Digital Humanities and Cultural Heritage in AI and IT-enabled Environments

Ciara Breathnach and Tiziana Margaria

University of Limerick, IE
HRI, Lero and CRT-AI
Ciara.Breathnach@ul.ie,
Tiziana.Margaria@ul.ie

Abstract. This track is the first output in Digital Humanities within AISoLA and, while its contributions span a range of diverse topics and approaches, it provides a good representation of the state of the art in the field, stemming from the interdisciplinary collaborations in the DBDIrl project and from the Great Leap COST Action that started in September 2023. It also underpins an ambitious research agenda arising from these collaborations, which aims to foster further international work on data interoperability.
The papers discuss the challenges faced by both computing and historical sciences when addressing on one side some of the most pressing issues of data access, preservation, conservation, harmonisation across national datasets, and governance, and on the other side the opportunities and threats brought by AI and machine learning to the advancement of reasoning, classification and rigorous data analytics.

1 The Topic: DH meets AI and IT

We are in the middle of an AI and IT revolution and at a point of digital cultural heritage data saturation, but humanities' scholarship is struggling to keep pace. In this Track we discuss the challenges faced by both computing and historical sciences to outline a roadmap to address some of the most pressing issues of data access, preservation, conservation, harmonisation across national datasets, and governance on one side, and the opportunities and threats brought by AI and machine learning to the advancement of rigorous data analytics. We welcomed contributions that address the following and other related topics:

- Advances brought by modern software development, AI, ML and data analytics to the transcription of documents and sources (Pedersen et al. [12], Mourits and Riswick [10] and O'Shea et al. [11]).
- Tools and platforms that address the digital divide between physical, analog or digital sources and the level of curation of datasets needed for modern analytics (Le Roux and Gasperini [8], Zafeiridi et al. [16] and Breathnach et al. [3]).

- Design for accessibility an interoperability of data sets, including corpora and thesauri (Le Roux and Gasperini [8], Mourits et al. [10] and Breathnach et al. [3]).
- Tools and techniques for machine-understanding form-based documents, recognition of digits and codes, handwriting, and other semantically structured data (Walsh and Clancy [14], Pedersen et al. [12], Mourits et al. [10] and O'Shea et al. [11]).
- Knowledge representation for better analysis of semi-structured data from relevant domains (diaries, registers, reports, etc) (Pedersen et al. [12], Mourits et al. [10], O'Shea et al. [11], and Breathnach et al. [3]).
- Specific needs arising from the study of minority languages and populations, disadvantaged groups and any other rare or less documented phenomena and groups (Fissore et al. [5]).
- Challenges relative to the conservation, publication, curation, and governance of data as open access artefacts (Ferilli et al. [4]).
- Challenges relative to initial and continuing education and curricular or extracurricular professional formation in the digital humanities professions (Le Roux and Gasperini [8] and Fissore et al. [5])
- Spatial digital humanities (Walsh and Clancy [14]).
- Digital humanities aspects concerning occupation, medicine and health (Pedersen et al. [12] and Breathnach et al. [3]).

2 The Background: Digital Meets Humanities

Owing primarily to over forty years of born digital content and also because of the digitisation of handwritten and printed documents, we are currently in an age of 'digital abundance' [9]. Digitisation can take many forms, if we take a broad view and include photography as a means of preservation (which developed into microforms including microfilm and microfiche) then it has been a critical component of records management in the finance and banking sector since the 1920s [13]. In more recent times, it has been the panacea for conservation and preservation in cultural heritage since the 1940s [2]. While such activity has resolved problems associated with storing newspapers and large datasets like census returns and has taken fragile manuscripts out of the handling environment, it has created several legacy issues associated with obsolescence and discoverability. Indeed Milligan [9] has elucidated the problems associated with the digitisation of newspapers from microfilm, where old problems of shading, depth of field and blurring, have simply been replicated in large scale digital archive projects. It seems that the end-user, or the researcher, has received little priority in the quest for 'digital abundance'. Greater accessibility has created new vulnerabilities that range from the whims of corporate decision-making to cyberattacks. For example, in the case of the former, executives at the MTV News Website wiped all archival content from the web in June 2024 [7], which has dealt a devastating blow to music history as there are no hard copies of this born-digital content and surrogate copies are limited to a partial capture

on the internet archive on archive.org. Cyberattack has wreaked havoc at the British Library since November 2023, and has negated much of its enormous efforts in the digitisation of manuscripts and other data types over the last century. Clearly there are many old and new problems that scholars must address and here we showcase a range of projects that exhibit the great potential when researchers collaborate in interdisciplinary ways.

