OCR and Digital Text Production: Learning from the Past, Fostering Collaboration and Coordination for the Future

January 29th-30th, 2020
University of Maryland, College Park

Conveners:

SHARIAsource
at HARVARD LAW SCHOOL

With generous funding provided by:

THE
ANDREW W.
MELLON
FOUNDATION

COLLEGE OF
ARTS & HUMANITIES

Welcome Workshop Participants!

The entire Open Islamicate Texts Initiative Arabic OCR Catalyst Project (OpenITI AOCP) team would like to thank you for agreeing to participate in this workshop and share your expertise and experience with all in attendance.

Before we delve into more specifics on the workshop itself, let's go over the logistical details. Samar Ata—the incredible coordinator of the Roshan Institute for Persian Studies at the University of Maryland—has already sent you your flight and accommodation information in a separate email (if for some reason you did not receive this email please let us know as soon as possible). Some of this information is repeated below. But be sure to review everything below in detail because there are additional details provided here that were not included in her emails.

Transportation
Please use Lyft or another car service to travel between the airport and the hotel. You can submit these receipts to Samar Ata for reimbursement. If using a taxi, an original receipt is required. If for any reason obtaining a Lyft or other car service is difficult for you, please let us know as soon as possible and we will schedule transportation for you. DC traffic can be quite bad, depending on the time your flight arrives. If you are departing Dulles International Airport (IAD) for The Hotel, especially near rush hour time (6:30am-9:30am and 4:00-6:30pm), please prepare for a lengthy car ride (sometimes as long as two hours) by using the restroom and grabbing something to eat in the airport before departing.

Accommodations
You will all be staying at The Hotel at University of Maryland (UMD). The Hotel is Located across from the university's entrance at 7777 Baltimore Ave, College Park, MD 2074.

Food
All meals during the workshop will be provided. We will also be hosting a (optional) pre-workshop dinner on Tuesday, January 28th at 7:30 at the Old Maryland Grill (located off of the lobby of The Hotel) for those who arrive early enough (and aren't too jetlagged to attend). For breakfast each morning you will receive $10 voucher that can be used at Bagels N Grinds, which is located off of the lobby of The Hotel. We will have some pastries, fruit, and coffee in the meeting room for snacking during the morning hours of the workshop, but if you want a sizable breakfast please eat breakfast before we depart The Hotel for the workshop each morning. Lastly, if you will be missing the

group dinner on the second day of the workshop due to an early flight time, please feel free to request a boxed meal to go, or you may purchase dinner at the airport and we will reimburse you for it.

Workshop Location

The workshop will take place in the Prince George's room at the Adele H. Stamp Student Union (3972 Campus Dr., College Park, MD 20742). Matthew Miller will meet you in The Hotel lobby each morning at 8:40 to walk with you from The Hotel to the workshop venue, but in case you need to navigate the campus on your own, here is a link to a campus map.

Reimbursement

Please see Samar Ata for reimbursement for any travel or food expenses.

Workshop Background

This workshop is a part of the *Open Islamicate Texts Initiative Arabic OCR Catalyst Project (OpenITI AOCP)*—a two-year project generously funded by The Andrew W. Mellon Foundation (full details on this project can be found here). OpenITI AOCP is led by Matthew Thomas Miller (Roshan Institute for Persian Studies at UMD), Maxim Romanov (University of Vienna), Sarah Bowen Savant (Aga Khan University), David Smith (Northeastern University), and Raffaele Viglianti (Maryland Institute for Technology in the Humanities at UMD). SHARIAsource, a project of the Program in Islamic Law (PIL) at Harvard Law School (both led by Intisar Rabb), provided significant support for the initial technical infrastructure upon which this project will build (i.e., CorpusBuilder 1.0) and they will also play a leading role in the technical development portion of OpenITI AOCP. We are proud to be co-convening this workshop with them. We are also delighted to announce that Jonathan Parkes Allen and Asad Zaman have joined the OpenITI AOCP team at UMD in the roles of Mellon Islamicate Digital Humanities Postdoctoral Fellow and Mellon Islamicate Digital Humanities Graduate Research Assistant respectively and Alejandro Toselli and Rui Dong have joined the CS team of OpenITI AOCP at Northeastern University in the roles of Mellon Computer Science Postdoctoral Fellow and Mellon Computer Science Graduate Research Assistant respectively.

The primary goal of OpenITI AOCP is to advance the digitization of the Persian and Arabic written traditions by addressing the central technical and organizational impediments stymying the development of improved OCR for Arabic-script languages. In this workshop we will discuss our technical development plan for how to achieve this goal (again, more details can be found here). However, building on the

recommendations in Smith's *A Research Agenda for Historical and Multilingual Optical Character Recognition*, we believe it is crucially important that our development work in the OpenITI AOCP is informed by the latest research, pursued with a variety of different users and projects' needs and past experiences in mind, and done in strategic collaboration with other groups working on related problems. Therefore, this workshop aims to bring together leading figures in the world of OCR research and OCR project development to both survey the state of the field and forge connections that will allow us all to more strategically collaborate on advancing the field of OCR for an increasingly large number of languages, use cases, and document types (e.g., manuscripts).

