first digital epigraphic databases from the region such as the *Telamon* collection of ancient Greek inscriptions from Bulgaria. EpiDoc projects, collaborations, and workshops have also taken place in recent years in Serbia and North Macedonia, for example, the *Digitizing Ancient Epigraphic Heritage from Serbia* Project. Such activities are necessarily being conducted in collaboration with various museums. However, the work done so far only emphasizes the need for closer ties between academic and GLAM institutions. Furthermore, the common activities inspire both academics and museum workers to get the wider public better acquainted with this rich and various branch of cultural and historical heritage in its diverse aspects and curious intricacies.

Thus, several academic and GLAM institutions from North Macedonia, Serbia, and Bulgaria have started the PROMETHEUS project in the framework of the Creative Europe Programme. This cooperation aims to bring together the existing successful practices in Digital Epigraphy to create a common digital collection of inscriptions. The collection will serve as a basis for the establishment of a storytelling platform. In an engaging and inspiring way, the storytelling would reveal, through the documents written on stone, the everyday lives, loves, deaths, beliefs and hopes of the inhabitants of the Balkans during Antiquity, stressing the common culture and the universal values they shared. Various workshops, conferences, as well as education and dissemination events, will take place in all of the countries involved in the project. Educational and dissemination materials will be distributed in English as well as in all of the languages of the region (Macedonian, Bulgarian and Serbian, as well as possibly

Turkish, Roma, Armenian, etc.). Training videos for scholars, field archaeologists, and GLAM institution employees will also be created, acquainting them with good practices in (Digital) Epigraphy and Public Outreach.

A Blueprint for Digital Work Practices in the Humanities

https://doi.org/10.5281/zenodo.12080000

Panos Constantopoulos[1,2], **Vicky Dritsou**[2,1], **Katerina Gkirtzou**[3], **Helen Goulis**[4], **Patritsia Kalafata**[4], **Irakleitos Sougioultzoglou**[4], **Yorgos Tzedopoulos**[4]

[1]Department of Informatics, Athens University of Economics and Business, Greece; [2]Digital Curation Unit, IMSI/Athena RC, Greece; [3]ILSP/Athena RC, Greece; [4]DARIAHGR/DYAS Academy of Athens; p.constantopoulos@athenarc.gr, v.dritsou@athenarc.gr, katerina.gkirtzou@athenarc.gr, egouli@academyofathens.gr, pkalafata@academyofathens.gr, isougioultzoglou@academyofathens.gr, gtzedopoulos@academyofathens.gr

We present a blueprint for representing digital work practices in the Humanities, capable of accommodating domain-specific variations while fostering conformance with standards, metadata schemas and good practice guides. Its purpose is to help digital humanities researchers learn, adapt, apply and critically assess digital work practices, not only in isolated tasks, but in entire workflows. Understanding situation- and domain- specific workflows as specializations of more general processes can enhance effectiveness by virtue of normalization. It is then easier to reuse tools, information structures and vocabularies, to apply standards, to adhere to FAIR data principles [1] and data openness.

The blueprint was defined in the framework of the Greek project "The emerging landscape of digital work practices in the Humanities in the context of the European projects DARIAH and CLARIN" [2]. It was informed by an extensive survey of work practices, comprising a widely answered questionnaire [3] and six focus groups, and a critical review of relevant metadata schemas, conducted as part of the project in 2023. Relevant, among others, also are research requirements [4] recently reported by Europeana, previous analyses of work practices in the Humanities [5,6], and the findings of a pan-European survey conducted in DARIAH [7].

The blueprint features a structural and a procedural part. The structural part has the form of a knowledge graph, based on the Scholarly Ontology (SO) [8] and implemented using an open-source graph database system [9].The SO provides a domain-neutral ontological model of research work. Specialized classes and domain-specific terminological resources are then employed to capture the work practices and support the requirements of different disciplines. In this way, processes, tools and information resources are contextualized in a manner allowing comparability and interoperability. In the work reported here, practices recorded in the survey have been represented in the knowledge base. Special effort was devoted to enriching the hierarchy of SO Activity Types to capture the scope of specific data seeking, production, organization and dissemination practices. A critical effectiveness factor is the alignment of metadata schemas and terminology with research practices. In our work we have identified metadata schemas appropriate to the domains and tasks recorded in the survey and mapped them to the relevant blueprint elements along with suggestions reflecting current best practice.

