DARIAH Annual Event 2023: Cultural Heritage Data as Humanities Research Data?

Book of Abstracts

Programme Committee

Sally Chambers (chair)
Palkó Gábor (ELTE/DH LAB)
Andrea Scharnhorst (co-chair)
Francesca Morselli (co-chair)
Laure Barbot
Alba Irollo
Kim Ferguson (BoA Editor)
Alessia Bardi
Erzsébet Tóth-Czifra
Edward Gray
Darja Fiser
Eliza Papaki
Milena Dobreva
Arnaud Gingold
Agiatis Benardou
Martin Lhoták
Ilaria Manzini
Vicky Garnett
Katrine Gasser
Sharif Islam
Marian Lefferts
Caleb Derven
Olga Holownia
Georgios Artopoulos
Tibor Kalman
Tomasz Parkoła
Adeline Joffres
Elena Gigliarelli
Tanja Wissik

Links:
DARIAH AE 2023 Website
DARIAH AE 2023 Zenodo Community

DOI: 10.5281/zenodo.8340671

Table of Contents

Conference Schedule

Session Overview

Date: Tuesday, 06/June/2023

8:30am - 9:00am	**Welcome coffee** Location: Gólyavár - Foyer A small coffee before we begin the day		
9:00am - 11:00am	**NCC - Internal** Location: Gólyavár - Small Auditorium Chair: **Nanette Rissler-Pipka**, Max Weber Foundation; nanette.rissler-pipka@gwdg.de Chair: **Martin Lhotak**, Library of the Czech Academy of Sciences; lhotak@knav.cz This is an internal meeting, invite only	**Working Group - DH Course Registry** Location: ELTE DH lab, Campus Chair: **Anna Woldrich**, Austrian Academy of Sciences; anna.woldrich@oeaw.ac.at Chair: **Iulianna Van der Lek**, CLARIN; j.vanderlek@uu.nl Meeting agenda: https://tinyurl.com/mve6p2de https://tinyurl.com/mve6p2de	
11:00am - 11:30am	**Morning coffee break** Location: Gólyavár - Foyer		
11:30am - 1:00pm	**NCC - Internal continues** Location: Gólyavár - Small Auditorium Chair: **Nanette Rissler-Pipka**, Max Weber Foundation; nanette.rissler-pipka@gwdg.de Chair: **Martin Lhotak**, Library of the Czech Academy of Sciences; lhotak@knav.cz This is an internal meeting, invite-only	**Working Group - Thesaurus Maintenance** Location: Extra meeting room (TBD) Chair: **Helen Goulis**, Academy of Athens; egouli@academyofathens.gr **Meeting agenda:** https://tinyurl.com/5b6e25ba	
1:00pm - 2:00pm	**Lunch break** Location: Gólyavár - Foyer		
2:00pm - 3:30pm	**JRC - Internal** Location: ELTE DH lab, Campus Chair: **Andrea Scharnhorst**, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Science; andrea.scharnhorst@dans.knaw.nl Chair: **Tibor Kálmán**, GWDG; tibor.kalman@gwdg.de This is an internal meeting, invite-only	**Research Infrastructure, DH Masters & Industry Workshop - Invite Only** Location: Gólyavár - Small Auditorium Chair: **Amelia Sanz**, Complutense University of Madrid; amsanz@ucm.es Chair: **Edward Joseph Gray**, CNRS; edward.gray523@gmail.com This is an internal meeting, invite-only	
3:30pm - 4:00pm	**Afternoon coffee break** Location: Gólyavár - Foyer		
4:00pm - 5:30pm	**Working Group - Bibliographical Data** Location: ELTE DH lab, Campus Chair: **Tomasz Umerle**, Instytut Badań Literackich Polskiej Akademii Nauk; tomasz.umerle@ibl.waw.pl Chair: **Vojtěch Malínek**, Institute of Czech Literature, Czech Academy of Sciences; malinek@ucl.cas.cz **Meeting agenda:** https://tinyurl.com/53yz6fh7	**Working Group - Multilingual DH** Location: Room 229 - Main Campus Building Foepulet Chair: **Aliz Horvath**, Eötvös Loránd University; aliz.horvath06@gmail.com **Meeting agenda:** https://tinyurl.com/2dmhw72h	**Research Infrastructure, DH Masters & Industry Workshop - Invite Only** Location: Gólyavár - Small Auditorium Chair: **Amelia Sanz**, Complutense University of Madrid; amsanz@ucm.es Chair: **Edward Joseph Gray**, CNRS; edward.gray523@gmail.com This is an internal meeting, invite-only
5:30pm - 7:00pm	**Working Group - Ethics and Legality in the Digital Arts and Humanities (ELDAH)** Location: ELTE DH lab, Campus Chair: **Koraljka Kuzman Šlogar**, Institute for Ethnology and Folklore Research; koraljka@ief.hr Chair: **Walter Scholger**, CLARIAH-AT; walter.scholger@uni-graz.at **Meeting agenda:** https://tinyurl.com/295euwht		

Date: Wednesday, 07/June/2023

8:30am - 9:15am	**Welcome coffee** Location: Gólyavár - Foyer A small coffee before we begin the day	
9:30am - 11:00am	**Opening Keynote Speech by Thomas Padilla** Location: Gólyavár - Main Auditorium Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be Chair: **Gábor Palkó**, Eötvös Loránd University; palko.gabor@btk.elte.hu "A Mutualistic View of AI in the Library or a Continuation of Craft" Thomas is Deputy Director, Archiving and Data Services at the Internet Archive. Throughout his career, Thomas has focused on the promotion of open knowledge and advancing responsible computational use of cultural heritage collections. He has extensive experience leading collections as data efforts (e.g., Collections as Data: Part to Whole, Always Already Computational: Collections as Data) as well as efforts that advance AI use in cultural heritage organizations (e.g., Responsible Operations). Thomas is Advisory Board Member, Roy Rosenzweig Center for History and New Media; National Advisory Board Member, Opioid Industry Documents Archive; and Advisory Board Member, The Sloane Lab: Looking back to build future shared collections. A Mutualistic View of AI in the Library or a Continuation of Craft Aspects of a long made AI promise suddenly seem possible. Discerning which aspects of that promise are actual...	
11:00am - 11:30am	**Morning coffee break** Location: Gólyavár - Foyer	
11:30am - 1:00pm	**Panel | The digital research axis at C2DH: sustainable workflows, data usability, and multi-layered publishing** Location: Gólyavár - Main Auditorium Chair: **Lorella Viola**, University of Luxembourg; lorella.viola@uni.lu	**Exploring Cultural Heritage in Research: Case Studies in Genealogy, Gaming, Language, and Historical Data** Location: Gólyavár - Small Auditorium Chair: **Agiatis Benardou**, DARIAH-EU; a.benardou@dcu.gr
1:00pm - 2:00pm	**Group photo outside, then Lunch break** Location: Gólyavár - Foyer	
2:00pm - 3:30pm	**Imagining Data Spaces** Location: Gólyavár - Main Auditorium Chair: **Edward Joseph Gray**, CNRS; edward.gray523@gmail.com	**Data quality and data management for CH in the context of open science** Location: Gólyavár - Small Auditorium Chair: **Adeline Joffres**, CNRS, UAR 3598 IR* HUMA-NUM; adeline.joffres@huma-num.fr
3:30pm - 4:00pm	**Afternoon coffee break** Location: Gólyavár - Foyer	
4:00pm - 5:00pm	**DARIAH Hungary Showcase: Road to DARIAH** Location: Gólyavár - Main Auditorium Chair: **Vicky Garnett**, DARIAH-EU; vicky.garnett@dariah.eu Chair: **Gábor Palkó**, Eötvös Loránd University; palko.gabor@btk.elte.hu	
5:00pm - 6:30pm	**Poster Session** Location: Gólyavár - Foyer	
6:00pm - 7:30pm	**Scientific Board - Internal** Location: Gólyavár - Small Auditorium Chair: **Panos Constantopoulos**, Athens University of Economics and Business & Athena Research Center; panosc@aueb.gr This is an internal meeting, invite-only	

Date: Thursday, 08/June/2023

8:30am - 9:00am	**Welcome coffee** Location: Gólyavár - Foyer A small coffee before we begin the day		
9:00am - 10:00am	**DARIAH Working Groups' projects** Location: Gólyavár - Main Auditorium Chair: **Francesca Morselli**, DARIAH-EU; francesca.morselli@dariah.eu		
10:00am - 11:00am	**DARIAH Directors' session** Location: Gólyavár - Main Auditorium Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org		
11:00am - 11:30am	**Morning coffee break** Location: Gólyavár - Foyer		
11:30am - 1:00pm	**New ways of accessing Digital Humanities data for GLAM** Location: Gólyavár - Main Auditorium Chair: **Georgios Artopoulos**, The Cyprus Institute; g.artopoulos@cyi.ac.cy	**Navigating the Digital Landscape: Data Practices and Scholarly Communication** Location: Gólyavár - Small Auditorium Chair: **Alba Irollo**, Europeana Foundation; alba.irollo@europeana.eu	
1:00pm - 2:00pm	**Lunch break** Location: Gólyavár - Foyer	**Poster Session** Location: Gólyavár - Foyer	
2:00pm - 3:30pm	**Exploring Digital Heritage: Innovations in Digitization and Data Services** Location: Gólyavár - Main Auditorium Chair: **Elena Gigliarelli**, ISPC CNR; elena.gigliarelli@cnr.it	**Cultural Heritage Data: Use Cases** Location: Gólyavár - Small Auditorium Chair: **Tanja Wissik**, Austrian Academy of Sciences; tanja.wissik@oeaw.ac.at	**Working Group - RDM (Internal)** Location: ELTE DH lab, Campus Chair: **Marta Blaszczynska**, IBL PAN; marta.blaszczynska@ibl.waw.pl Chair: **Erzsébet Tóth-Czifra**, DARIAH-EU; erzsebet.toth-czifra@dariah.eu internal meeting
3:30pm - 4:00pm	**Afternoon coffee break** Location: Gólyavár - Foyer		
4:00pm - 5:00pm	**Dariah.lab Poland - Together for Cultural Heritage** Location: Gólyavár - Main Auditorium Chair: **Vicky Garnett**, DARIAH-EU; vicky.garnett@dariah.eu 1. Introduction to Dariah.lab Poland 2. Non-invasive technologies for discovering cultural heritage 3. From Music Information Retrieval to neuroscience of sound 4. Harmonising cultural heritage: vocabularies, data quality and Linked Open Data		
5:00pm - 6:30pm	**Poster Session** Location: Gólyavár - Foyer	**Working Group Meeting Theatralia** Location: ELTE DH lab, Campus Chair: **Anamarija Žugić Borić**, Institute of Ethnology and Folklore Research; zugicboric@ief.hr Meeting agenda: https://tinyurl.com/5n86bvju	
7:00pm - 10:00pm	**Social Dinner in Budapest** Venue: Faculty of Law building, building "A" / ELTE-ÁJK "A" épület - we will be on the balcony The venue is a short walk (5-10 minutes) from the Gólyavár venue, those interested in walking to the venue together can meet at Gólyavár to leave for 18.45. Registration required. https://goo.gl/maps/dysFiJqfb4SGr1gP8		

Date: Friday, 09/June/2023

9:00am - 9:30am	**Welcome coffee** Location: Gólyavár - Foyer A small coffee before we begin the day
9:30am - 11:00am	**Fostering and improving DH data reusability** Location: Gólyavár - Main Auditorium Chair: **Arnaud Gingold**, OpenEdition-AMU; arnaud.gingold@openedition.org
11:00am - 11:30am	**Morning coffee break** Location: Gólyavár - Foyer
11:30am - 11:50am	**Open Access Book Bursary Prize** Location: Gólyavár - Main Auditorium Chair: **Erzsébet Tóth-Czifra**, DARIAH-EU; erzsebet.toth-czifra@dariah.eu Chair: **Toma Tasovac**, DARIAH-EU; ttasovac@humanistika.org
11:50am - 12:50pm	**Keynote Panel. DARIAH Data Spaces Dialogue: Imagining experimental data spaces for analysis of cultural heritage using digital methods** Location: Gólyavár - Main Auditorium Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be
12:50pm - 1:15pm	**Closing Remarks** Location: Gólyavár - Main Auditorium Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be
1:15pm - 2:00pm	**Brown Bag Lunch (take away with you)** Location: Gólyavár - Foyer

Keynotes

Opening Keynote Speech by Thomas Padilla

Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be
Chair: **Gábor Palkó**, Eötvös Loránd University; palko.gabor@btk.elte.hu

"A Mutualistic View of AI in the Library or a Continuation of Craft"

Aspects of a long made AI promise suddenly seem possible. Discerning which aspects of that promise are actually given breathless marketing and news coverage that pervades our environments can be challenging. GLAM professionals should be able to cut through that chaff – guided by a strong sense of professional history – marked by instructive successes and failures, responsibilities to our users, and discrete, well-known challenges in want of resolution. We can have a relationship with these tools akin to the mutualistic relationships that we have with any tool. We acquire tools for a certain purpose and we refine our purpose as we experience the world through them — modifying them, maintaining them, and investing in them. It's a sort of symbiotic relationship between human and tool that changes usefully over time. In many ways considering and contending with the potential of AI is just a continuation of craft. We change just as our tools change.
In this talk, a critical and ultimately generative position will be advanced that seeks to empower GLAM professionals as they embark on forming a mutualistic relationship with AI. This position will engage with individual, organizational, and community impacts of AI in the library.
Thomas is Deputy Director, Archiving and Data Services at the Internet Archive. Throughout his career, Thomas has focused on the promotion of open knowledge and advancing responsible computational use of cultural heritage collections. He has extensive experience leading collections as data efforts (e.g., Collections as Data: Part to Whole, Always Already Computational: Collections as Data) as well as efforts that advance AI use in cultural heritage organizations (e.g., Responsible Operations). Thomas is Advisory Board Member, Roy Rosenzweig Center for History and New Media; National Advisory Board Member, Opioid Industry Documents Archive; and Advisory Board Member, The Sloane Lab: Looking back to build future shared collections.

Recording: https://youtu.be/jvEMFY6b4mA

Keynote Panel: DARIAH Data Spaces Dialogue

Chair: **Sally Chambers**, DARIAH-EU; sally.chambers@ugent.be
Panelists: **Alba Irollo**, Europeana; **Steven Claeyssens**, National Library of the Netherlands, **Tomasz Parkoła**, PSNC and **Marta Błaszczyńska,** Digital Humanities Centre, Institute of Literary Research of the Polish Academy of Sciences

Imagining experimental data spaces for analysis of cultural heritage using digital methods

Within the context of the European Open Science Cloud (EOSC), the Social Sciences and Humanities Open Marketplace has already made substantial in-roads in aggregating and contextualising tools, services, training materials, datasets, publications and workflows for the SSH research communities. Within the cultural heritage sector, the development of a common European data space for cultural heritage and Collaborative Cloud for Cultural Heritage are both set to innovate the access to and sharing of cultural heritage data. In this keynote panel we will explore the practical challenges of imagining experimental data spaces for analysis of cultural heritage using digital methods, focussing on the five key themes: data; users/communities; technology; interoperability and the role of research infrastructures such as DARIAH.

Recording: https://youtu.be/o2HUKWBCTvM

Presentations

Panel session #1: The digital research axis at C2DH: sustainable workflows, data usability, and multi-layered publishing

Time: Wednesday, 07/June/2023: 11:30am - 1:00pm · *Location:* Gólyavár - Main Auditorium
Session Chair: Lorella Viola, University of Luxembourg

The digital research axis at C2DH: sustainable workflows, data usability, and multi-layered publishing

Lorella Viola, Sean Takats, Lars Wieneke, Petros Apostolopoulos, Anita Lucchesi
University of Luxembourg, Luxembourg; lorella.viola@uni.lu, sean.takats@uni.lu, lars.wieneke@uni.lu, petros.apostolopoulos@uni.lu, anita.lucchesi@uni.lu

For the past twenty years, digital tools, technologies, and infrastructures have played an increasingly determining role in framing how digital objects are understood, preserved, managed, maintained, and shared (Cameron 2021). Even in traditionally object-centred sectors such as cultural heritage, digitisation has become the norm: heritage institutions continuously digitise huge quantities of heritage material. Similarly, universities and higher education institutions resort more and more to digital infrastructures for research and teaching. As a consequence, disciplines across scientific domains have increasingly incorporated technology within their traditional workflows and developed advanced data-driven approaches to analyse ever larger and more complex data-sets (Viola 2022). At the same time, the digital transformation presents many challenges, such as sustainability, interoperability, accessibility, publishing, privacy and ethical concerns, to name but a few. In this panel, we want to focus on three major challenges that digital research in the humanities faces today: workflow sustainability, data usability, and digital research publishing.

The panel brings together the efforts of the Centre for Contemporary and Digital History (C2DH) at the University of Luxembourg to share how through its infrastructural initiatives, digital methods and tools answer urgent questions such as:

● How does a digital infrastructure relate to the tools we use, the sources we apply them to, the way in which researchers and users work in relationship with them, but also how they work in relationship with each other?

● How can we host a generalizable infrastructure that ensures privacy and portability; and it will be supported by a long-term sustainability plan and training opportunities?

● How can we allow users to build and rebuild, and to run and rerun their own workflows or those they have adopted and adapted whilst ensuring sustainability?

● How can a platform allow users to contextualize artifacts by capturing their own subject expertise?

● How can a platform enhance users' retrieval of meaningful content and maximise the potential of digital collections?

● How does an innovative publication platform promote a new form of data-driven scholarship and of transmedia storytelling in the humanities?

The panel will begin with a five-minute introduction to the C2DH and the different projects and then present three ten-minute test cases exploring the above questions. The panel's main objective is to explore the dominant strategies and methodological approaches that unveil the collaborations between arts and humanities researchers, cultural heritage professionals and computer, information and data scientists.

Throughout the three papers of this panel, we offer multi-layered approaches that, encompassing several methods, respond to the three themes of this year's DARIAH annual event. The panel will appeal to humanities scholars, cultural heritage professionals and computer scientists interested in exploring how the challenges brought by the current focus on data-driven research, data management plans and the research data lifecycle can be addressed in practice.

Please see the abstracts' description of individual papers on Zenodo

https://zenodo.org/record/7849383#.ZEY9K3ZBw2y.

Paper session #1 Exploring Cultural Heritage in Research: Case Studies in Genealogy, Gaming, Language, and Historical Data

Time: Wednesday, 07/June/2023: 11:30am - 1:00pm · *Location:* Gólyavár - Small Auditorium
Session Chair: Agiatis Benardou, DARIAH-EU

Bridging the gap between cultural heritage and research--a case study of the Chinese Genealogy Knowledge Service platform

Yaming Fu[1,2], Simon Mahony[3], Wei Liu[1]
[1]Shanghai Library/Institute of Scientific and Technical Information of Shanghai, China; [2]School of Information Management, Nanjing University, China; [3]Beijing Normal University at Zhuhai, China; ymfu@libnet.sh.cn

The Chinese Genealogy Knowledge Service Platform, one of the Shanghai Library Knowledge Bases, aggregates genealogy data and links family names to resources in other libraries in China and beyond. By extracting family names, social background, migration events from around 70,000 entries and developing an event ontology, more than 100,000 migration events spanning more than 3,000 years were identified. This project tells the story of the origins and development of families, using data which records significant migration and historical events in time and place. Using methodology from storytelling and Bakhtin (2001)'s theory of chronotope, our project places the narrative time-space continuum in the context of cultural heritage resources, using genealogy data to explore changes in family names using two main dimensions: chronological (temporal) and spatial.

Family names and related elements are rich data with narrative characteristics allowing analyses from diverse perspectives (Dunn & Schumacher, 2016). They help the designing of narrative experience for both users and researchers, allowing this data to tell the stories behind collections (Vrettakis et al., 2019).

These are crucial data to study different ethnic groups, their relationships, characteristics and lifestyles; explore the reasons behind population changes, growth and distribution; record family traditions that reflect the development of moral values; research socio-cultural theory (Rowe & Wertsch, 2002), how values and customs are passed through generations; geographical characteristics reflecting the migration of families; supplement research on family history with personal memories, and reflections from specific periods (Hershkovitz, 2016); combine with oral history to give personal reflections on family history.

Genealogy data are often passed down through many generations, not treated as official records nor professionally preserved; many are incomplete giving concerns about reliability in a research context. Our solution is for a framework evaluating reliability based on contributor, source, content, and social and historical background; assertions or factoids ("a source that says something about a person") rather than verifiable evidence (Bradley, 2016, 2021). In some contexts, such as sociology and ethnology, where breadth of sources is crucial, genealogy data serves as an important reference with unique advantages, providing perspectives from diverse groups with different social identities.

With migration and war, these materials become scattered globally across different institutions. Using linked data, we provide access to the metadata from partner libraries but do not have access to their originals or digital surrogates. This raises significant issues for the GLAM open data movement and how to provide access under FAIR principles (Beretta, 2021), despite our collections being accessible in pdf and IIIF for users to read, manage, and annotate freely. Our project has the interest of library users exploring family history, but the challenge is to facilitate collaborations with researchers, helping them use our data more effectively. The current phase develops systems to help researchers identify problems and solve specific research questions, and for the library to work closely with them supporting their research journey.

Digital Heritage Implementation and Diffusion in Commercial Digital Games

Gabriele Aroni
Manchester Metropolitan University, United Kingdom; G.Aroni@mmu.ac.uk

This paper discusses how digital heritage can be included and diffused through commercial digital games, and conversely, how digital heritage techniques can be embedded in entertainment products, beyond specifically designed educational games.

The success of the digital entertainment industry in the past decades has spurred interest in the field of cultural heritage (Cesaria et al. 2020; Bellotti et al. 2013; Connolly et al. 2012), where the benefits of digital tools for the acquisition, visualization, simulation and dissemination of heritage is a frequent subject of discussion (Ferdani et al. 2020; Bekele et al. 2018; Hua et al. 2018; Champion 2016; Valls et al. 2016). The current literature focuses on serious games, i.e. games specifically designed and developed for the cultural or educational sectors (Malegiannaki and Daradoumis 2017; Mortara et al. 2014). Yet, despite the rising diffusion of the latter, the vast majority of the public plays regular, commercially available digital games. Indeed, many of the most successful commercial games are rich in cultural heritage representations (Aroni 2019; Copplestone 2017), and a sign that cultural heritage is not limited to educational applications is evidenced by the popular series of games Assassin's Creed (Ubisoft 2007), which even added a Discovery Tour to its latest entries as an additional education tool. In fact, commercial games are used in educational settings as well (Aroni, Bregni, and McDonald 2019).

Which are thus the best practices followed by game developers in representing cultural heritage? What is the potential for digital games to preserve and diffuse digital heritage and of what kind? What benefits can the inclusion of digital heritage bring to game developers, and how can collaborations between institutions and the private sphere be developed?

This paper will present case studies of games developed in different countries (India, Japan, China, and Canada) and how they successfully - or otherwise - represent tangible and intangible cultural heritage. Moreover, it will analyse the technologies and tools that are common between game development and digital heritage, such as photogrammetry, 3D scanning, and 3D modelling, their different applications, how they can be shared between the two fields and enrich each other. Museums, in particular, can benefit from the use of digital technologies, both for the promotion of their content, and for their educational role, on-site and online (Wang and Nunes 2019).

At the same time, it is important to be aware of possible hurdles in the use of cultural heritage in entertainment products, both from a technical standpoint (Colson and Levente 2019), and how such heritage is portrayed (Dubois and Gibbs 2018). This paper argues that the governing bodies responsible for cultural heritage should establish guidelines to assist in the appropriate implementation of the cultural heritage in digital games, while promoting such use to game developers and enabling ease of access through a common database.

Cultural heritage data as sources for databases of historical language use of Hungarian

Adrienne Dömötör, Katalin Gugán, Mónika Varga
Hungarian Research Centre for Linguistics, Hungary; gugan.kati@gmail.com

The Middle Hungarian period, i.e. the interval time between the second third of the sixteenth century and the second third of the eighteenth century is less intensively explored so far. This also is the earliest period of the history of Hungarian for which an appropriate amount of extant text material is at our disposal for studying the language use of everyday private life with the necessary thoroughness (cf. Dömötör–Gugán–Varga 2021).

The present proposal focuses on two databases designed by the presenters and their team: The Old and Middle Hungarian corpus of informal language use (Történeti Magánéleti Korpusz, TMK) and The corpus of memoirs and dramas (Középmagyar emlékirat- és drámakorpusz). Both of the corpora contain texts representing important sources of the cultural heritage of Hungarian: ego-documents from noblemen and noblewomen, genres related to everyday language use involving speakers with lower social status as well, and constructed dialogs imitating everyday language use in fiction.

The Old and Middle Hungarian corpus of informal language use (tmk.nytud.hu) consists of private letters and records of witch trials from between the fifteenth-century beginnings and 1772, a total of 8 million characters. This presentation highlights some requirements and steps of the corpus building executed by historical linguists in a collaboration with a computational linguist. It includes manual normalization and disambiguation for diachronic adequacy, morphological analysis, and the setting up of a query interface. This database is the first fully normalized and annotated historical corpus of Hungarian supplemented with sociolinguistic information (Novák–Gugán–Varga–Dömötör 2018).

The other topic of the presentation is The corpus of memoirs and dramas, the building of which is in progress following the guidelines developed for the previous corpus (cf. Gugán 2020). The language use of memoires and dramas in Middle Hungarian proved to be suitable as an extension to the more directly speech-related sources of TMK. Memoires are ego-documents, yet they are still farther from informal language use than private letters. Dramas are constructed texts, however, they are speech-purposed as well. Therefore, the four registers to be included all share certain characteristics, but each differs in at least one feature.

In both corpora, all of the records are normalized and morphologically annotated. The new corpus is also planned to get a freely available user-friendly query interface, providing a valuable source of information for historical linguists and specialists or students of related fields.

References

Attila Novák, Katalin Gugán, Mónika Varga, Adrienne Dömötör: Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. Language Resources and Evaluation 52 (2018): pp. 1–28.

Dömötör Adrienne, Gugán Katalin, Varga Mónika 2021. Váltakozás, változás és nyelvtörténet: a variacionista megközelítésmódról – és a jelen kötetről [Alternation, change, and history: on the variationist approach and the present volume]. Versengő szerkezetek a középmagyar kor nyelvében [Competing Variants in Middle Hungarian]. Akadémiai Kiadó, Budapest. 5–28. https://mersz.hu/domotor-gugan-varga-versengo-szerkezetek-a-kozepmagyar-kor-nyelveben

Gugán Katalin 2020. Regiszterfüggő változatok a középmagyarban [Variation in Middle Hungarian: a register perspective]. (Kézirat [Manuscript].)

The slaughterhouse of science: from scientific leftovers to cultural heritage to historical data

Alina Volynskaya
EPFL, Switzerland; alina.volynskaya@epfl.ch

Hegel famously called history a slaughter-bench (Schlachtbank); Franco Moretti transferred this metaphor to literature, referring to the “great unread”, an archive of forgotten books to be rediscovered through distant reading . I will borrow it in order to allude to the extensive array of neglected and overlooked sources in the history of science, which could be brought back into view through datafication.

Science produces not only inventions, discoveries, data, but also a lot of leftovers: drafts, protocols of failed experiments, obsolete scientific instruments, photographs, reports and other outdated documentation of the laboratory routine. Such artifacts constitute by-products of knowledge production. Being discarded and left behind by science, they find their way into institutional archives and museum collections, where they suddenly become objects of gaze and interpretation.

Exhibiting such objects, institutional and museum repositories tend to rely on the logic of cultural heritage, describing the leftovers of science almost like objects of art, in terms of their media, size, material, etc. Yet for the history of science, such artifacts are of interest not as a patrimony, but as an object-knowledge, bearing testimony to certain facets of science history, participating in our understanding of “science in the making”. The question then is how to reintroduce them into the knowledge process. How to turn them into data readable and interpretable for the history of science?

My paper meets this challenge by proposing a new data model for the leftovers of science. I will introduce the ontology aimed at preserving and re-using such artifacts as objects of knowledge. The ontology offers three models for describing a scientific artifact: “biography” (tracing production, circulation and usage of an artifact), “assemblage” (highlighting how an object enters into relations with other object), “mediation” (focusing on the connection between the object and the phenomenon under study). Each of these models is grounded in certain theories in the STS and Material Studies and offers a distinct language for encoding the traces of outdated scientific practices.

I will exemplify how the ontology works through a series of digital exhibitions, each presenting a different model of contextualizing the thing within the digital archive. Using the example of one particular scientific device from the history of experimental psychology – the reaction key – I will show how the digital archival representations can be made into data available for distant reading.

Paper session #2: Imagining Data Spaces

Time: Wednesday, 07/June/2023: 2:00pm - 3:30pm · *Location:* Gólyavár - Main Auditorium
Session Chair: Edward Joseph Gray, CNRS

Bringing Research and Cultural Heritage together through Digital Spaces : 20 years of Open Policies at the INHA

Martine Denoyelle, Cécile Colonna, Federico Nurra
Institut national d'histoire de l'art, France; martine.denoyelle@inha.fr

The French National Institute of Art History (INHA), a research institution dedicated to art and cultural heritage established in the Digital Era, has had the production of born-digital datasets as a primary mission since its inception. Over the course of two decades, it has honed its expertise in descriptive vocabularies, harmonization, data export and reuse. Today, it is committed to promoting open access and meeting the demands of Digital Humanities. Its research programs encompass a wide range of subjects, including antiquity, contemporary art, archive studies, art inventory, and architectural heritage, and are developed in collaboration with cultural institutions such as archives, libraries, and museums.

Since 2006, INHA has developed an online system for research data management: AGORHA (https://agorha.inha.fr). Most of the INHA's research projects concern the study of heritage data from museum institutions, but not exclusively: there are databases linked to artworks, as well as prosopographical and event-oriented databases. These data are made available in interoperable formats (such as json-ld using CIDOC-CRM) under open licences (CC-BY 4.0). A new version of this web research data platform, went online in 2021, following the creation of the Digital Research Service. This service is responsible for managing INHA's documentary resources and steering all digital projects concerning the Institute's research programmes. These actions, along with many others, confirm the Institute's commitment to an open science approach.

The data generated by two research programs in the field of "History of Ancient Art and Archaeology" hosted in AGORHA, enable us to demonstrate two case studies and the methods used to construct new digital platforms in partnership with heritage institutions. The Répertoire des ventes d'antiques en France au XIXe siècle, conducted in collaboration with the Louvre Museum since 2011, has been experimenting with data visualization techniques since 2018 to present an interactive display of what the data reveals about the modern history of objects that have passed through the art market, the individuals involved, and their economic and museological impact (Sur la piste des œuvres antiques, https://ventesdantiques.inha.fr/). The second program, executed with the French National Library (BnF) from 2017 to 2022, facilitated the digital edition and exploration of the "Recueil des monuments antiques", a collection of plates that depict ancient objects preserved in private and public collections throughout Europe, which were drawn by Jean-Baptiste Muret in the 19th century. The digital outcome of this program was based on a collaboration with multiple French and international partner museums, leading to tangible museographic achievements, and was built using the Omeka S platform (Digital Muret, https://digitalmuret.inha.fr/).

From these premises and the case studies presented, we intend to demonstrate how to articulate a dialogue between research and heritage in a digital space. Using means such as datavisualization and storytelling allows to enlarge the audience of our resources in a true open science approach. Mapping data to shared and trusted ontologies enables interoperability of multiple sources and progress towards large-scale common digital spaces of Cultural Heritag

Integrating 3D Virtual Tours with Iconographical Art Digital Library in Old Orthodox Churches

Detelin Luchev[1], Desislava Paneva-Marinova[1], Maxim Goynov[1], Radoslav Pavlov[1], Lilia Pavlova[2], Zsolt László Márkus[3], Tibor Szkaliczki[3], György Szántó[3], Miklós Veres[3], Zsolt Weisz[3]
[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS),; [2]Laboratory of Telematics, Bulgarian Academy of Sciences, Bulgaria; [3]Institute for Computer Science and Control (SZTAKI), Hungary; markus.zsolt@sztaki.hu

The digital library "Virtual Encyclopedia of Bulgarian Iconography" (also known as BIDL, Bulgarian Iconographical Digital Library) was initially created to lay the foundations for the registration, documentation and virtual presentation of a potentially unlimited number of Bulgarian icons and iconographic objects. The current version of BIDL (Bulgarian Iconographical Digital Library, 2022) was finished in 2022 as an infrastructure component of CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in Favor of the Bulgarian Language and Cultural Heritage. It is a valuable part of the EU Infrastructures CLARIN and DARIAH and represents a web-based software environment that supports a variety of digital cultural units and rich functionality for interaction, with an accent on components providing storage, retrieval and intelligent curation of data and metadata.

