Bridging the Gap:
Digital Humanities and the Arabic-Islamic Corpus

Dafne van Kuppevelt, Patrick Bos, Melle Lyklema, Christian Lange, Umar Ryad, Janneke van der Zwaan

Introduction

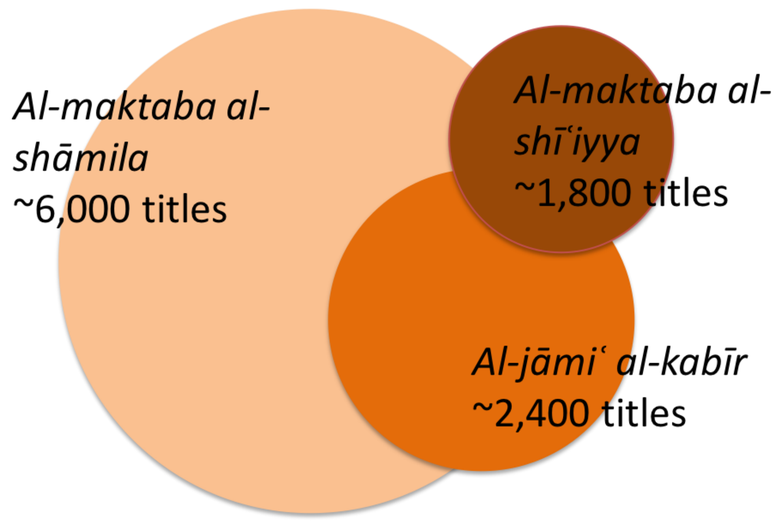
Despite some pioneering efforts in recent times, the longue durée analysis of conceptual history in the Islamic world remains largely unexplored. Researchers of Islamic intellectual history still tend to study a certain canon of texts, made available by previous Western researchers of the Islamic world. Indigenous conceptual developments and innovations are therefore insufficiently understood, particularly concerning the transition from premodern to modern thought in Islam.

What, then, are the silenced continuities, transformations and major fault lines in Arabic-Islamic discourses? The Islamic tradition offers a vast textual corpus for exploring this question from a longue durée perspective, but its very breadth poses substantial problems for the individual scholar seeking to survey the literature by traditional methods. This poster presents ongoing research to use state-of-the art Digital Humanities approaches and technologies to make pioneering forays into the vast corpus of digitized Arabic.

The Arabic Digital Corpus

Digitized materials

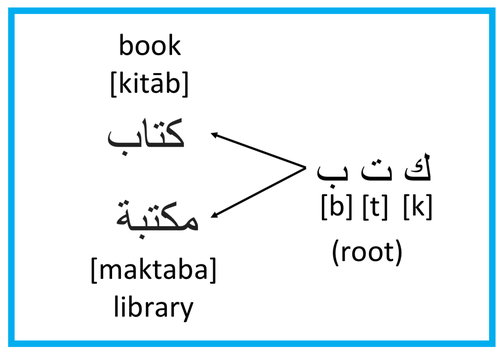
- ca. 7,900 unique titles
- ~1,1 billion words (vs. 150 million in classical Greek and Latin combined)
- ~4,000 unique titles pre-1900
- <10% of what may have been written



Challenges Arabic digital humanities

Many existing tools for digital humanities are **not adapted** for non-western languages. The **right-to-left** text direction is problematic for many web applications

Arabic is a **morphologically complex** language. 3-letter roots are combined with different prefixes, infixes and suffixes to construct semantically related words. Thus, **searching on roots** is valuable for research on concepts.