Workshop Participants

Hany A. Elsawy Abdellatif, Head Of Digitization Services, IT Operations and Infrastructure, Qatar National Library

Dr. Mansoor Alghamdi, University of Tabuk, Saudi Arabia

Dr. Jonathan Parkes Allen, Mellon Humanities Postdoctoral Fellow, Roshan Institute for Persian Studies, University of Maryland, College Park
Project Affiliation: OpenITI AOCP

Laurie Allen, Senior Innovation Specialist, Library of Congress Digital Innovation Lab

Dr. Hannah Alpert-Abrams, Program Specialist, Office of Digital Humanities, National Endowment for the Humanities

Dr. Taylor Berg-Kirkpatrick, Assistant Professor of Computer Science and Engineering, University of California, San Diego

Dr. Thomas Breuel, Distinguished Research Scientist, NVIDIA

Dr. Dale J. Correa, Middle Eastern Studies Librarian, The University of Texas at Austin
Project Affiliation: OpenITI AOCP

Dr. Gregory Crane, Professor of Classics, Tufts University; Alexander von Humboldt Professor of Digital Humanities, University of Leipzig, University of Leipzig
Project Affiliation: Perseus Digital Library

Rui Dong, Mellon Computer Science Graduate Research Assistant, Northeastern University
Project Affiliation: OpenITI AOCP, OCR-D

Kartik Goyal, PhD student, Carnegie Mellon University

Wayne Graham, Chief Technology Officer, Council on Library and Information Resources
Project Affiliation: Digital Library Federation, Digital Library of the Middle East

Dr. Nizar Habash, Associate Professor of Computer Science and Program Head, New York University, Abu Dhabi

Dr. Bushra Jaswal, Chief Librarian and Associate Professor, Forman Christian College University, Lahore, Pakistan

Dr. Jesse P. Karlsberg, Senior Digital Scholarship Strategist, Emory University
Project Affiliations: Readux, Sounding Spirit

Dr. Adi Keinan-Schoonbaert, Digital Curator, Asian and African Collections, British Library
Project Affiliations: Hebrew Manuscripts Digitisation Project, Two Centuries of Indian Print, British Library's Arabic OCR Competition

Benjamin Kiessling, Machine Vision Engineer, Université Paris Sciences et Lettres
Project Affiliation: eScriptorium/Scripta numérique project, Kraken

John Kiplinger, Director of Production, JSTOR-Ithaka

Dr. Matthew Thomas Miller, Assistant Professor of Persian Literature and Digital Humanities, Roshan Institute for Persian Studies, University of Maryland, College Park
Project Affiliation: OpenITI AOCP

David Millman, Assistant Dean for Digital Library Technology Services, New York University
Project Affiliations: Arabic Collections Online

Dr. Hussein Adnan Mohammed, Principal Investigator, Hamburg University

Project Affiliation: "Pattern Recognition in 2D Data from Digitised Images and Advanced Acquisition Techniques" within the Cluster of Excellence: Understanding Written Artefacts in Hamburg University

Dr. Günter Mühlberger, Senior Project Manager of the Digitisation and Digital Preservation, University of Innsbruck
Project Affiliation: Transkribus

Dr. Lorenz Nigst, Research Associate, Aga Khan University (London)
Project Affiliation: KITAB

Mark Patton, Senior Software Engineer, Johns Hopkins University
Project Affiliation: *Archaeology of Reading*

Dr. Intisar Rabb, Professor of Law and History, Director, Program in Islamic Law, Harvard University
Project Affiliation: SHARIAsource

Anne Ray, Senior Licensing Editor, JSTOR-Ithaka

Dr. Bruce Robertson, Professor of Classics, Mount Allison University
Project Affiliation: LACE: Visualizing, Editing and Searching Polylingual OCR Results; Open Greek and Open Latin

Dr. Maxim Romanov, Universitätassistent für Digital Humanities, Institut für Geschichte Universität Wien; Senior Research Fellow, KITAB, Aga Khan University
Project Affiliation: KITAB, OpenITI AOCP

Ahmed Samir, Director of the Institutional Repositories & Integrated Library Systems, Bibliotheca Alexandrina, Egypt

Dr. Sarah Bowen Savant, Professor of Islamic History, Aga Khan University
Project Affiliation: KITAB, OpenITI AOCP

Dr. David Smith, Associate Professor of Computer Science, Khoury College of Computer Sciences, Northeastern University
Project Affiliation: OpenITI AOCP

Dr. Nur Sobers-Khan, Lead Curator South Asia Collections, British Library

Project Affiliation: Two Centuries of Indian Print

Dr. Daniel Stoekl, Research Professor, École Pratique des Hautes Études
Project Affiliation: eScriptorium/Scripta numérique project

Sharon Tai, Deputy Editor at the Program in Islamic Law, Harvard Law School
Project Affiliation: SHARIAsource