This report aims to present a solution for integrating a digital library of iconographical art, panorama pictures and virtual walks to optimizing the exploration of one of the most richly decorated churches in the village of Arbanasi the Nativity of Christ, an extremely valuable monument of culture and art in Bulgaria.

The Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences (IMI-BAS) and the Institute for Computer Science and Control (SZTAKI) have been conducting an extensive joint research and development work for the innovative presentation of Bulgarian cultural assets (Márkus et al, 2016), museums, churches, historical buildings, etc. High-resolution spherical panorama pictures were created in Veliko Tarnovo and Arbanasi during the Bulgarian-Hungarian academic co-operation. They proved to be an attractive tool to present cultural and historical sites. The panorama pictures of Bulgarian churches were added to the Bulgarian Iconographical Digital Library (BIDL) and the related mobile application, BOOK@HAND BIDL (Márkus et al., 2015, Luchev et al., 2017).

SZTAKI implemented a proprietary tool to present panorama pictures. This SZTAKI panorama viewer offers the user the following functions: display of full 360° spherical panorama pictures, offline panoramas, hotspots, replay of background music, presenting panoramas from a list of high-resolution images, zoom, etc. The panorama pictures can be published on different platforms (Web, mobile and VR). Publishing panorama pictures on Google Maps allows to reach an especially wide audience.

As the continuation of the joint work, a virtual walk was created from the panorama pictures taken outside and inside of the Church of Nativity of Christ in Arbanasi where the user can explore the church by moving from one section of the church to another.

Virtual walks represent a service to present real spaces and special environments in an attractive way and support virtual navigation and interactive discovery of various locations. The virtual walk of the Church of Nativity was integrated with BIDL through hotspots. Hotspots were assigned to specific parts of the church where users can get additional information during the virtual walk. The information is available in multiple languages (Bulgarian and English). By clicking on a hotspot, the image and the description of the icon in BIDL are presented.

Storytelling with Linked open Data

Bart Boskaljon[1,2], Ruben Schalk[1], Joop Vanderheiden[1]

[1]Cultural Heritage Agency of the Netherlands; [2]Dutch Digital Heritage Network; b.boskaljon@cultureelerfgoed.nl

For decades we've seen heritage institutions as places to collect and preserve, display and research objects. And even though this is still a mayor task of GLAMs, an alternative approach in heritage studies and practices is more common these days. The relevance of the institutions is not in the objects they preserve, it is in the stories they have to tell. Without a narrative, there is no value in the object. We believe the same goes for data. Without the narrative, the data has no value.

Due to the growing availability of heritage datasets, we have more and more opportunities of telling a story that holds, is interesting and where the data can be of value.

We've been experimenting telling stories using Linked open Data with simple visualization tools. It learned us, with a focus on transparency and re-use of sources available by multiple institutions, the key components of a datastory.

One of the cases that we learned from is the datastory that is created by R. Schalk about the Spanish flu in the Netherlands , a research project that was followed by a publication. This experience tells us a lot about data quality, availability and processing data so that it can be used to tell a story.

Another case that is used, is a datastory for understanding the process of designating monuments in the Netherlands. Creating this story gave us many new insights about the symbiosis between data expert and the topic expert, the time you need to understand the data and the value it can deliver and the skill that is needed to tell a good story.

We share our experiences on several topics, including:

Symbiosis of the data expert and the topic specific expert

It doesn't happen often that we find both skills in the same person. Therefor we need combine knowledge. Creating a common ground is essential. On forehand the expert usually doesn't know what digital tools are capable of. On the other hand, the data expert usually doesn't know the details about the data.

Take your time

Creating a good story and finding the right data is a process that takes several iterations. Don't rush the process.

Data sustainability

Linking sources from outside your institute requires a certainty of sustainable access. Think about the sustainability of the story you tell.

Data is never perfect

The quality of the data that are available has a wide range of quality. We expect that processing of data are always necessary.

FAIR

Publishing in the digital realm makes it possible to publish include all available context and therefor you can be as transparent as possible to your public. Make sure you invite others to fact check your story and your data.

Telling a story is a skill

Not everyone can come up with a good story. This is a skill. Storytelling with data is a more difficult skill. Make sure that the implicit knowledge about the data becomes part of the story you tell.

From exiled cultural heritage to cultural resilience : bringing data home

Murielle Sandra Tiako Djomatchoua

Princeton University, United States of America; mt2200@princeton.edu

From exiled cultural heritage to cultural resilience : bringing data home

The colonial order, in its ideologies and its discourses, had recourse to techniques and strategies that were geared towards achieving cultural amnesia within territories and communities under their control. The destruction of cultural heritage and identities was a prerequisite for the total domination and exploitation of the people and the resources. One of the practices that preceded and outlived colonialism was the massive looting of artistic and cultural artifacts. Through commissioned and punitive expeditions, colonial empires fueled an active thirst for the exotic and a leading lucrative and prosperous arts dealing and public and private arts consumption.

The domain of public and private arts consumption was pioneered by museums (exhibitions and collections), galleries, and the art market (collectors). The movements of artifacts translocated African shrines to these new spaces.

The translocated shrines and sanctuaries of African arts in museums created in many communities a tradition of cultural amnesia that accounted for the dislocation of their worldviews and belief systems. As a form of pseudo-compensation for other imported forms of religious practices were imposed in the name of civilization. Some communities forsake their traditional magico-religious practices, connected to looted objects now abroad, in favor to Western religions (Christianism, Protestantism, Evangelism, Judaism, etc.) or Islam. However, other communities, like the Tibati and the Nso chiefdoms in Cameroon, held on to their intrinsic cultural identities. This connection to their cultural identities in the absence of the heritage objects and under the constant threat of politico-economic violence and of (post)colonialism crisis is what I consider to be cultural resilience. This cultural resilience generated mechanisms of cultural and collective survivals. The Nso and Tibati, are peculiar communities because they are set

at two different stages of the process of restitution and that of the re-insertion of absent/looted/stolen material heritage within their respective tradition circles.

The intensification of provenance studies and restitution debates at German cultural institutions produce data that mechanically inform politics and advocacies which does not appropriately serve the interest of the populations. Thinking about creative experimental data spaces for African cultural heritage as a meeting point between institutional resources and community needs is the key aspect of my argument. As political bureaucracies and individual interests slow down the process of restitution, alternative digital spaces need to be designed in order to re-connnect African societies/communities to their cultural heritage abroad.

A prototype model for alternative digital spaces will be built upon existing models in order to discuss the concept of virtual repatriation as a contested panacea to cultural amnesia and as a problematic solution reproducing imbalanced power dynamics in the field of cultural heritage that answer the question of co-creation of knowledge about material cultures at Western cultural institutions without serving the interests of community of origin to whom is imposed the constraints and dictates of the virtual age. Can the traditional and the digital cohabitate in reconstructing lost/hidden/forgotten cultural heritage is the main question of this presentation.

Paper session #3: Data quality and data management for CH in the context of open science

Time: Wednesday, 07/June/2023: 2:00pm - 3:30pm · *Location:* Gólyavár - Small Auditorium
Session Chair: Adeline Joffres, CNRS, UAR 3598 IR* HUMA-NUM

A new metadata schema about "Architectural Heritage in the Built Environment"

Marissia Deligiorgi, Valentina Vassallo, Anastasia Tsagka, Georgios Artopoulos
The Cyprus Institute, Cyprus; g.artopoulos@cyi.ac.cy

Historic urban environments are not given static formations disconnected from the contemporary fabric of a city, but rather a set of tangible and intangible assets subjected to dynamic pressures of economic, environmental, and social activities. The sustainable development of these environments is often threatened by urbanization, neglect and climate. The cross-disciplinary nature of the pressing challenges posed by these phenomena (Historic England 2020) requires the development of novel data-driven tools for agile safeguarding of our historic building stock. The adoption of holistic, integrated, multi-disciplinary methods can bridge technological innovation with the conservation and restoration of heritage buildings. The next step in the development of built heritage digitisation methods should focus in expanding the scope of study beyond the single built heritage structure and allowing deeper understanding and interdisciplinary interpretation of its condition and performance within its topographical context and the surrounding built environment. This could become a reality today by means of the advancements in digital tools, remote sensing, algorithms and computation prowess of hardware available to researchers (Mohamed et al. 2020). Data interoperability and re-use, in the context of the FAIR principles, can only be achieved through the semantic description, using metadata and ontologies (Messaoudi et al. 2018).

Arguably, the penetration of Building Information Modelling software in the building industry is enabled by semantic tools, such as the Industry Foundation Classes (IFC) data model (2022) or the Green Building XML schema (gbXML, 2023). Acknowledging this, the authors will present:

- Challenges and practical aspects of the creation of a new metadata schema that would enable linking the relevant to cultural heritage of our historic cities' multi-modal and multi-discipline datasets, e.g., 3D models, IFC classes, historical descriptions, and environmental data.

- How this multiscale data-driven study can facilitate a more holistic, integrated application of digital methods, such as BIM, modelling & simulation to cities, and better contextualise cross-disciplinary enquiries (Cursi et al. 2022).

The paper will assess a series of existing metadata schemas such as bibliographic documentation or geoinformatics, that are required to be added to the BIM model in order to meet the specifications of a CIDOC-CRM based on ISO standards ISO 21127:2014 (Acierno et al. 2017). The result of this action is the description of a new metadata schema, CIDOC-CRM compatible, as an extension of the CARARE 2.0 metadata schema (Fernie et al. 2013) for built heritage and monuments (English Heritage 2012).

The authors expect that the results of this activity will contribute to the long-standing discourse in the field (Ronzino et al. 2013; Ronzino, Niccolucci and D'Andrea 2013), to support in the future the big data management of historic cities. In this effort, a further step, currently under completion, is the alignment of the proposed metadata schema to the CIDOC CRM ontology and its extensions (Fig. 2). The final results of this activity will be published on Zenodo and also disseminated to the relevant digital humanities communities through the DARIAH ERIC WG on Digital Practices for the Study of Urban Heritage.

How to Create Knowledge in Cultural Heritage Documentation: The Importance of High Quality Paradata and Metadata

Maria Hadjiathanasiou, Elena Karittevli, Elina Argyridou, Iliana Koulafeti, Panayiota Samara, Ioannis Panagi, Kyriakos Efstathiou, Marinos Ioannides
UNESCO & ERA Chairs on Digital Cultural Heritage, Cyprus University of Technology, Cyprus; m.hadjiathanasiou@cut.ac.cy

The abstract demonstrates the importance of high quality paradata and metadata in 3D digitisation by illustrating the unique results of the recently published EU "Study on quality in 3D digitisation of tangible Cultural Heritage (CH): mapping parameters, formats, standards, benchmarks, methodologies, and guidelines - VIGIE 2020/654" (2022), commissioned by the European Commission (EC) to help advance 3D digitisation across Europe and thereby to support the objectives of the Recommendation on a common European data space for CH. The aim of the Study's mapping was to further the quality of relevant projects by enabling CH professionals to define and produce high-quality digitisation standards for tangible CH. The Study also estimated the relative complexity and how it is linked to technology, its impact on quality and its various factors, also identifying 3D digitisation standards and formats. The potential impacts of future technological advances on 3D digitisation were also forecasted. This Study demonstrates that complexity and quality are fundamental considerations in determining the necessary effort for a 3D digitisation project to achieve the required value of the output. The complexity of 3D data acquisition projects can be determined after assessing factors such as: stakeholder requirements, project specifications, personnel qualifications, object type and location, environmental conditions, equipment, real object conditions, and pre-processing software. On the other hand, determination of quality may comprise the degree of detail, precision, and resolution of the geometric accuracy of the 3D shape and the fidelity of capturing colour/texture. Seventeen heritage documentation case-studies were chosen to be studied and digitized under the EU-funded "Mnemosyne" project, of which three (Frescoes of the Saint Euphemianos (Lysi); Zoomorphic clay vases; The Church of the Virgin Karmiotissa), are going to illustrate the results of the Study and how this can fulfill the needs of the multidisciplinary community of experts. Progress in 3D digitisation has significantly improved the accessibility of the European CH for research, innovation, education and enjoyment. At the same time, unresolved issues remain, concerning aspects that may refer to the digital twin, short and long-term preservation, use/re-use, sustainability, return on investment and long-term cost. Such aspects relate to broader questions on the topics of accuracy, complexity and quality. If the aim is to achieve high-

quality results during the 2D and 3D recording process of CH tangible assets, what are the “standards” needed? For example, how much are they going to cost, how long will they take, and will they meet multidisciplinary needs? Which formats should be used to record the results, thus enabling long-term preservation? What kind of knowledge can/should be embedded in 3D records and how can models be shared interoperably? These are some questions of crucial importance to be explored via Mnemosyne’s case-studies which are going to serve as key-pilot cases presented for the validation of the Study’s results. In light of Europe’s need for cooperative, transnational, multilingual CH digitisation policies, the Study enables effective collaboration between Member States by democratising the benefits that could be reaped from high quality digitisation, thus creating useful and valuable knowledge in CH documentation.

Libraries as Data Infrastructures: Towards a CENL Dialogue Forum

Sally Chambers[1,2,3], Peter Leinen[4], Andreas Witt[5], Martin Wynne[8], Martijn Kleppe[6], Frédéric Lemmers[2], Hélène Bergès[7], Marie Carlin[7]

[1]DARIAH-EU; [2]KBR, Royal Library of Belgium; [3]Ghent Centre for Digital Humanities, Ghent University, Belgium; [4]German National Library; [5]Universität Mannheim; [6]National Library of the Netherlands; [7]Bibliothèque nationale de France; [8]University of Oxford; sally.chambers@ugent.be

National Libraries have not only been pioneers in the development of data infrastructures, they play an essential role in facilitating research in the arts and humanities. Likewise, the continual growth of digital (digitised and born-digital) cultural heritage is crucial for arts and humanities researchers, especially for analysis and interpretation using digital methods (Tasovac et al, 2020). The digital data infrastructure landscape is currently in considerable flux, both nationally and internationally (ESFRI, 2021). Existing Research Infrastructures, such as DARIAH and CLARIN, are joining forces to contribute to the European Open Science Cloud (EOSC). In the cultural heritage space, emerging initiatives such as the common European Data Space for Cultural Heritage and the European Collaborative Cloud for Cultural Heritage are set to disrupt this landscape further, providing both challenges, as well as unprecedented opportunities for both (national) libraries and research infrastructures alike. It is within this evolving context that the idea of a CENL Dialogue Forum on Libraries as Data Infrastructures was born.

CENL, the Conference of European National Librarians, brings together the National Libraries of Europe. It is a network of 46 national libraries in 45 European countries in the Council of Europe. Founded in 1987, the mission of CENL is to advance the cause of Europe’s national libraries through collaboration to preserve the continent’s cultural heritage and making it accessible to all, with a specific focus on skills and knowledge exchange. Collaboration between libraries and research Infrastructures such as DARIAH and CLARIN is not new. There is an active CLARIN and Libraries community, which holds workshops and contributes topics for debate. DARIAH has been exploring the inter-relationship between digital collections and digital scholarship together with library organisations such as LIBER and IFLA, as well as being an active participant in the International GLAM Labs Community.

To facilitate structural and strategic collaboration between Europe’s National Libraries and Research Infrastructures, the idea of a CENL Dialogue Forum was born. It provides an ideal opportunity to assess the landscape; identify and prioritise specific challenges and opportunities, and understand how (national) libraries could benefit from structural collaboration with, and active participation in Research Infrastructures such as DARIAH and CLARIN. A key issue for debate is the international accessibility of FAIR (Findable, Accessible, Interoperable and Reusable) datasets and related challenges in implementation. With the increasing emergence of ‘data labs’ throughout the library community, such labs could be an ideal point of intersection between the libraries, research infrastructures and digital humanities research communities. Not only could Dialogue Forum be the voice of libraries in this data space, at the same time, it would raise awareness of this crucial topic throughout the (national) library community.

Following a series of preparatory meetings, a short paper at the DARIAH Annual Event 2023 on “Cultural Heritage Data as Humanities Research Data?” will be an ideal opportunity to kick-start this debate at the intersection between the library, digital humanities and research infrastructure communities.

Rethinking access to aggregated cultural heritage data. User-centered restructuring at Deutsche Digitale Bibliothek

Dr. Martin Breuer, Dr. Lena Hennewig

Stiftung Preußischer Kulturbesitz (Prussian Cultural Heritage Foundation), Germany; l.hennewig@hv.spk-berlin.de

Platforms such as the Deutsche Digitale Bibliothek (henceforth: DDB) aggregate large-scale and cross-disciplinary digitized cultural heritage data. Conceived as a one-size-fits-all solution for the general public and specific target groups alike, cultural heritage platforms (CHPs) face a variety of fundamental problems to effectively structure, present, and offer access to the aggregated data. One central aspect is usability: Who uses the platforms for what purposes? This raises the question of how CHPs may actively design their data, search environments, and offered activities to become more user-friendly for a vastly heterogeneous user pool. The DDB has addressed these questions in the large-scale project "User-oriented restructuring of the DDB" (2020-2023), whose outcomes we critically analyse in our contribution. The project has endowed us with a unique data set of surveys and customer interactions that we exploit to derive key insights for user-centric design of CHPs.

Our paper highlights the usability of our data for academic research. We inquire into how the DDB, and CHPs more broadly, may become more relevant and sustainable partners for (digital) humanities researchers. In this, our paper focuses on four key points. First, we will give a brief introduction to the structure of the DDB and the data made available on it. Second, we report on the aims, activities, and achievements of DDB’s user-oriented restructuring. The project’s principal objective is to simultaneously improve the usability of both the web interface and the aggregated data for different user groups (researchers, cultural heritage professionals, teachers, general public). This implies re-designing the website and its search functionalities, creating specific content and theme-focused information for particular user groups, and fostering collaborations with partner institutions from education and research.

Third, we present our user-facing activities to enhance the usability of the DDB and its data for academic researchers. Here, we address and discuss the prospects and limits which we consider crucial for improving aggregating platforms like the DDB for the field of (digital) humanities research: the creation of specialized sub-portals, the gathering of interoperable, cross-sector datasets, and the development of customized outreach activities.

Fourth, we take a step back to ask which role CHPs can play in providing and adapting cultural heritage data for research in the humanities. In the fast-growing and changing field of digital humanities, cultural heritage data has to be FAIR. Here, we argue that CHPs can serve as a critical and sustainable interface between research projects and cultural institutions worldwide, opening new ways for inclusive collaboration.

References:

Baum, C.; Stäcker, T. (2015): “Methoden – Theorien – Projekte.“ In C. Baum, T. Stäcker (eds.) Grenzen und Möglichkeiten der Digital Humanities. Wolfenbüttel: Forschungsverbund MWW.

Berget, G. et al. (eds.) (2021): Linking Theory and Practice of Digital Libraries - 25th International Conference on Theory and Practice of Digital Libraries. Berlin: Springer.

Berry, D. M.; Fagerjord, A. (2017): Digital Humanities: Knowledge and Critique in a Digital Age. London: Polity.

Kremers, H. (2020): Digital Cultural Heritage. Berlin: Springer.

Owens, T. (2013): Digital Cultural Heritage and the Crowd, in: Creator. The Museum Journal, 56,1.

Paper session #4: New ways of accessing Digital Humanities data for GLAM

Time: Thursday, 08/June/2023: 11:30am - 1:00pm · *Location:* Gólyavár - Main Auditorium
Session Chair: Georgios Artopoulos, The Cyprus Institute

Unlocking the Network of Digital Cultural Heritage

Peter Verhaar, Karin de Wild
Leiden University, Netherlands; p.a.f.verhaar@hum.leidenuniv.nl

Libraries, museums and archival institutions increasingly make their collections available as data, and this makes it possible for students and researchers to engage with cultural heritage in new and exciting ways. The project “Connecting to the Network of Cultural Heritage” was carried out at Leiden University to encourage students and researchers to experiment with cultural heritage data. One of the project’s central results was a series of tutorials explaining the use and the creation of Linked Open Data. The tutorials also included explanations of some of the tools that can be used while working with LOD, such as the CLARIAH Data Legend tools GRLC and CoW.

Following examples set by the GLAM Workbenches initiative and the collections as data programme, the project also focused on the development of a series of interactive Jupyter notebooks exemplifying the research questions that can be addressed using computational analyses of digital heritage collections. These notebooks, which contain code snippets, data visualisations and explanations of research questions and quantitative methods, were all developed to enable teachers and researchers to develop a better understanding of the methods associated with data science, and to stimulate them to think about relevant applications of such methods within humanities research.

One of the interactive notebooks that was developed concentrated on the medieval books that have been added to Europeana as part of the project The Art of Reading in the Middle Ages, by Leiden University Libraries and the Bibliothèque Nationale de France, among other institutions. Using the data that were collected, interactive maps and timelines were developed to clarify the places and the years of creation of all the books in the collection (see images 2). Additionally, a network visualisation was created to examine the manifold connections between authors and manuscripts. This network graph can be used to examine the dominance and the authority of particular authors during the Middle Ages (see image 1). Similar types of notebooks were developed to explore the scholarly potential of the data made available by the Rijksmuseum, the Netherlands Institute for Art History and the Short title catalogue of the Netherlands.

During the last two years, the tutorials and the notebooks that were created in the project have been used in MA courses and at coding literacy workshops at Leiden Univerity’s Humanities faculty. Students and researchers who have worked with the learning materials often stressed that the concrete examples that were given in the notebooks genuinely helped to deepen and to enliven the learning process. The code that was provided could easily be tweaked to address new research questions. Interestingly, the results emerging from the research that was performed during such interactive training sessions often provoked valuable discussions about the limitations of computational research, and about the occasional inaccuracy and subjectivity of the data provided by heritage institutions.

In the coming years, we plan to develop our collection of interactive Jupiter notebooks further, based on the experiences of students and researchers, and in close collaboration with national and international partners.

Digital Humanities and Industry: identifying employment niches. A first overview on challenges and potential solutions

Amelia Sanz[1], Vicky Garnett[2], Tom Gheldof[3], Edward Gray[4], Adeline Joffres[4], Iulianna van der Lek[5]
[1]Complutense University of Madrid (Spain); [2]Trinity College, Dublin (Ireland); [3]KU-Leuven (Belgium); [4]Huma-num- CNRS (France); [5]Utrecht University (The Netherlands); amsanz@ucm.es

Postgraduate education in Digital Humanities (DH) has often led to careers for students in either the research or cultural heritage sector. Traditionally, the relationship between industry and Cultural Heritage institutions has typically been conceived as a collaboration to leverage funding mechanisms and develop projects to pursue a common interest, such as a technical innovation, or a knowledge sharing endeavour. The skills acquired within Digital Humanities (DH) taught postgraduate degrees are interdisciplinary and therefore transferable by their very nature, something that has been recognised among larger multinational companies. Indeed, a strong humanities education and familiarity with our methods can be an asset for business. Best practices for data stewardship and data management are similar whether one focuses on cultural heritage data, or business data, even if there are particularities. Yet among small and medium enterprises (SMEs) the proposition of employing a graduate from a field that is still in its relative infancy compared with more traditional disciplines can be seen as a risk. It therefore becomes necessary to identify the gaps, and indeed niches that rest between the current provision of training among DH scholars at a postgraduate (Masters) level, and the needs of the companies and future employers of DH graduates. Indeed, greater collaboration and fluidity between the cultural heritage and academic sphere, and that of business, via the DH alumni, can lead to greater outcomes for both, as these students can bring the best practices of both sectors in their future careers, thereby enriching both sectors and establishing interpersonal links (and the collaboration that grows from these links) via their networks. In light of this, it becomes necessary to foster internships that encourage and nurture experimental data spaces between cultural heritage, industry and academia

This paper will therefore share the conversation around the relationship between taught postgraduate DH programmes and industry by presenting the outcomes of a joint working-group workshop to be held on the periphery of the DARIAH Annual Event 2023. Furthermore, it will also include the results of preparatory surveys and interviews with directors and coordinators of various DH postgraduate programmes across Europe, specifically identifying the challenges and professional issues experienced by both DH Masters directors, and their alumni. This paper addresses the following key objectives:

- Identify the professional challenges and (new) employment opportunities of DH postgraduate taught programmes and their alumni at the European scale.

- Identifythe benefits such a collaboration and exchange between the two sectors can bring

- Identify opportunities and good practices of internships with industry and cultural heritage institutions, and their associated challenges.

- Strengthen the networking opportunities between master degrees, in such a way that expertise can be mapped at a pan-Infrastructural level to share and exchange trainers and trainees in the frame of Erasmus mobilities or Erasmus Mundus programmes.

Our presentation will give visibility to these outputs, as a first step in a long-term effort to improve collaboration between industry, cultural heritage institutions and academia (specifically taught postgraduate DH degrees) in the frame of research infrastructures.

European Literary Bibliography (literarybibliography.eu) – a Prototype for a New Dimension of a Cultural Heritage Data Space

Tomasz Umerle[1], Vojtěch Malínek[2], Róbert Péter[3]

[1]Institute of Literary Research, Polish Academy of Sciences, Poland; [2]Institute of Czech Literature, Czech Academy of Sciences, Czech Republic; [3]University of Széged, Hungary; malinek@ucl.cas.cz

European Literary Bibliography (literarybibliography.eu) is a joint initiative of Czech and Polish Literary Bibliographies which aims to create an international subject bibliography for European literatures by collecting, harmonising and providing a single interface for multilingual metadata. This project takes advantage of the automated extraction, normalisation and enrichment of open metadata (thus relying on FAIR and Open Science principles) and fills the gaps in the metadata landscape by harmonising cultural heritage (librarian metadata collections) and research metadata (scientific output metadata). In this paper we are presenting ELB as a prototype for the experimental new dimension of a cultural heritage data space – an initiative that is based on the reuse of key European open metadata collections.

Establishment of ELB is motivated by four factors. (1) Growing accessibility of open metadata especially from national libraries. Automated data extraction is facilitated by the librarian linked data services providing records under open licences and enriched in accordance with the Semantic Web principles and with relevant persistent identifiers, controlled vocabularies and ontologies (Siwecka 2017, Cagnazzo 2017, Gaitanou 2022). (2) Unsatisfactory alternative sources. Despite these efforts, there are no open metadata aggregators of satisfying scope and accessibility. The main existing alternative (WorldCat) does not allow unlimited access to the data under a free licence and is collecting mainly the metadata on the books. (3) Research community needs. On the part of researchers there is a growing interest in data collections as not only research data, but as "apparatus" (Bode 2020b). In recent years there has been a lively (Underwood 2019, Piper 2018, Bode 2020) discussion between leading scholars in the field of data-based research on the significance of the data infrastructure for modelling of research and assessing the representativeness of data. (4) European topics need international infrastructures. To develop research projects and create knowledge that relates to Europe as a whole there is a need for truly international data resources.

All metadata aggregators face data harmonisation challenges, of course, and these need to be taken into account while setting up solutions similar to ELB or scaling up this solution. Those are related mainly to the persistent identifiers (PID) and controlled vocabularies attribution. Firstly, it is necessary to choose proper PID solutions for the harmonised data (i.e. VIAF, ORCID). Secondly, realistic solutions for mappings of original datasets to the general classification have to be implemented. Thirdly, the constant harmonisation workflows have to be developed for differences in classifications (e.g. conflicting subject headings) as well as for the values missing in the central PID system.

Based on the experiences of the ELB and the assessment of the current metadata ecosystem for the humanities we argue that the digital humanities community – and DARIAH especially – should pay a specific attention to the issues of open data documentation and re-use. Projects that follow a general direction of ELB – automated aggregation and harmonisation of open metadata, especially originating from libraries – could create a new, key dimension of the cultural heritage data space with great value for the humanities.

Multisensory Representations and Immersive Experiences for Inclusive Cultural Heritage: The Case of MuseIT

Eleni Matinopoulou[1], Georgia Georgiou[1], Maria Kyrou[1], Panagiotis Petrantonakis[1], Eleftherios Anastasovitis[1], Spiros Nikolopoulos[1], Ioannis Kompatsiaris[1], Nasrine Olson[2]

[1]Centre for Research and Technology Hellas, Greece; [2]University of Borås; matinopoulou@iti.gr

The museum and cultural heritage fields, being the principal cultural and educational memory repositories of humanity, have exploited the digital transformation to embrace a more holistic approach to offer more emotional and personalised experiences [1]. The notion of digital transformation appears to take a shift from technology-centric concerns to one focused on a human resource approach. Creating a transparent, open, and inclusive digital heritage data landscape is more vital than ever. The MuseIT project aspires to co-design, develop, and co-evaluate a user-centred inclusion platform through multisensory representations of cultural heritage for enhancing engagement and providing equal opportunity for all as core principles. Centered mainly on democratisation and social inclusion, MuseIT enables digital transition to move beyond creating a knowledge repository capable of preserving the multisensory representations created within the project for interoperability, and interfacing with external systems. The multidisciplinary and transnational collaboration between humanities researchers and Cultural Heritage institutions brings valuable and complementary expertise to the project. This has allowed for a more comprehensive approach to the work as well as increased sharing and reuse of cultural heritage resources.

Immersive technologies such as virtual reality are capable of augmenting the cultural experience, including interaction, participation, and personalization, while ensuring accessibility for all [2]. Additionally, the embedding of gamification in the virtual experience triggers the user's curiosity, transforming the users from passive viewers into active participants. In this way, the foundation of the development of the immersive cultural heritage experiences in MuseIT has been based on data collected from various European cultural heritage organisations. Data of different material, date, context and culture of origin will harmonically co-exist to narrate their stories, and stimulate feelings through the multisensory interpretation approaches. Hence, an inclusive multisensory and multimodal metadata experience can be provided to a wider audience through the provision of data and the creation of a virtual environment. Additionally, in MuseIT, a variety of scientific domains of researchers, cultural heritage professionals, and computer, information, and data scientists have successfully collaborated to create a sustainable workflow for metamorphosing reusable, understandable, accessible, and sustainable data.

The proliferation of Machine Learning technologies in the Affective Computing area has raised new opportunities in the evaluation of cultural heritage experiences. The multisensory representations of cultural artefacts in an immersive virtual reality environment provide a potent and highly influential stimulus to platform users, which drives their physiological responses. In addition, the digital methods for transformation enable humanities researchers along with data scientists to explore the nature of a virtual reality experience of cultural aspects, imitating real-life scenarios. Building on the inclusion principle, an Affective Computing framework is developed in pursuance of tracking and accessing the emotional state and engagement status of individuals who interact with the cultural heritage resources. Physiological signals, such as electroencephalography, galvanic skin response and heart rate, are employed to train and test Machine Learning algorithms, providing objective evaluation of the overall enhanced experience. This approach also proffers new means for inclusion, accessibility and personalization for all.

Paper session #5: Navigating the Digital Landscape: Data Practices and Scholarly Communication

Time: Thursday, 08/June/2023: 11:30am - 1:00pm · *Location:* Gólyavár - Small Auditorium
Session Chair: Alba Irollo, Europeana Foundation

Cultural Heritage meets Arts and Humanities Research Data – Voices from the Community

Erzsébet Tóth-Czifra[1], Marta Blaszczynska[2], Rita Gautschy[3], Francesco Gelati[4]
[1]DARIAH-EU; [2]IBL PAN; [3]DaSCH - Swiss National Data and Service Center for the Humanities; [4]Universität Hamburg;
marta.blaszczynska@ibl.waw.pl

A frequently voiced concern about the Open Science paradigm is that it has not had the same impact on the different disciplines. Research areas and researchers who are heavily dependent on third-party data are still facing complex legal, ethical, technical, infrastructural and interoperability challenges and their needs form a rather underrepresented blind spot in the open research culture. A defining Open Science challenge in research data workflows in the Arts and Humanities domain (as well as in many other disciplinary areas, such as Heritage Science, Natural History, or Palaeontology) is their dependence on Cultural Heritage sources hosted and curated in museums, libraries, galleries and archives. The availability of digital (digitised and born-digital) Cultural Heritage is fundamental to research in many disciplines because without it, undertaking humanities research with digital methods would be impossible (Tasovac et al. 2020).