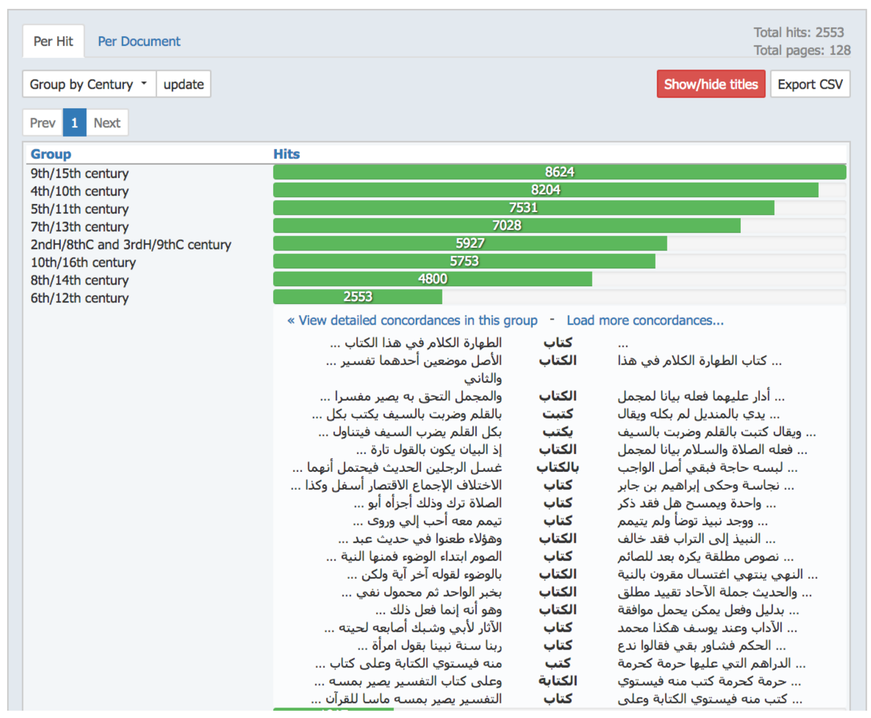


Digital solution

Combine and adapt existing tools: **nlppln** [1] for creating flexible pipelines for corpus processing, **SAFAR** [2] and **Al-Khalil** [3] for morphological analysis, **Blacklab** [4] as a search engine.

Try the demo!

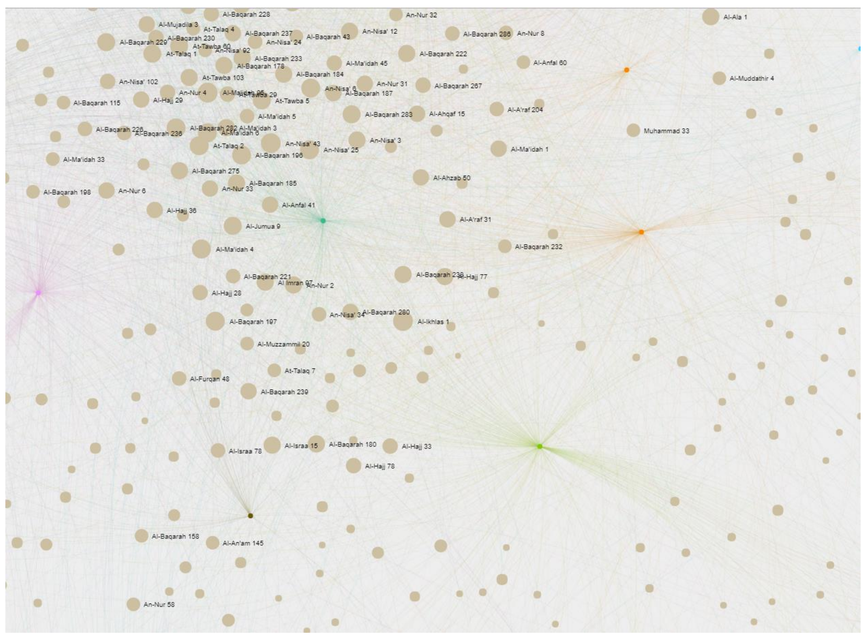
http://arabic-dh.hum.uu.nl/corpus-frontend



- Root search
- View local context of search results
- Aggregate results on metadata (e.g., per century)

Citation analysis

- Which Quran and Hadith verses are quoted in the corpus?
- Do we see patterns in the citations? E.g., schools of thought that cite similar sources



Visualization made with Gephi – Sigma js plugin

References

[1] Janneke van der Zwaan, & Dafne van Kuppevelt. nlppln http://doi.org/10.5281/zenodo.1185197

[2] Jaafar, Y. and Bouzoubaa, K. (2015) 'Arabic Natural Language Processing from Software Engineering to Complex Pipeline', in 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE, pp. 29–36

[3] Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A. and Boudlal, A. (2017) 'AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer', Journal of King Saud University - Computer and Information Sciences. Elsevier, 29(2), pp. 141–146.

[4] http://inl.github.io/BlackLab/

Want to learn more?

Contact us! d.vankuppevelt@esciencecenter.nl

Utrecht University

netherlands eScience center