The procedural part has the form of a workflow generator, i.e., a generic workflow of research work, in particular the information processes involved. Formulated as a BMPN-style diagram, it represents an archetypical overall process specified in terms of high-level SO activity classes. Through a specialization parallel to that performed in the structural part, contextualized workflows are generated to represent the flow of work in specific domains and tasks. The workflow has evolved from that in [10]. For validation, it has been used to generate specific workflows corresponding to practices identified in the abovementioned survey, in addition to previously having generated the workflow of unifying historical archives [10]. The generic workflow is shown in the Figure below. For illustration purposes, it also includes certain subordinate activity types, quite widely used.

Pipelines, workflows, work packages: what's in a word? A reflection on metaphors used to design interdisciplinary projects in Digital Humanities

Anna-Maria Sichani, Caio Mello, Kaspar Beelen, Kunika Kono

University of London, United Kingdom; annamaria.sichani@sas.ac.uk, caiomellodh@gmail.com, kaspar.beelen@sas.ac.uk, kunika.kono@sas.ac.uk

This presentation seeks to offer a critical reflection on the currently available metaphors for describing project design, implementation and planning in interdisciplinary Digital Humanities projects, by using existing DH projects as case studies.

With the "project" being the core component of Digital Humanities scholarship, a number of concepts and methodologies emerged to describe the operational structure and processes taking place in DH interdisciplinary projects. 'Pipelines', 'workflows', 'work packages', all the above mentioned concepts capture aspects of the technical, methodological and infrastructural organisation and collaboration in interdisciplinary Digital Humanities projects.

'Pipelines' foreground the computational aspects of DH projects by depicting themes as a linear series of data-driven tasks and describing a plan and order of execution of methodologies. 'Workflows' suggest an organisational and structural focus, and refer to a structured plan of work, allowing more flexibility and modularity of actions. Lastly, 'work packages' are a concept suggested by funding bodies' requirements, delineate a series of quantifiable tasks (in terms of cost, quality, and time) and are associated with outcomes such as milestones and deliverables.

What are the differences among these concepts and how do these distinctions apply to realistic research scenarios? Do these concepts incorporate the principles of interoperability and reproducibility? Do they allow room for (or enable) efficient and productive interdisciplinary collaboration among team members, throughout the project lifecycle? What other values or requirements do we think are necessary to address within these structures? And finally, how is such a (metalevel) study about the discourse we use to describe interdisciplinarity in the DH projects helping us to improve the organization of and collaboration within DH projects?

An ethical data practice for intangible cultural heritage: a practical guide for collecting and linking data on immaterieelerfgoed.be

https://doi.org/10.5281/zenodo.11949992

Sofie Veramme

Workshop Intangible Heritage Flandres, Belgium; sofie@werkplaatsimmaterieelerfgoed.be

Intangible cultural heritage (ICH) is very much alive: practices evolve over time and adapt to a fast-changing world.

Communities, groups and individuals are the prime custodians of ICH. It's them who safeguard this living heritage for future generations, through a myriad of actions. As of today, this inherent participatory approach to heritage care is still not customary in relation to many other types of heritage. Article 15 of the 2003 UNESCO Convention for the Safeguarding of ICH clearly states that the safeguarding of ICH cannot be achieved without the widest possible participation of (heritage) communities. This implies that existing approaches and methodologies for linking data that are commonly applied to tangible heritage cannot be directly practised in the context of (meta)dating ICH, since most of them don't imply the outspoken involvement of the communities concerned.

