Int. J. Middle East Stud. **50** (2018), 103–109

ROUNDTABLE

Digital Humanities in Middle East Studies

Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans

MATTHEW THOMAS MILLER, MAXIM G. ROMANOV, AND SARAH BOWEN SAVANT
Roshan Institute for Persian Studies, University of Maryland, College Park, Md.; e-mail: mtmiller@umd.edu; Department of History, University of Vienna, Vienna, Austria; e-mail: maxim.romanov@univie.ac.at; Aga Khan University, Institute for the Study of Muslim Civilisations, London; e-mail: sarahsavant@aku.edu

doi:10.1017/S0020743817000964

The varied textual traditions of the premodern Islamicate World represent an opportunity and a problem for the Digital Humanities (DH). The opportunity lies in the sheer extent of this textual heritage: if we combine the textual output of premodern Persian and Arabic authors (not to mention Turkish and other less well-represented Islamicate languages), this body of texts constitutes arguably the largest written repository of human culture. Analytical methods developed for other linguistic heritages can be repurposed to make use of this wealth of texts, and efforts are now underway to apply to them a series of computationally enhanced methods that derive from a variety of disciplines (e.g., corpus linguistics, computational linguistics, the social sciences, and statistics). The application of these forms of analysis to these large new corpora promises new insights on premodern Islamicate cultures and the improvement of existing digital tools and methodologies.

The sheer extent also constitutes a problem—specifically, its wide dispersion across national and linguistic borders and the uneven state of digitization efforts. The existing repositories of digital Persian and Arabic texts suffer from three main critical problems: they are not compliant with any international data standards; they typically lack scholarly metadata; and they do not adequately represent the diversity of the historic traditions. This final issue of "representativeness" only partly arises in the digital era, however. Selection biases throughout history have limited which texts are passed down and, more recently, which texts are selected for printing and how they are employed for the creation of editions. Existing open access digital corpora reflect this broader problem, too, most notably with distinct biases towards Sunni or Twelver Shiʿi traditionalism and against scientific and philosophical texts, texts written by non-Muslims, and, in the case of Persian, prose works of all kinds. The result is that our ability to do innovative research with promising new digital methods has been limited by our source base.

Here, we would like to propose principles to guide corpus building, and then to introduce what the Open Islamicate Texts Initiative (OpenITI)—as a collaborative corpus-building effort—is doing to help build the digital infrastructure for the computational study of the premodern Islamicate world.

First we present the principles. These are partly conceived with reference to work in wide-ranging DH subfields and partly with reference to the Middle East. The Middle East Studies Association has recently published guidelines for evaluating digital scholarship for the purpose of hiring, tenure, and promotion, and some of what follows echoes points made there.[2]

RESEARCH RELEVANCE AND SCHOLARLY STANDARDS, NOT TECH FOR TECH'S SAKE

Priority must be given to scholarly corpus building. That is, our focus must be on constructing digital corpora that address the research needs of scholars and meet academic standards in terms of the quality of their texts and metadata. This work should be understood and financially supported by the field as a foundational form of intellectual labor that is akin in many ways to more traditional forms of humanistic data acquisition, such as archival research and other forms of fieldwork.

COLLABORATION, NOT GOING-IT-ALONE

Any specific project, digital or analogue, struggles with the state of our texts, and cannot address, much less remedy, the issues alone. Although humanities disciplines have sometimes resisted collaborative modes of intellectual production, the vast scale of Islamicate textual traditions and the widely diverse types of expertise required to build digital corpora make collaboration essential. It is critical that review committees consider the collaborative nature of much of this work and develop standards for evaluating it in the hiring and promotion processes. We also need to develop guidelines appropriate to Middle East studies for the proper recognition of individual contributions to such projects, from the undergraduate student to the senior professor (including, especially, technical team members), and rules/mechanisms to ensure the equitable treatment of junior members of large research groups. (Several initial attempts to sketch such "Collaborators' Bill of Rights" have already been attempted and they should be closely studied and enacted by all digital projects.)[3]