Still, Cultural Heritage collections are usually not made available digitally with research/academic reuse in mind. A major difficulty when scholars interact with heritage data is that Cultural Heritage institutions, universities and other research-performing organizations are embedded into very different legal, funding, structural and organizational frameworks. From a data curation perspective, the different layers of analysis (from the ways in which Cultural Heritage resources are made digitally available, the ways in which these digitized resources are corrected and enriched for computational analysis, the ways in which they are used to answer research questions) form a natural continuum of knowledge creation.

Bearing this fundamental challenge in mind, one of the core aims of the DARIAH Research Data Management Working Group is to connect and mitigate these silos by bringing together data support professionals (data stewards, subject librarians, open science officers, etc.) from both the Arts and Humanities research/academic domain and the Cultural Heritage domain. During the presentation, we aim to bring highlights of the Working Group's first longer publication, "Research Data Management for Arts and Humanities: Integrating Voices of the Community" to the DARIAH community. It brings together case studies and consolidated experience from the members of the group about:

- How certain Arts and Humanities and Cultural Heritage institutions, workplaces of the authors, developed capacities for data support;

- Ways in which Cultural Heritage professionals can be efficiently involved in open and sustainable Arts and Humanities data workflows;

- How to facilitate the reuse, dissemination and solidification of researcher-friendly, FAIR-by-design curation practices of Cultural Heritage data as research data, including also sensitive data;

- How multilingualism can be supported throughout this work;

- How to solve the problem of long-term digital preservation.

The Open Access publication (to be published in 2023) has been made possible by the third Working Groups (WG) Funding Scheme Call for the years 2021-2023 and the resulting grant administered by the Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN).

Digital Scholarly Editions: from the desktop to the semantic web. A workflow to follow.

Zsófia Fellegi[1], Kata Dobás[1], Gábor Palkó[2]
[1]Research Centre for the Humanities, Institute for Literary Studies, Hungary; [2]Eötvös Loránd University of Budapest, Department for Digital Humaities; fellegi.zsofia@abtk.hu

In recent years, the phenomenon of the semantic web and linked open data has become a hot topic in digital philology. The volume Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing, based on a workshop held in 2019, is entirely dedicated to this particular issue.

We have seen several examples of modelling the data of scholarly editions using semantic web technology or graph data models. There are also examples of abandonment of the TEI XML format, which was designed specifically for the needs of digital scholarly editing.

Our talk will attempt to present an architecture that avoids the shortcomings of some recent experiments. The development of custom ontologies and data structures, as well as custom software for writing and reading them, is seen as isolating digital philology from the possibilities of open scholarship, and increasing rather than solving the closedness of each edition.

We start from the premise that abandoning widely accepted recommendations and standards works against the practical implementation of linked open data. Our infrastructure is therefore based on TEI XML, which can be described as the "lingua franca" of digital philology. In order to make the published texts accessible to practitioners in other disciplines (historians, linguists), not only for close reading but also for computer-assisted remote reading, we are developing codes that can easily produce other data formats from XML (e.g., plain text, CSV, LaTeX).

Although TEI XML is itself semantic, a semantic document description language, it does not in itself cross the boundary between a network of documents and a network of data as defined by Berners-Lee. Previous attempts to link TEI XML to the semantic web have not been successful, and have not contributed to the use of philological data outside the narrow framework of its creation, at a global cultural and scientific level.

The architecture is based on the interconnection of services and software such as WikiData, GitHub and Zenodo.org, as well as Wikibase and Invenio RDM. These are perhaps the most important depositories of the open data philosophy, but their interconnection is far from self-evident. In this talk we will describe the whole workflow from the first steps of editing scholarly texts to publishing the texts and their semantically data-rich linked data. We will describe the relational network of metadata describing larger text units (the work as a digital object) and the practice of semantic annotation and linking of smaller text units.

The architecture is illustrated through the concrete publishing practice of the DigiPhil project.

The Archiving Reproductive Health project as a FAIR data resource for humanities researchers

Clare Lanigan, Lorraine Grimes, Preetam Singhvi
Digital Repository of Ireland, Ireland; c.lanigan@ria.ie

This presentation describes how the Archiving Reproductive Health project at the Digital Repository of Ireland can be used as a FAIR data resource for humanities researchers. We summarise the project progress to date and explain how ARH's digital collections can be used by researchers to build databases or data tools, can be searched using standardised vocabularies, and its outputs shared as openly licensed publications.

Archiving Reproductive Health (ARH) is a Wellcome-funded project coordinated by the Digital Repository of Ireland (DRI), which is working to preserve digital material created by grassroots organisations working for reproductive justice in Ireland between 1983 and 2018.

The DRI has pioneered the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) data principles in Ireland in the context of a trustworthy and certified repository for social sciences and humanities data.

Some FAIR data policies used by DRI and ARH include: consistent application of persistent identifiers including DOIs and ORCIDs, rich and standardised metadata that is indexed and searchable, promoting the use of vocabularies, and licences for reuse.

Using ARH as a FAIR resource:

1. Creating databases and queries.

One of ARH's collections, In Her Shoes: Women of the Eighth, is a collection of text files and XML metadata, making up a collection of personal stories shared to a Facebook page during the 2018 referendum campaign. The data format of this collection means it can be structured and used in sophisticated ways by researchers with some developer experience.

Data from the XML files and txt files can be extracted and re-purposed by a developer for the purpose of creating, maintaining and modifying a database. This process does not require modification of the data set and can be used as it is.

A developer can also use these files to query data. The XML files can be used as a parameter to extract data from the set as per requirement. This does not involve additional coding, just intermediate knowledge of the query language being used.

A humanities researcher with basic knowledge of coding can use this data set to view, extract, modify and manipulate the data set for academic and research purposes.

2. Searching material using standardised vocabularies.

All metadata for ARH collections uses consistent and standardised subject terms.

For example, the stories in the In Her Shoes collection were reviewed and sociologically coded using a grounded theory approach. This was augmented by mapping the themes identified and matching them to 'subject terms' used in the HASSETT (Humanities and Social Science Electronic Thesaurus) and the Library of Congress Subject Headings vocabulary.

3. Open access research protocols

ARH have documented our workflows, processes and outputs to date, and are continuously making these findings available to others in the form of publications, with the aim of sharing our findings, successes and challenges with the wider digital preservation community. Our current publications include a guide to archiving digital records for activist organisations; a project Ethics Protocol; and a Self-Care Protocol for Researchers.

Recognising Digital Scholarly Outputs in the Humanities

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

The system of research assessment and evaluation is often at odds with the practices of the scholarly community. Consider the example of Polish Literary Bibliography (PBL), a research tool providing access to over three million records on Polish culture, aggregated with resources from other countries, which could be used both in traditional scholarship as well as in data-driven approaches. Yet, every time when research unit assessment begins, the new records from the database have to be converted into a PDF monograph – which probably nobody will read – only because this is how the 'accepted' scholarly output in the humanities looks like.

In recent years we are observing an increase in digital practices and outputs in scholarship which should be understood as a standard evolution of scholarly practices which take the advantage of digital technologies. And although written genres, such as a monograph or essay, remain dominant in the humanities, the range of technological possibilities allow scholars to redefine those forms of expression and enrich them with other media or genres. However, as the opening example showed, the innovation is not supported by the assessment system, or even sometimes takes place happens in spite of it. The change of this attitude requires a recognition of three key aspects of digital humanities work: (1) its interdisciplinarity in borrowing tools and methods from ICT or social sciences; (2) the new research practices which should be recognised as valid scholarly work; (3) innovative scholarly outputs that go beyond the traditional genres but provide valid research results.

This paper presents the recommendations of the ALLEA E-Humanities Working Group with regards to the assessment of novel scholarly communication genres in the humanities. The work is based on the group's previous report, Sustainable and FAIR

Data Sharing in the Humanities, which provided recommendations on data practices in the humanities. The current focus is on attuning institutional policies to emerging scholarly needs in connection to current research assessment reform (COARA). The new recommendations will be published in 2023 and we would like to use this presentation to gather DARIAH community feedback on the document.

The ALLEA E-Humanities WG recommendations are meant to serve as a guide for institutions and evaluators to embrace innovative outputs in the humanities and thus create space for their development. The Working Group has prepared tailored recommendations which could be divided into two main groups. First, the group focuses on the cross-cutting issues pertinent to digital practices in the humanities, which are (1) linking studies with underlying data, (2) updating and versioning of the outputs, (3) collaboration and authorship, (4) training and competence building, and (5) reviewing. Next, we are discussing particular case studies of innovative outputs where cross-cutting issues manifest themselves, such as digital scholarly editions, extended publications, databases, visualisations, code and blogs. The overall conclusions provide some general remarks on recognising and evaluating digital practices in the humanities.

Paper session #6: Exploring Digital Heritage: Innovations in Digitization and Data Services

Time: Thursday, 08/June/2023: 2:00pm - 3:30pm · *Location:* Gólyavár - Main Auditorium
Session Chair: Elena Gigliarelli, ISPC CNR

The Writing on the Wall: Digitally Rediscovering Bulgaria's Post-Byzantine Heritage

Dimitar Ilkov Iliev
St. Kliment Ohridski University of Sofia, Bulgaria; dimitar.illiev@gmail.com

Text-bearing objects are among the cultural heritage items most thoroughly studied with digital methods and most widely encoded and published with digital tools. The TEI subset for inscriptions and papyri known as EpiDoc (https://sourceforge.net/p/epidoc/wiki/Home/) is constantly developed by an active community of contributors. It is applied to a number of online epigraphic and papyrological collections, including Bulgaria's Telamon database of the Greek inscriptions (https://telamon.uni-sofia.bg/) whose stages of development have been presented before the DARIAH community on different occasions. There is also the SigiDoc subset for seals and stamps (http://sigidoc.huma-num.fr/).

Among the most interesting historical inscriptions hitherto poorly covered by Digital Epigraphy are those accompanying church murals and icons, especially from the (post)-Byzantine world. Byzantine religious art left a rich heritage of artistic canons and conventions that lived on long after Byzantium itself was gone, sometimes in quite different linguistic and cultural contexts. There is a great number of texts written in (sometimes substandard) Byzantine Greek accompanying various religious scenes in churches and monasteries throughout Bulgaria from the period of Ottoman rule (XV-XIX c.). The forms and functions of such inscriptions have rarely been an object of research beyond the scope of art-historical publications where they are usually described as a part of the image. However, such texts, even when they appear extremely short and standardized, enter into a whole range of relations that require further study and proper representation.

Firstly, there is the question of the place of the inscription in the context of the entire visual composition: a relation that can vary from the simple explanation of a scene to the subtle interplay between a saintly character and the quotation contained in the book or scroll (s)he holds. Then, the issue arises about the intertextual relation of such quotations and the scriptural or liturgical traditions of which they were instances, echoes, etc. As well, the roles of the inscriptions in the larger framework not only of the particular religious building but also of the whole culture of the respective period and the linguistic competence of its audiences needs examination.

The present paper will describe the methodology and the workflow of a new project at the University of Sofia, Bulgaria, which aims at resolving such complex issues through the EpiDoc-based tools elaborated during the work on the Telamon collection. For the processing, indexing, and publication of the texts and the images, our own AIAX EpiDoc front-end service is customized and applied. For the accurate and user-friendly representation of the different connections of the texts with their wider iconographical, intertextual, and cultural contexts, a conceptual model is currently being created. According to this model, different authority files will be linked to the metadata of the particular texts in ways that will allow for the searching and organizing of the monuments according to intersections of different classifying criteria. A demonstration of the online collection will accompany the presentation.

Collections as Data at the KB, the National Library of the Netherlands: Redesigning Data Services for the Future

Steven Claeyssens, **Mirjam Raaphorst**
KB, the National Library of the Netherlands, Netherlands, The; steven.claeyssens@kb.nl, mirjam.raaphorst@kb.nl

In less than 20 years' time, the collections of digitized materials from the KB, the national library of the Netherlands, have grown into fully-fledged large-scale national collections, actively maintained and well established. They are supplemented on a regular basis and access to the collections is facilitated according to the contemporary generally accepted primary and secondary access methods of digital cultural heritage: with an online graphical search interface (Delpher) and with a suite of services, in line with the 'Collections as Data imperative' first elaborated by Thomas Padilla and colleagues (2019). Based on ten years of experience the KB is now in the process of rethinking and redesigning these Data Services. In this paper we will offer a concise analysis of our experiences so far and discuss the plans we have to get Data Services ready for another ten years.

Data Services was launched in 2012, based on a set of API's, to give access to the KB collections as data. It basically consists of a coordinator and a handful of manuals on how to search and harvest our collections. Upon request a license may be granted to get access to parts of the copyright protected collections for research purposes. Data Services has been successful in opening up the Delpher collections for a variety of national and international research projects (Polimedia, Translantis, Nederlab, Media Suite, Impresso, to name a few) as well as many individual researchers in the Netherlands, and beyond. In tandem with the KB Lab (lab.kb.nl), launched in 2014, it has served as an inspiration for several European national libraries to give access to digital library collections as data.

Today we are reorganizing Data Services. The remake will put a stronger emphasis on the importance of FAIR data principles, documenting data provenance, transparency of selection workflows and user-friendliness. On a practical level it will include the introduction of a data registry to make the data more easily findable for both humans and machines, and a series of data sheets and/or data cards as standardized documentation, encouraging transparency and mitigating potential bias. We also aim to build a corpus selection tool, offering advanced functionalities of data discovery and selection to support the creation of research corpora, as such functioning as a more intuitive user interface to the existing API's. Finally we want to address the need for giving access to _all_ in-copyright materials by creating an onsite mining facility and an online tools-to-data solution, providing ways to mine our collections without violating the rights of copyright owners.

The VAST methodology and workflows for experience digitisation

Maria Dagioglou[1], Dora Katsamori[1], Georgios Petasis[1], Alfio Ferrara[2], Stefano Montanelli[2], Theodore Grammatas[3], Maria Dimaki Zora[3], Aikaterini Diamantakou[3], Marco Berni[4], Elena Fani[4], Carla Murteira[5], Alba Morollón Díaz-Faes[5], Elena Aristodemou[6], Marko Kokol[7]

[1]National Centre for Scientific Research "Demokritos", Greece; [2]Università degli Studi di Milano, Italy; [3]National and Kapodistrian University of Athens, Greece; [4]Museo Galileo - Istituto e Museo di Storia della Scienza, Italy; [5]NOVA University of Lisbon – School of Social Sciences and Humanities, Portugal; [6]Fairy Tale Museum, Cyprus; [7]Semantika Research, Slovenia;
mdagiogl@iit.demokritos.gr

The interaction of aesthetic and moral values [1], as well as the role of art in moral education [2] are topics of debate. Nevertheless, artifacts often cannot escape but embody the values of their times, their creators or of the stories/people they talk about, at least in the perception of the audiences. Reactions with respect to transmitted values can get as strong as toppling statues, like that of Edward Colston in Bristol, UK in 2020. At the same time, observing and understanding the values of citizens and stakeholders becomes increasingly important in many fields, from sociology and policy-making to technology and Artificial Intelligence. Taken together, the systematic collection of the values born in cultural artifacts, as perceived by audiences, can offer a 'valuable' source of research data in various fields.

Unlike measurable or objective properties of artifacts and historical metadata, people's experience during their exposure to an artifact is highly subjective. From spontaneous emotional reactions to personal biases, capturing the audience's perception of values is a challenging task. VAST is a European H2020 research and innovation action that has been developing methodologies, tools and data infrastructure to capture and digitise the values of intangible cultural heritage (CH) artifacts, including narratives of ancient Greek theatre, 17th century scientific texts and European fairy tales. From text and visual content annotation, to applications and educational activities, VAST has structured a methodology so that any data collected are scientifically sound, can populate VAST's ontology, and stimulate research in humanities. VAST methodology, instilled in the ontological specifications, deals with three major aspects of a person's interaction with an artifact or experience: a) the participant's individual characteristics (e.g. demographic information, philosophical beliefs), b) the description of the artifact/ experience and c) the collection of the perceived values based on the artifact/ experience.

The 'born-digital' datasets, collected through the above workflow, populate VAST's ontology and will be available through the VAST platform for various uses. Besides supporting research, from Humanities to AI (e.g. providing 'unbiased' datasets for values mining and Natural Language Processing), VAST datasets can be used for citizen-informed curation, as well as for promoting new ways of engagement with CH, including digital apps and post-activity interactions. Again, these interactions with the VAST platform can further be used for augmenting existing datasets or producing new ones.

Example of the VAST workflow based on an educational activity. An excerpt of Sophocles' Antigone is used to trigger a discussion about the concept of 'values' and the values perceived in the play. Pre-activity, participants provide demographic information and fill-in the Personal Values Questionnaire. After exposure to the play (central experience), the value perception is collected through text annotation and mind maps. The collected data are organised into the VAST ontology and then become available for post-activity use. Importantly, the methodology of each activity becomes available, allowing the reproducibility of the experience and the relevant dataset augmentation.

[1] Ravasio, M., 2021. What is the Connection Between Art and Morality?

[2] Hospers, J., 2022. Art as means to moral improvement.

HTR in BnF DataLab : first steps with researchers

Marie Carlin, Sébastien Cretin

Bibliothèque nationale de France, France; marie.carlin@bnf.fr, sebastien.cretin@bnf.fr

The BnF DataLab opened in October 2021 to welcome researchers working on the BnF's digital collections. It has been designed to facilitate access to the collections by providing physical and computer work spaces but also support from a wide range of experts. More than a data supply service, the BnF DataLab is a coordination of services and expertise that accompanies research projects from the constitution of corpora to the valorisation of research results. It is a place of co-construction, but also a laboratory for experimenting the tools of the future library, not only in the service of research projects but also in business applications. To enable this dialogue between researchers and the library and to ensure the follow-up of projects, partnerships with major players in the research world such as the CNRS via the IR* Huma-Num or the ObTIC project-team of Sorbonne-Université have been set up. These partnerships allow researchers to be hosted in residence at the BnF DataLab, supply services of staff and co-sponsored projects.

The challenges facing heritage institutions and academic research are similar when it comes to FAIR data. We believe that the DataLabs flourishing in different national libraries can serve to forge relationships between patrimonial institutions, between patrimonials and research, to provide researchers with usable and interoperable data. By offering a place for experimenting with new technologies, but also ensuring best practices, methodological and technical standards, to enhance the value of the datasets and tools.

This harmonization and exchange between researcher and cultural insitution can be shown through HTR programs, hosted in the Bnf DataLab.

The transcription of handwritten documents and the provision of robust models is a subject that mobilizes both researchers and heritage institutions. It is one of the major projects of the BnF. Indeed, the state of HTR technologies has improved a lot during the last few years. It seems the moment has come for cultural heritage institutions to take advantage and start considering processing large sets of documents. But after an important period of using OCR softwares as a routine, the project of developing HTR brings some fundamental questions.

None of the HTR engines that have emerged from the scholar research field is capable of processing efficiently a huge variety of writings, from widely different times and scriptors. And nothing indicates that such a tool or generic model will be soon available. So if we don't want to use an API from some of the major IT actors, and if we need to set up the same kind of industrial workflow that we use for OCR, how do we do it? How do we manage to retrieve all the models that are available ? How do we test them to certify their efficiency? How do we interrogate their description to relate them to the relevant documents to process? And how do we feedback the creators of the open-source models to help them improve their work?

Paper session #7: Cultural Heritage Data: Use Cases

Time: Thursday, 08/June/2023: 2:00pm - 3:30pm · *Location:* Gólyavár - Small Auditorium
Session Chair: Tanja Wissik, Austrian Academy of Sciences

Creating and Using a National Linked Open Data Infrastructure for Cultural Heritage Applications and Digital Humanities Research: Lessons Learned

Eero Hyvönen
Aalto University and University of Helsinki, Finland; eero.hyvonen@aalto.fi

This paper presents lessons learned in Finland for creating a national ontology and Linked Open Data (LOD) infrastructure for the Cultural Heritage (CH) domain and Digital Humanities (DH) research. To test the infrastructure, a series of twenty LOD services and semantic portals in use have been created in 2002–2023. This work is unique due to its systematic national level nature and long-time span. The infrastructure is in use on the Semantic Web (SW) and has attracted up to millions of users suggesting feasibility of the proposed approach. The presentation is an overview of the three new articles [1-3] available on the Web for more details.

The elements needed for a national infrastructure include, firstly, metadata models populated by resources taken from shared domain ontologies published on ontology services for interoperability. Most of our applications are founded on extending the event-based CIDOC CRM to areas such as history and biography/prosopography. Our prototype ontologies and services were deployed by the National Library of Finland and are used today by many museums and other memory organizations.

Secondly, data services are needed based on the FAIR principles that are compatible with the linked data standards and best practices of the W3C. Our work focuses on enhancing the re-usability and quality of LOD by extending Tim Berners-Lee's 5-star model to a "7-star" model. Thirdly, tools are needed for aggregating distributed heterogeneous siloed data, for extracting entities, relations, and events from texts, and for data publishing, analysis, and user interfaces (UI). Finally, a human infrastructure is also needed for educating people about the new SW approach.

Applications for testing and demonstrating the usability of the infrastructure were developed that evolved gradually into the "Sampo" model and series of ca 20 semantic portals and LOD services for research in DH and application development. The novelty of the Sampo model lays in its attempt to formulate a set of consolidated design guidelines for developing CH applications and DH research.

The LOD approach presented has its own challenges. More agreements on data models and ontologies are needed between the data producers, which complicates the publication process. Integration of SW technologies with legacy systems may be challenging, and there is lack of IT personnel competency in using SW technologies. Creating LOD manually is costly, but automation lowers data quality. Issues of Big Data interpretation, quality, completeness, veracity, skewness, uncertainty, fuzziness, and errors arise. However, enriching data carefully with semantics is in my mind a promising way for creating a more and more intelligent web services in a cost-efficient way.

[1] Eero Hyvönen: Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. Semantic Web, vol. 11, no. 1, pp. 187-193, 2020.

[2] Eero Hyvönen: Digital Humanities on the Semantic Web: Sampo Model and Portal Series. 2022. Semantic Web, in press.

[3] Eero Hyvönen: How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web. 2022. Semantic Web, under review.

Urban History and Heritage Data for Learning by Gaming

Rosa Tamborrino, Pelin Bolca
Politecnico di Torino, Italy; pelin.bolca@polito.it

Digital collections together with digital approaches and tools have the potential to improve the analysis of urban history and heritage by providing new opportunities for researchers to explore and interpret historical data in ways that were previously not as much as effective. Through digital libraries, museums, and archives, digital collections offer an effective access to a wide range of information and resources that would otherwise be difficult or impossible to consult. Cultural institutions (the GLAM) are creating and organizing their digital collections to preserve the data, share knowledge, and create digital exhibition formats on their websites. However other potentialities can be explored. They can also frame an ample opportunity for data to be reused to facilitate research, disseminating, and learning by promoting public engagement and fostering collaboration between institutions and individuals.

However, accessing data collections conserved in different archives, institutions, and libraries and their guidance has always been problematic for researchers. Managing such multiple data usually creates complexity in terms of identification, categorization, and verification of reliability as well as their organization and standardization to construct digital narratives that are effectively understandable for public users. To deal with this challenge, effective collaboration and continuous interaction between institutions and individual researchers are crucial.

In 2019, with the occasion of International Summer School ''Cultural Heritage in Context. Digital Technologies for the Humanities'' with the topic of ''Learning by Game Creation'', the collaboration between institutional and academic research has been experienced in Turin, Italy. This was actually 3rd edition of the yearly organized summer school which is a joint project of Politecnico di Torino and the University of California, Los Angeles. The main aim was to deal with the creation of games with the use of digital historical narratives as a tool for learning and disseminating knowledge. In this aim, the involved researchers collaborated with the National Cinema Museum of Turin, the Historical Archive of Turin, and MuseoTorino with particular attention to experimenting with the potentiality of digitalized historical data for promoting the urban history and heritage of Turin and for widespread use of digital collections. A huge amount of historical sources including videos, movie clips, historical photos, cartographies, and maps were accessed. These multiple data were collaboratively re-used with international experts and participants to prototype games as outcomes of the experiment. Each game envisaged the usage of digital devices such as smartphones, computers, or tablets. A variation of gamer targets was considered from the newcomers of the city with few information to the citizens with a lack of local knowledge. These features have resulted in different game solutions.

This paper presents and discusses the diversity of prototyped outcomes by exploring some potentialities. Data acquisition, curation, historical interpretation, and usage of digital collections to structure the gaming strategy are considered. The particular interest is to discuss the pros-cons manners of the data governance process to co-built critical games linked to the urban history of heritage by working strongly together with cultural institutions and academy as a research framework.

Parcels of Venice: A Platform for Indexing Cultural-Historical Data in Space and Time

Paul Guhennec, Didier Dupertuis, Isabella di Lenardo

Digital Humanities Institute, EPFL, Switzerland; paul.guhennec@epfl.ch

This contribution is based on an integrated platform, developed in the Parcels of Venice SNF project, using the concept of urban footprint as a pivotal information unit for organising contemporary and historical datasets. The geometry of footprint can be used both (1) for segmenting larger representation of the city like cloud points, (2) to index historical documents like photographs, engravings or paintings and make a direct link with historical parcels documented in cadastres, (3) to structure smaller urban units, like individual facades or information indexed on specific addresses. More precisely, the platform takes as a backbone the current day vectorial footprints, published openly by the city of Venice. Each footprint is enriched with the result processing of different sources:

- a point cloud covering the full city and split into individual point clouds, using the vectorial geometries of each building;

- vector polygons and text fields of the 1808 Napoleonic cadastre of the city, and which were extracted automatically [1];

- data points and text entries of the 1741 Catastici, a proto-cadastral survey of the city's dwellings;

- corresponding Wikidata, Wikipedia, and OpenStreetMap entities, matched on the geographic coordinates of both the entity and the footprint;

- photographs and paintings from the Venice-based Fondazione Cini Fototeca, and realigned through matching place names [2];

- textual descriptions from architectural catalogues, and realigned on the Venetian address system.

This integrated dataset can be used to automatically compute information of each of the facade of Venice. an algorithmic method of automatically producing orthophotos from the point cloudì was devised. Through a geometrical intersection with the vector edges of each building, each 3D point is labelled with a facade id. The points of each facade are then projected into two-dimensions through the Principal Component Analysis of their 3D position [3].

This abstracted planimetric representation allows the extraction of the constitutive elements of an architectural grammar (proportions, forometry and rhythm, colour) as well as to identify local specificities [4]. For example, the homogeneity in the facade dimensions and style in the housings of Santa Marta (Fig. 2a) is strong evidence of a construction phase coherent both in terms of chronology and function, as opposed to the diversity of Fondamenta de Cannaregio (Fig. 2b). Computing the standard deviation in facade heights for each street in our information system thus yields a cartography of this height heterogeneity (Fig. 3).

The ability to cross-interrogate different sources, covering different pieces of information and time frames (as is illustrated in Fig. 4, where the orthophotos of all 2022 canal- and square-side facades belonging to a member of the Grimani family in the 1808 Cadastre are shown) is a powerful tool, provided it is accompanied by a critical reading of the limits of the processing and of the data itself.

Building an infrastructure for cultural heritage of the present

Juliette Vion-Dury[1], Vyacheslav Tykhonov[2], Andrea Scharnhorst[2], Yves Rozenholc[3]

[1]l'Institut de recherche interdisciplinaire sur les enjeux sociaux, Université Sorbonne Paris Nord; [2]Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences; [3]Faculty of Pharmacy, Université Paris Cité; juliette.vion-dury@univ-paris13.fr

A lot has been written on the internet enabling us to preserve cultural artifacts; the traces of our society like no other medium. Equally numerous are the proposals on how to best organize this heritage of the future.

This paper takes as an example the recent experience of the Covid-19 epidemics and introduces into an agile infrastructure which supports the collection, annotation and preservation of testimonies during this period. The idea has been developed in France under the name 'Covid-19 Museum'. (Rozenholc 2021) What distinguishes this initiative from others addressing the same phenomena, lies in the width of the content imagined to be collected (from news items over scientific debates to individual narratives) and the underlying infrastructure which obliges FAIR principles of data preservation and respects ownership of content at the same time (Tykhonov 2021), closely related to current debates on structuring the EOSC.

For the DARIAH Annual Event, we zoom in on the technological challenges to create an open, still protected memory space in collaboration with cultural heritage data providers (in this case the Archives Départementales) and linking it to other large Pan-European initiatives on Digital Humanities and Cultural Heritage (such as the Time Machine Europe). More particularly; we zoom into the epistemological framework behind the Covid-19 Museum which is particularly rooted in oral history, cultural studies, and creating musea and encyclopedia as a sharable body of knowledge. At the same time, by creating such a heritage space of the recent past, the project aspires to understand the effect of such a singular event on the future of a society system which is intrinsically anticipatory. The paper is mostly related to the stream 'Sustainable workflows for data management and curation' of the conference.

We are also open to present this work as a poster.

Paper session #8: Fostering and improving DH data reusability

Time: Friday, 09/June/2023: 9:30am - 11:00am · *Location:* Gólyavár - Main Auditorium
Session Chair: Arnaud Gingold, OpenEdition-AMU

Sharing and Sustaining Digitisation Knowledge: a White Paper on written cultural heritage digitisation

Gustavo Candela[1,8], Mirjam Cuper[2], Ines Vodopivec[3], Katrien Depuydt[4], Christel Annemieke Romein[5], Sally Chambers[6], Jesse de Does[4], Isabel Martínez-Sempere[1], Tomasz Parkoła[9], Apostolos Antonacopoulos[7]

[1]Universidad de Alicante, Spain; IMPACT Centre for Competence in Digitisation; [2]KB, National Library of the Netherlands; [3]National and university library of Slovenia, Slovenia; [4]Instituut voor de Nederlandse Taal/Dutch Language Institute; [5]Huygens Institute for History and Culture of the Netherlands; [6]Ghent Centre for Digital Humanities, Ghent University; KBR, Royal Library of Belgium, IMPACT Centre for Competence in Digitisation and DARIAH-EU; [7]PRImA Research Lab, University of Salford, United Kingdom; [8]Centro Biblioteca Virtual Miguel de Cervantes; National Library of Scotland; [9]Poznan Supercomputing and Networking Center; sally.chambers@ugent.be

Since the foundation of the IMPACT Centre of Competence in 2012, the digitisation landscape has changed significantly. Examples include: developments in digitisation technologies (AI), digitisation practices (cultural heritage institutions, research institutes, researchers) and digitisation quality expectations (mass digitisation, digitisation-on-demand, researcher-led digitisation). After more than 10 years of exploring new ways to make digitisation more efficient, it has become even more challenging to integrate state-of-the-art digitisation research into day-to-day digitisation practices and operations.

The IMPACT White Paper, which resulted from a collaborative workshop and writing sprint held in Alicante in December 2022, documents the key themes and challenges for the international digitisation community and paves the way for establishing a common interdisciplinary research agenda for cultural heritage, digital humanities, as well as computer and data science, on digitisation. Its main focus: sharing and sustaining digitisation knowledge. It explores four key themes: digitisation quality and standards; datafication of digitised cultural heritage; digitisation workflows and tools and digitisation community management.

This contribution presents prioritised challenges addressed by the White Paper, including best practices for sustainable workflows in quality and maintenance. It emphasises standardisation of metadata in documentation and description of datasets, along with measuring quality per resource type and interoperability.

Based on previous collaborative editing approaches, the identification of the main challenges for sharing and sustaining digitisation knowledge was performed using a framework that was agreed by an international team consisting of representatives of the heritage, university and research institutions involved with the IMPACT Centre of Competence. Five steps were applied to the identified key themes: i) identification of themes; ii) selection of participants; iii) creation of working groups; iv) identification of challenges in themes; and v) ranking identified challenges.

The intended audience of the White Paper is the international digitisation community, which IMPACT inclusively defines as anyone who self-identifies through personal interest, professional practice or research expertise in digitisation of historical (textual) materials. To date, the IMPACT community consists of cultural heritage professionals; (digital) humanities researchers and computer, information and data scientists. Furthermore, national, European and international policymakers and funding agencies, may also find this White Paper of interest as it provides a foundation to determine community-endorsed priorities and recommendations, for policy priorities and investment opportunities in the coming years.

The aim of presenting the IMPACT White Paper to the DARIAH Community is not only to validate the key challenges identified but also to explore how partner institutions in IMPACT and DARIAH could collaborate to propose and implement solutions for the both benefit of the wider Arts and Humanities research community as well as fostering increased collaboration with cultural heritage institutions.

