Authors / Co-PIs* (alphabetically): Matthew Thomas Miller*, U Maryland–College Park; Maxim Romanov*, U Wien; Sarah Bowen Savant*, Aga Khan U (ISMC)–London; Masoumeh Seydi, U Leipzig

CREATING eTEXTS: CorpusBuilder 1.0

In order to achieve comprehensiveness, we are developing a digital text production pipeline. Called CorpusBuilder, this user-friendly, web-based, open-source application will significantly lower the technological barriers to entry and help involve people from around the world in our project of corpus creation. Supported by a Mellon grant since 2019.

TAGGING: OpenITI mARkdown

Texts have been automatically converted into our custom format — OpenITI mARkdown (AR stands for Arabic). Facilitating conversion of raw texts into machine-actionable formats, this flavor of markdown 1) simplifies work with multivolume texts; and 2) helps to avoid problems that one faces when paired symbols (e.g., angle brackets), LTR and RTL languages, and connected scripts occur in the same document. mARkdown will also facilitate conversion into TEI XML. For description: https://maximromanov.github.io/mARkdown/.

STUDYING: Research Projects

ROLLING STYLOMETRY TEST. Three samples of 10,000 words were taken from the beginning (red), middle (green), and end (blue) of al-Dhahabi’s Taʾrikh al-islam and used to test to what extent the “style” of these samples is similar to the rest of the book: the “early style” (red) dominates the language of the 1st Islamic century, disappears completely by the end of the 3rd Islamic century, not reaching even the middle of the book. The style in the end of the book is completely different from that of the beginning of the book.

MANAGING: a Python Library

We are developing a specialized Python library that will make working with the corpus easier for scholars.

COLLECTING eTEXTS: Open Libraries

1,859 authors; 7,144 texts (4,288 unique); 1,5 billion words

ORGANIZING: CTS & URNs


1: Namespaces 2: AuthorID 3: BookID 4: VersionID 5: Lang 6: Passage

The OpenITI corpus is organized in compliance with Canonical Text Services (CTS) guidelines as implemented in the CapiTainS Suite (Cléric et al., 2017), with two exceptions made for practical purposes. [1] We use human-readable URNs as this allows for easier subsetting of the corpus and helps to engage non-DH specialists in collaboration. [2] We are postponing conversion to TEI XML, as it poses a number of challenges for RTL languages, connected scripts, and extensive texts.

CTS URNs offer a powerful mechanism for building expandable and interoperable corpora (figure above), which can be easily expanded in a decentralized manner accommodating as many texts and their versions as might be necessary.

ADAPTING: Instantiations

In order to increase the usability of the corpus, we generate instantiations which include all texts adapted for use with specific existing tools. For example, in https://github.com/OpenITI/i.stylo, texts are renamed and reformatted as required by the stylo package for R (Eder et al., 2016).

PUBLISHING: DOI 10.5281/zenodo.3082464