TRANSPARENT INTELLECTUAL SUPPLY CHAIN, NOT USE WITHOUT ATTRIBUTION

As digital tools and texts have rapidly expanded over the last several decades, we have often failed to properly highlight the ways and places in which these new developments have facilitated our research and corpus building. More specifically, how many times have we as scholars utilized digital text repositories such as Shamela or Ganjoor to find citations or conduct a quick review of relevant passages in the classical texts without highlighting the role these digital repositories played in enabling, or at least dramatically expediting, our research? The method of locating passages is not irrelevant; the producers of such digital text repositories have a right to be recognized for the increasingly important role they play in facilitating the construction of knowledge in our field, just as, for example, research assistants are acknowledged for their intellectual contributions to our work. The ethics of "use without attribution" are made even more problematic by the unequal power relationship involved—that is, between the researcher (often situated in a US or European university) and the Middle Eastern producers of these digital text repositories (e.g., Shamela and Ganjoor).

OPEN ACCESS AND OPEN SOURCE, NOT PROPRIETARY

Data, code, and publications must be made freely available for reuse. Building upon each other's work will save much-needed time and resources, and it will enable the general public, especially in the Middle East, to benefit and engage in citizen-science research projects without the traditional hurdles to access that individuals outside of university networks experience. These hurdles include, most notably, the financial and technical barriers associated with using (often prohibitively expensive) proprietary software and accessing subscription-based journal databases. Moreover, by making data, code, and research results freely available, we promote a transparent model of research that will foster healthy criticism of all steps in the research process and thereby improve the field as a whole. Pressures limiting open access are frequently commercial, but they are sometimes driven by lack of follow through and/or insecurities within the research community itself.

COLLECTIVELY CARING FOR OUR SHARED DIGITAL COMMONS, NOT FREE RIDING

Open access and open source projects are not free; resources—both labor and financial—are required to build, maintain, improve, and expand them. All users of open access and open source materials should remember this point: their free use is a type of free riding. Now, we certainly do not want to urge open access and open source projects to privatize and monetize their work! But we do want to encourage users of these resources to think of them as a form of digital commons that must be collectively cared for through the small contributions of each user. In the vast majority of cases such contributions would be in kind (e.g., help with postcorrection of an OCRed text or training data generation, adding scholarly metadata for some authors' works, fixing typos in the existing texts of the corpus), but there may be some cases in which larger institutions or grant-funded project teams actually contribute financially to help ensure the maintenance and sustainability of the open source or open access digital resources they are using. This ethic of collective care for our shared digital commons will help ensure that it is of the highest quality and remains under the control of and freely accessible to our intellectual community and society more broadly.

ADOPTION OF INTERNATIONAL DATA STANDARDS AND BEST PRACTICES, NOT SILOED PROJECTS

It is critical to regularly seek the advice of other projects on a wide range of issues, including those relating to research, ethics, and international data standards. Practically speaking, this means that Islamicate DH practitioners must work to cultivate

FIGURE 1. (Color online) Open Islamicate Texts Initiative (OpenITI).

relationships with their DH colleagues in other fields and participate in DH professional organizations and conferences. Major corpus-building projects require a substantial investment of time, energy, and money, so it is important to seek out all (ethical) ways of avoiding unnecessary costs and inefficient use of time and to ensure that a wide range of users can build on one's own work (e.g., by conforming to international data standards and frameworks, such as Text Encoding Initiative [TEI], XML, RDF, and IIIF).

