Towards a Dynamic Knowledge Graph of a Non-Western Book Tradition

Kinitz, Daniel

kinitz@saw-leipzig.de
Saxon Academy od Sciences and Humanities, Germany

Efer, Thomas

efer@informatik.uni-leipzig.de
Leipzig University

Introduction

How can we generate and integrate data on a pre-modern, Arabic book tradition in such a way that research can gain new insights? In the long-term project “Bibliotheca Arabica” (Brinkmann/Löhr 2021), we are creating an agile knowledge graph (Kinitz/Klemm 2020ff.) integrating a wide range of data on (handwritten) Arabic manuscripts and their historical context. The main sources for the long-term data integration are more than 250 volumes of manuscript catalogues, Linked Open Data as well as tens of thousands of manuscript documentary notes by readers, owners etc. Our aim is to create a digital research environment to investigate the production, transmission and reception of Arabic manuscripts and their social environment as clusters of linked entities: scholars linked to works with students as readers, reproduced by scribes in manuscripts, linked by ownership notes and combined to historical libraries, etc.

With this contribution, we aim to share our digital approaches that can promote the ongoing evolution in the field of contextualised distant readings of post-colonial intellectual history.

Challenges and Approaches

What are the challenges in developing such an integrated knowledge graph, and how can digital methods address them?

The **first challenge** is the integration of large resources that are heterogeneous in multiple ways regarding their

- file formats (binary: scans, pdf, xlsx, db formats; text: xml, json, txt),
- structuredness (e.g. full text, semistructured, tables),
- languages (Arabic, Persian, Hebrew, English, German; partially mixed),
- alphabets (Arabic, Latin), transcription standards and intermixed writing directions (right to left, left to right), as well as
- ontologies – from undeclared catalogue ontologies to library-driven standards such as MARC21.

We address this challenge through agile data modelling/integration and a source-based, multi-layered ontology. This includes the distinction between raw data (catalogues, reference works, etc.), its representation within the project (atomised, quality checked) and its integration into the actual graph (entities connected through associations). While keeping identifiers persistent, an adjustable, permanently improved processing chain for every source with flexibility regarding changing research questions is enabled.

Modelling complex knowledge from manuscript catalogues while preserving provenance chains of fact(oid)s (Bradley/Short 2005) is the **second challenge**. Building on a Property Graph framework, we use associations inspired by the Topic Maps data model (ISO 2006) to connect entities in n-ary relations with an unlimited number of role players. These associations provide anchor points for the reification of relationships, much like RDF-Star does for RDF data. Building on this analogy, we implememt factoidal provenance chains (Efer 2019) for associations in a similar fashion to the PROV Ontology.

The **third challenge** lies in a flexible authority control of entities, managing conflicting statements and authorship attributions. We do not have central authority records, but linked entity attestations. Identity claims are virtually merged; contradictory statements and authorship attributions are expressed as relations (associations) with their specific factoidal provenance.

Our **fourth challenge** is postcolonial data infrastructures, such as international authority files and library systems, which ignore Arabic idiosyncrasies like complex person names and the non-Western Hijri calendar. Once imported, this kind of westernised data can be disambiguated, but not returned due to lack of exchange standards.

The size and complexity of the graph and the evidence-focused-data modelling lead to scalability and performance issues, which is our **fifth challenge**. We use caching mechanisms to reduce requery runtime, a Lucenebased full text index for node discovery, and are working to further refine our graph queries. These are generally fast for small result sets and slower for large sets. We are experimenting with interactive queries that give the users feedback on the expected query time.

A research platform may have conflicting objectives defined by potential target groups and respective functionalities – in our case a database as a robust, user-friendly reference work versus a research-driven prototype, integrating latest technologies. We have not yet developed an optimal solution to this **sixth challenge**. One possible approach is to use an experimental graph database prototype and, additionally, to run a simple but stable database using technically mature components.

Conclusion

A factoid-provenance driven, graph-based approach to intellectual history has the “revolutionary potential” (ADHO 2022) to challenge analogue historical narratives by enabling data-centric multidimensional and contradictory distant reading approaches. Within our domain, Arabic Studies and neighbouring fields, we can expect traditional narratives to be challenged – such as the subordinate role of the cultural “periphery” (versus the established centres) and the apparent decline of Arabic literature after its “classical” period.

Bibliography

Alliance of Digital Humanities Organizations (2022): "Digital Humanities 2023: Collaboration as opportunity", Conference Call, https://dh2023.adho.org/?page_id=308.

Bradley, John / Short, Harold (2005): “Texts into Databases: The Evolving Field of New-style Prosopography” , in: *Literary and Linguistic Computing* 20, Suppl Issue, 1-24.

Brinkmann, Stefanie / Löhr, Nadine (2021): “Bibliotheca Arabica – Towards a New History of Arabic Literature”, in: *Comparative Oriental Manuscript Studies Bulletin* 7, 197-206, https://www.fdr.uni-hamburg.de/record/9871.

Efer, Thomas (2019): “Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen “, in: *Zeitschrift für digitale Geisteswissenschaften*, special issue 4, https://zfdg.de/sb004_011.

International Organization for Standardization (2006): “Topic Maps – Part 2: Data model”, ISO/IEC Standard No. 13250-2:2006.

Kinitz, Daniel / Klemm, Verena (2020ff.): KHIZANA. Bibliotheca Arabica’s Reference Work on the Arabic Manuscript Tradition, https://khizana.bibliotheca-arabica.de.