Engaging academic and research communities in the common European data space for cultural heritage: a DARIAH and Europeana partnership

Sally Chambers[1,2,3], Toma Tasovac[1,4], Alba Irollo[5]

[1]DARIAH-EU; [2]Ghent Centre for Digital Humanities, Ghent University; [3]KBR, Royal Library of Belgium; [4]Belgrade Center for Digital Humanities; [5]Europeana Foundation; sally.chambers@ugent.be

The common European data space for cultural heritage is an initiative of the European Union, funded under the Digital Europe programme. Conceived with Europeana at its core, the service will be deployed by a consortium of 19 partners, including DARIAH-EU, and will become concrete through four key areas: a) development and operation of the data space infrastructure; b) Integration of high-quality data; c) capacity building and fostering reuse and d) digital services for the public.

This paper will analyse the role of the data space in the broader research infrastructure landscape in Europe and focus on DARIAH's activities aimed at strengthening the academic and research communities' reuse of digital cultural heritage. In close collaboration with the Europeana Foundation and the Europeana Research community, DARIAH will contribute to gathering requirements from these communities and to linking the data space to platforms specifically conceived for them such as the SSH (Social Sciences and Humanities) Open Marketplace within the EOSC (European Open Science Cloud). In particular, DARIAH will produce two key outcomes: a) SSHOC Marketplace Workflow on "Collections as Data" and b) a DARIAH-Campus course on "Using Europeana APIs in Research and Higher Education".

A SSH Open Marketplace Workflow on "Collections as Data"

A workflow is a content type in the SSHOC Open Marketplace which describes a series of steps necessary to apply a method, use a tool or answer a research question in the humanities context. For instance, see "Extract textual content from images" as a workflow composed of 13 steps. Inspired by the 'Collections as Data' movement as an approach for cultural heritage institutions to prepare their digital collections for analysis and reuse, DARIAH will design a workflow on creating Collections as Data by identifying, providing access and retrieval, cleaning and reformatting, enriching and packaging the datasets for digital humanities research, taking into account previous work on the Heritage Data Reuse Charter. Each dataset will be linked to various tools in the SSH Open Marketplace as well as the relevant literature.

A DARIAH-Campus Course on “Using Europeana APIs in Research and Higher Education”

DARIAH-Campus is both a discovery framework and a hosting platform for DARIAH and DARIAH-affiliated offerings in training and education. The goal of DARIAH-Campus is to widen access to open, inclusive, high-quality learning materials that aim to enhance creativity, skills, technology and knowledge in the digitally-enabled arts and humanities. Europeana APIs are a set of services calibrated to the needs of professionals interested in the reuse of digitised or digital-born cultural heritage in education, research and the creative industries. Within the context of the data space, DARIAH will develop a course which will introduce researchers and academics to both Europeana APIs and the process of creating and using Jupyter notebooks.

The DARIAH Annual Event presents an ideal opportunity to raise the visibility of the data space within the wider Arts and Humanities research community, provide an update on the current status of activities and stimulate an interactive dialogue with the DARIAH community to ensure that their needs are heard.

Oral history data as linguistic data: a case study from a semi-digitised collection

Vicky Garnett
Trinity College Dublin, Ireland; vicky.garnett@dariah.eu

Oral history recordings are often used for dialectological studies, but this reuse of data can present challenges as well as opportunities for the linguist. On the one hand, recorded oral histories can provide a reasonably large data sample of a particular linguistic feature, and often in the natural spoken language of the informant. On the other hand the reuse of data that was originally captured for a different purpose can lead to a mismatch between the data sample and the dialectological analysis. The researcher is limited in their control over aspects of the data, such as quality of the recording, the structure of any interviews conducted, and the supplementary materials that might accompany the resources. Coupled with technical issues of the data, the archive itself can present a further challenge. While larger government-supported archives have the resources to both preserve the recordings and provide access to them, smaller local archives often struggle for human as well as technical resources. Moreover such local archives are often volunteer-led, leading to limited opportunities for formal training or skills development among the custodians of the collections for often very practical reasons.

This paper will present a case study from my own experiences of reusing partially-digitised oral histories from a small local archive. The project investigated a particular sound change in Somerset and, alongside original data collected by this researcher through linguistic interviews between 2015 and 2017, also made extensive use of existing data from two different datasets: the Survey of English Dialects (Orton et al., 1967) and a small collection of oral histories from a rural part of the county (this collection will remain nameless to preserve anonymity of both the archive and the historian who collected them). The Survey of English Dialects (SED) was collected by a team of linguists from the University of Leeds in the mid 1950s with specific linguistic intent and the original and digitised materials are now held both within the University of Leeds and the British Library. The oral history archive recordings are held in CD and paper format only in a small local archive with a duplicate available via the county archive. Short audio clips from the oral histories were also available online between 2002 and 2022, but the archive website itself is no longer accessible.

The experience of attempting to use recordings from the oral history archive is indicative of the challenges faced by researchers, as well as practitioners, when accessing and reusing data from smaller cultural memory institutions. Ultimately this dataset was rejected from the final analysis for practical reasons, in part related to issues of access and reusability. In this case study, the steps taken in attempts to access, process and analyse these two secondary datasets will be discussed in light of FAIR data principles, and the practical needs of the researcher. The experiences and challenges to the researcher in using these oral history datasets can tell us much about reuse of semi-digitised legacy cultural data in the increased demand for remote digital access.

Bridging Islands: Interoperation between DH tools

Maarten Janssen
Charles University, Czech Republic; janssen@ufal.mff.cuni.cz

There are many tools in DH that combine several text-oriented functions into an integrated platform. For instance TEITOK has modules to search, edit, tokenize and annotate, display, visualise frequencies, links with sound, video, facsimile images, etc. Apart from fully integrated tools, there are collections of collaborating tools, such as those built around the FoLiA file format, or the Universal Dependency tools around the CoNLL-U format.

Creators of such tools want people to use their tools and facilitate importing data from elsewhere, but not focus much on export. And when there is a need for more functionality, the tendency is to build local solutions. But there are several reasons why providing the option to work with external tools is beneficial. Below are examples involving TEITOK, but similar reasons apply to any tool (infrastructure).

Functionality: TEITOK is a corpus tool that provides document visualisation, but does not have the same range of functions for that that a digital edition environment like TEIPublisher provides. And it provides efficient document editing functions, but does not provide the type of versioned, inter-annotator type of stand-off annotation that for instance INCePTION provides.

Quality: TEITOK provides an NLP pipeline, to automatically annotate with NeoTag or UDPIPE. But due to the rapid development in AI, it is impossible to provide state-of-the-art NLP for all circumstances, and there will always be situations in which external tools simply work better.

Therefore, it would be beneficial to the field if tools would embrace the fact that other tools are better in certain respects, and facilitate full-fledged exports or ports. In TEITOK, we have modules to generate a full Kontext corpus from TEITOK, are working on a module to generate corpora in TXM and TEI Publisher, and this on top of a wide range of export scripts to generate OCR type files in PageXML or ALTO, sound-based files in ELAN, dependency files in CoNLL-U, and multi-layered files in IUMA CAS and TCF (WebLicht).

For the use of external annotation tools, export alone is not sufficient, the results need to be reincorporated. The new TEITOK API (programming interface) provides the option to export all the relevant data in the most appropriate format, and reintegrate the annotation results back into their original TEI/XML documents. So documents in the corpus can be exported to the IUMA CAS format that is used in INCePTION, then adorned with annotations in INCePTION. Or the documents can be exported to

CoNLL-U, run through any NLP pipeline. And once the external annotation is completed, the results can be loaded back into the TEITOK file using the API.

The option to port data also provides a guarantee that even if the tool would stop working, the data can still be exploited. And by providing ways to port data from one (corpus) tool to another, data can be used for additional types of research.

Poster Session

Time: Wednesday, 07/June/2023: 5:00pm - 6:30pm · *Location:* Gólyavár - Foyer

Can yesterday's data fulfil today's researchers' needs? Crafting additional ways to improve the fitness for use

Sitthida Samath

CNRS Persée UAR 3602, France; sitthida.samath@persee.fr

For 20 years, Persée has been the French operator for scientific literature heritage digitization and online dissemination. Persée currently has a 34 people team, involved in publishing, technical and scientific projects, and working over a fully-fledged paper-to-digital document processing chain. So far Persée has produced :

- a web portal (standard digital library) ;

- a triplestore (linked open metadata outlet) ;

- 5 'perséides' (project-specific online corpora).

Alongside the perseides which are the result of a co-design process, Persée has been providing legacy (meta)data to digital humanists for whom data science and computational methods, not to mention AI/ML, have become standard practice. Since 'Collections as data designed for everyone serve no one' (The Santa Barbara Statement on Collections as Data, 2017), Persée has implemented various methods to ensure the fitness for use, including quality checks, user feedback analysis, co-design. Yet, the same questions eventually arise again: 'Can yesterday's data fulfil today's researchers' needs? What are these needs? How can we ascertain quality?'.

Taking our part in the collaboration of library science, computer science and humanities working together with digital cultural heritage (Oberbichler et al., 2022), we saw fit to extend customer study. Drawing up explicitly the Persée FAIR Implementation Profile (Schultes et al., 2020), mapping visually user types and data dissemination channels, then applying graph analysis, makes it possible to highlight critical nodes, popular use cases and intensive use areas. Moreover, modelling a typical data reuse process makes it possible to (re)locate occurrences of data friction (Edwards et al., 2011). Besides, to our knowledge, few are the summary papers about OCR quality for so called 'downstream tasks' (van Strien et al., 2020) in an otherwise abundant literature since Holley's OCR accuracy categories (Holley, 2009). Setting up a dedicated process of literature review, we found that :

- some tasks, like part-of-speech tagging and NER, seem more affected by OCR quality than others ;

- the performance of a task might relate to specific ranges of OCR quality metrics, although the matter, still open to question, largely depends on the chosen task, language, metrics, and deserves to be explored further.

The introduction of graph study, process modelling and literature review has proved useful to qualify usage more accurately and to get a better scope of the corresponding quality dimensions and metrics. We argue that, for lack of a one-stop standard, such endeavours could help to :

- formulate project-specific data reuse strategies : using data 'as-is', counterbalancing defects with smarter processing, applying post-production corrections ;

- tackle the ubiquitous topic of data quality more effectively : finetuning supply and services, avoiding the 'garbage in, garbage out' effect in research projects, closing the data lifecycle by re-ingesting the research output.

ExploreSalon: Unveil Hidden Stories from the Past - Concept and Outcome of a Digital Humanities and Cultural Heritage "Hackathon

Vera Maria Charvat[1], Matej Ďurčo[1], Elisabeth Königshofer[1], Sylvia Petrovic-Majer[2], Anna Woldrich[1]

[1]Austrian Academy of Sciences (OEAW), Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), Austria; [2]OpenGLAM.at, Austria; veramaria.charvat@oeaw.ac.at, matej.durco@oeaw.ac.at, elisabeth.koenigshofer@oeaw.ac.at, sylviainpublic@gmail.com, anna.woldrich@oeaw.ac.at

For the duration of one week (22nd – 26th May 2023) the so-called ExploreSalon will offer a collaborative space to explore digitized memories, in the form of curated biographic, spatial and temporal datasets with focus on Vienna 1900. The event will be organized by the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) and supported by the CLARIAH-AT national consortium as part of its knowledge sharing activities. The ExploreSalon has two goals: Bring people of diverse backgrounds together and provide them with the opportunity to discover innovative ways of data-based storytelling – exploring stories hidden in the data, presenting ideas and sharing findings.

This poster will present the underlying concept and general workflow of this form of cultural hackathon, elaborate on challenges such as bringing together people with different backgrounds and motivate them to get creative and exchange their ideas. The results and outcomes of the individual groups formed during the ExploreSalon event will enrich and underline the presentation.

aLTAG3D: A User-Friendly Metadata Documentation Software

Bruno Dutailly[1], Sarah Tournon-Valiente[1], Mehdi Chayani[1], Valentin Grimaud[2], Xavier Granier[3]

[1]Archéosciences Bordeaux; [2]LARA, CReAAH; [3]IOGS (Institut d'Optique Graduate School); mehdi.chayani@u-bordeaux-montaigne.fr

The role of metadata in scientific research cannot be overstated as it helps to organize, preserve and enhance the value and usability of vast amounts of documents, thereby improving their traceability and understanding.

Despite its importance, documenting metadata is often a time-consuming task for researchers, especially for 3D projects that contain, beyond the 3D data, all the related documents.

To prevent the loss of traceability of these resources, it is essential to simplify the metadata documentation process. The consortium "3D for Humanities", labeled by Huma-Num since 2014, has taken major steps to tackle this challenge by researching the most effective way to document massive 3D data and simplify the 3D deposit project, both in the French National 3D Data Repository in the fields of Humanities and Social Sciences which offers publication services, and at the C.I.N.E.S. (National Computer Center for Higher Education) which offers long-term archiving of electronic data.

The tedious task of documenting metadata for 3D projects highlights the need for simplification. To effectively document, it is important to not only provide information about the document itself through its metadata, but also to link project files together. Traditional web-based interfaces cannot achieve this simplification, but it can be done through specialized software.

With this in mind, the consortium "3D for Humanities", has developed aLTAG3D, an open source software aimed to be as automated as possible to document 3D research projects and thereby reducing the amount of information that the researcher will have to fill in manually.

The software uses a graphical user interface to make documenting data for archiving and publication more user-friendly. Users simply drag and drop to construct the package, object-by-object, by attaching sources and filling in the requested metadata at all levels. Some information will be extracted automatically, while others will be filled in manually.

At the end of the process, on one hand, we get a hierarchical folder containing all the information and comprehensible data for future users and on the other hand, an XML file of the metadata, a file compatible with the archive at C.I.N.E.S. More metadata schemes may be introduced in a xsd format.

First steps towards a workflow for 3D-models based on IIIF

Rita Gautschy

DaSCH, University of Basel, Switzerland; rita.gautschy@dasch.swiss

Images play an important role in archaeological research and the 3D medium is becoming increasingly popular: 3D scans of archaeological objects, geo-referenced 3D scans of structures, e.g. tombs, or reflectance transformation imaging (RTI) are increasingly likely to form one part of the digital data created in research projects or by cultural heritage institutions. 3D offers many application possibilities that can contribute significantly to knowledge acquisition. Very often, however, the resulting files are extremely large and not really suited for smooth display. The situation is further complicated by the fact that there is still no real standard for both long-term storage and dissemination of 3D images.

For a data repository such as the Swiss National Data and Service Center for the Humanities (DaSCH) which provides a “living archive” where data can be accessed and searched directly by humans as well as by machines, 3D images pose a series of questions and challenges.

• The original 3D-models are often several gigabytes large and their key features, the geometry, texture and possibly also animation, may come in separate files. How can such data be archived properly? Which non-proprietary formats can or should be accepted?

• To achieve interoperability and enable annotation, the goal is to have a type of IIIF dissemination. How can this be achieved from the original 3D-models?

• What open source 3D-viewer should be implemented in web applications?

In order to achieve the long-term goal of proper archiving of the original file(s) on the one hand, and the provision of a lightweight, interoperable, viewable and annotatable version of the 3D-model in the web application on the other hand, a process needs to be established that can be automated. I will present our first attempts of establishing such a workflow.

References

DaSCH Service Platform (DSP): https://www.dasch.swiss/platformcharacteristics

IIIF: https://iiif.io/

IIIF 3D Technical Specification Group : https://iiif.io/community/groups/3d/tsg/

IIIF-Prezi3: https://github.com/iiif-prezi/iiif-prezi3

Sustaining Digital Scholarship: keeping research data alive

Damon Strange, Megan Gooch

University of Oxford, United Kingdom; damon.strange@humanities.ox.ac.uk, megan.gooch@humanities.ox.ac.uk

The need to sustain data and outputs from research projects well beyond their initial grant-funded period is commonplace within academia, particularly so within the Humanities. We often find that these semi-active or warm data collections have value and require care beyond the life of the original hosting, system, or platform they were conceived upon. With the increasing demand to make data more open, and research funding bodies increasing requirements and time periods for which research data must remain available, institutions need to be ready to offer researchers the tools and platforms to comply. Compliance is one lens through which to view this particular challenge. Yet we believe that institutions should and can be motivated to sustain data by celebrating the research they develop, and through reaping the continued scholarly benefits and impact gained through research data being hosted and shared for as long as possible. Our poster will present practical insights into the methods employed at the University of Oxford to support digital humanities scholars (and others) safeguard their digital legacies.

We will showcase three key focus areas of ‘People’, ‘Process’ & ‘Technology’ to highlight how the Sustainable Digital Scholarship (SDS) service at Oxford is helping researchers and their projects secure the long-term future of their research data outputs:

People – Researchers are at the heart of research, as obvious as that sounds! But our data repository service has been developed and led by the needs of academics and this has helped us build a flexible approach to supporting a variety of research projects from may disciplines.

Process – The SDS service offers both a repository and also free researcher consultations, with an aim to embed better research data management processes and practices within research projects with a significant digital component at Oxford. We also have developed a methodology for assessing suitability for supporting research projects on a common / shared infrastructure.

Technology – We offer access to a 'Software as a Service' (SaaS) Open Access Repository for the curation and publication of Digital Humanities research project data and offer flexibility within our standard infrastructure that many benefits to research data hosting, compared to more archival repository solutions, or the custom-built databases and websites developed for each individual research projects.

EHRI in TEITOK: reusing well-structured DH data for corpus exploration

Maarten Janssen

ÚFAL, Charles University, Czech Republic; janssen@ufal.mff.cuni.cz

In this demo, we will demonstrate how TEITOK [1] makes it possible and hopefully even easy to reuse cultural heritage data as data for linguistically driven research. As an example of this, we will showcase the conversion of textual data from the European Holocaust Research Infrastructure (EHRI) [2] into a linguistic corpus in TEITOK [3].

TEITOK is an online environment for annotated corpora - it is used to create, annotate (manually or automatically), display, and search corpora in which each corpus document is a richly annotated TEI/XML document. It can handle linguistic annotation, but also other textual metadata such as manuscript alignment, multimedia alignment, geolocation data, linked open data, etc. It has a modular design, in which each module explores the TEI/XML document(s) in a different fashion. For instance, there is a module that can exploit geolocation data to map search results onto the world map, and a module that can exploit multimedia alignment to display the annotated transcription below a waveform display of the audio.

EHRI is a European research infrastructure aiming to connect sources, institutions, and people connected to the Holocaust around the world. It hosts several digital editions, or collections of source materials, including the Begrentzte Flugt [4] collection, which brings together source material related to people fleeing across the border between Austria and Czechoslovakia after 1938.

We will show how the conversion of the Begrentzte Flugt to a TEITOK corpus [5] opens up this digital edition to a different kind of DH research. The TEITOK version of the corpus makes it possible to quickly get access to passages in the corpus mentioning specific people, places, or organisations, independently of the way they are mentioned or the languages they are mentioned in. And it makes it possible to obtain statistical data about documents with different textual metadata, for instance to compare the kind of words or linguistic constructions used in Czech texts written by women, against similar texts written in German.

The conversion of EHRI to TEITOK is part of an ongoing effort at LINDAT [6] to make relevant project data available as searchable (linguistic) corpora, which includes also ELTeC, DRACOR, and generic schemes to reuse for data from several popular platforms, such as spoken data created with ELAN [7] and manuscript data created with Transkribus [8].

References

[1] https://www.teitok.org/

[2] https://www.ehri-project.eu/

[3] https://blog.ehri-project.eu/2021/03/17/ehri-in-teitok

[4] https://begrenzte-flucht.ehri-project.eu/

[5] https://quest.ms.mff.cuni.cz/teitok-dev/teitok/ehri/index.php

[6] https://lindat.mff.cuni.cz/

[7] https://archive.mpi.nl/tla/elan

[8] https://lite.transkribus.eu/home

You've Been Framed - Partial Audio Matching Functionality to Support Framing Analysis

Philo van Kemenade, Rasa Bocyte, Johan Oomen

Netherlands Institute for Sound & Vision, Netherlands; joomen@beeldengeluid.nl

How can humanities scholars study framing across audiovisual collections? Segments of audiovisual content are being constantly reinterpreted as they are reused and repeated in new contexts. For instance, a clip from a politician's speech from an evening news programme might be repeated in the context of a talk show where it is critiqued by experts; that same segment might be repeated in a radio programme where it serves to support a completely different narrative. When applied across large collections, framing analysis can reveal patterns and biases in the way audiovisual content is reused and topics are represented in media. Some clips are being reused so often that they become canonised in our media memory. Researchers in the field of Media Studies are particularly interested in studying such phenomena. At the moment this type of research is performed manually by watching and/or listening to large quantities of audiovisual content or by combing through transcripts. In the context of large audiovisual archives, this anecdotal approach is not scalable.

Questions around framing are highly pertinent in the context of research environments, such as the CLARIAH Media Suite - a virtual research space for humanities scholars that enables exploration and analysis of audiovisual collections. Currently, the Media Suite provides tools for exploring data and collections based on existing metadata and automatically generated transcripts. In the context of the European research project AI4Media, the authors of this paper have been exploring how state-of-the-art Artificial Intelligence (AI) technologies can further enhance these research possibilities while being aligned with researcher expectations for explainability and transparency when using digital tools.

The focus of this demo paper is on Partial Audio Matching (PAM) functionality developed to support framing analysis in the Media Suite. PAM is a technique that can identify segments in one 'source' audio file that are identical to segments in another 'target' audio file. This matching process is performed on specifically generated audio fingerprint files, which are compact representations of the audio signal. By exporting audio fingerprints for a range of audiovisual items, it becomes possible to trace reuse at archival scale in an automated fashion. The authors have worked on the underlying infrastructure to process both fingerprint extraction and matching at scale, as well as a User Interface that lets Media Suite users select the 'source' and 'target' programs to perform matching on and get results which link back to the matched segments in the originating audiovisual programs.

This demo is a work-in-progress and we are keen to showcase the potential of this functionality as well as discuss its limitations and open questions that arose during the implementation and user evaluation.

Metadata Schema for 3D Data Publication and Archiving

Valentin Grimaud[1], Sylvie Eusebe[2], Sarah Tournon-Valiente[3], Mehdi Chayani[3], Matthieu Quantin[5], Loic Jeanson[6], Olivier Marlet[7], Sylvain Rassat[8], Xavier Granier[4]

[1]LARA, CReAAH; [2]INRAP; [3]Archéosciences Bordeaux; [4]IOGS (Institut d'Optique Graduate School); [5]LS2N; [6]INHA; [7]CITERES / MSH VdL; [8]LAMPEA; mehdi.chayani@u-bordeaux-montaigne.fr

The French National 3D Data Repository, the C.I.N.E.S (National Computer Center for Higher Education) which offers long-term archiving of electronic data, and aLTAG3D - their companion software to document the deposits, use a unique and dedicated metadata scheme based on the work from 2014 to 2017 of the national consortium “3D for Humanities” from Huma-Num infrastructure. This first version is currently evolving to take into account the evolution of deposited data and research domains.

Scientific Domains. The first scheme was focused on archaeology - the main domain in digital humanities that make use of 3D models. One of our goals is to offer such previously cited services to the entire Humanities and Social Sciences. For this purpose, we divide metadata into generic and domain specific ones, such as 3D restitutions or scans of “physical object” in archaeology.

Automatic Extraction. The metadata are also categorized in two groups: those which can be extracted automatically for the data themselves (such as number of points in a 3D file, creation date from EXIF), and those which have to be filled manually. The automatic process simplifies the user work, and the remaining information volume is still quite large, and we are working on either making it smaller or easier to fill. For this purpose, we are pushing forward the experience gained in developing and experimenting the first scheme.

Referencing data. While data redundancy is a guarantee of sustainability, this has an economic and environmental impact. This is even more important for our 3D deposit since, for an open science perspective, we recommend integrating all the documentation, independently of the format and the origin. To tackle this problem, we are integrating into the metadata scheme the possibility to outsource data curation, by using URI to refer to any data present of a sufficiently perennial and reliable infrastructure. We favor institutional hosting over private enterprise. Corresponding requested metadata has also to be harvested. This process is becoming easier with the development of FAIR and institutional infrastructures.

Standards. The FAIR principles also impose the interoperability of our scheme with standard formats. For this purpose, we ensure the compatibility with Dublin Core, CIDOC CRM and with DOI. We also integrate the use of standard thesaurus such as PeriodO, GeoNames, PACTOLS.

3D data definition. The structure of our metadata scheme has evolved due to our experience in understanding 3D data and their usage. Initially, our 3D object was only defined by a colored or textured mesh. However, it seemed important to include any kind of 3D, from volumetric data to CAO (used in the history of technologies). We need also to consider how to take into account the object's structure, appearance properties (reflection, etc.), animations, annotations, and everything that can be attached to a 3D object.

Furthermore, the 3D data are intrinsically linked with the software and the hardware that have been used to process it and/or to interact with (e.g., in VR usage). We thus extend the metadata scheme in order to include them.

Pioneering data stewardship in the humanities: one year of experience at the University of Vienna

Monika Bargmann

University of Vienna, Austria; monika.bargmann@univie.ac.at

The University of Vienna, Austria, supports its researchers in managing their research data in various ways. In autumn 2021, the research data management (RDM) policy was adopted. It codifies RDM support and clarifies responsibilities and opportunities for the institution and the researchers. Summer 2022 saw the establishment of a data stewardship programme. It is based on the “data steward network“ model developed in the FAIR Data Austria project: There is one coordinator at the University Library and embedded data steward:esse:s at faculties/centres – in addition to the central services that are offered by various departments and coordinated by a university-wide strategy board. Two faculties and one centre participate in the pilot phase set on three years.

Data steward:esse:s – in contrast to the data managers in the central services – are the central go-to persons with discipline-specific know-how and have a bridging function between the people who use RDM services with those who design and develop them.

The first faculty to employ a data steward:ess was the Faculty of Philological and Cultural studies. Being the largest faculty of the university, it has about 800 researchers in fourteen departments of philologies and area studies. As part of the humanities, the faculty addresses the world’s cultural heritage in its material and immaterial form – languages, literature, music, religious practices, as well as artefacts and media of every type – and has a strong focus on Digital Humanities.

The poster will report on the experiences as a data stewardess in the humanities - the first of its kind in Austria – and address the following aspects:

* specific requirements of humanities scholars: The first few months in this role have shown that there is a high demand for support in archiving and/or keeping online web applications such as digital editions. Establishing a sustainable work-flow from designing the application to long-term preservation is necessary. When cultural heritage institutions provide material for digital editions, they rely on the research institutions securing long-term access.

* establishing the role within the faculty: defining workflows; self-positioning as a go-to person for all things RDM; introducing discipline-specific training; connecting with other DH or CH institutions; balancing cooperation and boundaries with our “sister faculty”, the Faculty of Historical and Cultural Studies

* possible differences between being a data steward:ess in the humanities and having this role in other disciplines: This aspect includes experiences from the international participants of the Data Steward certificate course that the University of Vienna established in autumn 2022 and from the other data steward:esse:s at the university.

* perception by researchers: Many data managers in the central services also have a humanities background – is there a difference in how researchers perceive them versus the data stewardess? Does it make a difference to have a professional understanding as a librarian and not one as a researcher like many other data stewards have?

This first-hand presentation of experiences with humanities-focussed data stewardship helps others when establishing this role in their institution.

The 50 technological platforms of the RnMSH, a national research infrastructure in Humanities and Social Sciences

Nicolas Thély[1], Serge Wolikow[2], Chiara Chelini[3], Jean Vigreux[4]

[1]Université de Rennes 2, France; [2]Université de Bourgogne-Franche-Comté, France; [3]CNRS (National center of scientific research - France), France; [4]Université de Bourgogne-Franche-Comté, France; Jean.Vigreux@u-bourgogne.fr

The RnMSH is a national research infrastructure in humanities and social sciences which was created in 2006. It federates the 22 Maisons des sciences de l'homme (MSH) located in the major university sites in France.

Each of the 22 MSHs hosts technology platforms with staff (researchers, engineers) and equipment that support large-scale projects in digital humanities, cognitive sciences, statistical surveys, geomatics, 3D, and scientific audiovisual.

Today, there are 50 platforms that bring together 80 qualified personnels (study engineers, research engineers). These are exceptional devices where interdisciplinarity between the humanities and social sciences, and other scientific sectors (including digital sciences) is practiced.

Since 2016, the RnMSH has federated all of these 50 technological platforms into five categories that take into account the nature of the data processed and the methods used. Thus, the 11 Scripto platforms concern textual data, the 8 Audio-Visio platforms deal with audio and visual corpora, the 9 Spatio platforms deal with spatial and archeological data, the 7 Cogito platforms are facilities for experimental research in behavioral sciences, and the 15 Data platforms are devoted to quantitative data.

Our proposal is in line with axis 3 "Advancing digital methods for the analysis of cultural heritage". It can be presented either as a paper or as a panel (with the presence of European colleagues who participate in the RnMSH Scientific Council).

We would like to present the activities of these 5 platform networks, focusing on exemplary programs in the field of heritage resources (funds, collections, databases), methods, and major principles (Open Science, FAIR, IIIF, CIDOC-CRM). We will also discuss strategies for sharing skills and know-how among engineers, and the forms of dissemination and valorization of research results.

The first objective of this proposal is to present the RnMSH, its role in coordinating and supporting scientific projects in the field of cultural heritage.

The second objective is to initiate collaborations and partnerships with other European actors (infrastructures, research programs, etc.). It is indeed the first time that we present at the European level this research infrastructure which is registered in the National Strategy of Research Infrastructures of the ministry of Higher Education and Research in France (MESR).

At the crossing of patrimonial and scientific methods: Methodology and digital tools development

Céline ALAZARD, Camille BERTHON, Jean VIGREUX, Serge WOLIKOW

Maison des Sciences de l'Homme de Dijon (UAR 3516), France; jean.vigreux@u-bourgogne.fr

Since 2004, the Maison des Sciences de l'Homme de Dijon (MSH) has been developing a scientific and methodological method to deal with the heterogeneity and the exponential volume of archives and documentation (from the 19th century to the present day). Through the combined work of researchers and scientific information engineers, the MSH is creating and developing standardized digital resources. All stages of the processing chain are ensured: from the digitization of traditional media (such as paper, microfilm, photographs, etc.) to the processing of native digital documents (text, images, audiovisual for example) and from the conversion of data into intelligent resources thanks to indexing, data mining, etc. until the online release and provision of new digital resources.

To make these data accessible to the public, the MSH has set up its web portal PANDOR (https://pandor.u-bourgogne.fr/) which is a powerful tool for the querying and promoting of digital resources. The objective is to enable a unified access to the corpus created in the framework of research programs carried out or supported by the MSH and to provide, via the OIA-PMH (Open Archives Initiative Protocol for Metadata Harvesting) interoperability protocol, an access to FAIR data stored in different locations. By publishing archival finding aids, library catalogs, digitized and OCRized documents, PANDOR meets the national standards for data processing.

The MSH of Dijon possesses an original approach which mixes patrimonial and scientific approach with two collections labeled "Vine, wine and gastronomy" and "Social criticism and movements". The development of digital collections on PANDOR website is based on the conservation of documentary and archival collections through high-definition digitization (400 DPI in TIFF format), but also by an OCRization process making it possible to consult these documents, which are collected in separate finding aids, in full text search. This patrimonial approach is paired with a complete scientific treatment of the resources, from the constitution of thematic focuses to supplying metadata, including the creation of finding aids.

The choice of archival and documentary material is determined by the MSH's carried research topics in conjunction with the scientific committees that are linked to them. The production of finding aids is done according to ISAD(G) standard while their encoding is made in XML-EAD. Metadata complement this system, this way enriching the general indexing. For instance, persons names can be accurately specified (author, preface, photographer, etc.) as well as those of geographical locations or keywords, and this depending on the relevance for each collection. The interoperability of these multiple description tools makes it possible

to cross-reference full-text searches and searches within notices and, on a macro level, searches within the corresponding thematic axis. An experiment is currently being carried (CONVEX 2 “Collection numérique vitivinicole d’excellence 2” Program : https://pandor.u-bourgogne.fr/fr/convex-2) whose purpose resides in establishing a documentary database in which every publication is provided a scientific notice mentioning the background of writing as well as its historical interest with the ultimate goal of proposing a more comprehensive picture in the corpus-making process.