The foregoing principles have been the guiding framework for our collaborative Islamicate DH project, the OpenITI (Figure 1). OpenITI was established in the summer of 2016 to create a machine-readable and scholarly metadata-enriched corpus of digitized premodern Persian and Arabic texts. It consolidates the individual efforts of the three authors, who previously had begun working on Persian and Arabic corpus projects through the KITAB Project (Sarah Bowen Savant and Maxim Romanov), OpenArabic (Maxim Romanov), and the Persian Digital Library (Matthew Thomas Miller). The authors realized that they share a wide range of common concerns that could best be addressed through a collaborative partnership, including increasing the size and representativeness of their collections and making them useful for new modes of computational textual analysis.

Our work can be broken down into the following four areas.

COLLECTING AND FORMATTING ALL TEXTS AVAILABLE ON THE WEB IN VARIOUS OPEN ACCESS REPOSITORIES OF TEXTS

For Arabic, the primary sources of these texts—over 4,200 unique titles—were al-Jamiʿ al-Kabir (HDD), al-Maktaba al-Shamila (Shamela),[4] and al-Maktaba al-Shiʿiyya (ShiaOnlineLibrary),[5] but we also collected texts from a couple of small but unique projects which bring the Graeco-Arabic texts into our corpus—namely, A Digital Corpus for Graeco-Arabic Studies[6] and Arabic Commentaries on the Hippocratic

Aphorisms.[7] The Persian texts were all gathered from Ganjoor.[8] As asserted earlier, it is important that we recognize the foundational efforts of these projects, even if we have to do further work to forge their texts into scholarly corpora. We have put these texts into structured data formats (OpenITI mARkdown for Arabic and Persian texts, and CapiTainS CTS-complaint TEI XML for the Persian texts). (On the Persian side, Elijah Cooke has completed most of this technical work.) We have also begun the long process of adding scholarly metadata to each work.[9] Our texts are available in the new OpenITI text repository on Github, and organized for easy access (specifically, with metadata-enriched, machine-readable, CTS-compliant Uniform Resource Identifiers/Uniform Resource Names [URIs/URNs]).[10]

DEVELOPING ARABIC-SCRIPT OPTICAL CHARACTER RECOGNITION (OCR)

OCR allows us to convert scans (i.e., images) of printed text into machine-readable and editable text, and thereby to address critical gaps in our existing collections. We have demonstrated that we can consistently achieve OCR accuracy rates in the high nineties—which is significantly higher than the primary commercial solutions for Arabic—with a new, open source, neural network–powered OCR solution called Kraken, which was developed by OpenITI team member Benjamin Kiessling (Leipzig University).[11] Our results appear in the current edition of the online peer-reviewed journal *al-ʿUsur al-Wusta*.[12] Our success with Arabic led us to undertake—with a larger team of collaborators—follow-up trials on Persian, Hebrew, and Syriac typefaces, and the early results are promising (full results will be released in winter 2017/spring 2018). We are also beginning trials on Persian and Arabic manuscripts. For the Arabic printed texts, our computer science team members—Benjamin Kiessling and Elijah Cooke—have been experimenting with the creation of generalized models (so that users can achieve better results for new typefaces without generating their own training data),[13] improving Kraken's ability to handle complex layouts (e.g., newspapers), and integrating the microtask/crowdsourcing platform Pybossa into the Nidaba OCR pipeline, which will enable more user-friendly production of training data.

CREATING A CORPUSBUILDER

Through a collaboration with Intisar Rabb and Sharon Tai's SHARIAsource project at Harvard University Law School, we are building a "corpus builder" that integrates Kraken (OCR software), Nidaba (OCR pipeline), and Pybossa (OCR microtask/crowdsourcing platform) with a powerful new version-controlled database, user management system, and flexible reading, editing, postcorrection, and annotation environment (Figure 2). The goal is to create a user-friendly platform that will allow projects anywhere to use OCR to build their own standards-compliant corpora through their organization's customized user interfaces. The corpus builder—tentatively named, Digital Islamicate CorpusBuilder (DICb)—will allow users to correct OCR output and to annotate the newly OCRed texts. It will also gather and make freely available training data from multiple projects (this training data is extremely valuable—Google, for example, shares its OCR code but not its training data).