During the project Towards Persistent URIs for immaterieelerfgoed.be: Vital Building Blocks for Digital Transformation in the Cultural Heritage Sector, NGO Workshop Intangible Heritage Flandres researched how to develop a persistent and sustainable data practise for ICH. The project addressed questions relating to the tension between the principles behind open data on the one hand, and the stewardship of communities over their living and evolving ICH on the other. Immaterieelerfgoed.be is the platform for ICH in Flanders and Brussels. It includes inventories of ICH. The content of these inventories mainly take on the form of descriptive data and are collated through co-creation.

To achieve an ethical data practice, the rights and well-being of the ICH community involved in all phases of the data life cycle on - and beyond - immaterieelerfgoed.be have to be prioritised. The aim is to reduce the potential harm that data sharing may cause to the communities involved, e.g. sharing images of protests by socially vulnerable people or the misuse of heritage that has customary restrictions on access and use). By building upon known principles for data sharing, combining FAIR (open and persistent data sharing) and CARE (people and purposes oriented data sharing), a methodology to map the current data practice applied on immaterieelerfgoed.be has been developed. Additionally a series of actions to establish an ethical data strategy, ensuring both clean and open data and agency and advocacy of the ICH communities, has been discerned.

During the course of the project, we scrutinised the current registration module for the ICH inventories, applied user agreements and privacy policies and the attribution of rights statements and reuse licences (e.g. Creative Commons). We researched how we can implement persistent URI's and install an opt-out option for them without breaking their persistence. In addition we found that it is indispensable that the communities concerned have to be sensitised about the impact of sharing their data.

In doing so the data on immaterieelerfgoed.be will not only become linked open data, but also LOUD: linked open usable data. This opens the door for further applications of the data and maximises the impact of ICH for the communities.

Supporting digitally enhanced scientific workflows in a clustered Infrastructure: H2IOSC

https://doi.org/10.5281/zenodo.12073492

Alessia Spadi, Emiliano Degl'Innocenti, Federica Spinelli, Francesco Coradeschi, Irene Falini, Michela Perino, Lucia Francalanci, Francesco Pinna

Istituto Opera del Vocabolario Italiano del Consiglio Nazionale delle Ricerche (OVI-CNR), Italy; alessia.spadi@cnr.it, emiliano.deglinnocenti@cnr.it, federica.spinelli@cnr.it, francesco.coradeschi@cnr.it, irene.falini@cnr.it, michela.perino@cnr.it, lucia.francalanci@cnr.it, francesco.pinna@cnr.it

H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud) is a project funded by the National Recovery and Resilience Plan (PNRR) in Italy, aiming to create a federated cluster comprising the Italian branches of four European research infrastructures (RIs) - CLARIN, DARIAH, E-RIHS, OPERAS - operating in the "Social and Cultural Innovation" sector of ESFRI (European Strategy Forum for Research Infrastructures).

H2IOSC facilitates collaboration among researchers from various disciplines in the social sciences and humanities fields. Its primary objective is to conduct data-driven research activities, by supporting several scientific workflows. Around the workflows selected by each participating infrastructure H2IOSC will build as many Scientific Pilots supported by digital data, tools and services, in the form of Virtual Research Environments, by using 8 high-performance computing data centers built by the cluster. Additionally, it aims to enhance the adaptation, implementation, and efficiency of existing facilities, based on the needs identified and expressed by the respective communities. H2IOSC will establish an onboarding process for each federated infrastructure within its framework, that will ensure the designated maturity threshold. Moreover, H2IOSC promotes training and dissemination activities to foster proficiency in the FAIR (Findable, Accessible, Interoperable, Reusable) and Open Science domain. Ultimately, H2IOSC seeks to drive the digital transformation of cultural and creative industries.

The role of DARIAH.it in the project is to build the distributed infrastructure and federate its 8 nodes, elaborate the semantic framework to represent the knowledge gathered in H2IOSC in an interconnected way and develop the Scientific Pilots supporting the digital philology workflow. To do so, DARIAH.it will undertake a set of preparatory activities, including collecting and evaluating existing tools, datasets, and services relevant for the above task. A crucial part of this process is the semantization of selected data and metadata to promote interoperability of diverse resources; this involves the definition of standardized vocabularies and ontologies alongside the deployment of tools and workflows which actually implement the semantic transformation. The Pilots, presented as platforms or hubs, integrate domain-specific services, workflows, and interfaces. They operate on meticulously parsed data-subsets to facilitate the development of prototypes. Connected to this research and development activities, DARIAH.it will promote training, outreach, and dissemination efforts, as the funding of a Doctoral Scholarship in Digital Philology.