Sustainable Practices for the Large-Scale TEI Editions at the School of Salamanca Text Collection

Polina Solonets[1], Maxim Kupreyev[2]

[1]Max Planck Institute for Legal History and Legal Theory, Germany; [2]Goethe-Universität Frankfurt; solonets@lhlt.mpg.de

The project "The School of Salamanca" is creating a freely-accessible online collection of texts produced in the intellectual centre of the Spanish monarchy during the XVI and XVII centuries. Currently 33 works have completed the production cycle (out of total 116) which includes TEI XML encoding, HTML export for the online access and full-text search, IIIF presentation APIs, PDF, and RDF export. The development of a sustainable workflow for the project has been influenced by the massive size of our textual collection and its unique features. In preparing editions of Early Modern Latin and Spanish texts it is crucial to take into account their inherent instability, i.e. heterogeneous structures, orthographic, typographical, and punctuation variations etc. Our editorial principles were therefore shaped by the necessity to trace and reproduce the development steps at any given moment; to reuse tools independent of the context and individual texts; to scale the complex processing tasks; to perform constant data quality checks, and to document the requirements and the results.

Salamanca’s workflow shares common ground with both Waterfall and Agile development techniques. The concept of pipeline, inherent to Waterfall, is in the centre of Salamanca’s editorial technique: the production on the edition consists of a number of steps executed sequentially, where the output of each stage serves as the input for the next one.

Digitization → Transcription in TEI TITE → Structural annotation → TEI transformation → Manual and automatic corrections → HTML, PDF, IIIF, and RDF generation.

The advantage of such a predefined sequence is its reproducibility, where each part of an editorial process can be restored at any time. It makes individual work steps traceable and enables comprehensive documentation in form of the program code and editorial guidelines.

In Agile, the software product is built in small chunks, and each of the development cycles includes feature clarification, design, coding, testing and deployment. For this purpose Agile integrates the software development, testing and operations teams in a single collaborative iterative process. In Salamanca’s adoption of Agile practises each of the above-mentioned development stages contains the definition of the requirements, development and quality assurance. This also means that each stage of the production of the digital edition delivers a part of the overall product features, which can be accessed and disseminated.

For example, the QA routines after step 1 - digitization of the print originals - allow us to publish the IIIF presentation manifests even before the TEI transcription starts. IIIF manifests are later enriched with additional data, pertaining to chapters and pagination. The same applies to PDF generation – it was initially intended to be one of the export methods, located at the end of the workflow. Yet, when implemented, it exposed a number of semantic and structural inconsistencies of the source XML. We therefore decided to use PDF earlier in the pipeline as a diagnostic tool and data quality service. In our talk we will show how the sustainable editorial workflow, adapted for processing of large-scale textual sources, translates into the delivery of high-quality data.

An open workflow for the digitisation of built heritage

Stefano Cursi, Letizia Martinelli, Michele Calvano, Filippo Calcerano, Luciano Cessari, Elena Gigliarelli

Institute of Heritage Science - CNR, Italy; stefano.cursi@ispc.cnr.it

Built heritage, in addition to being an element of cultural identity, is to be regarded as a source of inspiration and creativity for present and future generations. It is an ever-evolving repository of knowledge about culture and society, which, due to its importance and uniqueness, must be preserved. Historical buildings are the expression of the traditions and techniques of people and societies that made them and thus must be studied and documented to prevent loss or damage, also ensuring that any restoration, maintenance, and reuse activities are undertaken consciously. To date, existing digital methods are not always appropriate for the built heritage sector. At the same time, older domain experts and physical textual archives collected a large amount of unique knowledge which is only partially transferred to digital resources. Examples are historical construction techniques, terminology for damage pathologies and heritage elements, etc. Transferring this knowledge into new applications and open data spaces could prevent its loss and improve its sharing.

In recent years, the digitalisation of existing knowledge has been widely debated; there are currently many digital resources available to improve modeling capabilities and solve complex problems in well-defined subject areas of analysis, conservation and restoration. However, this research faces the scattering of a wide variety of digital data, that is distributed in numerous independent archives and possesses a great heterogeneity in the type of media and formats used. Among them, the gradual introduction of Building Information Modelling in the field of built heritage - also referred to as Heritage BIM - HBIM - has only partially mitigated the criticality; especially on aspects concerning representation through standard formats, transmission, and consistency of modeled data, especially in open collaborative formats and platforms. A selection of multiple technologies should be made on a case by case basis.

To solve some of the aforementioned problems, this paper proposes an innovative workflow to integrate technologies related to Open Data, Semantic Web, Linked Data, and HBIM , to manage all the knowledge collected, used and shared for the conservation and enhancement of historical buildings in a collaborative platform, including .geometric-constructive aspects from direct analysis up to intangible data from indirect sources and related to traditions and knowledge of arts and crafts. Such an approach, and the models based on it, would support the ideal collaborative space for the management of open built heritage data, in the framework of the European Open Science Cloud (EOSC); thus, they can take on a richer meaning and value, becoming an integral part of an information network that enables their correlation with different disciplines.

The adoption of this open data space workflow, within EOSC, will allow experts and stakeholders in the field to gather the knowledge about construction techniques, spread in different resources around the world. This will enable interdisciplinary exchange and interoperability, as well as promote a culture of data sharing on the permanence and mutations of architectural typologies built over the centuries in Europe, helping to foster a deeper engagement of heritage in modern society.

Cultural Heritage data and digital workflows in the SSH Open Marketplace

Laure Barbot[1], Edward Gray[1], Alexander König[2], Irena Vipavc Brvar[3,4], Mari Kleemola[3,5], Martin Kirnbauer[6], Klaus Illmayer[6], Cesare Concordia[7], Michael Kurzmeier[8], Clara Parente Boavida[9], Stefan Buddenbohm[10], Elena Battaner Moro[11], Christian Schuster[12], Magdalena Wnuk[13], Cristina Grisot[14], Barbara McGillivray[15], Maja Dolinar[3,4]

[1]DARIAH; [2]CLARIN; [3]CESSDA; [4]ADP; [5]FSD; [6]OEAW; [7]CNR-ISTI; [8]University College Cork; [9]Iscte-Instituto Universitário de Lisboa; [10]Göttingen State and University Library; [11]Universidad Rey Juan Carlos; [12]Babeș-Bolyai University Cluj-Napoca; [13]Institute of Literary Research of the Polish Academy of Sciences; [14]University of Zurich & DaSCH; [15]King's College London; laure.barbot@dariah.eu

The Social Sciences and Humanities Open Marketplace (SSH Open Marketplace) - marketplace.sshopencloud.eu - is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: tools, services, training materials, datasets, publications and workflows. The SSH Open Marketplace showcases solutions and research practices for every step of the research data life cycle. In doing so, it facilitates discoverability and findability of research services and products that are essential to enable sharing and re-use of workflows and methodologies.

The SSH Open Marketplace might be regarded as an innovative service. It is, of course, a directory in a classic way, indexing information and allowing researchers to navigate their way to new tools and methods. But it is something new, as well: through the creation and marking of relations between indexed items in a semantically much more meaningful way than the limited textual relatedness of keyword or category classes, it creates contextualization which increases the serendipity in the search of resources.

As a discovery service providing a platform where interactions between arts and humanities researchers, cultural heritage professionals and the computer, information and data scientists enable a collective documentation of uses, sources and tools, the SSH Open Marketplace presents a networking opportunity for both GLAM and academic communities.

The institutional context of the European Open Science Cloud (EOSC) on which the SSH Open Marketplace has been built and is sustained - via the SSH Open Cluster - allows not only for a stable maintenance environment, but also for significant collaborations with Cultural Heritage institutions and infrastructure initiatives such as the European Data Space for Cultural Heritage for example. Within the DARIAH context, it is also a powerful service allowing member countries and cooperating partners to disseminate and promote their collections and the tools and services built on them. Beyond human use, the SSH Open Marketplace also allows for added-value scenarios relying on the machine accessibility of its content via API. The audience of the DARIAH Annual Event can bring their own ideas and discuss them with the SSH Open Marketplace representatives.

This poster and demo will present how the SSH Open Marketplace can provide insights into the use of tools, methods and standards related to cultural heritage collections 'as data', and how it can increase serendipity in the discovery of new methods and standards, by interlinking the resources and describing workflows. Some efforts have already been made to curate relevant resources, for example the "Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools" training (https://marketplace.sshopencloud.eu/training-material/duVII1), but participants of the DARIAH Annual Event 2023 will be invited to contribute to the creation of new or enrichment of existing records guided by the members of the SSH Open Marketplace Editorial Board presenting the poster and leading the demo during the event.

The Association for Research Infrastructures in the Humanities and Cultural Studies – an interface between national and international research infrastructures

Nanette Rißler-Pipka[1], Lukas Weimer[2]

[1]Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen; [2]Göttingen State and University Library, Germany; nanette.rissler-pipka@gwdg.de

Science and research have always benefited from exchange and openness – both within their own discipline and community and beyond. This principle is the basis of the Association for Research Infrastructures in the Humanities and Cultural Studies (Verein Geistes- und kulturwissenschaftliche Forschungsinfrastrukturen e.V., GKFI) which is presented on this poster.

Linking and visualizing cultural heritage data for humanities research

Eva Mayr[1], Gregor Pobezin[2], Matthias Schlögl[3], Johannes Liem[1], Florian Windhager[1]

[1]University for Continuing Education Krems, Austria; [2]Research Center of the Slovenian Academy of Sciences and Arts (ZRC SAZU), Slovenia; [3]Austrian Center for Digital Humanities and Cultural Heritage (ACDH-CH), Austria; eva.mayr@donau-uni.ac.at, gregor.pobezin@zrc-sazu.si

This set of posters brings together complementary disciplinary perspectives on working with cultural heritage data from the European project "In/Tangible European Heritage – Visual Analysis, Curation and Communication" (InTaVia).

GREGOR POBEZIN: OPEN CHALLENGES IN TRADITIONAL BIOGRAPHICAL RESEARCH: NOVEL APPROACHES TO BIOGRAPHICAL RESEARCH WITH INTAVIA

Among the many factors that distinguish biography, we can perhaps most readily point to two that, while usually helpful and indeed essential to biographical studies, can also be important obstacles: tradition and structure. As a noble historiographical genre with a long tradition, biography inevitably produces its texts along structural guidelines. This structure may not be a problem for scholars of cultural or political history - theirs is the task of querying, questioning and imagining: in various languages and across multi-varied texts. But even historians can sometimes ask too little and imagine too much - or vice versa.

For example, three biographical entries (Dizionario Biografico degli Italiani, Deutsche Biographie, Slovenska biografija) about Pier Paolo Vergerio the Younger (1498-1565) all cover the all major events along an almost identical timeline. However, while

one omits almost all of Vergerio's literary production, the other concentrates on his contacts with protestants, as if life before his apostasy never existed. Tradition gets in our way.

Technological advancements have had limited success (searching for Vergerio in Europeana will yield little success) - so far. For wider biographical research, a comparative platform seems a tool sorely needed, since it will finally render the comparative reading of numerous multivariate texts, the visualizing of essential data and asking relevant questions beyond the limitations of narrative structure.

MATTHIAS SCHLÖGL: ON BUILDING A TRANSNATIONAL, MULTIMODAL KNOWLEDGE GRAPH

InTaVia aims at bringing together different kinds of cultural data. To do so, the InTaVia Knowledge Graph (IKG) is built using structured biographical data on cultural actors from four European National Biographies as a base layer (Austria, Finland, The Netherlands, and Slovenia) – which have been converted using the IDM-RDF (https://github.com/InTaVia/idm-rdf).

Second level layers enrich the biographical base layer with data on related cultural heritage objects, institutions, and places from reference resources such as Wikidata and Europeana. Currently the IKG contains data on around 300.000 distinct persons from various periods who had an impact on the history of European countries in one way or another.

However, the IKG has to tackle several problems: for example, the source data is quite unbalanced with respect to the level of detail and structure; the sources contain biased data in various forms (e.g. gender bias, "national myths", …).

This poster gives an overview of the InTaVia System Architecture, presents some statistical analyses on the structure and problems of the IKG and shares the "lessons learned" when integrating large historical secondary data sources.

FLORIAN WINDHAGER, JOHANNES LIEM & EVA MAYR: VISUAL-ANALYTICAL PERSPECTIVES ON CULTURAL INFORMATION

The cultural knowledge graph IKG can be a rich integrative resource for multi-faceted research but is challenging to access for people without technical skills or prior knowledge on the structure and kinds of information available. To overcome these barriers, InTaVia develops a visualization-based interfaces for the visual search, analysis and communication of linked cultural information.

Different visual analytical perspectives (for temporal, spatial, but also relational and categorical information related to the cultural entities in the IKG) support researchers to gain an overview on the structured data available, to see patterns from a distant reading perspective, and to identify interesting pieces of cultural information for close reading. For example, a humanist researcher can search for a cultural actor and will easily find a range of relevant cultural resources: biographical texts, related actors and related objects. They can then select entities of interest and analyze their movements over time, see cultural objects related to biographical events. By making transparent also the provenance of the information, this interface allows to compare and synoptically see different biographies on the same actor.

Our poster walks visitors through the architecture of the visualization-based interface and we offer a demo how to use the developing platform it for searching, analysing, and communicating cultural information.

Challenges and solutions of database archiving of born-digital research data

István Alföldi, Anna Laura Dióssy, Gábor Palkó, Dorottya Szemigán
ELTE, Hungary; szemigan.dorottya@btk.elte.hu

Hungarian cultural heritage experiences a significant loss of valuable research data due to a lack of knowledge, infrastructure, and often even plans by academic institutions to preserve born-digital information. National Laboratory for Digital Heritage (DH-LAB) has launched a project, supporting the open academic research initiative, to archive and provide open access to born-digital research data. DH-LAB plans to achieve a set of digital archiving pilot scenarios starting with a database archiving pilot. This presentation provides a summary of the pilot activities and results.

In the scope of the pilot, three active research databases, created by the Institute for Literary Studies of Hungarian Academy of Sciences, have been archived: Literary and Scholarly Correspondence in the period of enlightenment, Popular prints in the 17-19th century Hungary and Novels in Hungary from the period 1730–1836.

During the pilot we have been testing the results of the E-ARK program by the European Commission. The E-ARK program aims to facilitate interoperability among archival institutions and data producers and impact the development of internationally accessible archives through the provision of technical specifications and tools. E-ARK components are based on international standards and comply with the Open Archival Information System (OAIS) reference model for digital archiving.

All three databases are archived as SIARD files. SIARD is a format designed specifically for database archiving by the Schweizerisches Bundesarchiv and further developed in the scope of the E-ARK program. We have used the E-ARK submission information package (E-ARK SIP) specification and the CITS-SIARD content type specification. From the E-ARK toolset the Database Preservation Toolkit was responsible for creating the SIARD files, the RODA-in tool for creating the information packages and RODA Repository for the long-term preservation and access of the archived information.

All three databases are in the phase of their active lifecycle (e.g. they are still in use by the research community). They follow different internal logic and have different data structure, but they all contain front and inside images of the publications, external links, larger OCRed text, and formatted HTML documents. They are also extended with diverse user interface screens (written in PHP language) to provide user-friendly navigation and access of data. These special features of the databases needed to be considered in our archival and long-term preservation strategy.

We have decided to create one-time snapshots of the databases and consider them as the primary archival data for the pilot. Three different large object (LOB, BLOB) archiving strategies have been tested to collect information to find an optimal solution for storing images or other large objects of the databases in SIARD files. We have also analysed possible options to somehow archive external links and the PHP-based user interface within the E-ARK SIP structure.

In our presentation we discuss the known challenges and best practises of database archiving along with the special considerations required in our pilot. We also highlight the large object (LOB, BLOB) handling options provided by the SIARD specification as used in three different archiving scenarios.

Interactive history - georeferenced and connected archival documents

Agnes Telek
Budapest City Archives, Hungary; teleka@bparchiv.hu

One of our online surfaces for our databases is called “Digital Archives Portal” which is a joint surface for many Hungarian archives for online administration and to reach databases and requests for physical documents. Some digitized documents are available here as well.

The other platform, called Hungaricana is especially for online research. This jointly, the huge database serves a representative number of the Hungarian GLAM sector, integrating multiple databases and document types, offering not only searching tools but even accessibility to the scanned documents themselves:

- more than 100 data provider collections

- multiply types of sources and collections

- integrated and sophisticated search

- georeferenced metadata and access via interactive maps

Here we offer databases and metadata structure also, but this surface’s strongest skill is that a lot of digitized documents are easily available and readable here. This is also the site where we manage to solve the challenge of making connections between multiple various document types and the key connection is their geographical aspect: the lot numbers.

The numbers are showing the amount of the lot numbers on the map, in other words: all the accessible documents through the site, are geo-referenced.

Each and every document can have a geographical aspect. Architectural plans, notarial deeds, land registry files, vintage photographs and postcards. They all can be connected to each other and further, create a real data network. Having quite universal datasets about the documents, even the actors are visible

In the past few months, the biggest development has been achieved on the site: the Budapest Time Machine was released on two different maps back in 2015: one was georeferenced and the other is vectorized. Both offered different types of documents to research.

Recently we renewed the site by unifying the two maps into one: integrating all the georeferenced metadata and documents into the page: all the maps, all the documents, the transparent present maps, added newly digitized datasets and documents - and we also have a groundbreaking new feature.

As a result of our contribution to the Óbuda University Ybl Miklós Faculty of Architecture and Civil Engineering, architecture students have learned how to create 3D models from historical blueprints. They received the digitized drawings from us and delivered the finished building reconstructions in exchange.

Even though, those born digital data types face challenges as they need a completely different handling and interface as the digitized one – and this was one of the reasons that boosted the unification of the different map layers.

Another project is just in the pipeline now: in contribution to the Contemporary Architecture Centre, we will receive data sheets of houses that participated in the annual event Budapest100, and researched mostly in our archive. In this case, our digitally published and physical documents are processed and turned into new documents which are also valid and worthy to preserve in an archival manner.

These two examples mentioned above can cause a data-circulation in the archive.

Introducing the dHUpla (Digital Humanities Platform)

Eszter Mihály
National Széchényi Library, Hungary; mihaly.eszter@oszk.hu

Digital Humanities Platform: Publishing manuscripts using digital humanities tools

The dHUpla, developed by the Digital Humanities Centre of the National Széchényi Library, is a state-of-the-art online platform for publishing written sources held in public collections, providing a unified research environment for the Humanities. The main objective of this poster-presentation is to describe the platform specialized in publishing manuscript collections and the overall infrastructure behind it. At the same time we will be presenting the workflow of publishing a digitized manuscript, which, as a result becomes a digital text in the online environment.. The tools we employ at each step of the publication process will also be described, including our use of the comprehensive platform of Transkribus, the XML editing software framework of Oxygen, and the different stages of manuscript editing.

The uniqueness of our system lies in the GIT-based version tracking tool and the associated set of automated operations, which altogether replace the need for a traditional database. The described workflow offers a solution for the mass processing of manuscripts but is also suitable for the in-depth philological analysis and publication of smaller corpora.

Before publishing manuscripts, it was also necessary to develop a complete editorial system, a framework that enables editing TEI XML formats in a user-friendly way. In addition, such an editorial shell has many advantages. It supports the encoding of texts in a markup language and makes machine validation possible according to specific rules. In the framework, we can link the text to external databases and perform artificial intelligence-based operations as well.

With the framework, the manual and automated sub-processes go hand in hand, always complementing each other, thus creating a high-quality manuscript publication method that exploits the potential of the digital medium.

The primary collection of proper names identified in text editions is the dHUpla Entity Repository, where entities are associated with fundamental data such as a person's occupation, place and date of birth, and death. In some cases project-specific notes are also provided alongside the core data. The editors are assisted in the data input by automatic, AI-based name entity recognition and also have the possibility to link the recognised entities directly to external namespaces.

In the dHUpla publishing environment the text-image linking technology makes it possible to the user to examine the position within the text or the search result in both the transcribed text and the original facsimile. TEI encoding allows the annotation of basic textual phenomena such as insertions, deletions, illegible or missing passages, etc.

The text is also annotated with factual and textological-philological notes and bibliographical references. As a result, the text itself functions as a sort of database, and the data it contains can be queried, filtered, searched, displayed and processed in various ways. And thanks to the GIT-based version tracking, the already published source(s) can be easily corrected if necessary.

In developing the framework, we have prepared a Hungarian translation and a detailed specification for the TEI coding of Hungarian texts, which we publish as a national recommendation.

Visibilities and accountability of contributing institutions to research infrastructures - the case of the DARIAH Service Portfolio

Edward J. Gray[1,2], Laure Barbot[2], Agiatis Benardou[2], Sally Chambers[2], Matej Ďurčo[2], Nicolas Larrousse[1], Francesca Morselli[2,3], Arnaud Roi[2], Nanette Rissler-Pipka[2,4], Simon Saldner[2,3], Andrea Scharnhorst[2,3], Toma Tasovac[2], Erzsébet Tóth-Czifra[2]

[1]CNRS; [2]DARIAH-EU; [3]DANS-KNAW; [4]GWDG; edward.gray523@gmail.com

ERICs (European Research Infrastructure Consortia) are European legal entities to facilitate the establishment and operation of Research Infrastructures with European interest. ERICs are increasingly asked to account for their operations (see debate on KPI in the ERIC Forum). At the same time, web services such as the EOSC Resource Catalogues, the SSH Open Marketplace, and OpenAire, provide possibilities to register and disseminate research outputs, making them ready for automatic harvesting. While the need for standardization of the description of ERIC provided services and content is still very much debated, there are also conceptual dimensions we would like to address in this paper.

ERICs, by definition, create and operate in a landscape of distributed, co-created, network of processes. To summarize, services in the large domain of SSH, are increasingly 'shared' resources. In the spirit of Open Science and the further consolidation of the European Research Area, this interconnectedness can only be applauded. This is how the integrating function of ERICs materializes, and how we avoid the duplication of effort. At the same time, ERICs are also defined organizations with boundaries, and subject to individual assessments. To some extent ERICs are even in competition with each other, for visibility, user groups, in-kind contributions, funds allocated nationally and from the EC.

This paper makes a first attempt to reflect about intrinsic challenges in accounting single nodes in a network of collaborating nodes. While each ERIC already defines its own policy of how to deal with this complexity according to their needs, there is a debate needed as to how to assess contributions to distributed infrastructures in a way which gives credit to all participating while at the same time avoid mis-allocation, double counting and so on, in particular in the context of the development of the EOSC. The matter of ethical, balanced, effective research assessment which has been discussed extensively in the realm of scholarly communication (Leiden Manifesto) also needs to be addressed in the area of research infrastructures. Such a reflection will inform the current discourse of further standardizing the metadata schema which are behind the different catalogues; the search for meaningful Key Performance Indicators (KPIs) for ERICs and the further shaping of the collaboration and division of labor between ERIC's at national and European level.

This paper takes the current discourse around the DARIAH service portfolio as an example/case to highlight these more generic challenges. It also attempts to sort out various sources of 'assignment uncertainty' of contributions (e.g. differences in controlled vocabularies, incompatibilities of metadata schemes, user-generated content, metrics of assignments, and policy decisions).

While we focus the discussion on ERICs, our motivation to present this at the DARIAH Annual event is that it might be of particular interest to see how cultural heritage institutions (libraries, archives, museums…), in their roles as content providers also for ERICs are credited for their contribution to ERICs, and beyond.

Interfacing the BookSampo Knowledge Graph of Finnish Literature for Data Analyses in Digital Humanities

Annastiina Ahola[1], Telma Peura[1,2], Eero Hyvönen[1,2]

[1]Aalto University Finland; [2]University of Helsinki; eero.hyvonen@aalto.fi

In Digital Humanities (DH) research one often needs tools for searching and browsing semantically complex interlinked big data and, at the same time, tools for visualizing and analyzing the data found. This paper first overviews the "Sampo model" [1] for publishing linked data and for integrating faceted search and browsing of linked data seamlessly with data-analytic tools, and then applies the model, as a case study, to design a new DH user interface for the in-use Booksampo knowledge graph. It contains semantically rich comprehensive information about literature published in Finland, based on various databases of the Public Libraries of Finland. The data considered in our work includes 1) novels (80 718 pcs), 2) nonfiction books (1956), 3) authors and other persons (62 207), 4) book covers (112 971), and manifestations of the works (213 877), nearly 9 million triples in total. [2]

We report lessons learned on 1) adapting the open source Sampo-UI framework [3] for searching, browsing, and analyzing the data using faceted search as well as on 2) challenges related to data literacy and quality in using linked data for Cultural Heritage (CH) applications. [4] We also present how the Booksampo knowledge graph has been used in DH research, using the SPARQL API of the BookSampo linked data service and the new portal interface, for analyzing trends in developments in Finnish fiction literature. Analyses are presented about how the geographical attention and diversity of Finnish novels have developed over the past decades. [5,6] The analyses suggest, e.g., that the diversity has increased since the 90's and that the interest of Finnish authors towards foreign countries is increasing, although a majority of Finnish novels still seem to be set in Finland.

[1] Eero Hyvönen: Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web – Interoperability, Usability, Applicability, in press, 2023.

[2] Eero Hyvönen, Annastiina Ahola and Esko Ikkala: BookSampo Fiction Literature Knowledge Graph Revisited: Building a Faceted Search Interface with Seamlessly Integrated Data-analytic Tools. In: Theory and Practice of Digital Libraries (TDPL 2022), Accelerating Innovations Track, Padova, Italy, Springer, 2022.

[3] Esko Ikkala, Eero Hyvönen, Heikki Rantala and Mikko Koho: Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability, vol. 13, no. 1, pp. 69–84, January, 2022.

[4] Annastiina Ahola: Utilizing BookSampo data to develop tools for information retrieval and research purposes. Master's Thesis, Aalto University, Department of Computer Science, forth-coming, 2023.

[5] Telma Peura, Petri Leskinen and Eero Hyvönen: What Linked Data Can Tell about Geographical Trends in Finnish Fiction Literature - Using the BookSampo Knowledge Graph in Digital Humanities. 2022. Abstract under peer review.

[6] Telma Peura: Suomeksi yli rajojen. Kvantitatiivinen tutkimus suomenkielisten romaanien monimuotoisuudesta 1970–2020. Master's Thesis, University of Helsinki, Department of Digital Humanities, Helsinki Centre for Digital Humanities (HELDIG).

Multimodal video annotations as metadata for performing arts documentation

Carla Montez Fernandes
Universidade NOVA de Lisboa, Portugal; carla.fernandes@fcsh.unl.pt

This talk will focus on the affordance of idiosyncratic video annotations to be understood as indispensable metadata when digitally documenting performing arts materials or visual intangible heritage in general.

If the use of annotations is a common practice in several daily tasks, it is because making a note of our thoughts is essential to retaining important information in our memory and to help reconstitute and transmit the knowledge contained in that specific event or activity after they have happened.

Annotating and tagging directly over video in real-time have started (Cabral et al. 2011) to be mostly a personal practice used by those involved in the footage or analytical processes of video documentation for dance, but they have not yet been perceived as complementary metadata by cultural heritage archivists.

Based on our previous experience with the TKB project, from 2009 to 2013, my team has since then defined as one of its objectives to research on the multimodal annotation practices as part of the creative process for dance and choreography. The innovation at that time was to accept individual tagging as indexation cues and means of interaction amongst the artists who were self-curating their works, therefore generating dynamic and idiosyncratic "archives of processes". During TKB, several artists have expressed the need for a handy digital tool to assist them in rehearsal periods where they could take notes on what was being filmed.

Therefore, when joining the subsequent EU projects Culture Moves and Weave, we decided to develop a friendly web-based tool to work as a video annotator in real-time, in fact as a digital notebook to replace paper notes.

MotionNotes is now freely available in the web and is being used by very different kinds of users, from choreographers to ethnographers and CH professionals. Most of the users report that their digital notes (drawings with touch pen, text, marks, links and sound) have become inseparable of the video footage at the end of the creative process, and suggest imaginative ways to embed those personal annotations in future presentations of their work, as well as to include them as an integral part of the documentation and metadata when the work is digitally archived.

Based on this strong and unexpected interest, we would like to discuss with the audience the future usability of idiosyncratic multimodal annotations as great potential for new approaches to documenting, archiving and sharing intangible cultural heritage video content. Moreover, it will be interesting to discuss the possibilities for reusing those annotated files in other contexts and future platforms for transmission and preservation of embodied tacit knowledge that would otherwise be lost in times.

Streamlining poetry research with Averell

Álvaro Pérez, Javier de la Rosa, Aitor Díaz, Salvador Ros, Elena González-Blanco
UNED, Spain; alvaro.perez@linhd.uned.es

Averell is a tool designed to ease working with annotated poetry repositories. It allows users to download and reconcile different annotated corpora, providing a unified JSON output at the desired granularity. This can be useful for researchers who need to work with substantial amounts of poetry data and need to be able to specify the level of detail they want in their final dataset.

One of the key features of Averell is its ability to reconcile different annotations from different corpora. This can be particularly useful when working with multiple annotated corpora that use various TEI entities, as Averell is able to integrate these entities into a single JSON output, or a single TEI-annotated collection. The importance of being able to use different corpora cannot be overstated, as it allows researchers to expand the scope and depth of their investigations. By using Averell, researchers can easily access a wide range of annotated poetry, making it easier to compare different works and authors.

Averell also enables researchers to customize the final generated dataset by specifying the desired granularity. This allows researchers to tailor the output to their specific needs and research goals. For example, a researcher studying the use of imagery in poetry might choose to work with the entire poem as a unit, while a researcher analysing the structure of poems might prefer to work with lines or even individual words. The flexibility offered by Averell enables researchers to extract the maximum amount of information from annotated poetry corpora, making it an invaluable resource for a wide range of research projects.

In addition to facilitating access to existing corpora to researchers, Averell also offers an additional automatic annotation feature for Spanish poetry. This allows researchers to annotate large volumes of poetry quickly and easily, saving time and effort in the process.

To use Averell, each corpus in the catalogue must specify the parser to produce the expected data format. This ensures that the tool will be able to handle and integrate data from a wide range of sources, making it an ideal tool for researchers working with poetry repositories.

Furthermore, another of the strengths of Averell is that it is an open-source platform, allowing anyone to contribute their own corpora to the catalogue. To do so, users simply need to annotate basic information such as the corpus name, poem author, poem title, and the poem text previously split by stanzas. This means that Averell is constantly evolving and expanding, making it an invaluable resource for researchers working with annotated poetry.

Overall, Averell is a valuable tool for researchers working with annotated poetry data. Its ability to reconcile different annotations and allow users to specify the granularity of the final dataset make it a valuable resource for anyone working with large of poetry collections. By providing a unified JSON output, Averell makes it easy for researchers to work with and analyse complex datasets, enabling them to make more informed and accurate conclusions about the poetry they are studying.

Connective explorations with a digital infrastructure for corpus-based research on heritage-centred social media interactions

Costis Dallas[1,2], Ingrida Kelpšienė[1], Rimvydas Laužikas[1], Justas Gribovskis[1]
[1]Faculty of Communication, Vilnius University, Lithuania; [2]Digital Curation Unit, Athena Research Centre, Greece;
konstantinos.dallas@kf.vu.lt

Social media networks have become an increasingly important domain for community interactions and institutional interventions entangling heritage with the formation of contemporary identities and attitudes. In the context of CONNECTIVE Digital Memory in the Borderlands Project, we have been working with digital data to reveal how memory practices on Lithuanian Social Network Sites (SNS) mediated by contested heritage shape cultural identities. Key questions in establishing a digital infrastructure suitable for our evidence-based study were: (a) what should its structure and affordances be? and, (b) what should the properties of entities and relationships representing heritage-related SNS interactions be? Drawing from cultural semiotics and activity theory, we established an approach to research data constitution and infrastructure design based on an event-centric formal conceptualisation of SNS semiotic activity on heritage, memory and identity. This was further used to construct semi-open interviewing scripts and a provisional code system for qualitative data analysis. It also informed the schema of a Neo4j graph database, which we employed to establish a research repository composed of:

(a) digital facsimiles and serialised data streams of SNS conversations (in OAIS terms, Submission Information Package),

(b) graph nodes and relationships of deposits, agents, threads, and messages involved in SNS conversations and represented in message content, enriched with research-driven classifications enabling meaningful queries on SNS interactions related to material and intangible heritage and the difficult past of Lithuania, as well as the people and institutions involved in such interactions (as Archival Information Package), and

(c) export files suitable for researcher-driven qualitative coding and analysis using the MaxQDA Qualitative Data Analysis software, as well as knowledge graphs, tabulations, data summarisations, and outputs suitable for further computational analysis (as Dissemination Information Packages).

