MEHDIE: The Middle East Heritage Data Integration Endeavor

Rusinek, Sinai

sinai.rusinek@mail.huji.ac.il Haifa University, Israel

Sagi, Tomer

tsagi@cs.aau.dk Haifa University, Israel

Zaga, Moran

moran82@gmail.com Haifa University, Israel

Lev, Efraim

elev@univ.haifa.ac.il Haifa University, Israel

Moshe, Lavee

mlavee@univ.haifa.ac.il Haifa University, Israel

MEHDIE, the Middle East Heritage Data Integration Endeavour is a project at Haifa University and Aalborg University, dedicated to facilitating semantic integration of knowledge sources related to the history of the Middle East, using innovative data integration methods.

The multi-modal entity resolution tool we are developing will utilise spatial (geo-coordinates), temporal (dates/periods) and textual information about a pair of place/person records to help identify the semantic relationship between them. In the demonstration we present tools related to spatial data, which we hope to integrate into the *World Historical Gazetteer* (Grossner and Mostern 2021) system to enhance its capabilities for inter-dataset matching. The poster will address our workflows for data extraction and assembly, as well as the results of matching experiments between various datasets . We describe how we leverage the linguistic affinities and continuities between the domain languages and various textual matching techniques.



In figure 1 we see the interface of the MEHDIE clone of the World Historical Gazetteer, with the added option to match two

datasets against each other, in this case a dataset of the Hebrew toponyms from Benjamin of Tudela's travelog collected in the framework of the *TraveLab* project (Rusinek and Shalit 2022), and *Damast* Multilingual Dataset from the Dhimmis and & Muslims project (Franke and Kock 2023). Also visible in the figure are the confidence measure and the explanation of the matching methods that yielded the candidate.

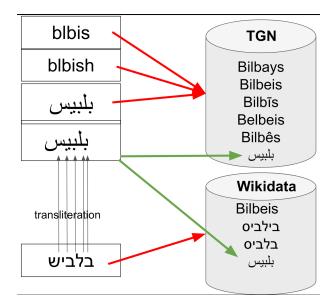
The mission to connect data on the Middle eastern history domain is faced with special challenges: the scarcity of vowels in Semitic scripts, the indeterminacy of pronunciation and transliteration of written toponyms, and varying practices of naming and classifying places.

A common way to identify places, their locations and other information about them is to reconcile the source place names with unique identifiers from an external Gazetteer. *Geonames* and the *Getty Thesaurus of geographic names* are two such resources, but they contain few, if any, toponyms in the Semitic scripts relevant to the sources of Middle Eastern History. Even Wikidata, which is essentially multilingual, shows low recall when matching against them. Such general gazetteers may include contemporary Arabic, Hebrew or Syriac forms that denote major, well known cities, but very little historic variants of the lesser known settlements.

With this obstacle in mind, one may turn to transliteration. The similarity between semitic languages, however, means that direct transliteration between semitic languages would be much more efficient in creating candidates for matching that transliteration into Latin script languages.

The example in figure 2 is a telling illustration for this:

The Hebrew toponym for the city of Bilbeis on the Nile's delta (spelled "Bilbays" in Adler's translation of Benjamin of Tudela) was mentioned in the 19th century edition of Benjamin of Tudela's itinerary in the Hebrew form שיבלב. In contemporary Hebrew it is spelled סיבליב, with an additional vowel and a different letter to represent the final consonant. Wikidata currently includes a He-brew alternative variant without the additional vowel, but not avariant with the more archaic letter 'ש' (Wikidata contributors, "Q393495").



Using the transliteration tool in our system to match against Ara-bic datasets such as *Usaybia*, *Damast*, 'Al-Turayya gazetteer (Ro-manov and Seydi, 2019) and from a dataset of toponyms from

Yaqut al-Hamawi's Mu 'jam ul-Buldān, we could easily find the match thanks to the transliteration of the Hebrew form to the Ara-bic form سيبلب, one of two possible transliterations. This shows the power of transliteration between semitic languages that redu-ces, rather than adds ambiguity: Neither Arabic nor Hebrew ex-plicate the sound that is expressed in English by the first i and the e in the name. Therefore, transliteration to latin script could not deduce an acceptable transliteration.

Several projects have been dedicated in recent years to the aggregation of geographic and prosopographic data on Middle Eastern History. The *Historical Middle East Data Alliance (Hist-ME)* was initiated by Nathan P. Gibson at the *Linked Pasts 2021* confe-rence with a joint vision of connecting intra-religious and mono-lingual silos of disconnected data into a linked contextual wealth of knowledge. The system we are developing at *MEHDI* facilita-tes collaboration between *Hist-ME*'s partners towards this vision.

The advantages of intra-domain transliteration for matching opens a meta-discussion on approaches for linked open data and the preference of a "bottom-up" intra-domain strategy over a centralised one. The results of our work with the different datasets of Hist-ME member projects may also constitute an argument in favour of intra-domain collaboration.

Bibliography

Carlson, Thomas A (2022). Historical Index of the Medieval Middle East (HIMME), a digital history project. https://medieval-mideast.org

Franke, M., & Koch, S (2023). "Damast: A visual analysis approach for religious history research". In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023)* - Volume 3: IVAPP, pages 40-52. DOI: 10.5220/0011609700003417

Gibson, Nathan P (Ed.). (2020). Communities of knowledge: Interreligious networks of scholars in Ibn Abi Usaybi#a's history of the physicians. Version 0.6.0-dev. Munich: Usaybia.net. https://usaybia.net. DOI: 10.5281/zenodo.3975505.

Grossner, Karl / Mostern, Ruth (2021). "Linked Places in World Historical Gazetteer". In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities '21)* (pp. 40-43). Association for Computing Ma-chinery. DOI: 10.1145/3486187.3490203

Romanov, Maxim / Seydi, Masoumeh (2019). "Al-Turayya, the Gazetteer and the Geospatial Model of the Early Isla-mic World". In *Digital Humanities 2019 Conference Papers* (9–12 July 2019), Utrecht University. clariah.nl/files/dh2019/boa/0909.html

Rusinek, Sinai / Shalit, Gil (2002). *TraveLab: A multilingual mapped edition of Benjamin of Tudela's travelogue*. https://teipublisher.info/exist/apps/TraveLab/index.html

Wikidata contributors (2002). "Q393495. Wikidata". Retrieved November 4, 2022, from https://www.wikidata.org/w/index.php?title=Q393495&oldid=1710404536