3 The Contributions in Context

In this Track, we welcomed contributions from interdisciplinarians who work in the field of Digital Humanities, broadly defined. It includes contributions that discuss how scholars can reuse legacy data, the tools and methods they employ to conduct that work, and the role of AI and Machine Learning can have in future research. Further contributions include the potentials of wearable tech and Large Language Models (LLMs) in mental health and in gamification in mathematics pedagogy. Ostensibly what emerged from the conference is a volume about data, its generation in various forms and how we, as interdisciplinarians, exploit its use.

This volume is arranged chronologically in accordance with the evolution of digitization processes and the advancement of analytical tools.

P1: It begins with **Le Roux and Gasperini**'s paper [8] discussing the problems that past decision making in digitisation projects create for current research. Taking medical literature pertaining to child health published between 1850 and 1914 in England, France and Italy as their sample they explore the problems that Optical Character Recognition (OCR) generates for digital humanists and how computer science methods might help to resolve them. Adopting a corpus linguistics approach they argue that discoverability remains a problem and that the retrospective application of Free Accessible Interoperable and Reusable (FAIR) [15] principles will require closer collaboration between stakeholders as AI and ML technologies take further root.

Historical Demography features heavily in this track as historical 'Big Data' provide a useful comparative case study in old and new research methods. It showcases some of the work of collaborators in the COST Action CA22116 - The Great Leap. Multidisciplinary approaches to health inequalities, 1800-2022 (GREATLEAP) [1]. Tim Riswick is Chair and Breathnach is Vice Chair of this international network. Locating its work in historical population data The Great Leap examines structural inequalities in health and how they emerged over time, primarily across Europe. It is an interdisciplinary network of scholars and it counts contributors Clancy, Ferilli, Garret, Margaria, Mourits, Reid, Sommerseth and Walsh among its wider membership. Further to experts in history, social sciences, life sciences, computer sciences and epidemiology, it also involves government agencies and data owners such as statistical offices and national archives. Like the contributions to this volume, the Great Leap focuses

on individual-level metadata but it has the ambitious aim of creating interoperability between national datasets. The ultimate aim of the network is to push the capabilities of historical data to have impact in current public health policy and practice.

The Great Leap associated papers in this volume capture a timeline in the development of transcription and machine learning methods over the past 40 years. The earliest contributions to the field were constructed in North America and across Europe (for example, in the Netherlands) but with old technology and its associated methods, comes new challenges.

P2: **Mourits, Riskiwck, and Stapel** [10] deal with the matter of interoperability and common languages in historical demography, they note how modern ontologies are insufficient and cannot be neatly applied to the past. They have gathered data from several international projects to assess its FAIR compliance and found that occupational data was the most standardised category. They point to the success of past research projects like HISCO [6] and its work on standardisation and the encoding of occupations, and present some of their findings here.

P3: Building on Mourits et al., **Pedersen, Islam, Kristofferson, Bongo, Garrett, Reid and Somerseth** [12] discuss the problems surrounding data encoding and how useful LLMs might be in the retrospective application of ICD10 to historical cause of death. Experiments using three AI-models on a random data sample of death registrations from three areas in the UK (that were fully transcribed and encoded by domain experts) showed a varying success rate of correctness. While the authors found the exercise worthwhile they caution that significant levels of fine-tuning (perhaps using Retrieval-Augmented Generation) would be necessary to correct the error rate.

P4: Although they address a different problem, **Fissore, Floris, Marchiso Conte and Sacchet** [5] examine the transformative potential of gamification, LLMs and AI to shift away from didactic approaches to teaching mathematics and create more opportunities for co-production in education. Identifying the specialised language of mathematics as a major obstacle to learning, they recommend a Data-Driven Learning (DDL) methodology to help teachers and students identify and work through difficulties in a scaffolded digital learning environment. They note that while further in-service training is necessary gamification offers massive opportunities to individualise learning while also supporting the educators.

P5: **Breathnach, Murphy, Schieweck and Margaria** [3] describe the difficulties of interoperating various Irish historical datasets uncovered during Death and Burial Data: Ireland 1864-1922 (DBDIrl) a project funded by the Irish Research Council (2018-2023). Taking old age as a case study the authors who are drawn from history and computer science describe the pipelines necessary to automate the process of data linkage using a low-code no-code approach. While age heaping (rounding to nearest 0 or 5) is a peren-

nial problem in census data and disrupts efforts to create linkages between decennial data, marital status was also found to be inconsistently recorded.

P6: In another paper arising from DBDIrl activities, **O'Shea, Krumrey, Mitwalli, Teumert and Margaria**'s work [11] follows the technical problems that limitations in OCR technology poses and the necessity for precise responses tailored to the problems of each data type. It discusses the benefits and limitations of an AI-ML Data Analytics Pipeline designed as an automated solution to transcription of historical handwritten death registers. In many respects this work shows how countries without longstanding traditions in crowdsourced transcription data (like those described by Mourits and Riswick [10] and Pedersen et al. [12]) can play catch-up through pipelines of segmentation, word detection, and data synthesis, classification and linkage.