A transdisciplinary team of researchers was actively involved in data conceptualisation and corpus constitution. Operating as an SNS counter-archive, the infrastructure offers flexibility to experiment with semantic structure and power to visualise patterns involving conversations, users and communities they belong to, as well as posts or comments they authored, commented on or reacted to, and the historical events, actors, places, and other salient entities these messages refer to. Qualitative data analysis of exported data supports research team members to conduct corpus-based qualitative research on topics as diverse as the Soviet monuments war in Lithuania, the cultural memory of Lithuanian independence in the 1990s, post-World War II memories of resistance, the memory of the Holocaust, metaphor use by SNS mnemonic actors, and contemporary memory wars around Soviet monuments. The experience allows us also to address broader issues related to digital infrastructures suitable for memory- and heritage-related humanities research.

In this presentation we outline the theoretical framework, processes and digital representations established for our research infrastructure, as well as our exploration of its use for computer-assisted research on the mnemonic work of communities and institutions related to the Lithuanian difficult past and heritage. We also offer insights on the significant properties and critical affordances of digital humanities infrastructures as counter-archives capable of supporting the pragmatic curation and fruitful analysis of heritage-related social media data.

COVID-19 Digital Archives in the Latin America

Ian Kisil Marino[1,2]
[1]University of Campinas, Germany; [2]Leibniz-Institut für Europäische Geschichte, Germany; iankmarino@gmail.com

The writing of history refers to remains of the past. Not only the time when documents are produced affects the work of historians; the conditions in which documents are preserved and the way they are accessed and displayed also play a role. These aspects invoke the archive, that is, the entity responsible for the custody and disclosure of documents. Therefore, although sometimes disregarded in theoretical-methodological reviews, they are determinant for the writing of history. Historians devoted to writing the story of COVID-19 pandemic should question the involved archival conditions, which involves a deep presence of digital media. With the spreading of personal recording devices, the number people that can produce and archive digital documents by using hard drives and online storage platforms have multiplied. These numerous collections, usually set on social media, compose archives that are significantly different from formal initiatives of public archives or consolidated personal archives. As suggested by Dipesh Chakrabarty, COVID-19 gives rise to a new experience of global historical event inasmuch as all world was affected, which could raise one's awareness on the transnationality of this tragedy. If the history of COVID-19 pandemic is undoubtedly global, then the places where it was more intensely experienced need to be the object of priority attention for researchers interested in having a thorough understanding of this event. In this paper, I approach the COVID-19 pandemic digital archives with emphasis in Latin America. The main goal is to discuss the impact of digital transformation within archives and sources on a transnational approach, highlighting issues such as crowdsourcing, born-digital documents, inequality, and memory. Firstly, I discuss the emergence of digital memory initiatives focused on COVID-19, showing typological relations that may arise from transnational analyses. Secondly, I dive into some Brazilian archival initiatives with major ethnographic rigor in order to provide more accurate immersion into the agents, platforms, and challenges of this pandemic digital undertaking. Accordingly, I point out the complex situation of public archives amidst the mass of documents resulting from the pandemic.

The DH Course Registry: A bridge between Digital Humanities and Cultural Heritage

Anna Woldrich[1], **Iulianna van der Lek**[2]
[1]Austrian Centre for Digital Humanities & Cultural Heritage (ACDH-CH, OeAW); [2]CLARIN ERIC; anna.woldrich@oeaw.ac.at

The Digital Humanities (DH) are very closely connected to the Cultural Heritage Sector. The development of Information Technology and Digital Humanities have had a great impact on the Cultural Heritage domain (Marsili, G. & Orlandi, L. 2020). DH can provide methods, tools and infrastructures to help cultural heritage institutions digitise their materials and make the data accessible and reusable by scholars in the Social Sciences and Humanities.

In the poster session at DARIAH 2023 we would like to give an update about the latest developments in the DH Course Registry and highlight those courses and programmes that teach digital techniques, methods and tools for the extraction and visualisation of information from cultural heritage objects and creation of of cultural heritage datasets and corpus building.

Utilitarian jack-of-all-trades? Digital research methods in historiographical analysis

Marko Milosev
Central European University, Vienna, Austria; milosev_marko@phd.ceu.edu

As a historian of the 20th century, one of the main challenges is the quantitative analysis of contemporary sources. With the abundance of written material, the use of programming languages like Python, and libraries like NLTK or RegEx has become a new reality for the discipline. Recent advances in ML and AI technology have only stressed the need of digital methodology in research. However, the availability and quality of digital data in larger quantities is one of the difficulties for many humanists. For example, my geographical focus of research mostly rests in the interwar Balkans, and the varying degrees of digitalization and the quality of "digitalized" material is often problematic and may vary greatly depending on the country (advanced on the case of Slovenia, and far from satisfying in the case of Serbia).

Therefore, my paper will focus on the ways in which gaps in the span and type of digitalization can be tackled by using various methods connected to the use of Python and many of its libraries. The work will mostly focus on digital repositories of Serbian national institutions like the National Library or regional and state archives, and their digitalized source material which is often in the form of a simple .jpg or .png file without metadata or any readily extractable text. In regard to that, I will discuss some simple methods which can be employed in order to extract text (web scraping, OCR) and analyze it. The analysis is problematic in terms of languages, since e.g.

Serbian is not supported by the current NLTK version, so I will point to some useful codes for simple programs, like custom sentiment analysis or authorship establishment by using vectorized bag-of-words models.

This approach should serve scholars in similar positions (limited or incomplete digitalized data) and with a modest understanding of Python and programming in general. Also, the examples should aid the partial automation of research and an easier grasp of vast amounts of text in those situations where researchers need to take on multiple roles.

Everyone, everywhere, all at once: transforming cultural heritage data into research data in the Greek National Aggregator

Agathi Papanoti, Elena Lagoudi, Georgia Angelaki, Haris Georgiadis, Evi Sachini
National Documentation Centre (EKT), Greece; apapano@ekt.gr

SearchCulture.gr is the Greek national cultural aggregator, providing access to over 800k items. Data ingested come from a variety of collections of archeology, folklore, history, arts and crafts and span more than 7,000 years of Greek history. SearchCulture.gr is the accredited national aggregator for Europeana.

Addressing metadata heterogeneity in order to be able to provide advanced search, browsing and filtering functionalities to users has been a key target from the start. Controlled linked data vocabularies for item types, historical periods and themes were developed over the course of the past years and are being used for the semantic enrichment of the items' metadata.

In the current paper we present the methodology used over the past 3 years for the process of enriching the aggregated items' metadata with person entities and geographical information from two Linked Data vocabularies that we created for that purpose, one with over 8.200 entities (Greek persons) and another with 11.900 geolocations.

The persons' enrichment process drastically improved SearchCulture.gr by allowing users to easily find all works of a creator and all the items referring a person. We enriched and disambiguated the names in agent fields (creator and contributor) and we extracted personal names from descriptive fields (subject, title and description).

As a result our enrichments are used as search and browsing criteria.The functionalities we developed allow users to retrieve items created by or referring specific persons. The facets can be used in combination, so one can search, for example, for letters exchanged between two individuals or for the "protagonists" of a historical event. The dedicated Persons browsing page provides a list of person entities that include a thumbnail depicting the person, key biographical information, the link to the person's entity in our vocabulary and external links (VIAF, Wikipedia etc). Users can also see the number of items the person has created, those that depict/refer to him/her and the total number of items relevant to this person.

For the geographical enrichment of our content we retrieved the relevant information from place fields (coverage and spatial) and we extracted additional placenames from descriptive fields (subject, title and description). We used a LOD vocabulary with terms linked to the GeoNames geographical database. We painstakingly added alternative names in our vocabulary entries in order to increase the discoverability and to identify and showcase the way a place changed throughout the centuries (ancient greek, byzantine, turkish etc).

As a result all the enriched items are featured on an interactive map that can be configured by types of items, historical periods or persons a.o.

Enabling person-driven and geographical search can be a excellent wayfinding strategy that widens access to cultural content aggregated for communities such as teachers, students, researchers etc. Our ultimate goal is using contemporary semantic technologies in order to increase content discoverability and to provide the means of extracting new leads, connections and insights.

Linking Tangible to Intangible: a Sustainable Workflow for Cultural Heritage and Humanities Data Integration

Emiliano Degl'Innocenti, Francesco Coradeschi, Leonardo Canova, Federica Spinelli
OVI CNR, Italy; spinelli@ovi.cnr.it

The aim of this paper is to present the work carried out by the DARIAH-IT team at CNR-OVI (Consiglio Nazionale delle Ricerche - Opera del Vocabolario Italiano) within the context of several national and international projects such as the SSHOC thematic cluster, the IPERION-HS H2020 project and the Italian Roadmap for the development of the DARIAH national node. The paper will focus on issues related to the collection, analysis and digital processing of resources and collections provided by memory and cultural institutions (GLAMs) and Research Organisations. In particular, we will present and discuss the most relevant findings and experiences and present a sustainable workflow for ingesting, mapping, modelling and visualising data collections from the aforementioned research domains.

The discussion will build on the experience of the RESTORE (smaRt accESs TO digital heRitage and mEmory, http://restore.ovi.cnr.it/) project, focussed on providing access to a vast number of non-interoperable digital resources produced by Archives, Museums and Research Institutes during a couple of decades of work. Another goal of RESTORE was to create the right conditions to promote the interpretation, contextualisation and understanding of digital cultural heritage objects by representing their semantic depth and thus reconstructing the complex network of interconnections with other entities (persons, objects, places, concepts) by restoring the links existing between the tangible (intellectual aspects) and the intangible (physical aspects) dimensions of Cultural Heritage.

The paper will focus on different aspects related to the development of the RESTORE platform, including: the architectural design and the implementation of the semantic knowledge base; the integration of custom components to support domain specific standards and needs; the elaboration of specialised interfaces to display information according to the different partners' needs and resources characteristics; the deployment of tools allowing the users to access the resources in their original contexts or in a new - highly integrated - environment, allowing the transition from a conventional display (lists of data) to a data-graph representation with semantic implications.

Particular relevance will be given to:

Issues in describing and representing tangible and intangible aspects of CH in the digital domain and their implications for the development of innovative, cross-domain, multilingual research tools;

Design and implementation of workflows and tools supporting requirements gathered from the scientific reference communities (SSHOC, IPERION-HS);

Population of the RESTORE infrastructure and its scalability towards the Italian (i.e.: H2IOSC National Recovery and Resilience Plan) and European contexts (i.e.: EGI-ACE) within the DARIAH framework;

best practices to foster interoperability and to ensure reuse and sustainability of data collections and tools.

Development of a prototype supporting specific scientific workflows related to linguistic data (i.e. digital textual corpora and digital lexicography);

OPERAS Innovation Lab: Workflows for Innovative Outputs in Social Sciences and Humanities

Maciej Maryl[1], Marta Błaszczyńska[1], Erzsébet Tóth-Czifra[2]
[1]Institute of Literary Research of the Polish Academy of Sciences, Poland; [2]DARIAH ERIC; maciej.maryl@ibl.waw.pl

Scholars see innovative means of disseminating their work and data as a chance to improve the process of sharing ideas with different audiences, thanks to technological affordances. They understand innovation either in terms of form (novel means of communicating ideas through different media), or access (opening up outputs and making them easily accessible) (cf. Maryl and Błaszczyńska 2021, p. 34). What we consider a "scholarly text" has thus become understood as an expression that can employ different media and engage creatively with underlying data.

However, engaging with novel forms of communication poses new challenges to scholars, who may lack competencies, know-how, or adequate resources to take full advantage of innovative outputs (Tasovac et al. 2018). This presentation will outline the means of support provided by OPERAS Innovation Lab in establishing interdisciplinary, collaborative workflows for supporting innovative outputs in social sciences and humanities (SSH) throughout their lifecycle. The Lab provides guidelines on how to create and sustain FAIR innovative outputs in the SSH.

The poster will showcase three diverse case studies analysed in the ongoing OPERAS-PLUS project. They all aim at addressing the actual needs faced by SSH researchers who decide to use innovative forms of disseminating their output:

(1) The novel publication of project outputs: a scholarly toolkit (SHAPE-ID toolkit) and anthology (Projet Savoirs);

(2) Linking text and data in an interdisciplinary online journal (Journal for Digital History);

(3) Prototyping software services for open science on the example of a recommender system for open access books based on text and data mining (Snijder 2021).

A case study workflow is firstly discussed with the scholars responsible for the project to identify 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. This process is based on a principle of open collaboration whereby different users will be able to engage with the process and thus the workflow itself remains open for contributions from the wider community through feedback and consultation events, such as this presentation.

Each case study will be presented as a workflow, containing solutions and consecutive steps that should be followed by researchers willing to engage with similar formats and face related challenges. It will also feature the support options that could be solicited from various e-infrastructures. The goal of the presentation is to validate the project results by engaging with practitioners at DARIAH Annual Event and receiving feedback on the work.

The Present Future of the Past – Convergent archiving workflow for digitalized print and archived web sources for research continuity

Balázs Indig[1,3], Zsófia Sárközi-Lindner[1,2,3], Mihály Nagy[1,3]

[1]Eötvös Loránd University, Department of Digital Humanities; [2]Eötvös Loránd University, Modern Hungarian History Doctoral Programme; [3]National Laboratory for Digital Humanities; lindner.zsofia@btk.elte.hu

Journalism in the internet age has considerably transformed from what it was in the 20th century, and so have the methods used for its archiving. The credible long-term preservation of printed cultural artifacts and those born on the web are separate research areas. However, maintaining interoperability between source formats and ensuring their continuous researchability has become a major challenge for researchers. We present how two projects, one for digital conversion of prints and one for archived web sources, can mutually enhance each other by using a shared workflow based on similar philological methodology and principles.

Researchers experimenting with distant reading methods are often forced to adopt transient procedures when facing temporal, technical, and resource dependent limitations, that can impair the accuracy of their research results. Press sources, archived either as images or text (extracted with OCR), or as collected web data do not automatically become material ready for research and analysis. The solution is to create uniform, standardized, transparent, and machine-readable versions using a credible methodology in order to produce flexible and extendable collections that can serve multiple research projects.

We developed our web archiving project with extensive philological requirements and the above principles in mind; the project consists of collecting and managing material from online Hungarian news portals (currently ~3 million articles). Our workflow of generating clean and unified TEI versions for initial HTML documents can be altered to enable processing digitized print sources extracted with OCR. We combine the experiences and the toolset from our workflow with standards of digital philology (TEI XML, Schema.org) to build a hand annotated pilot corpus in collaboration with history students, consisting of thematically selected articles of the daily newspaper Szabad Nép, from the second half of the 20th century (currently 98 articles). With the further help of querying and visualization facilities of our ecosystem even non-technical researchers can conduct analysis on these collections.

We store and make publicly available the created datasets in a repository (Zenodo.org) and mark semantic relations between different versions.

Both our original web archive and the new corpus of women's day articles could be analyzed using our trend-viewer that allows filtering and visualization according to both timescales and any other recorded metadata. Our initial findings include that some expressions that were used on women's day in the 1950's are still prevalent today, such as 'mother' (anya), but others, specifically those related to the ideological environment of the time such as 'work' (munka) or 'productivity' (teljesítmény) were used more frequently than they are today.

The ratio of print and digital sources has consistently shifted towards favoring the latter, therefore current methods should be developed to remain applicable in the predictable future. Our uniform, sustainable and scalable workflow demonstrates an example that could fulfill these criteria. For future progress, source (data) preparation must be counted as a valued research practice and collaboration between researchers should become more explicit, efficient, and common place.

Wikibase as an environment for harmonisation of data about past: the example of WikiHum

Adam Zapała, Tomasz Królik

Polish Academy of Sciences, Poland; adam.j.zapala@gmail.com

The increasing number of digital projects in the field of history and cultural heritage over the last decade has led to a situation where very extensive information about the past became accessible online. Unfortunately, most projects use their own data model and do not refer to any reference databases. For this reason, researchers who intend to utilize combined resources are often forced to merge the data by themselves. Hence, data produced by projects often do not enter into general use and are not used optimally. The solution of this problem is the creation of an infrastructure that harmonizes and combines different data sets. An environment perfectly suited for this task is Wikibase.

As part of the DARIAH.Lab project, an instance of Wikibase (called WikiHum) has been developed. First of all, it serves as an interface for the automatic delivery of permanent identifiers (each item will automatically receive a Handle.net identifier). Secondly, it enables data from different projects to be added, stored and harmonized with tools compatible with Wikibase (e.g. Open Refine). Thirdly, the infrastructure also makes data available, both through the SPARQL- endpoint, and also through plug-ins to other software (e.g. TEI Publisher). Furthermore, the database will contain not only various external identifiers, but also the most important data to capture the relations between entities and its change in time.

As part of the project, the most important Polish historical resources for places and people in the past will be added to the WikiHum database and mapped onto the most important international resources (VIAF, Wikidata). These include resources already available online (Historical Atlas of Poland, The Historical-Geographical dictionary of the Polish Lands in the Middle Ages), as well as resources that were previously available only in printed form (Polish Biographical Dictionary, Lists of Officials of the Polish-Lithuanian Commonwealth). Ultimately, the database in question will harmonize a much larger scale of resources. The database will be expanded both with new resources related to people and places, but also with new types of entities (e.g. intellectual or artistic works as well as seals). The aim of the proposed presentation is to share the experiences gathered during the creation of WikiHum.

Diefenbach, D., Wilde, M.D., Alipio, S. (2021), Wikibase as an infrastructure for knowledge graphs: The EU knowledge graph, in: International Semantic Web Conference, Springer, pp. 631–647

Hyvönen, E. (2012), Publishing and using cultural heritage linked data on the semantic web. Synthesis lectures on the semantic web: theory and technology, 2(1), pp. 1-159

Hyvönen, E. (2022), Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web Preprint, pp. 1-16

Scholz, M., & Goerz, G. (2012). WissKI: a virtual research environment for cultural heritage, in: ECAI 2012, IOS Press, pp. Spp. 1017-1018

Smith-Yoshimura K., Washburn B., et al. (2019), Creating library linked data with Wikibase: Lessons learned from project passage, DOI: 10.25333/faq3-ax08.

Shimizu C., A. Elles, S. Gonzales, et al., Ontology Design Facilitating Wikibase Integration – and a Worked Example for Historical Data, arXiv:2205.14032 , DOI: https://doi.org/10.48550/arXiv.2205.14032

Reframing the Italian cultural heritage collections in the era of digital acceleration: MNEMONIC the Italian digital Hub of cultural resilience.

Rosa Tamborrino, Giulia Mezzalama

Politecnico di Torino, Italy; giulia.mezzalama@polito.it

Italy has been the first European country to declare national lockdown being affected by the COVID-19 virus. In March 2020 the need for a digital cultural offer arose. While cultural institutions were closing, some of them started offering new formats to encourage people to stay at home experiencing virtual entertainment. Despite a lack of digitization of the Italian cultural institutions, they 're-opened online' in response to the dramatic situation. As a consequence, an amount of digital, collaborative and creative cultural initiatives appeared.

The crisis has clearly demonstrated the vulnerability of some relevant Italian museums and cultural institutions due to the national digital backwardness (in term of lack of digital services, scarcity of digital collections, delay in digitalisation, lack of experts in digital data management). On the other hand the crisis has also produced a digital acceleration that has inverted the tendency of "normalize" the diverse cultural institutions (museums, libraries, archives) by simply digitalizing their collections without truly producing innovative curatorial digital formats. That is particularly relevant in Italy where unlike elsewhere in Europe, cultural heritage is spread among a multitude of diverse cultural institutions. In this respect the pandemic has fostered innovative curatorial formats envisaging new ways to manage and to democratize cultural heritage, and offering new ways to provide data accessibility (enhancing for instance the social media to reach a broader audience).

In order to better investigate, analyse and understand this phenomenon, in June 2020, during the COVID-19 pandemic, the Politecnico di Torino, undertook a multidisciplinary research project entitled MNEMONIC: the Italian Hub of Cultural Resilience combining different expertise including digital history, digital humanists, information technology experts, communication experts. Using a GIS system of data collection, the MNEMONIC online platform allows the networking of the Italian Cultural Heritage data by mapping the exceptional cultural production during the Coronavirus pandemic.

Conceived as a broad curatorial hub of Italian Cultural Heritage, MNEMONIC makes possible connecting data provided by the marginalized and non-centralised small Italian cultural institutions especially those located in the non-digitalized areas. As a digital database on Italian Cultural Heritage during Covid Pandemic, MNEMONIC not only makes data easily findable, but it also allows multiple qualitative and quantitative analysis.

If MNEMONIC mainly shows how digital technologies can help to reach a broader and marginalized audience, it also reveals that the same digital data can be reuse to fulfil different targets, from people with special needs to scholars and researchers. In some cases, the institutions take advantage of digital technologies, to enrich their own collections by overcoming physical barriers, reaching others global collections. For instance the #BotticelliSpringMarathon launched by the Uffizi Galleries was an international campaign on Twitter entitled #BotticelliSpringMarathon allowing followers and the most important museums in the world sharing related digital works on the Uffizi profile).

By analysing the results of the MNEMONIC research project, the paper presents how the Italian cultural heritage institutions (museums, libraries, archives) took advantage of digital acceleration to entails new curatorial practices, to differently manage their collection and to reframe their idea of accessibility.

Connecting documents - experiencing surfaces

Agnes Telek

Budapest City Archives, Hungary; teleka@bparchiv.hu

The fundamental structure and perspective of archives are on a turning track recently. A classic, physical document only has one possibly right place and order in the archival structure, although the digitized version can be accessed from many different angles, reflecting all its possible connections.

The Budapest City Archives positions itself as an "open-archives" and therefore we operate with a very user-friendly, open-minded attitude on a daily basis. Beyond the physical research room which is wide-open not only for professional/scientific visitors but for one-time visitors/citizens as well, we also welcome school groups to get familiar with the institution itself, with the collection. We have recurrent research groups from universities and NGOs, and those are great examples of long-term and mutually advantageous contributions.

Oftentimes they don't even have to visit us since many documents (blueprints, land registry files, address registries, notarial deeds, postcards and photographs) are already available through the Budapest Time Machine. The surface is integrated into the Hungaricana joint Cultural Heritage Portal of Hungarian public collections — moreover, most of them are on a georeferenced surface, connecting the documents and the metadata through lot numbers.

We have a very efficient contribution with the Óbuda University Ybl Miklós Faculty of Architecture and Civil Engineering, where architect students learn how to create 3D models using our archive blueprints of the houses of Budapest. We also receive their digital house reconstructions and a series has already been uploaded to the BpTM.

The Budapest100, organized annually by the Contemporary Architecture Centre coordinates voluntary researchers and encourages them to use our online databases already in the early phases of the research. As a result of the work of their researchers, data sheets of houses are provided, with metadata, photo series of the present stage of the houses and sometimes historical research and description – mostly based on our sources. We have made an agreement that for using our documents at their events (e.g., installing exhibitions about the stories of the buildings) we receive these data sheets in return and also

integrate them into the Budapest Time Machine. This method ensures the sustainability and the continuity of the mutually rewarding partnership.

The success of the Budapest Time Machine program reached a new level of international professional attention recently. The team of Bp City Archives, Stockholm City Archives and Copenhagen City Archives united in the framework of the program 'CityMemories' to exchange practical experiences. We consider co-working and co-creation essential tools in developing a 'code of practice' in the field of data management.

The challenge in our data-management system was from the beginning to find connections between multiple various document types. The drawback in the universal vision of standardization also lies in the fact, that the objects we strive to describe cover a colourful palette of sources.

Digital thematic research collection - the case of ethnological Collection of research reports

Andrej Gogora, Tomáš Kubisa

Institute of Ethnology and Social Anthropology Slovak Academy of Sciences, Slovak Republic; andrej.gogora@savba.sk

It is well known that in the last two decades, the incidence and usability of digital thematic research collections in the humanities, mainly in top research institutions, has been constantly growing (Palmer, 2004, 348-365). Furthermore, it is increasingly confirmed that the active use of well-built digital research collections should facilitate research practice and support the further development of humanities including ethnology and anthropology (Flanders, 2014, pp. 163-174; Hughes, 2011). The motivation of this paper (and the previous work behind it) is based on the intention to reflect on the persistent unsatisfactory condition of Slovak digital thematic research collections in ethnology (Zajonc, 2006, pp. 30-47) and to fulfill the need for a more systematic and sustainable building of this kind of collections.

The content of the paper can be disciplinary defined at the intersection of archival studies, digital curation, digital humanities as well as the domain of documentation practice in ethnology and anthropology. It follows that it represents a balanced mixture of theoretical, methodological, practical, organizational, and partly institutional knowledge and approaches.

The primary aims of the paper are 1. to analyze the selected significant principles of conceptual preparation and practical building of digital thematic research collections, and 2. to exemplify these principles on the concept and strategy of digitizing and computer processing of the ethnological Collection of research reports located in the Institute of Ethnology and Social Anthropology Slovak Academy of Sciences. The Collection of research reports has been built by the institute staff since 1953 and contains over 1.500 documents recorded in the ethnological field research (a total of 120.000 items in various formats). This unique collection was systematically built and documented by experts in the field of ethnology and comprehensively covers traditional cultural heritage throughout Slovakia (Gogora – Kubisa, 2018).

The scope of the first aim includes, besides principles analysis, highlighting the specificity of thematic scientific digital collections (compared to digital collections as such); determining the differences between key terms (to archive vs. to collect, to preserve vs. to curate, to adapt vs. to standardize). The second aim consists of these investigations: an explanation of the theoretical, methodological, and practical consequences of the particular case that meets the requirements of the digital thematic scientific collection; and an examination of what the effective use of such a collection in domestic ethnological research means for its next life cycle and sustainability.

One of the main contributions of the paper will be to improve the knowledge of professionals from the field of ethnology and anthropology (especially in Slovakia and the central-European community) about the theoretical foundations and practical management of digital thematic research collections.

Creating Digital Assets Which Can Be More Than Just Research Data

Vera Chiquet[1], Marian Manz[2]

[1]Virtual Culture GmbH, Switzerland; [2]University Basel, Switzerland; vera.chiquet@icloud.com

By providing simple instructions on how GLAM institutions can create good 3D models themselves, we enable them to create digital assets that can be used for multiple applications, meet scientific requirements, but also be easily integrated into marketing and other communication contexts.

So instead of each department in the museum managing its own digitized material, they can all reuse the same thing and integrate it into different contexts. This initiative in terms of empowerment and capacity building of the small CHI is what we are doing in the photogrammetry project, a collaboration between the Digital Humanities Lab of the University of Basel (https://www.https://dhlab.philhist.unibas.ch), Virtual Culture (https://www.virtualculture.ch) and the Digital Museum of Learning (https://www.museumoflearning.org, an Initiative of the Jacobs Foundation).

Building an automatically generated rhyming dictionary of Hungarian canonical poetry

Péter Horváth

Eötvös Loránd University, Hungary; horvathpeti99@gmail.com

The poster presents the development of an automatically generated rhyming dictionary of Hungarian canonical poetry. While the creation of explanatory dictionaries presenting the senses of words is a time-consuming task that requires the manual work of many lexicographers, the creation of rhyming dictionaries can be completely automated if you have a sufficiently large corpus of poems. Besides the specific features of the rhyming dictionary, the poster highlights a typical scenario of creating new data from literary corpora. The example of the rhyming dictionary of Hungarian poetry can illustrate the successive stages of such projects, which start with corpus creation, continue with the automatic generation of the domain-specific database and its conversion into formats suitable for different tasks, and end with a web application and with research results.

The rhyming dictionary of Hungarian canonical poetry was generated from the ELTE Poetry Corpus (Horváth et al. 2022). ELTE Poetry Corpus is a database containing all the poems of 50 canonical Hungarian poets with annotations of grammatical properties and poetic features related to sound devices. On the one hand, the poster shows a new algorithm analyzing the rhyme patterns of the poems in the corpus. While the old algorithm analyzed the poems' rhyme pattern on the basis of one rule set, the new algorithm uses multiple rule sets and selects the consistent output. The script generating the rhyming dictionary is based on the result of the automatic analysis of rhyme patterns.

On the other hand, the poster briefly presents the different formats of the rhyming dictionary and the different characteristics included in the dictionary. The rhyming dictionary contains three types of features: (1) features of the words in the rhyme pairs, (2) features of the rhyme pairs themselves and (3) bibliographic information of the rhyme pairs. The features of the rhyming words were extracted from the ELTE Poetry Corpus. The grammatical features were annotated using the NLP toolchain e-magyar (Váradi et al. 2018, Indig et al. 2019) and the phonological features were annotated using a program developed for the corpus. The features of the rhyme pairs were generated by the script creating the rhyming dictionary. These are the following: distance between the two rhyming words, order of the rhyming words, number of lines rhyming with the rhyming pair between the two words of the rhyming pair.

A web application using an SQLite database of the rhyming dictionary is also under development. The poster briefly presents the test version of the query interface. It also gives some examples of research results based on the data of the rhyming dictionary. These results reveal a peculiar pattern in the length of the first and second words of rhyme pairs and illustrate the decline of inflectional rhymes in Hungarian canonical poetry in the mid-19th century.

Archiving a Mailing List. A Case Study of the Katalist

Gyula Kalcsó
National Széchényi Library, Hungary; kalcso.gyula@oszk.hu

Of all born digital objects, email is one of the most challenging to preserve in the long term. Internationally, there is already experience with archiving even large-scale correspondence, but in Hungary the development of a (public) collection-level practice is still to be seen. In order to implement international good practices and results, the Digital Humanities Centre of the National Széchényi Library has launched a pilot project to archive the entire material of the Katalist mailing list (more than 40,000 emails from the start until August 2022), which could serve as a model for further e-mail archiving tasks. Katalist is a list of library and library informatics topics, in operation since the early 1990s. The letters on the list have been preserved since 1997. It has thousands of members, its contents are publicly available and it is an important source of the history of Hungarian librarianship.

In the framework of the process, the entire archive was first discovered using the ePADD software. EPADD is free and open source software developed by Stanford University's Special Collections & University Archives that supports the appraisal, processing, preservation, discovery, and delivery of historical email archives. EPADD incorporates techniques from computer science and computational linguistics, including machine learning, natural language processing, and named entity recognition to help users access and search email collections of historical and cultural value. The archive was then packaged with the Mailbagit software into an OAIS-compliant AIP package, which includes, in addition to the standard EML, other formats suitable for long-term preservation (HTML, TXT, PDF, WARC) and the extracted attachments from the emails, in accordance with the BagIt package format specifications, together with the collection-relevant metadata for the emails. The Mailbag project is a specification with an open source tool for preserving email archives using multiple formats, such as MBOX, PDF, and WARC developed by a consortium lead by M.E. Grenander Department of Special Collections & Archives, University at Albany, SUNY. The Mailbag proposal is an extension of the Library of Congress Bagit specification. The archived material will be searchable through a Solr-based search engine.

During the demonstration, one can test the discovery and processing capabilities of the ePADD software on the Katalist mailing list, inspect the OAIS-compatible packages created with Mailbagit software (including testing Mailbagit on the fly, e.g. creating standard derivatives like WARC, PDF etc. from emails), and try out the Solr-based search engine.

The influence of editorial work in authorship studies of 19th century Hungarian short stories

Mária Regina Timári
ELTE, Hungary; mariazimanyi@gmail.com

While it is still widely accepted in the field of computational stylistics that there are unique patterns of individual language use, so-called authorial "fingerprints". However, the metaphor of this term can falsely suggest that patterns specific to a particular author can be objectively extracted from texts. The construction of this authorial fingerprint is a much more complex, creative digital humanities task. The main issue is finding a method that can detect 'patterns', always interpretable only in comparison with other authorial texts, on the basis of a selection and combination of linguistic markers that can be interpreted statistically in the text and then of various similarity calculations based on these markers.

Given the size of the corpus of texts under study and the complexity of the linguistic markers and similarity calculations, this is now unthinkable without the use of computer algorithms. However, even though stylistometric and computer-based authorship analyses are becoming more and more widespread, and that the last two decades have witnessed intensive technical and methodological changes in these fields, they are still not widely used in Hungary, so further studies are needed to determine which methods, distance measures and stylometric tools would be most effective in clustering Hungarian texts.