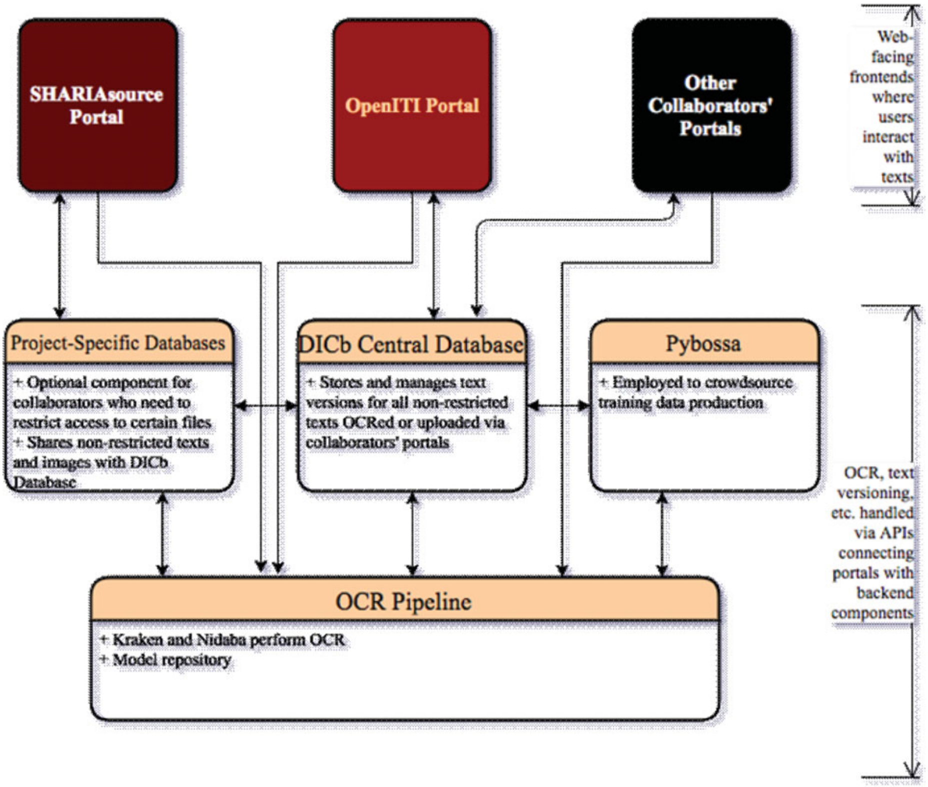
SHARIAsource Portal
OpenITI Portal
Other Collaborators' Portals
Web-facing frontends where users interact with texts
Project-Specific Databases
+ Optional component for collaborators who need to restrict access to certain files
+ Shares non-restricted texts and images with DICb Database
DICb Central Database
+ Stores and manages text versions for all non-restricted texts OCRed or uploaded via collaborators' portals
Pybossa
+ Employed to crowdsource training data production
OCR, text versioning, etc. handled via APIs connecting portals with backend components
OCR Pipeline
+ Kraken and Nidaba perform OCR
+ Model repository

FIGURE 2. (Color online) Digital Islamicate CorpusBuilder (DICb).

ASSEMBLING PRIORITY WORKS FOR DIGITIZATION

In our own research projects we are working to identify the most important books for OCRing in the future. One of our aspirational goals is to organize the field to enhance the OpenITI's collection by collectively choosing the highest priority texts for digitization.

We would like to conclude, however, with a cautionary note. Despite some important advances in the last several years, much work remains to be done, including the development of scholarly corpora, modification of tools and methods for Islamicate languages, and field-wide training to fully leverage and properly evaluate the potential of these new methods. The first studies based on Arabic and Persian are still being conducted, written, and revised. This means that it will be many years before we can responsibly pass intellectual judgement on this new enterprise we have termed *Islamicate DH*.

