CHNT 29

Session 217: Cultural Heritage and the new future of new technologies

Semantics and the integration of diverse data sets

The integration of rich, diverse data sets is a skill that is central to archaeological practice. It lies at the core of the archaeologist's role as a synthesist. Currently it relies on the immersion of the scholar in the pool of data that is to be integrated. This effectively places an upper limit on the quantity and diversity of data sources that can be integrated. The creation of synthetic data repositories tailored to a particular research question has long been a strategy to extend the size of the pool but it has some serious shortcomings. It requires a significant intellectual effort to design the repository and a large logistical effort to compile the contents. When the work is completed the repository may be of little use when addressing new questions that it was not originally designed to accommodate.

The brave new world of semantically marked-up Linked-Open-Data (LOD) will allow a boundless lake of data to be integrated. New skills will be required both to prepare the data for integration and to query it: these are relatively trivial to acquire and understand. However, a new layer of skills is also required: these are the creation, assessment and use of para-data to understand if the, now easily accessible, data is fit for your purpose.

Current generative AIs (relying on Stochastic Large Language Models (LLMs)) are poor at synthesising from raw data (Saba 2023). However, as large quantities of richly marked-up LOD become available, together with the machine-readable semantic ontologies used to mark them up, this will become more mainstream. New AI tools that moderate and support long traversals in LOD will also become available. However, the questions will still have to come from us.

The paper aims to provide a number of points of interest for the audience to consider:-

- What is para-data?
- What is the role of para-data in future research practice?
- How do we cross disciplinary boundaries in LOD?
- What support will we, as scholars, need to exploit these new, rich resources?
- What support will we, as scholars, need to create these new, rich resources complete with their essential para-data?
- How should we represent this rich data and its context?
- How do we represent differences of opinion and poly-vocality?

This paper will demonstrate these areas with data that uses the new RDF* representation of LOD (Hiebel et al 2024). It will show some of the rich and long

traversals that can be explored using existing tools. The examples include the representation of data, meta-data and para-data from the archaeological and oral history domains.

Heibel, G., Friedburg, P.F., Stead, S.D., 2024 *RDF creation pipeline*. (available at https://zenodo.org/records/11092544) (Accessed 2024/07/01)

Saba, W.S. (2023). Stochastic LLMs do not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs. In: Almeida, J.P.A., Borbinha, J., Guizzardi, G., Link, S., Zdravkovic, J. (eds) Conceptual Modeling. ER 2023. Lecture Notes in Computer Science, vol 14320. Springer, Cham. https://doi.org/10.1007/978-3-031-47262-6_1