It is also unclear, for example, how the linguistic condition of the texts under study may impact the results of the research. We do not know, for example, whether the result of clustering 19th century Hungarian literary texts is distorted or not, if we examine the texts not in their current state, according to the newer edition, but using their old editions, where the orthography was not yet consolidated.

In my research, I am therefore investigating whether the normalisation of texts or the correction of their grammar affects the results of the clustering in any way. Of course, this would require having both versions in computer-readable format, which is

problematic in the case of older editions, because most of the 19th century Hungarian literary texts are only available in digitized form in their recent editions, and older editions are generally not accessible. To overcome this challenge in my research, I am using the versions of older editions of short stories which I originally accessed in PDF format, and then digitised, and corrected using standard spelling rules.

Cultural heritage geodata: The Warsaw Statement on the provision of geographical data

Francis Harvey[1], Tomasz Panecki[2], Marta Kużma[3], Wiesława Duży[2]

[1]Leibniz Institute for Regional Geography, Germany and University of Warsaw; [2]Polish Academy of Sciences and University of Warsaw; [3]University of Warsaw; f_harvey@leibniz-ifl.de

Cultural heritage geodata presents an important resource for cultural heritage institutions. It helps people connect places of the past with locations in the present, understand different ways of knowing where culture happens, and relate their experiences to the experiences of others and the past. Beyond scanned artefacts that show historical and cultural geographies, the many possibilities and vibrant interests in places make historical mapped data a central information resource to connect experiences and knowledge of the present with evidence of the past. The workshop, held from 23 to 25.4.2022 with the support of an Europeana Research grant, brought together people involved in collecting, curating, presenting, and researching historical digital geographic data from digitized maps and other sources held and curated by the GLAM (galleries, libraries, archives and museums) sector. The outcome is a document, the "Warsaw Statement", with points that offer guidance on making cultural heritage geodata more accessible through open access. The statement can help inform programmatic endeavours to provide digital geographic data and covers three topics: 1) Citizen science as a guiding concept; 2) The importance of institutional infrastructure; 3) Spatial data in institutions and research in alignment with European activities. The presentation covers these points in detail for cultural heritage institutions and researchers. The Warsaw Statement on Spatial Data in Cultural Heritage is available online at https://zenodo.org/record/6814297#.Y-DySC8w2MA.

From Musical Notes to Medieval Codex: What the Open Research Case Studies Reveal about Humanities Data at the University of Leeds

Dorottya Tamás

University of Leeds, United Kingdom; d.tamas@leeds.ac.uk

This paper presents the outcomes of the Open Research Case Studies conducted at the University of Leeds over the period of May 2022 to early 2023. Funded by Research England, I have conducted interviews with colleagues from different faculties, schools and services across the University of Leeds to raise awareness of open practice across disciplines and career stages. Since I am a researcher in the Humanities, I have been specifically interested in the way in which Arts and Humanities scholars think about research data, use data, and use open and FAIR data as part of their open research practices.

My presentation brings attention to the specificities of Humanities and practice-research data, particularly on how digitised and born-digital data is used by researchers at Leeds. I use the Open Research Case Studies to demonstrate the complexities Arts and Humanities researchers need to navigate from copyrights to 'data' being an insufficient terminology that does not describe the vast type of evidence they use. I bring in the perspective of the Library and Research Support Services, including the Digital Content and Copyright Manager at Leeds who described the digitisation strategy of Leeds Special Collections. I will also introduce the SPARKLE (Sustaining Practice Assets for Research, Knowledge, Learning and Engagement) project-in-development, which wants to find ways to better manage and curate practice-research data 'which may include text, but also image/video/audio/software, and other less common mediums'.

My paper investigates how researchers in the Arts and Humanities can benefit from open research, particularly practices around data sharing and management for a sustainable, digital, engaged, and equitable knowledge production. Ultimately, this presentation will showcase the Humanities perspective from the Open Research Case Studies and demonstrates the challenges academics and practitioners face as well as the benefits of rethinking what data means in the Arts and Humanities and how can we manage and share the multitudes of evidence researchers work with.

Digitising the values of cultural artifacts

Maria Dagioglou[1], Dora Katsamori[1], Georgios Petasis[1], Alfio Ferrara[2], Stefano Montanelli[2], Theodore Grammatas[3], Maria Dimaki Zora[3], Aikaterini Diamantakou[3], Marco Berni[4], Elena Fani[4], Carla Murteira[5], Alba Morollón Díaz-Faes[5], Elena Aristodemou[6], Marko Kokol[7]

[1]National Centre for Scientific Research "Demokritos", Greece; [2]Università degli Studi di Milano, Italy; [3]National and Kapodistrian University of Athens, Greece; [4]Museo Galileo - Istituto e Museo di Storia della Scienza, Italy; [5]NOVA University of Lisbon – School of Social Sciences and Humanities, Portugal; [6]Fairy Tale Museum, Cyprus; [7]Semantika Research, Slovenia; mdagiogl@iit.demokritos.gr, dkatsamori@iit.demokritos.gr

Have you ever explored how your visitors perceive values associated with the collections of your museum?

Have you tried to rethink your collection from a 'values-perspective'?

Do you want to create novel value-centric activities?

VAST is a research and innovation action in the context of Horizon 2020 Curation of digital assets and advanced digitisation actions that aims to study the transformation of moral values across space and time and bring them to the forefront of advanced digitisation.

The project will trace and inter-link the values…

…of the past through the analysis of collections of narratives, such as theatrical plays, fairy tales, and scientific documents, that come from different places and from significant moments of European history.

…of the present through the collection and digitision of how values are conveyed today and of how the audiences experience and perceive the communicated values.

At the core of the project lies the VAST Digital Platform. What can VAST Digital Platform offer to you?

1. A methodology for capturing you visitor's experience
2. A toolkit of value-based educational activities
3. Tools for conducting online user surveys
4. Tools for annotating artefacts with values

This poster proposal aims at demonstrating the offerings of the VAST platform, along with some exemplar activities that has helped us in capturing and digitising values, as perceived by audiences engaging in activities like visiting a museum, or participating in an educational activity.

Lessons Learned from History of Sofia's Street Names Project

Ivaylo Nachev

nachev@balkanstudies.bg, Bulgaria; nachev@balkanstudies.bg

The paper will present an ongoing project exploring street names changes and history of the streets in the Bulgarian capital city of Sofia since the late nineteenth century. The project that is affiliated within the E-infrastructure CLaDA-BG aims at disseminating its results through digital tools to both professional and wider non-expert public. Exploring a specific case but also suggesting the applicability of the model in other contexts, the project treats frequent street name changes in the Bulgarian capital city as a complex phenomenon reflecting major political and cultural shifts that allows simultaneously to present these broader processes in an accessible and meaningful way. The main product is a web site that is based on an extensive data set featuring streets names and other data throughout history. The onomastic data set is supplemented with other information including pictures, map fragments and original archival documents. The paper will elaborate on the opportunities provided by digital tools for presenting research results, visualization, reaching new audiences etc.

In between data dimensions - data for, in, and after research

Liisa Näpärä

National library of Finland, Finland; liisa.napara@helsinki.fi

As the volume of data increases and research methods develop, there are becoming more possibilities to do research based on cultural heritage data. The environment is still evolving, and many questions are being debated and some of them are just beginning to emerge for discussion. This paper presents the complex role of data and analyses what kind of role cultural heritage organizations can play across the different dimensions of data now and in the future. The analysis is based on the collaboration experience with researchers, other cultural heritage organizations, and data stewards. The perspective is provided by the National Library of Finland (NLF).

Creating a sustainable workflow from cultural heritage master data to research data requires understanding both data itself and the researchers' needs. It also requires understanding the direction in which the archiving of research data in humanities and social sciences is going and what kind of infrastructure is available.

The NLF offers researchers many of its collections as data. The purpose is that the data is managed by following the FAIR (findable, accessible, interoperable, and reusable) principles. There is an ongoing FAIR project with an aim to assist the personnel of the NLF to understand how FAIR principles can help improve the usability of different data types such as data produced in the digitization process, licensed, or harvested as born-digital to the web archive. This also means that the FAIR principles should be adapted wholesomely considering systems and users, not only data.

Identifying all available data sources, and how they can be used, is not done without active collaboration with the researchers of humanities and social sciences. In practice, the NLF participates actively in research-driven projects to be able to develop its data services to meet the needs of researchers. This is a mutual way of closing the gaps and bridging the sense-making of data as approaches towards data vary between participants. So far, the collaboration has increased knowledge about data usage possibilities and raised discussion about data description and research data metadata considering both cultural heritage data for research and research data after the actual research. At this end of the data life cycle, the research data should be archived in accordance with FAIR principles. However, not all products of research projects have a suitable repository to ensure sustainable FAIRness. This has emerged pilot collaboration to integrate a research database into the library's systems. Further collaboration and discussion are still needed as well as resources to meet the researchers' needs and to serve in the current and future data-driven research environment.

Miklós Bethlen (1642-1716) and the "Early Modern Hungarian Political Dictionary"

Bence Vida

Eötvös Loránd University ; Faculty of Humanities, Hungary; vida.bence@btk.elte.hu

Anyone who decided to enter Hungarian politics in the 17th century had to face a fragmented environment in the sense of power, culture, and religion as well, which included many interest groups that were sometimes loosely allied, sometimes more closely allied, or even in conflict with each other. Miklós Bethlen, Chancellor of Transylvania (1642–1716), was aware of the power of the written word, and from his peregrination (a.k.a. student years) he consciously took every opportunity to enrich his education, expand his political vocabulary and become able to speak in Hungarian, Latin, French and German, aiming to spread his agenda to different political groups whether it be Catholic Habsburg, Turkish, Protestant-European or Hungarian. He used his network of correspondence not only to promote his private interests effectively, but also to serve the political interests of Transylvania and Hungary—although he acted most effectively when these interests coincided.

In Bethlen's case, the shaping of collective memory and public opinion was transformed into a political tool, and he sought to influence his political partners by presenting historical facts as arguments and by effectively adapting the views of contemporary political thinkers. All this, together with the religious arguments that were also effectively mobilised, formed part of his political toolkit, the outline of which forms the basis of this presentation. In my presentation, I will identify the basic concepts he used to win over his audience in different situations and outline the possibilities of how to map and, in a longer-term context, encode a political language containing several symbolic expressions and metaphors.

In my research, I analyse Bethlen's activities in the cause of the galley-slave Protestant ministers, i.e. the Epistola Nicolai Bethlen ad Ministros Exules tam Helveticae luam Augustanae Confessionis, ex Hungaria per hodiernam persecutionem ejectos, and his related private letters, notes and pamphlet with the title Apologia ministrorum. I will investigate the types of arguments and metaphors that can be identified in these private and public political writings, and I attempt to build a data model based on them that can be used as a starting point for topic modelling or for building a metaphor database based on semantic data techniques. My aim is to create a data model that can be incorporated into the analysis tools of historical documents published digitally on ELTEdata.

Sources (Published and Unpublished):

Bethlen, Miklós: Apologia ministrorum Evangelicorum Hungariae… 1677 (1678) Tiszántúli Református Egyházkerület Nagykönyvtára, Debrecen, Rmk929

Jankovics, József (ed.): Bethlen Miklós levelei I – II. [The letters of Miklós Bethlen] Akadémiai Kiadó, Budapest, 1987.

References:

Brugman, Britta C., Burgers, Christian, Steen, Gerard J.: Recategorizing political frames: a systematic review of metaphorical framing in experiments on political communication, Annals of the International Communication Association, 41:2 (2017), 181-197, DOI: 10.1080/23808985.2017.1312481

Dorst, A. G.: Metaphor in Fiction: Language, Thought and Communication. Semantic Scholar, 2011. URL: https://www.semanticscholar.org/paper/Metaphor-in-Fiction%3A-Language%2C-Thought-and-Dorst/1f68ed74c9287e2e93fad4341a493f460e6fb532

Hampe, Beate ed.: Metaphor. Embodied Cognition & Discourse. Cambridge University Press, Cambridge, 2017.

Semino, Elena – Demjén, Zsófia eds. The Routledge Handbook of Metaphor and Language. Routledge, New York, 2017.

Historical sources and semantic database development

Ádám Sebestyén

Eötvös Loránd University, Hungary; akkon88@gmail.com

The semantic database, ELTEdata, developed by the National Digital Heritage Laboratory of Eötvös Loránd University, aims to organize the sources of prosopographical, bibliographical and other historical research groups into a semantic data network. Following the data structure of Wikidata, items consist of semantic statements, which can be described in the form of property-value pairs. The SPARQL semantic query language enables complex searches on the database and offers various ways for visualizations. It is important to know that by means of entity linking, it is possible to connect semantic entities to the text fragments of a digitalized source edition. We used a semi-automatic name entity linking method in order to identify the personal and geographical names in the texts and add namespace identifiers to them. Currently, ELTEdata contains more than 15000 items. Besides the possible data enrichment of the three original collections, the forthcoming work focuses on the integration of new projects in our database. Within the framework of a longer cooperation, ELTEdata relates closely to the database developed by the Institute for Literary Studies of the Research Centre for the Humanities (ITIdata). Both projects use the wikibase-software and the semantic statements, entities and properties are also connected. My presentation is centred around the mapping of large datasets, first of all repositories, regarding the history of education. Although the sources are diverse (datasets of the early modern schooling history and prosopography of university teachers from the late 19th, early 20th century period), the similarities of the data structure and the possible common patterns of the sources facilitate the mapping process. Besides the current state of the project, my presentation also pays attention to visualization (in order to analyze geographical and social mobility and the development of social networks based on correspondence) and the ways of a semi-automatic data entry. Because the data import from other namespaces into the aforementioned ITIdata was successful, similar data imports can be realised in ELTEdata as longer-term developments. I will also present, how to run complex queries in SPARQL connecting diverse platforms. For instance, users can search in Wikidata and ELTEdata at the same time, in order to collect items, corresponding the conditions of the query. This way, the separate databases can be connected in the queries as well.

Encoded Archival Description (EAD) and Records in Contexts (RiC): a Cultural Heritage Data Ecosystem for Humanities Research

Francesco Gelati

Universität Hamburg, Germany; francesco.gelati@uni-hamburg.de

The Encoded Archival Description (EAD) [1] has been promoted since the 90'ies by the International Council on Archives (ICA) as the only international digital standard for describing archival holdings. It is XML-based, i.e., it is written in the utmost preferred language for cultural heritage data. EAD files can be automatically harvested by means of e.g. the Open Archival Initiative Protocol for Metadata Harvesting (OAI-PMH). EAD metadata fields can be used to populate e.g. a SQL as well as a NOSQL database. EAD is successfully mapped to among others MARC21 and Dublin Core in order to enhance Galleries, Libraries, Archives and Museums (GLAM) intercommunication and data FAIRness. Free open-source tools are available for managing, online-publishing and visualising EAD datasets, which can be ad libitum gathered, processed, queried and stored by digital humanists.

EAD is a data model which procreated in 2016 to children: Records in Contexts Conceptual Model (RiC-CM) [2] and Records in Contexts Ontology (RiC-O) [3]. These two semantic artifacts encompass EAD's siblings Encoded Archival Context (EAC) for archival authority records and Encoded Archival Guide (EAG) for institutions with archival holdings and make EAD compatible to Linked Open Data.

Should we still agree that “les archives [sont] [l]es greniers à faits [des historiens]” [4], than we should also see in archival digital standards and particularly in EADs the foundations for the management of historical research data, together with the more famous standards TEI, FRBR and CIDOC-CRM. Archival holdings are cultural heritage, whereas archival descriptions are cultural heritage / humanities research data. Archival metadata exemplary shows how deep research data and cultural heritage fusion.

EADs and RiC are an utmost powerful tool for research in the humanities, whose wide and rich diffusion cannot be doubted: the French National Archives make e.g. online available 8 millions [4] descriptions of archival units that can all be exported as EADs; the same does the European Holocaust Research Infrastructure (EHRI) portal with its 325’000 archival units [5] from all over the world. Many smaller archival institutions too work hard in order to be able to generate their own EADs. Yet, EAD and RiC have got so far little attention in digital humanities scholarship, with notable exceptions [7].

I would like to hold this communication in order to raise awareness and interest among cultural heritage professionals for EAD and archival metadata standards.

Notes

[1] https://www.loc.gov/ead/

[2] https://www.ica.org/en/records-in-contexts-conceptual-model

[3] https://www.ica.org/en/records-in-contexts-ontology

[4] Lucien Febvre. (1948). Sur une forme d’histoire qui n’est pas la nôtre. Annales. Histoire, Sciences Sociales. 3-1, 21-24. See also: Etienne Anheim. (2019). Science des archives, science de l’histoire. Annales. Histoire, Sciences Sociales, 74(3-4), 507-520. doi:10.1017/ahss.2020.56

[5] See: https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/recherche/ir/rechercheGeneralisteResultat.action?formCaller=GENERALISTE&searchText=

[6] See: https://portal.ehri-project.eu/

[7] I am thinking to Georg Vogeler. See Georg Vogeler. (2019). Das Digitale Archiv: Der Computer als Mediator, Leser und Begriffsbildner, in Klaus Kastberger, Stefan Maurer and Christian Neuhuber (Eds.), Schauplatz Archiv : Objekt – Narrativ – Performanz. DOI: 10.1515/9783110656725-006

The First Line of Digital Humanities

Ari Häyrinen, Irene Ylönen

Open Science Centre/University of Jyväskylä, Finland; ari.hayrinen@jyu.fi

There is a long tradition of solitary scholarship in humanities and many humanities researchers are trained to work independently. This mindset is often reinforced by academic reward structures that value individual achievement and original ideas. This poses a challenge for digital humanities. Utilizing the tools used for processing and analyzing materials –including cultural heritage data - in digital humanities requires a certain level of technical expertise. Without a research group with different kinds of experts in it, individual researchers must be able to run these tools themselves.

The objective of the first line of digital humanities is to offer a web-based tool that facilitates entry into the field also for individual students and researchers. The tool serves two purposes. Firstly, it acts as a generic processing tool for levels commonly used in digital humanities, such as extracting text and images from PDFs, utilizing OCR for image sets, removing stop words, and detecting the language of a document.

The second objective is educational. By allowing students and researchers in the field of cultural heritage to easily test different ways of processing materials with their own data, users become familiar with these tools, and they can determine if these tools will be useful for their research or not.

One of the key requirements for this type of tool is transparency. Instead of functioning as a black box, the tool must clearly display how, and which tools were utilized at each stage. For instance, in the case of OCR, the output should include the version of the OCR engine and the parameters used. This transparency guarantees that the processing of material can be accurately documented in research papers, enabling the analysis to be replicated in other environments. Additionally, transparency permits researchers to surpass the limitations of the first-line tool by utilizing command line tools.

One of the necessary features of this type of tool is the ability to perform multiple variations of the same process. Optical character recognition is a prime example of this, as the typical OCR process often involves multiple attempts with different settings. The tool should enable the creation of multiple processing pipelines with different configurations and allow for easy comparison of the results to determine the best option.

To fulfil the need of accessible tools for digital humanities, we have initiated the development of a first-line tool called Messydesk. It features a graph-based approach with node-based user interface where import nodes, file nodes, processing nodes, and output nodes form a visual graph. The graph-based approach enables the visual chaining of processing stages while maintaining a clear understanding of the data flow and the origin of the data.

Integrating Museums Activities on Intangible Cultural Heritage with Data-Driven Research on Early Modern Scientific Texts

Natacha Fabbri[1], Francesco Barreca[2], Stefano Montanelli[2], Marco Berni[1]

[1]Museo Galileo - Institute and Museum for the History of Science, Florence, Italy; [2]University of Milan, Italy;
f.barreca@museogalileo.it

Since 2020, Museo Galileo of Florence and University of Milan have been involved as partners in the VAST (Values Across Space and Time) project. VAST has received funding from the European Union’s Horizon 2020 research and innovation programme, and is led by NCSR Demokritos, Athens, Greece. The main aim of VAST is to investigate the spatio-temporal transformation of moral values, and how they are tracked in different mediums, in order to exploit the digitized tangible and

intangible cultural assets thus supporting and empowering European cultural heritage institutions and practitioners. This is done in two ways: 1. By externalizing values implicitly present in resources from the past (Past of Values); 2. By digitizing values as they are perceived by the general audience today (Present of Values). VAST focuses on three main sources of values: 1. Greek tragedies (Pilot 1); 2. Seventeenth century works of natural philosophy or strongly influenced by natural philosophy (Pilot 2); 3. Traditional European storytelling (Pilot 3). Within VAST, University of Milan and Museo Galileo are together in charge of Pilot 2 (Past of Values and Present of Values, respectively). Drawing on the Museo Galileo's twenty-year experience in history of science educational activities – which rely on the strong interoperability between its collections of scientific instruments, library, and digital archives – and on the University of Milan's expertise in semantic web technologies and ontology design, this joint effort produced an effective, data-driven approach to research, digitization, and (co)creation of educational programs for museum visitors. This approach has been developed with insights from many perspectives, and features data gathering through annotation, the creation and implementation of educational activities based on the gathered data, and an iterative-integrative ontology design approach to ensure that different and possibly conflicting interpretations of values are represented and coexist in a semantic graph. The approach is scalable, sustainable, and reusable. It relies on advanced digitization and enables collaborative study, co-creation, and continuous capture and digitization of experience.

For Museo Galileo, which has already begun to move toward Open Data and Open Science, the VAST project represents a step further: the annotation system of seventeenth century scientific texts and images provided by University of Milan makes these data sets available to a public of non-scholar and fosters the dissemination of cultural heritage, thus facilitating the dialogue between past and present. The creation of specific interfaces for an integrated system of annotations and educational programs needs an informatic infrastructure capable of optimizing the browsing experience and giving the audience the possibility to interact and provide feedback. Referring to a specific educational activity that was designed for the VAST project, we shall clarify how and to what extent the VAST project has been integrating the Museum educational program, by implementing the possibility of making cultural heritage data easily accessible to a wider range of audiences, findable and interoperable, reusable during both physical and virtual visits, and sustainable.

Exploring interview collections with the help of named entity linking and topic classification

András Micsik[1], Balázs Pataki[1], László Kovács[1], Júlia Egyed-Gergely[2], Judit Gárdos[2], Anna Horváth[2], Róza Vajda[2], Enikő Meiszterics[2]

[1]ELKH SZTAKI, Hungary; [2]ELKH TK, Hungary; micsik@sztaki.hu

In this paper we present an approach to support the processing of long in-depth social scientific interviews and to enable these texts for secondary research. However, the problem area and the implemented solution are generalizable to many other tasks where long texts have to be explored without reading them thoroughly, and relevant text parts have to be found based on complex criteria.

In our approach, the processing of texts starts with digitization (if needed, in case of analogue, heritage interviews) and then they are cut into manageable sized blocks such as pages, which will serve as units for analysis and reuse in other research. These text blocks get automatically assigned keywords, which are consequently mapped to topics using NLP tools. The topic thesaurus specifically created for this task is a hierarchical structure of terms based on the ELSST (European Language Social Science Thesaurus by CESSDA), which contains all major sociological research areas and fields of inquiry. Furthermore, named entities are extracted from texts and, where possible, linked to Wikidata (wikification). Through Wikidata we also collect links to other important registries such as GeoNames, ISNI or VIAF.

Based on these preprocessing steps, an exploratory user interface was built to facilitate searching for pages/blocks in the interview corpus related to the given topics. After selecting a page/block, the researcher sees the text with named entities, keywords and topics highlighted. This will help her decide whether to use the text in her research.

We have also experimented with different methods of mapping the contents of the archives we are working with. We have built various diagrams to characterize interviews or interview collections, and implemented an interface where researchers can prepare these diagrams themselves, for their own use, without any programming.

The Intersection of Digital Libraries and Digital Humanities: the role of the embedded librarian for multifarious DH needs

Anna Maria Tammaro[1], Klaus Kempf[2], Márton Németh[4], Plamen Miltenoff[3]

[1]Università di Parma, Italy; [2]Bayerische Staatsbibliothek, Germany; [3]University of Minnesota, Duluth, United States; [4]Blinken OSA Archivum, Budapest, Hungary; nemethm@gmail.com

For more than a decade the digitization of libraries and their development of digital services for users have been growing up and developing parallel to the evolution of digital humanities (DH) research, or the application of digital research methods to humanities disciplines, and its growing establishment as a scientific field. More recently on both sides of the pew a partnership between digital humanities (DH and digital libraries (DL) is being forged, coagulating in a demand for DH centers within academic libraries and an increase in the call for "Embedded Digital Librarians". The purpose of the Poster is to present socially grounded approaches to understanding DLs. The objectives are: to identify and discuss major issues that arise from the intersection of DLs and DH and, more generally, from the social nature of DLs; and to consider implications for the design and evaluation of DLs.

The Poster's theme aligns well with the Dariah 2023 Annual Meeting theme 1 "Sustainable workflows for data management and curation" and offers an opportunity to focus scholarly attention on the social, cultural, political, and economic shaping of digital libraries as sociotechnical systems and their consequences.

The authors discuss their current positions and how their experiences have had an impact on DH research. They also suggest directions for future work among researchers and practitioners.

CLARIN Resource Families for Oral History

Jakob Lenardič[1], Silvia Calamai[2], Stefania Scagliola[3], Henk van den Heuvel[4]
[1]Institute of Contemporary History, Slovenia; [2]Università di Siena , Italy; [3]Independent researcher; [4]Radboud University;
jakob.lenardic@inz.si

The CLARIN Resource Families (CRF) initiative provides manually curated overviews of prominent language resources and technologies deposited across the distributed CLARIN infrastructure (Lenardič and Fišer 2022). The main aim of CRF is to support other core services of CLARIN from the perspective of the FAIR principles (Wilkinson et al. 2016). CRF enhances the findability and accessibility of CLARIN resources by collating them under their most common typological characteristic. The initiative facilitates re-use by providing comprehensive descriptions tailored to the unique technical features of each of the families, as well as their qualitative characteristics. Furthermore, CRF provides a funding instrument for external projects to contribute new overviews.

Though originally focused on written corpora (e.g., corpora of parliamentary proceedings, corpora of academic texts), in 2022, CRF was expanded to include corpora of oral history. At present one collection is currently featured – the Ravensbrück corpora (Calamai et al. 2022a) – whose creation was supported by the aforementioned CRF funding instrument. This corpus family contains 8 collections of recorded interviews with survivors of the female concentration camp Ravensbrück, conducted in different languages, such as English, German, Hebrew, and French. See https://www.clarin.eu/resource-families/oral-history-corpora. One collection is available for download (Collection Bruzzone; see Bruzzone and Beccaria Rolfi 1976) while the others can be streamed online.

The inclusion of the Ravensbrück corpora in CRF represents an illustrative example of how the CLARIN infrastructure incorporates and provides documentation for complex objects like oral history sources whose provenance and metadata documentation widely differ from standard written corpora and even from contemporary interviews born digitally. The team working on the Ravensbrück resource family (see Calamai et al. 2022b) availed themselves of CLARIN's Component Metadata Infrastructure (CMDI), which is a framework for metadata description that "supports flexible definitions of metadata structure and semantics" by allowing researchers to "create and use their own [metadata] schema tailored specifically towards the requirements of [their] project" (Windhouwer and Goosen 2022: 194 and 199). All the 8 collections within the Ravensbrück family are accompanied by extensive CMDI metadata, prepared by Calamai et al. (2022a,b).

The peculiarity of the interviews in the Ravensbrück family is that they were mostly recorded on an analogue carrier (i.e., audio cassettes), so a new CMDI metadata profile was created that is tailored to such legacy interviews not born digitally. This metadata profile has additional components describing "information about the context in which the interviews were conducted" as well as "information about the process of digitisation" (Calamai et al. 2022a: 3). Being thus digitised, comprehensively described, and carefully curated, the Ravensbrück corpora present a unique opportunity to study and compare these historical interviews. To facilitate their use in research, CLARIN offers through its Speech data and Technology network (Draxler et al. 2020) an open-source web application called TranscriptionPortal (https://speechandtech.eu/transcription-portal), where certain audio recordings (e.g., Collection Bruzzone, United States Holocaust Memorial Museum) can be uploaded and then orthographically transcribed on the fly, with manual phonetic and word alignment for a variety of languages.

The Text+ interface to NFDI and the ERICs: Task Area Infrastructure/Operations as Assembly Tool

Stefan Buddenbohm, Lukas Weimer
Göttingen State and University Library, Germany; buddenbohm@sub.uni-goettingen.de

Text+ is part of the National Research Data Infrastructure (NFDI) and addresses the needs of long term preservation and interoperability of text- and language-based research data. Within Text+ the task area Infrastructure/Operations is responsible for the integration and coordination of the overall infrastructural developments with regard to Text+ as a whole - integrating the three data domains collections, lexical resources, and editions - and with regard to Text+ as part of the NFDI big picture, and the European infrastructure level as well, particularly DARIAH-EU. In this regard IO can be described as an assembly tool for sustainable data management and curation workflows.

The "Digital Landscape in Greece" Web Survey

Maria Ilvanidou[1,2], Vicky Dritsou[1,2], Maria Gavriilidou[3], Kanella Pouli[3], Yorgos Tzedopoulos[4], Irakleitos Souyioultzoglou[4]
[1]Digital Curation Unit, Information Management Systems Institute, Athena RC, Greece; [2]Department of Informatics, Athens University of Economics and Business, Greece; [3]Institute for Language and Speech Processing, Athena RC, Greece;
[4]Academy of Athens, Greece; m.ilvanidou@dcu.gr

The way of working in the Humanities undergoes a digital transformation. The Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation "APOLLONIS" and the European research infrastructures DARIAH for the Arts and Humanities and CLARIN for Language Resources and Technology are important facilitating agents of this transformation in Greece and in Europe respectively. The transformation has been forcefully accelerated by the recent pandemic and, as some early research initiatives are indicating [1], readiness for acceptance of digital work practices appears to rise. Also rising is the need to delve deeper into the nature of digital methods and practices and their impact on research.

The COVID-19 outbreak and the restrictions it imposed brought about new needs in the Humanities research and teaching, introducing new modi operandi, accelerating a digital transformation that has been under way for decades in the Digital Humanities, but also demonstrating the gaps that remain. Funded by the Hellenic Foundation for Research and Innovation (HFRI), the project "Digital Landscape" (The emerging landscape of digital work practices in the Humanities in the context of the European projects DARIAH and CLARIN, 2022-2024), is capturing the impact of the pandemic in the way we work in the Humanities and Social Sciences in Greece, to better understand current trends and needs, and plan effectively for the future. To this end, we are employing a mixed methodology. The tools to collect evidence-based information are: a) a communities web-

survey and b) Humanities and Social Sciences focus groups. The report resulting from the analysis of these data will be used to support the strategic vision and inform the future planning of the APOLLONIS research infrastructure.

The proposed poster will present the findings of the web survey, in which we analyzed data on digital research practices, tools and interaction with digital content gathered from the humanities and social sciences research community in Greece, paying special attention to the impact of COVID-19 on digital humanities research and practice.

Based on past experience and data collected in the context of European research infrastructures and projects in the Arts, Humanities and Language Technologies and Social Sciences, such as Preparing DARIAH, DARIAH-EU, CLARIN-EU, EHRI (the European Holocaust Research Infrastructure), ARIADNE (Advanced Research Infrastructure for Archaeological Data Networking in Europe), SSHOC, SoDaNet and Europeana Cloud, we designed, disseminated and implemented the web survey which gathered almost 400 responses, subsequently analysed and discussed in comparison with past trends in Greece.

In this poster we propose that we present these findings and prompt discussions with our European colleagues.

[1] E.g., “DH in the time of Virus: a Twitter Conference”, 02.04.2020, "DHgoesViral" workshop, 26.04.2021.

Named Entity Recognition and Knowledge Extraction from Spanish Golden Age theatre

Gustavo Candela, Juan Carlos García, Pilar Escobar, Dolores Sáez, David Quintela, Manuel Marco-Such
University of Alicante, Spain; gcandela@ua.es

Libraries and cultural heritage institutions, commonly known as Galleries, Libraries, Archives and Museums (GLAM), have traditionally provided access to contents in digital format. New methods of accessing to the digital collections in innovative ways have recently emerged in order to facilitate the application of computational methods based on Data Science, Machine Learning (ML) and Artificial Intelligence (AI).

