# Connecting the dots: the case of Omnipot

Matthias Schlögl*[1], Matej Durco*[1], Ingo Börner*[1], Peter Andorfer[1], and Klaus Illmayer*[1]

[1]Austrian Centre for Digital Humanities and Cultural Heritage (Austrian Academy of Sciences) – Austria

**Abstract**

Recent years have seen the rise of linked data in the digital humanities. Alongside with these ideas of "[...]using the Web to create typed links between data from different sources."[1] this development has driven a broad discussion about interoperability of data and standardisation[2]. One consequence of these discussions was the increased adoption of commonly used high level ontologies such as CIDOC CRM[3].

Data interoperability and integration into the linked open data cloud however comes at a cost. RDF is still not a mainstream technology, nor are triplestores which are needed to store and process RDF. While there exist several web development frameworks and content management systems for traditional tech stacks (e.g. SQL db, PHP, html) that ease the development burden for CRUD (create, update, delete) applications, similar tools are widely missing for RDF based tech stacks. To get the best of both worlds - ease of development and stability from the "relational tech stack" and linked open data and flexibility from the "RDF tech stack" - we manage entities data (such as prosopographies etc.) in applications built on relational databases and serialize the data later into RDF.

The ACDH-CH was founded to foster the use of digital methods in the humanities in Austria. The institute is involved in dozens of very diverse DH projects, generating large quantities of heterogeneous datasets.
Conceptually many connections/relations between these datasets can be identified. They may be representing similar kinds of entities or cover the same spatial or temporal scope. However, given the diverse contexts of the projects the datasets are not compatible, nor easily harmonizable into a common data model that would allow to use/explore them across project boundaries.

We have therefore decided for an opportunistic approach: Under the codeword "Omnipot" we put data from several projects into one common triple store, accepting any underlying ontology as well as only partially mapped data. We then use a customisation of the semantic knowledge platform ResearchSpace[4] to explore the links between these datasets and to improve the mapping to CIDOC CRM. We keep the data of the projects and the metadata in separate named graphs. To make the transformations and harmonizations explicit new named graphs are created that contain consolidated data.

*Speaker

This workflow brings several challenges along that are only partly solved so far. Omnipot is meant as an entity hub that should also be usable as a reference resource. To make that process efficient there need to be i) automatic scripts for the serialization of data from the source applications. These scripts make it not only easy to keep Omnipot up to date, but also make the mapping of the data explicit and allow us to keep provenance records. When the original named graphs are updated the ii) consolidated named graphs need to be updated too. As a consequence there is a need for an "orchestrator service" that runs the necessary updates on the triple store as well as the data stores of the various projects. It is good practice to avoid blank nodes wherever possible, URIs are needed for Researchspace to show rdfs:labels and they allow to diff changes from one serialization to the next[5]. While these advantages are obvious it can be iii) challenging to produce unique - but reproducible - URIs for e.g. attributes in relational databases. These rather technical issues aside there are still iv) modeling issues that - to our best knowledge - haven't been solved in CIDOC CRM or its extensions yet.

We believe that the workflow outlined here takes the realities of many DH research institutes - legacy data, legacy code and limited resources - into account and could act as a blueprint for further discussions in the Data for History (D4H) community. Such a workflow also has a positive effect on (meta)data quality issues. As one crucial key factor of Omnipot is the implementation of manageable mapping processes, there is the need to develop common agreements between projects, e.g. using the same authority files to identify entity matches. Additionally, this stimulates convergence of data models at the level of projects. Ideally, such a feedback loop does not only lay the foundation for entanglement of data but also for better machine-processable data as it is envisaged by the FAIR data principles[6].

The presentation at the D4H conference in Berlin will focus on i) the workflow used including examples from digital editions, prosopographies and cultural heritage collections, ii) limitations encountered while modeling our data in CIDOC CRM and iii) advantages and disadvantages of the ResearchSpace platform in a workflow like ours.

---

Christian Bizer, Tom Heath, and Tim Berners-Lee, 'Linked Data - The Story So Far', 26 .

For the library world e.g. see Bernhard Haslhofer, Antoine Isaac, and Rainer Simon, 'Knowledge Graphs in the Libraries and Digital Humanities Domain', ArXiv:1803.03198 [Cs], 2018, 1–8 . For historical research e.g. Alison Langmead and others, 'Towards Interoperable Network Ontologies for the Digital Humanities', International Journal of Humanities and Arts Computing, 10.1 (2016), 22–35 .

Patrick Le Boeuf and others, Definition of the CIDOC Conceptual Reference Model, 2017.

https://www.researchspace.org/

Most libraries/tools used for creating RDFs create random hashes that change in every serialization.

https://www.go-fair.org/fair-principles/

**Keywords:** data, modeling, RDF, researchspace, named graphs