DARIAH.it assumes the responsibility of designing, developing, and populating AEON, a service-provision oriented infrastructure interacting with the H2IOSC Marketplace and aligned with SSHOC and EOSC platforms. Ultimately, by participating in the project,

DARIAH.it endeavors to provide an indispensable service that ensures the development and sustainability of humanities and human sciences research, as well as cultural heritage preservation, by facilitating the transition to digital approaches in research activities, overcoming the obsolescence of current sector-specific research product systems, and maximizing the potential impact of research communities by developing innovative approaches.

Moreover, DARIAH.it establishes a foundation for interactions with the Italian Cultural and Creative industry and GLAM sectors, fostering collaborative research and development endeavors to enable the adoption of cutting-edge technologies. Simultaneously, it strengthens training programs and enriches content through valuable knowledge transfer activities.

Workflow Dynamics on the Mnemosine Academic Digital Library: Integrating Data and Expertise

https://doi.org/10.5281/zenodo.11221739

Dolores Romero-López[1], **Joaquín Gayoso-Cabada**[2], **José Miguel González-Soriano**[3]

[1]Universidad Complutense de Madrid, Spain; [2]Universidad Politécnica de Madrid, Spain; [3]Universidad Complutense de Madrid, Spain; dromerol@ucm.es

The primary aim of the Mnemosine Academic Digital Library, which focuses on rare and overlooked Spanish literary texts published between 1868 and 1936, is to curate, classify, and digitally showcase works that have fallen into oblivion. This endeavor facilitates a retrospective exploration of the period, highlighting texts and authors who were eclipsed by the prominent literary figures of early 20th-century Spanish culture. The accomplishment of this mission has been a synergistic effort among Spanish literary scholars, library data experts, and computer developers.

We aim to introduce Clavy, an experimental research platform that has enabled specialists to salvage, purify, enrich, and systematically organize the data intrinsic to this library. The central goal for computer scientists has been the creation of Clavy, a platform designed for the administration of highly specialized and adaptable digital collections. Clavy boasts an advanced navigational system capable of browsing, filtering, and searching, thus endowing users with the ability to perform in-depth analyses of collections replete with substantial data volumes and complex interrelations.

For experts in Spanish literature, the goal has been the recovery and expansion of the corpus of 'rare and forgotten' artists, texts, sources, and genres attributed to The Other Silver Age. Librarians have focused on importing data and generating metadata from national and international public libraries, refining and interlinking them to cultivate specialized collections within the database, and utilizing tools that bolster interoperability with other systems and platforms. With Clavy's assistance, metadata for over six thousand digitized objects from institutions such as the Biblioteca Nacional de España, the HathiTrust Digital Library, the University of California Library, and the Bibliothek des Ibero-Amerikanischen Instituts in Berlin have been assimilated into the Mnemosine database.

The results of our research have paved the way to the theoretical realm of digital cultural history concerning the Silver Age of Spanish Literature, as evidenced by the publication of a monograph in the Q1-ranked journal Signa, and another by Peter Lang Publishing. The data and metadata of Mnemosine are readily accessible on platforms such as Zenodo, Github, and Wikidata. In May 2022, the Mnemosine Digital Library was distinguished with the award for the Best Tool, Resource, or Infrastructure developed in Spain in 2021 by the Spanish Association of Digital Humanities, and it has received commendations from the Office of Transfer of Research Results at the Complutense University of Madrid. Globally, the Mnemosine Digital Library is incorporated within the Time Machine Europe Project.

We are very interested in presenting our research results to DARIAH Annual Event to show our research data and our Platform tool to converge with the workflow of the European structures such as DARIAH-EU, CLARIN-ERIC and SSHOC. We need to show and learn now that Spain takes part of them.