Recent trends, such as Collections as data, promote the publication of digital collections amenable for computational use. In this sense, and as a relevant example of making available and reusing digital collections in innovative and creative ways, GLAM Labs have gained popularity among the community. However, the adoption of the Collections as data principles has increasingly become a challenge in GLAM institutions due to different reasons including copyright issues, data quality, and lack of personnel or IT skills needed.

Advanced techniques based on ML and AI can be applied to digital collections such as Named Entity Recognition (NER) to identify geographic locations, dates or names of people, and entity linking to create links to external repositories. The richness of the collections hosted by GLAM institutions can play a crucial role in the training process of ML and AI models that can help to maintain their relevancy. Previous studies have focused on literature reviews concerning approaches as well as the developments based on ML models to extract key insights from textual data in a structured way.

In this context, the study of the Spanish Golden Age theatre has traditionally gained attention. Some examples include an analysis on authorship, modelling the social networks underpinning the early modern publication industry and a computational linguistic approach applied to Spanish Golden Age Sonnets.

However, the scarcity of annotated benchmark datasets for particular approaches and languages, such as Spanish Golden Age theatre, has resulted into relatively less progress in these domains. In addition, many of the approaches are offline, computationally expensive and mainly based on predominant languages such as English, French and German.

The objective of the present study was to introduce a method to train a NER model based on Spanish Golden Age theatre made available as datasets by GLAM institutions, in particular the Biblioteca Virtual Miguel de Cervantes (BVMC) in Spain. The results of this study are publicly available at the BVMC Lab (https://data.cervantesvirtual.com) and can be used and applied to other domains such as historical documents and newspapers.

The main contributions of this work are as follow: (a) a detailed method to train a NER model for Spanish Golden Age theatre; (b) the model generated; and (c) the results obtained after the analysis and assessment. These are relevant to encourage researchers to use the datasets published by GLAM institutions.%In addition, a collection of Jupyter notebooks is provided to reproduce the results.

Expanding DARIAH Teach with seven OERs from the Dimpah project

Marianne Ping Huang[1], Koraljka Golub[2]
[1]Aarhus University, Denmark; [2]Linnaeus University, Sweden; koraljka.golub@lnu.se

DiMPAH aims to aggregate, connect and make widely available novel Open Education Resources (OERs) on selected digital methods, apply these to interdisciplinary contexts and

foster novel creative learning experiences by taking data from the past into future stories.

DiMPAH has three objectives:

Create novel OERs on digital methods and associated tools for the construction of new knowledge on A&H research questions and for audience engagement in a suite across the complementary areas of qualitative and quantitative digital research tools and methods. The DiMPAH suite, channeled by the #dariahTeach platform, will be openly available across European institutions to support current and future professionals from cultural heritage sectors as well as academia in improving competencies, connecting best practices and applying spearheading technologies, to enable collective efforts towards future solutions.

The DiMPAH-selected methods are to be applied and tested via case studies in three prominent European digital heritage contexts: a) digitised newspaper collections; b) built heritage environments and their digital twins; and, c) performing arts collections.

DiMPAH will move this ‘towards new stories for Europe’: the selected new methods and technologies, and cultural heritage case studies, will be deployed in learning scenarios to localise and show possible solutions and potential impact on social equity, transnational and cultural diversity, gender equality, good health and well-being.

Through OERs developed in this project, the European student and researcher will gain access to a one-stop shop for OERs on digital methods for A&H, as well as related disciplines in the interdisciplinary sphere of DH. In addition, teachers in HE institutions will be able to train their students by building on material created by highly experienced professionals in the field. Cultural heritage professionals will learn on how they could best apply digital methods such as Augmented Reality for engagement with audiences online and on-site. Ultimately, long-term benefits can help Europe (and beyond) reach certain sustainability goals by allowing scholars to research digital content and datasets, such as cultural data from cultural heritage and art institutions, which are central to studying identity formation and social cohesion. Complex societal challenges can be addressed through international, cross-disciplinary, collaborative research into human conditions, societies and cultures, and through comprehensive studies using relevant digital methods and datasets throughout Europe and beyond.

The following seven courses are planned:

Introduction to Digital Methods

Text Analysis: Linguistic Meets Data Science

Digital Historical Research on European Historical Newspapers with the NewsEye Platform

Netnography

E-Spectator Digital Tool for Analysis of Performing Arts

Design, Development and Deployment of Augmented Reality Applications

Introduction to Knowledge Organisation Systems for Digital Humanities

The Dimpah suite of courses is organized into two parts, A. Introducing Digital Humanities Methods with six OERs, and B. Organising Digital Humanities Documents with 1 OER.

Digital Displacement: Responses to the Cultural Heritage Crisis in Ukraine

Jessie Labov

Corvinus University, Hungary; jessiemargaret.labov@uni-corvinus.hu

This presentation is part of an ongoing project to record and analyze different digital solutions to the humanitarian crises generated by the Russian invasion of Ukraine and the months of war that have followed it. In particular, the project focuses on issues of cultural heritage and attempts to both digitally preserve existing cultural heritage, and archive digital cultural heritage.

In November of 2022, we organized a session on this topic at the ASEEES convention in Chicago ("Digital Advocacy: Aggregation and Networking for Displaced Scholars"). The goal was slightly different: to discuss some of the ad hoc solutions that appeared in February and March 2022 to the challenge of immediate and efficient aggregation and distribution of information to help displaced academics, cultural workers, and other scholars. Scholars who represent different approaches to this issue, including Steven Seegal (creator of the Feb 24 archive on Twitter) to Dorine Schellen (who created and maintained the University of New Europe database of resources for displaced scholars). I also presented the work of Science4Ukraine, SUCHO (Saving Ukrainian Cultural Heritage Online) and the Lviv Center for Urban History (https://www.lvivcenter.org/en/). What emerged clearly from our discussion was the need to coordinate and compare the manifold efforts.

This is why I have begun a series of long-format interviews with people in Ukraine involved in cultural heritage institutions, as well as those abroad who have been actively involved in digital preservation efforts (Taras Nazaruk and the Lviv's Center's Telegram archive, Anton Mudrak, Maciej Maryl, Lars Wieneke, Qunn Dombrowski and Andy Janco), and in June I would like to present the first results of this work, with what will amount to a white paper on where things stand on digital preservation of cultural heritage, and what remains to be done.

Building a Gigacoprus for Language Model

Botond Szemes[1], Dávid Nemeskey[2], Balázs Indig[2], Gábor Palkó[2]

[1]Research Centre for the Humanities Institute for Literary Studies, Hungary; [2]National Laboratory for Digital Heritage, Budapest; boboszemes@gmail.com

In this presentation, we will outline the first phase of our consortium project aimed at creating a Hungarian language model, focusing on the creation of a large corpus of teaching material for the model. This phase is necessary for the effective functioning of the model: The quality and variety of the text collection play a crucial role in determining the performance of the model. The corpus is divided into three subsections.

Common Crawl, which is based on the textual material of the Hungarian Internet, constitutes the largest part of the corpus. The corpus, based on the Common Crawl collection, currently consists of 9 billion tokens and can be further extended (cp. Nemeskey 2020). The best Hungarian contextual word embedding, huBERT (Nemeskey 2021) was trained on this corpus and serves as a basis for other models. The size of this corpus is a major improvement, but the quality needs to be improved given that it is web data.

The second sub-section involves material collected from the Webaratás project (Indig et al 2020). A large amount of high quality text can be extracted from the project material, which collects the texts of Hungarian news portals and organises them in TEI-XML format, supplemented by metadata. To date, all the texts of 18 news portals, forums, and blogs have been collected and converted into TEI XML format.

The third subset consists of non raw text sources. On the web, not only is there raw text on news sites/forums, but there are also a large number of PDFs with live texts, e.g. through digitisation of public collections and publication in research spaces. There are many such texts that are openly accessible: scientific publications, scientific and non-scientific books, and other texts. In this presentation, we will demonstrate our method for the extraction and OCR of textual resources stored in Hungarian repository systems.

All three sources require manual annotation, either due to the nature of the task or for producing training material for machine learning models, such as filtering boilerplate or performing OCR analysis. Building on the three sources outlined above, our goal is to construct a larger Hungarian corpus than ever before by incorporating additional text sources, such as social media postings, which is crucial for a highly efficient language model.

Computational Drama Analysis: Genre identification

Botond Szemes

Research Centre for the Humanities Institute for Literary Studies, Hungary; boboszemes@gmail.com

In the presentation I aim to demonstrate what kind of structural data could be extracted from encoded drama texts, and how can we use these data to describe the characteristics of dramatic genre. This characterization could shed light to genres, which are traditionally described by thematic features (such as the extent to which historicity determines the setting and the period the events take place in, the emphasis on moral issues, the social status and speech patterns of the characters, or the conclusion of the plot.)

The presented method suggests that we can distinguish comedies and tragedies from each other primarily based on the character networks of plays, which allows the (1) plot, (2) interpersonal relations and societal world, (3) and/or dramatic form of a given play to be viewed simultaneously on the plane of a surface instead of in the process of reading a play or watching a performance, which inevitably unfold over time. In the study I use the concept of structure, by which I mean the system of relations between characters in the broadest sense: among the metrics of character networks (density, diameter, clustering coefficient, etc.) I also rely on the distribution of words/stage time between characters to reveal structural differences between genres. This conceptualization is closest to form, but it is also related to (and thus brings together) the other approaches—e.g. the analysis of structure also deals with the relationships between characters as a result of the progression of the plot; or following the considerations and methods of SNA (Social Network Analysis), character networks can also be seen as structural representations of the contemporary cultural-historical context of a given drama; and thus plays can be understood as the author's – not necessarily intentional – attempt to model a society.

Earlier similar research suffered from two crucial problems that call the usefulness of their results into question. First, most papers rely heavily on the size of the networks to classi-fy genres, which instead of capturing the structural design of networks, simply compares the number of characters in the plays. In what follows, an attempt will be made to present a meth-od that does not take the size of the networks into account. The other pitfall to be avoided is the application of overly complicated mathematical procedures to investigate similarities be-tween networks. The problem with this approach is that the results obtained this way cannot be attributed to real properties of the plays; such results would only show whether two works have a similar structure. By contrast, my goal is not simply to confirm that there is such a simi-larity but also to explain what this similarity consists of.

Finally, I will also discuss how these characteristics change over time in two ways: 1. from the perspective of literary history; 2. in term of the development of dramatic plot.

Toward FAIR Data Practices with the French National 3D Data Repository

Sarah Tournon-Valiente[1], Mehdi Chayani[1], Xavier Granier[2]

[1]Archéosciences Bordeaux; [2]IOGS (Institut d'Optique Graduate School); mehdi.chayani@u-bordeaux-montaigne.fr

The French National 3D Data Repository (CND3D) is a repository dedicated to the scientific publication and curation of 3D models together with the related data. Deployed in 2020 by the consortium "3D for Humanities" of the research infrastructure Huma-Num, it is currently tested on a large scale by the community of humanities and social sciences. There are currently more than 1100 deposits, 570 viewable 3D models, 7 collections and more than 180 documented sites and over twenty social sciences labs depositing. This secure repository relies on Huma-Num infrastructures, and it is pushing the integration of FAIR principles.

Findable

The National 3D data repository stores 3D models along with all the accompanying resources used in their creation (text, images, publications, etc.). A dedicated metadata scheme is continually augmented to accommodate the different data and scientific domains. With the right amount of metadata, a DOI is generated.

The CND3D relies on standardized and multilingual vocabulary, such as PeriodO, Geoname and PACTOLS thesaurus that covers archaeology and the sciences of antiquity from prehistory to the present day.

Accessible

The metadata is always accessible. Research with a REST API is also available and can be saved as a url link.

However, access to the actual data stored in the repository is subject to the depositor's decision and may be restricted.

Interoperable

To enhance its interoperability, the repository is actively working on compatibility with other European initiatives.

With the French MASA consortium, we made the metadata scheme compatible with the CIDOC CRM, enabling connection with the European research portal ARIADNEplus.

The repository's metadata also complies with Dublin-Core standards. It integrates an OAI-PMH harvester used by the French search engine Isidore dedicated to humanities and social sciences. This will also allow data to be harvested by TRIPLE from the European OPERAS project.

Reusable

The 3D data repository promotes the use of PLY and DAE formats for 3D models, which have been standardized by the consortium for archiving at CINES. We are also exploring other formats for various data such as scanned point clouds, tomography, and modeled 3D scenes. However, for the sake of simplicity and for promoting publication of scientific data, the repository accepts any type of data, regardless of format.

As for licensing, the repository favors the use of ETALAB license (equivalent to Creative Commons CC-BY license) or CC-BY-NC.

Deposit

According to the depositor, the access to the stored data can be done either by a direct download directly, or indirectly by contacteur the right owner. Thanks to the integrated 3D viewers, visualization and interactions (cuts, lighting, measures, …) with a dedicated version of the 3D model may be chosen.

Filling metadata can be done on-line. However, to facilitate deposits, an open source software has also been developed. This software, named aLTAG3D, allows a semi-automated deposit through a simple user interface and offers a solution for managing 3D data with their metadata and specific formats for use in the humanities and social sciences community.

From Dataset to Knowledge Graph: The "Chronology of Events 1940-1944" at the Academy of Athens

Maria Spiliotopoulou[1], Giorgos Stamou[2], Alexandros Chortaras[2], Yorgos Tzedopoulos[1]

[1]Academyof Athens, Greece; [2]National Technical University of Athens; mspil@academyofathens.gr

The paper deals with the curation of a collection created by the Modern Greek History Research Centre of the Academy of Athens in Greece. The collection is based on an archival series of the British Foreign Office from the period between the outbreak of the Greek-Italian War (October 1940) and the end of the Nazi occupation in Greece (October 1944). Based on this material, the researchers have compiled a detailed chronology, which records specific events, but also information exchange, proposals, plans and reports by Greeks, British, Allies and resistance organizations. The Chronology was published in two volumes. The entries were also captured in a database that allows multiple searches and extraction of data beyond the indexes contained in the publication. Yet this digitized collection had remained a closed corpus of information, without any interconnection with relevant collections of other institutions.

Three years ago, the Chronology was incorporated in the action "The 1940s in Greece" of the project APOLLONIS, in the framework of which the partners of the project, based on the metadata of various collections pertaining to that period, created common indexes and knowledge bases to increase the collections' accessibility and interoperability, and to familiarize researchers with the possibilities of datasets curation.

One of the outcomes of this action was a linked data-based knowledge graph representation of the collections to support expressive semantic queries. The first step for the construction of the knowledge graph was the production of linked data representations of the collections using a mapping tool. The mappings used both standard and custom vocabularies to cover collection-specific modeling needs. The second step was the automatic enrichment of the items in the knowledge graph with annotations, which were produced for three semantic dimensions (place, time, person/organization) by applying NERD tools on the relevant fields. For the Academy of Athens Chronology, these tools produced Geonames, TimeLine and Wikidata annotations. The resulting knowledge graph was further extended with relevant vocabularies (eg. DBPedia, Greek Historical Periods) by including cross vocabulary equivalence alignments for identical terms and containment alignments for temporal terms.

Through these annotations, the knowledge graph allowed unified multi-vocabulary searches over all collections, expressed in any of the supported vocabularies. To exploit the knowledge graph, a search application was built that supported combined cross collection queries over the three dimensions. Through limited reasoning support, the application was able to also answer semantic queries by exploiting the vocabulary hierarchies and other term categorizations.

Practical evaluation showed that, although the quality of the results depended on the quality of the automatically produced annotations, and the complexity of the queries affected performance, knowledge graph technologies can enable researchers to locate relevant material through expressive queries exploiting the levels of abstraction provided by the underlying knowledge. For the Academy of Athens, the outcome was particularly fruitful: firstly, its collection participated in a rich knowledge graph that facilitated research on the specific period; secondly, interconnectivity highlighted documentation inconsistencies that have to be adjusted to improve the quality of semantic queries.

Collaboration of Cultural Heritage Institutions, Researchers, Information Scientists, and Citizens in Sustainable Workflows for Data Management, Curation, and Communication

Goran Zlodi, Tomislav Ivanjko, Zoran Horvat

Faculty of Humanities and Social Sciences, University of Zagreb, Croatia; gzlodi@ffzg.hr

Cultural heritage institutions deal with functions of protection, research, and communication of material and intangible cultural heritage related to their collections. When we talk about research in cultural heritage institutions, it sometimes refers to scientific research in the narrower sense of the word (encompassing construction of research questions and hypotheses, and publication of results in scientific journals and monographs etc.), nonetheless it is always about continuous collection and study of information about cultural objects heritage. That information is both intrinsic information, that results from the materiality of the object (properties such as materials, techniques, dimensions, and condition, that are inquired through observations and predominantly natural science methods) and extrinsic information that is analyzed by studying the history of the objects and the wider context (title, creator, provenance, etc.). That information is recorded through cataloging by established metadata standards and mostly by audio-visual documentation, including 3D models, etc.

But is that enough if we want cultural heritage data to be interoperable and reusable for various future research topics and heterogeneous approaches (multidisciplinary, cross-disciplinary, interdisciplinary studies) in digital humanities? Digital humanities research is a complicated endeavor that could involve many diverse, connected and unconnected, sometimes recursive, human and machine activities of the extraction of evidence and knowledge from cultural heritage objects. That includes but is not limited to digitization, OCR, transcribing, NER techniques, AI methods, documentation and metadata recording, data curation and preparing for statistical analysis and visualizations, mapping and integrating to semantic sources and data publication as Linked Open Data.

All of the above implies continuous communication, dialogue, and cooperation between different experts - renaissance teams of social science and humanities researchers, cultural heritage experts, and computer, information and data scientists.

Furthermore, we believe that in such teams, whenever it is possible and makes sense for research, there is a place for citizens. Cooperation between cultural heritage institutions, researchers, and citizens is important for many reasons. As the number of materials in the collections is huge (most of them are not digitized, nor documented by metadata and contextual information), the engagement of citizens is a matter of feasibility and sustainability. Furthermore, cooperation with citizens implies a better insight into users' needs, as well as the social relevance of research and the democratization of science in general.

The presentation will report the workflow of the project based on the cooperation of academia (university teachers and students), citizens (retired sports journalists), and heritage experts in documentation, research, and communication of knowledge related to the collection of more than 16.500 digitized photographs from the Croatian Sports Museum (the photographs are documenting local and international sports events in Croatia and Yugoslavia from 1970 to 2000).

Die Gemeinsame Normdatei (GND) - The Integrated Authority File in Text+ as semantic link to DARIAH-EU

Stefan Buddenbohm[1], Barbara Katharina Fischer[2], Marie Annisius[2]

[1]Göttingen State and University Library, Germany; [2]German National Library, Germany; buddenbohm@sub.uni-goettingen.de

Among other goals, the German national research data infrastructure (NFDI) aims to provide an overarching knowledge graph across the domains and subjects. In the field of humanities, arts and culture the two consortia NFDI4Culture and Text+ have chosen the Integrated Authority File (GND) as an initial backbone structure. Within Text+, the NFDI consortium for the language- and text-based research data, a major task is devoted to initiate a GND- Agency revolving around the Integrated Authority File. Inevitably, this task takes interoperability into account, for instance towards Wikidata or VIAF or other approaches in use within DARIAH-EU or EOSC.

Working with Flow: Workflows for Accessing 4CH Services

Vicky Dritsou[1,2], Achille Felicetti[3], Panos Constantopoulos[2,1]

[1]Digital Curation Unit, IMSI/Athena RC, Greece; [2]Department of Informatics, Athens University of Economics and Business, Greece; [3]PIN, Università degli Studi di Firenze, Prato, Italy; v.dritsou@dcu.gr

In order to facilitate institutions and researchers in exploiting collections to the greatest possible extent and in describing best practices, it is essential to develop sustainable workflows for the creation of and access to cultural heritage data. At the same time, the common European dataspace for cultural heritage[1] outlines the critical need also to share, integrate and provide such high-quality cultural data through the development of respective services. In this paper we describe our ongoing work in developing workflows for accessing the services of the Competence Centre (CC) for the Conservation of Cultural Heritage in the context of the 4CH project [2], thus advancing and contributing to the establishment of best practices in accessing cultural heritage services.

Based on previous experience and considering workflows of known research portals and major infrastructures, e.g. the ARIADNE Portal[3], the PARTHENOS Portal[4] and Europeana[5], we studied and identified the sequence of actions required for accessing CC services. Firstly, an abstract model capturing the sequence of actions required by the user to access CC services in general has been developed, establishing the parameters that need to be defined when applying it. These include, but are not limited to, the type of user requiring access (individual or institution, data creator, data manager, data consumer, researcher, educator etc.), the type of service (access to data, access to tools, training, consultancy, etc.), the type of access (remote or on-site) and charging issues if applicable. This abstract model can then be adapted to each different case by providing specializations of the generalized actions included and leading to variations of the abstract model.

By way of example, we provide in Fig.1 a specialization of the workflow expressed as a UML activity diagram with swimlanes, where specific values have been applied to a subset of these parameters. In particular, we describe the sequence of actions required to provide access to an indivual user that requests remote access without charging fees. The three lanes employed distinguish the actions required to be performed by i. the user, ii. the CC in an automatic way, and iii. the CC interactively. The user formulates a specific need against the 4CH knowledge base, which contains a list of CC services along with other resources like digital tools, training material, documentation files, etc., and which is semantically organized. Several decision points (diamonds) determine in the sequel the exact flow of activities that will be performed each time, ranging from executing a totally automatic workflow to requiring human communication. Application of different values to the same parameters would lead to a different specialization of the workflow. Finally, the activity colored in blue represents a complex activity that requires the execution of a service-specific workflow for its accomplishment.

Upon completion, the proposed workflows will facilitate CC in organizing its services and in guiding users to access them. Going one step beyond, these could also be exploited to facilitate accessibility of services of cultural heritage data spaces, providing the means to formulate and promote common procedures and best practices.

The WorldFAIR Project: Making Cultural Heritage Data FAIR at the Digital Repository of Ireland

Joan Murphy, Beth Knazook

Digital Repository of Ireland, Ireland; j.murphy@ria.ie

The FAIR principles were published in 2016 to provide a baseline strategy for all domains of research to make their data broadly reusable by others. They describe approaches to findability, accessibility, interoperability and reproducibility for both data and accompanying metadata and, although they are not prescriptive in what may be considered FAIR, the metrics which have been derived to assess FAIRness have largely focused on a limited number of machine actionable criteria applied uniformly across research areas. The WorldFAIR Project, through 11 case studies of different disciplines led by global research partners, aims to broaden our understanding of how FAIR may be interpreted within these disciplinary contexts.

The Digital Repository of Ireland (DRI) is leading the Cultural Heritage case study, which seeks to explore how image sharing platforms in the cultural heritage landscape already facilitate the interoperability of both image data and associated metadata. While the overall outcome of the case study will be the production of a model for implementing FAIR recommendations at the Digital Repository of Ireland, it is also an opportunity to highlight achievements in the sector which may usefully inform work in other domains. In that respect, the work is both a valuable opportunity to review and situate the DRI's position as an image-sharing platform in the cultural heritage landscape, as well as a chance to raise questions about how FAIR may be perceived in that landscape. This presentation will discuss the results of a mapping report documenting policies and practices in image-sharing platforms, as well as a set of recommendations for improving FAIR assessment of cultural heritage image data.

Implementing the infrastructure for Dimitris Papaioannou's archive: approaching the degrees of separation in his work

Vassilis Poulopoulos[1], Elena Papalexiou[2], Costas Vassilakis[3], Avra Xepapadakou[4], Valia Vraka[5]

[1]Knowledge and Uncertainty Research Laboratory, University of Peloponnese, Greece; [2]Dept. of Theatre Studies, University of the Peloponnese; [3]Dept. of Informatics and Telecommunications, University of the Peloponnese; [4]Dept. of Languages and Literature, University of Nicosia; [5]Music Library of Greece "Lilian Voudouri" of The Friends of Music Society; vacilos@uop.gr

Creating the infrastructure for an archive is considered, nowadays, to be a trivial procedure. Several tools (e.g. Omeka [2], CollectiveAccess[3], CollectionSpace[4]) and prototypes (eg. Dublin Core [5], VRA CORE [6]) are available for use in order to, either directly install and use a CMS for cultural related data (first case), or build a system "from scratch" based on a common prototype (second case). This procedure would lead to the digitization of at least an archive's metadata and to the organization of the archive itself; it would offer the ability to search for information in the archive in an easy way, it would allow people to access the information of the archive, and it could also be part of a larger cultural data lake (Europeana). The solutions for these use cases, are extant, present and available.

On the other hand, occasionally one comes across archives of artists that have a special view of art, are ahead of their time, design their own universe, and they keep evolving it throughout their life. In this paper, we discuss the implementation of the archive for Dimitiris Papaioannou, a diverse archive which includes several different types of data. In this case, we have selected a special approach, in order to fully capture the artistic aspects of the archive's content, which was analyzed in the context of the GENESIS project [1]. Throughout this project, the procedure through which the artist approaches the performance arts and generates ideas was recorded and analyzed.

Following the aforementioned analysis, the archive material provides insight for the artist's visual arts, his dancing roads and his performances, and unveils his approach to performance arts (or art, in general). Then, the archive is structured in a way that would allow all archive visitors, including the general public as well as researchers, to obtain information about the individual elements of the work, but also locate the "genetic material" of the artist into the archive.

To achieve this goal, interdisciplinary research involving the cooperation of scientists and researchers from both the humanities and the technology domains was required. The research team designed an information layer entailing objective information (e.g. works, performances, participants and their roles etc.) and subjective information (e.g. tagging or free text descriptions), and accommodating semantic links between them. Metadata were also associated to information items, where applicable.

The information layer fully serves the needs for information search and retrieval, while it additionally provides the underpinnings for data mining procedures that can possibly reveal patterns or data correlations that have remained unobserved.

Cultural natural heritage data in creative mapping rural landscape for the RURITAGE ATLAS

Rosa Tamborrino, Mesut Dinler, Alessandro Aliberti

Politecnico di Torino, Italy; mesut.dinler@polito.it

In heritage research, the use of data requires a careful definition of research methodology with a multidisciplinary approach both due to the very nature of the cultural heritage which is wide and continuously evolving with new conceptual and practical developments, and due to the issues related to heritage data. An example where these issues have been confronted with digital methods is the RURITAGE ATLAS created in the scope of the EU-funded Horizon 2020 project RURITAGE as a component of the RURITAGE Resources Ecosystem. The ATLAS is the output of the innovative rural landscape creative mapping, and it has been conceived and designed for surveying and making available the multiple and rich functions of rural areas and human-landscape interactions, including heritage values. For this purpose, an integrated digital environment has been constructed on a WebGIS environment and integrated in a digital platform. The ATLAS provides data on cultural and natural heritage (CNH) including land uses and human interactions with CNH assets, but also on the rural regions as a whole by considering urban and rural interactions, issues of connectivity, accessibility, and mobility. These data are also collected by considering a larger framework of comparability, according to the research project methodology.

The data collected and presented on the ATLAS includes economic, demographic, and social factors that are able to underline those economic and social potentials to push resilience and social cohesion in the RURITAGE territories. Data are both extracted by desk research and by botto up processes that engage local stakeholders.

The availability of different kinds of visualization of the stored information, which is managed through databases, offers the possibility to focus either on the single region and human/landscape interactions.

The issues confronted during the ATLAS can be discussed under these themes.

- Identification of heritage data: ATLAS gives the possibility of interacting with data under four sections Places, Experiences, Rural Territory, Stories giving information on both tangible features and intangible futures.

- Collecting, co-creating, co-mapping the data: A co-creation approach was adopted with each rural area stakeholder for better understanding the characteristics of each area to be represented on Atlas.

- Integrating external data: In addition to data collection, further analysis has been carried out for including economic and social data, as well as making maximum use of Copernicus and Galileo data. This has required a continuous update of the back-end structure.

- Reuse of data: Data collected and co-created throughout the project has been reused for different purposes such as creating digital narratives for better presenting the intangible heritage of these areas or deriving performance indexes to better analyze local information and create benckmarks for deeply comparison.

- Linking heritage data to spatial information: cultural natural heritage data among other humanities data characterising spatialised information especially considering the case of data referring to intangible features

- Interoperability: RURITAGE ATLAS is a part of the larger, flexible e modular RURITAGE Ecosystem. Atlas ensures interoperability with existing platforms by using web-service based access and open APIs exploiting a spatial approach supported by a webGIS tool.

Cultural Collections as Challenging Research Data in Small States: the Case of Latvia

Jānis Daugavietis, Sanita Reinsone

Institute of Literature, Folklore and Art - University of Latvia, Latvia; janis.daugavietis@gmail.com

Research (and statistics) indicate that despite recent improvements the Latvian humanities and arts community in general still lacks awareness and understanding of contemporary open science and is not particularly supportive of open science practices (Bite et al 2020, Daugavietis et al 2022, Reinsone et al 2023). This phenomenon can be attributed to a multitude reasons, including one of the lowest levels of R&D funding in the European Union and the lag in the implementation of national science policy in this domain. The academic community in the field of humanities is just starting to grasp the notion of "research data" and necessary management it requires, while it remains unfamiliar to most cultural heritage institutions and public cultural administration. The development of digital research infrastructures continues to pose a persistent challenge. Although some progress has been made and there are a sparse examples across several research disciplines, including operation of national CLARIN ERIC repository in recent years and the joining of one institution as a cooperating partner in DARIAH-EU last year, the overall development of digital research infrastructures in Latvia remains deficient. This holds true for the social sciences as well.

In comparative political science, the notion of small states is often employed to argue that the process of socio-economic development in small countries differs from that of larger countries (Katzenstein 1985, Chodak 1989). Latvia, with a population less than 2 million and an even smaller Latvian-speaking community, is considered a small country. The attributes of a small and economically challenged country may shape other optimal scenarios for the development of research data repositories and research infrastructures in general. The research community in such states is commonly limited in size, with a scarcity of research institutions in each field, and characterized by a lack of inner competitiveness. Furthermore, the geographical distances within the country are relatively modest, academic research is predominantly concentrated in the capital city. The administration of science and culture is centralized but both fields are operating independently from one another. The main question this paper seeks to address, with Latvia being the main focus and other small countries being used for comparison, is: which of the above mentioned and other characteristics of a small state facilitate or hinder the EU's science and culture policy objectives, specifically, the implementation of open science principles and the complete digitisation and accessibility of cultural heritage?

REFERENCES

Bite, K., Daugavietis, J., Kampars, J., Kreicbergs, J., Kuchma, I., Ločmele, E., Ostrovska, D., Vecpuise, E., Veisa, K., & Želve, M. (2020). 'Pētījums par atvērto zinātni un rīcībpolitikas ceļa kartes izstrādi'. LNB.

Chodak, S. (1989). The New State. In The New State. Lynne Rienner Publishers.

Daugavietis, J., Karlsone, A., Kunda, I., & Kristāla, A. (2022). 'Latvijas digitālo humanitāro zinātņu rīku un resursu izstrādāšanas prakses'. Letonica, 47, 12–51.

Katzenstein, P. J. (1985). Small states in world markets. Cornell University Press.

Reinsone, S., Matulis, H., & Daugavietis, J. (2023). 'Digitālie resursi un rīki humanitārajām zinātnēm. Rekomendācijas politikas veidotājiem un digitālo resursu un rīku izstrādātājiem' [MANUSCRIPT]. LU LFMI.

São José: a COESO project for citizen sciences with a multimodal and transmedia approach toward exploring tourists' experiences in Lisbon

Véronique Bénéi[1], Camilo Leon-Quijano[2]

[1]CNRS (CNE); [2]CNRS (CNE/La Fabrique des Ecritures/COESO); veronique.benei@ehess.fr

This poster presents the multimodal experimental project carried out from 2021 to 2023 as part of the European project Collaborative Engagement in Societal Issues (COESO). The research project explores citizens' relationship to tourism in Sao José, Lisbon. Based on a triple collaboration between CRIA, ZERO association, and CNRS (La FEE - France), we created a transmedia website to depict the sensory experiences of tourism in a neighborhood of Lisbon: https://saojose.huma-num.fr/

Following a citizen science approach, this multimodal anthropology offers a journey into sensory experiences on an open-access visualization medium. In doing so, this project aims to develop a multi-modal anthropology that reaches out to new audiences. Following a participative and critical approach, it hopes to contribute to a better understanding of sensory experiences of tourism in the city.