NOTES

[1] In alphabetical order.

[2] For the guidelines, see http://mesana.org/resources/digital-scholarship.html. See also Todd Presner, "How to Evaluate Digital Scholarship." *Journal of Digital Humanities* 1 (2012), accessed 18 September 2017, http://journalofdigitalhumanities.org/1-4/how-to-evaluate-digital-scholarship-by-todd-presner.

[3] See, for example, the "Collaborators' Bill of Rights" and the "Student Collaborators' Bill of Rights" for important efforts to lay out foundational principles for equitable collaboration: Tanya Clement and Doug Reside, "Off the Tracks: Laying New Lines for Digital Humanities Scholars," Media Commons Press, accessed 15 September 2017, http://mcpress.media-commons.org/offthetracks/part-one-models-for-collaboration-career-paths-acquiring-institutional-support-and-transformation-in-the-field/a-collaboration/collaborators%E2%80%99-bill-of-rights/; Haley Di Pressi, Stephanie Gorman, Miriam Posner, Raphael Sasayama, and Tori Schmitt, with contributions from Roderic Crooks, Megan Driscoll, Amy Earhart, Spencer Keralis, Tiffany Naiman, and Todd Presner, "A Student Collaborators' Bill of Rights," UCLA Center for Digital Humanities, accessed 15 September 2017, www.cdh.ucla.edu/news-events/a-student-collaborators-bill-of-rights/.

[4] See al-Maktaba al-Shamila, accessed 15 September 2017, http://shamela.ws/.

[5] See al-Maktaba al-Shiʿiyya, accessed 15 September 2017, http://shiaonlinelibrary.com.

[6] See A Digital Corpus for Graeco-Arabic Studies, accessed 15 September 2017, https://www.graeco-arabic-studies.org/.

[7] See Arabic Commentaries on the Hippocratic Aphorisms, accessed 15 September 2017, http://cordis.europa.eu/project/rcn/100847_en.html.

[8] See Ganjoor, accessed 15 September 2017, https://ganjoor.net/.

[9] For more on OpenITI mARkdown schema, see Maxim Romanov, "OpenITI mARkdown," al-Raqmiyyat, accessed 15 September 2017, https://alraqmiyyat.github.io/mARkdown/. For more on CTS and specifically CapiTainS, see CapiTainS, accessed 15 September 2017, http://capitains.org/. For more on TEI, see Text Encoding Initiative, accessed 15 September 2017, http://www.tei-c.org/index.xml.

[10] The OpenITI repository is available at https://github.com/OpenITI/, accessed 15 September 2017. For more on OpenITI CTS URNs, see Maxim Romanov, "OpenITI," al-Raqmiyyat, accessed 15 September 2017, https://alraqmiyyat.github.io/OpenITI/.

[11] Traditional OCR approaches work by segmenting page images into lines, then each line into words, and then each word into characters. Since segmentation is extremely problematic when it comes to connected, ligature-rich scripts, performance is consistently poor on the last two steps. In contrast to this approach, Kraken completely eliminates the issue of word/character segmentation by instead employing a form of machine learning called a neural network. Neural networks mimic the way we learn, enabling Kraken to "learn" from transcriptions (training data) to recognize letters in the images of entire lines of text. This new approach to OCR makes Kraken uniquely able to handle the wide variety of ligatures in connected scripts such as Arabic and Persian.

[12] Benjamin Kiessling, Matthew Thomas Miller, Maxim Romanov, and Sarah Bowen Savant, "Important New Developments in Arabographic Optical Character Recognition (OCR)," *al-ʿUsur al-Wusta*, accessed 20 November 2017, http://islamichistorycommons.org/mem/wp-content/uploads/sites/55/2017/11/UW-25-Savant-et-al.pdf.

[13] Generalized models incorporate script features from multiple typefaces and thus are less typeface specific and better able to handle typefaces for which we have not trained a specific model.