Links

http://repositorios.fdi.ucm.es/mnemosine/

http://clavy.fdi.ucm.es/

https://www.ucm.es/loep

http://ilsa.fdi.ucm.es/

http://repositorios.fdi.ucm.es/mnemosine/

https://www.wikidata.org/wiki/Property:P10373

https://zenodo.org/communities/biblioteca_digital_mnemosine

https://github.com/Biblioteca-Digital-Mnemosine

https://www.ucm.es/otri/noticias-biblioteca-digital-mnemosine-dia-libro-ucm

http://dhawards.org/dhawards2021/results/

https://www.timemachine.eu/ltm-projects/mnemosyne-project-digital-library-for-rare-works-and-forgotten-authors-of-silver-age-spain/?token=621d7c44b24d2d962be621c5da3a5

https://test.timemachine.eu/a-laboratory-in-process-the-mnemosyne-project/

From the field to the lab and beyond: archaeological and bioanthropological applications of digital technology

https://doi.org/10.5281/zenodo.12007980

Francisca Alves Cardoso[1], **Nicholas Marquez-Grant**[2], **Carlos Moreira**[1], **Anne Malcherek**[1], **Steffi Vassallo**[1]

[1]CRIA-Center for Research in Anthropology, Portugal; [2]Cranfield University; francicard@fcsh.unl.pt, carlos.moreira@cria.org.pt, annemalcherek@campus.fcsh.unl.pt, svassallo@campus.fcsh.unl.pt

The use of digital technologies to record and reconstruct archaeological excavations is becoming increasingly popular. This is also true for reconstructing fragmented and fragile human remains, which aims to present, assess and disseminate them. The data collection and models produced during the process are also becoming research objects and creative outputs in their own right, adding to the narrative of the original excavation and digitalization efforts.

Incomplete and fragmented human remains pose a particular challenge in bioanthropological and bioarchaeological research. The more complete a skeleton is, either in the field or laboratory, the more accurate the analysis and subsequent interpretation can be. Moreover, human remains can be very fragile, requiring extreme care when handling. Traditional physical reconstruction techniques can be time-consuming, laborious, invasive, and sometimes lead to limited success in the long term, adding risk of further damage to the remains. Digital modelling offers an opportunity to address these concerns by allowing for the preservation and in-depth study of the remains safely and with a certain degree of automation. The same is true for the assessment of human remains in the field as well as in the laboratory.

By constructing a narrative and associated digitally-enabled workflows from the field to the laboratory and beyond, this panel seeks to investigate the different applications of digital technology related to archaeological contexts and the study of human remains. The panel will consist of three sessions, each framing a specific topic: the field, the laboratory and beyond - the latter here used very loosely. Collectively the sessions will explore ways of integrating digital workflows in 3D documentation, opening new avenues for scientific enquiry and expression in the arts and humanities.

During the panel, the audience will follow fragmentary human remains from being documented in the field to being reconstructed in digital models. Attendees can directly engage with 3D reconstructions on their devices and compare digital and 3D-printed models.

The panel comprises three presentations that, although autonomous, are interlinked.

The first presentation will focus on photogrammetric documentation in the field. Archaeological excavation is an inherently destructive process, and as such, meticulous documentation of in situ contexts is crucial. Hence, field excavation increasingly relies on photogrammetry and 3D models for documentation and post-excavation analysis. The benefits and challenges of using these digital tools will be explored in a case study by highlighting the importance of in situ models for subsequent interpretation of the mortuary and funerary environment. Moreover, post-deposition events that could not be achieved by decontextualized laboratory analysis of human remains will also be addressed.

The second presentation will take the workflow to the laboratory. It will demonstrate the process of capturing images/photographs of fragmented and incomplete human remains and their subsequent 3D modelling and reconstruction of the incomplete human remains into an accurate, complete digital model of a bone. This is done through photogrammetric digitization of each fragment, followed by the digital recombination of these fragments into a composite model. Validation of the accuracy of the reconstruction will also be critically discussed. Such reconstructions enable better assessments of form of singular bones, and evidence of taphonomic, trauma and pathological changes that would be impossible based solely on fragmentary remains and in situ models.