P7: Spatial epidemiology underpins **Walsh and Clancy**'s paper [14] on the Irish District Lunatic Asylum system. They illustrate the potential that the mapping of individual level data can bring to our understandings of mental health problems in both urban and rural Connaught (a large province in Ireland). The removal of patients from their households of origin was a key component of the 'confinement' model and in this chapter Walsh and Clancy interoperate committal data with civil registration records to plot the lives and movement of patients.

P8: Albeit in a modern context, **Zafeiridi, Qirtas, Bantry White and Pesch** [16] explore how machine learning models and wearable tech can act in a preventive capacity in detecting depression. They found that GPS should not be used in isolation and that passive sensing, which includes activity, sleep data and personal communication (texts, calls, Apps and proximity to other smart devices), offered a more holistic approach to the study of mental health and well being.

P9: With a focus on discoverability **Ferilli, Bernasconi, Di Pierro and Redavid** [4] contributes the final paper in this volume. It goes beyond the analytical limitations of the digital archive and shows how the GraphBRAIN framework for knowledge graphs operates in the domain of cultural heritage. In this project they show the realms of possibility for conservation, preservation and, more importantly, discoverability and interoperability.

4 The Next Steps

This track is the first output in Digital Humanities for AISoLA and, while its contributions span a range of diverse topics and approaches, it provides a good representation of the state of the art in the field. Several exciting prospects emerge from the use cases presented here, for example the capacity of LLMs and GIS to reconfigure legacy data and offer new insights from old 'Big Data', and the use of tools like gamification drawing on corpus linguistics and DDL to transform

mathematics pedagogy. Most importantly this volume shows the importance of collaborative and interdisciplinary approaches to complex research questions, and the enormous potential for future innovative scholarship. We expect further collaborations to arise from the Great Leap context, which started in September 2023 and will provide opportunities of collaboration with its network for the next three years.

References

1. The Great Leap. Multidisciplinary approaches to health inequalities, 1800 - 2022. EU COST Action CA22116. https://greatleap.eu/, [Online; accessed 28 July 2024]
2. Archives, U.S.N., Service, R.: National Archives and Records Service Microfilm Publications. General information leaflet, U.S. General Services Administration, National Archives and Records Service (1975), https://books.google.it/books?id=W6sXiHcrHdsC
3. Breathnach, C., Murphy, R., Schieweck, A., Margaria, T.: Interoperating Civil Registration of Death and Census Data: Old Age and Marriage as Categories of Analysis. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
4. Ferilli, S., Bernasconi, E., Di Pierro, D., Redavid, D.: The GraphBRAIN Framework for Knowledge Graph Management and its Applications to Cultural Heritage. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
5. Fissore, C., Floris, F., Marchisio Conte, M., Sacchet, M.: Teaching the specialized language of Mathematics with a data-driven approach: what data do we use? In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
6. International Institute of Social History: History of Work - HISCO. https://iisg.amsterdam/en/data/data-websites/history-of-work, [Online; accessed 28 July 2024]
7. Janine, D.: MTV Deleted! The Iconic Music TV Channel of the 80's & 90's is Gone. https://stagelync.com/news/mtv-deleted, [Online; accessed 28 July 2024]
8. Le Roux, M., Gasperini, A.: Digitised historical sources and non-digital humanists: an interdisciplinary challenge? In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
9. Milligan, I.: The Transformation of Historical Research in the Digital Age. Elements in Historical Theory and Practice, Cambridge University Press (2022)
10. Mourits, R.J., Riswick, Tim amnd Stapel, R.: Common Language for Accessibility, Interoperability, and Reusability in Historical Demography. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
11. O'Shea, E., Krumrey, M., Mitwalli, D.S., Teumert, S., Margaria, T.: From Data Science to Modular Workflows - Changing Perspectives from Data to Platform: DBDIrl 1864-1922 Case Study. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)

12. Pedersen, B.R., Islam, M., Kristoffersen, D.T., Bongo, L.A., Garrett, E., Reid, A., Sommerseth, H.: Coding historical causes of death data with Large Language Models. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
13. Teague, S.J.: Microform, Video and Electronic Media Librarianship. K. G. Saur Publishing, Inc., USA (1985)
14. Walsh, O., Clancy, S.: Mapping Madness: HGIS and the granular analysis of Irish patient records. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)
15. Wilkinson, M.D., Dumontier, M., et al.: The fair guiding principles for scientific data management and stewardship. Scientific Data **3**(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18, https://doi.org/10.1038/sdata.2016.18
16. Zafeiridi, E., Qirtas, M.M., Bantry White, E., Pesch, D.: Using Passive Sensing to Identify Depression. In: Proc. AISoLA 2023, Special Track *Digital Humanities and Cultural Heritage in AI and IT-enabled Environments*. p. (this volume). Springer Nature (2024)