Unifying Legacy Online Data with Knowledge Graphs: Archiving and Preserving Digital Projects of the Max Planck Institute for the History of Science

Hassan El-Hajj[1,2*], Steffen Hennicke[1], Pascal Belouin[1], Robert Casties[1], Robert Egel[1], Wishyut Pitawanik[1], and Kim Pham[1]

[1] Max Planck Institute for the History of Science, Boltzmannstr. 22, Berlin, 14195, Germany
[2] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, 10587, Germany
* hhajj@mpiwg-berlin.mpg.de

The term Digital Humanities (DH) has become established in most humanities disciplines in recent decades; numerous projects have since then collected and curated digital objects in extensive databases, applied digital methods to interpret and analyse them, and used novel ways to present their findings and hypotheses on the Web. In view of the substantial digital heritage of the digital humanities, one of the central challenges is now the question of how the long-term reuse of these valuable research data can be made possible in the face of outdated technology and increasingly expensive maintenance.

The Max Planck Institute for the History of Science (MPIWG), founded in the mid 90s, has been heavily engaged in digital scholarship from the very beginning. The result of this development is an extensive portfolio of digital offerings, ranging from organising and serving digitized Galileo manuscripts,[1] to virtual exhibitions[2] and influential databases.[3] In this presentation, we address the question of how we can preserve the MPIWG's digital legacy and how we are working towards a more sustainable institutional management of our research data so as to enable better long-term findability and re-usability. As one of the cornerstones of our institutional research data management strategy we are working on a graph database, the Central Knowledge Graph (CKG), where we plan to publish the project's collected research data as Linked Open Data. Most project data is mapped and converted into a common target model based on the CIDOC CRM formal ontology.[4] The CIDOC CRM plays a significant role in enabling FAIR[5] data representation, in particular by providing a semantically well-defined vocabulary for describing cultural heritage data.

Additionally, in the CKG we also document the scientific context of research projects, such as their research topics, the project members, and, in particular, their digital outcomes, i.e., the databases, datasets, and software they have produced and where they have been archived or published. For this purpose we have developed the *Project Description Layer Model* (PDLM) which is based on the Parthenos ontology[6] that we adapted and extended to our use case. We plan to publish the PDLM as an extension to the CIDOC CRM. The dedicated front end to the CKG is built with ResearchSpace[7] through which we aim to unify the different digital objects and their metadata in a single platform allowing MPIWG researchers to access and reuse data that has long been offline.

In our talk, we will focus on the different steps towards building a Proof of Concept (PoC) for the more than hundred legacy projects that have accumulated on the Institute's servers.

[1] https://www.mpiwg-berlin.mpg.de/Galileo_Prototype/MAIN.HTM published 1998
[2] http://pratolino.mpiwg-berlin.mpg.de/ published 2007
[3] https://dmd.mpiwg-berlin.mpg.de/home published 2006
[4] https://cidoc-crm.org (06.02.2024)
[5] https://www.go-fair.org/fair-principles/ (06.02.2024)
[6] https://www.parthenos-project.eu/Download/Deliverables/D5.1_Common_Semantic_Framework_Appendices.pdf (06.02.2024)
[7] https://researchspace.org (06.02.2024)

These legacy projects constitute a major part of the institute's digital history, and hold a treasure trove of digital and historical information that is in danger of becoming inaccessible online due to the increasing obsolescence of their technological infrastructure. To achieve this PoC, we addressed issues including project sampling, building a simple yet expressive ontological model, and building a pipeline that assembles all these different components, tests for data correctness, and serves it to its users, i.e. MPIWG researchers. We focus on reconciling and aligning the different entities representing the same historical document, actor, or event coming from different legacy projects. The problem of entity alignment is well known within the Knowledge Graph community, with numerous papers addressing the problem using simple entity embedding [4], Graph Neural Networks [3], and more recently multi-modal transformers aiming to include images (and other modalities) to enhance the entity alignment results [1]. Such approaches often lead to sub-par results when faced with highly detailed CIDOC CRM modelled Knowledge Graphs, with the added problem of data incompleteness that often results from salvaging older projects on historical topics. To remedy this, and to tailor this approach to our specific needs, we experiment with a pipeline that starts with 1) compartmentalizing our knowledge graph, 2) reducing the complexity of these compartmentalized graphs by mapping them to an abstract layer of simpler relations and produce derivative products, 3) inferring entity relation on the reduced graph product and finally 4) mapping these relations back to the original knowledge graph.

We believe that it is no longer possible to think about digital humanities without thinking about the preservation of its results. No longer an abstract concern, the capturing and the preservation of digital and web materials as data is recognized to have an immense research value in the future, as well as providing a institutional record. Preserving the research products alongside mapping networks provides the social and cultural context for richer historical analysis in the future [2]. In this sense, our work underscores the important relationship that exists between the act of creating digital knowledge and preserving it, and proposes a sustainable vision for maintaining and extending the 'shelf-life' of digital heritage at the MPIWG and beyond.

References

[1] Chen, Z., Chen, J., Zhang, W., Guo, L., Fang, Y., Huang, Y., Zhang, Y., Geng, Y., Pan, J. Z., Song, W., and Chen, H. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia* (New York, NY, USA, 2023), MM '23, Association for Computing Machinery, p. 3317–3327.

[2] Milligan, I. Lost in the infinite archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing 10*, 1 (2016), 78–94.

[3] Tam, N. T., Trung, H. T., Yin, H., Van Vinh, T., Sakong, D., Zheng, B., and Hung, N. Q. V. Entity alignment for knowledge graphs with multi-order convolutional networks. *IEEE Transactions on Knowledge and Data Engineering 34*, 9 (2022), 4201–4214.

[4] Wang, Z., Zhang, J., Feng, J., and Chen, Z. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), AAAI'14, AAAI Press, p. 1112–1119.