The final presentation will go beyond the field and laboratory to explore the potential of 3D modelling to valorize heritage and artistic endeavours. Entitled Digital Twin, this talk has as a starting point the artwork "Digital Twin". It combines the presence of interactive projections of real-time three-dimensional shape-shifted particles – an AI-computed vision entity – mixed with real-world artefacts (human skull). The piece refers to and positions itself on a luminous plane, building a non-linear bidirectional narrative as a product of the relationship between the audience and the interactive entity – a futuristic volumetric and metaphorical digital Artificial Intelligence (AI) inspired by the evolution of the current concept of Digital Twin.

This project reflects both the author's individual research around mixed realities and his musings around reality and mortality. It also intersects, uses, and abuses concepts such as information and databases, highly involved in the contemporary data-centred context, particularly regarding health issues. The concept's intent is to contribute to a promising debate, open to questions portraying the relationship between new technologies and their use in creating narrative strategies based on public action.

Identifying bias in cultural heritage descriptions: impact on and approaches for research

https://doi.org/10.5281/zenodo.12007758

Orfeas Menis-Mastromichalakis[1], Kerstin Arnold[2], Maria Teresa Natale[3]

[1]Thinkcode, National Technical University of Athens, Greece; [2]Archives Portal Europe Foundation, Netherlands, The; [3]Michael Culture Association, Italy; menisorfeas@gmail.com

DE-BIAS is a European project that, thanks to an AI-powered tool and specialised vocabularies developed in collaboration with marginalised communities, will help researchers detect harmful language in cultural heritage descriptions. DE-BIAS aims at equipping researchers and other users of cultural heritage online with the necessary skills to understand the various types of bias that they might find in the cultural heritage collections they work with and to consciously approach such existing bias in their own research and research output. In this, the DE-BIAS partners in collaboration with relevant communities explore three general themes: migration and colonial history; gender and sexual identity; ethnicity and ethno-religious identity.

This poster session will consist of two main strands of introductions and presentations, interwoven with the opportunity for on-the-spot show casing of the DE-BIAS tool as well as some opportunities for questions and answers with the presenting author.

Following a brief introduction of the project itself, participants will first discover and learn about types of bias identified by the DE-BIAS project and how these are reflected in the vocabulary and the tool developed by the DE-BIAS project partners: we will cover a variety of topics from generally raising awareness about the issue of bias in cultural heritage descriptions to presenting some of the project's use cases to illustrate the types of bias encountered as well as the language used in these contexts. We will also give a preview of evaluation, testing, and capacity building activities around the use of the vocabulary and the tool, which are planned throughout the next few months and allow for further interaction with the project and its outcomes.

The second strand of the session will focus on the functionalities of the DE-BIAS tool and interpreting its outputs as well as at the main concepts present in the DE-BIAS vocabulary and applying it to cultural heritage collections. Participants will be granted a look behind the scenes of the tool, when the session touches upon the topic of the opportunities and limitations of applying

natural language processing (tokenization, lemmatization, named entity recognition) for detecting bias, a highly language-, context- and time-dependent concept.

For both strands, use cases from the project will build a basis for the conversations in this session, but participants can also bring their own use cases to the table, whether these are specific to the three themes explored by DE-BIAS or touching on other aspects of diversity and inclusion in cultural heritage collections. Led by the presenting author, this session will provide opportunities to engage directly with the DE-BIAS tool and to analyse the tool's output, e.g. in how its results might be surfaced to end-users and how end-users might want to act upon the bias detected information as part of their own research.

Participants will learn that many object and collection descriptions that once fit into popular social narratives now convey outdated views that not only ignore and therefore alienate a wide range of people, but in some cases also use language that is offensive, inappropriate or even harmful. They will learn how to apply this awareness and knowledge when researching cultural heritage collections and that a more inclusive and respectful approach to telling the stories and histories of minoritised communities is possible thanks to the analyses of use cases and the adoption of the methodologies proposed by DE-BIAS.