Dr. Alejandro H. Toselli, Mellon Computer Science Postdoctoral Fellow, Northeastern University
Project Affiliation: OpenITI AOCP

Dr. Raffaele Viglianti, Research Programmer, Maryland Institute for Technology in the Humanities, University of Maryland, College Park
Project Affiliation: OpenITI AOCP

Kay-Michael Würzner, Research Software Engineer, Saxon State Library and University Library Dresden
Project Affiliation: OCR-D

Asad Zaman, Mellon Humanities Graduate Research Assistant, Roshan Institute for Persian Studies, University of Maryland, College Park
Project Affiliation: OpenITI AOCP

Schedule

Pre-workshop Event (Optional)
Tuesday, January 28th, 2020

7:30-9:30 pm: Welcome dinner (Old Maryland Grill located off of the lobby of The Hotel) for those who arrive early enough (and aren't too jetlagged to attend)

Workshop Day #1: OCR Research and Project Lightning Talks
Wednesday, January 29th, 2020

7:30-8:30am: Breakfast at *Bagels N Grinds* located at The Hotel (not a group event—please just grab breakfast at your leisure)

8:40am: Walk with Matthew Miller to Adele H. Stamp Student Union for workshop day #1

9:00-9:15 am: Opening remarks by Dr. Bonnie Thornton Dill, Dean of University of Maryland's College of Arts and Science, and Dr. Fatemeh Keshavarz, Director of the School of Languages, Literatures, and Cultures and Director of the Roshan Institute for Persian Studies

9:15-9:30 am: Workshop participant introductions

9:30-9:40 am: Overview of workshop goals (Matthew Miller)

9:40-10:00 am: *OCR-D* (Kay-Michael Würzner)

10:00-10:30 am: *Two Centuries of Indian Print, British Library's Arabic OCR Competition* (Dr. Adi Keinan-Schoonbaert and Dr. Nur Sobers-Khan)

10:30-10:50 am: *Deep Learning for OCR, Document Analysis, Text Recognition, and Language Modeling* (Dr. Thomas Breuel)

10:50-11:05 am: Morning break

11:05-11:25 am: *eScriptorium* (Dr. Daniel Stoekl)

11:25-11:45 am: *Kraken* (Benjamin Kiessling)

11:45-12:05 am: *JSTOR's Arabic digitization study* (John Kiplinger and Anne Ray)

12:05-12:25 pm: *CorpusBuilder and OpenITI AOCP* (Matthew Miller, Sharon Tai, Kamil Ciemniewski, David Smith, Raffaele Viglianti)

12:25-12:45 pm: *Visual Analysis of Early Modern Printed Documents* (Dr. Taylor Berg-Kirkpatrick)

12:45-2:00 pm: Lunch (catered in Prince George's room)

2:00-2:20 pm: *The Past and Future of OCR in the Perseus Digital Library* (Dr. Gregory Crane)

2:20-2:40 pm: *Arabic Collections Online* (David Millman)

2:40-3:00 pm: *Qatar National Library's OCR Workflow* (Hany A. Elsawy Abdellatif)

3:00-3:20 pm: *Pattern Recognition in 2D Data from Digitised Images and Advanced Acquisition Techniques* (Dr. Hussein Adnan Mohammed)

3:20-3:30 pm: Break

3:30-3:50 pm: *Transkribus* (Dr. Günter Mühlberger)

3:50-4:10 pm: *Bibliotheca Alexandrina's OCR Workflow* (Ahmed Samir)

4:10-4:30 pm: *Archaeology of Reading* (Mark Patton)

4:30-4:50 pm: *LACE: Visualizing, Editing and Searching Polylingual OCR Results* (Bruce Robertson)

4:50-5:10 pm: *OCR in Sounding Spirit Project* (Dr. Jesse P. Karlsberg)

5:10-6:30 pm: Break

6:30 pm: Dinner (Burtons Grill—we will take hotel shuttle from The Hotel)

Workshop Day #2: OCR Product Demos and Strategic Planning January 30th, 2020

7:30-8:30 am: Breakfast at *Bagels N Grinds* located at The Hotel (not a group event—please just grab breakfast at your leisure)

8:40 am: Walk with Matthew Miller to Adele H. Stamp Student Union for workshop day #2

9:00-9:30 am: *OCR and Emerging Research in Arabic NLP and Computational Linguistics—A Response to Day #1's Presentations* (Dr. Nizar Habash)

9:30-11:00 am: OCR Product Demos (OCR-D, Transkribus, LACE, eScriptorium, CorpusBuilder)

11:00-11:15 am: Break

11:15 am-12:15 pm: Discussion #1: Identifying the issues—what are the persistent unmet needs and barriers in OCR projects?

12:15-1:15 pm: Discussion #2: Defining the agenda—what are the most important areas for research and development in the coming years?

1:15-2:45 pm: Lunch (catered in Prince George's Room)

2:45-3:00pm: *Urdu OCR* (Dr. Bushra Jaswal)

3:00-5:00 pm: Strategic planning discussion: building partnerships and fostering collaboration—how can we work together to address these OCR challenges?

5:00-6:00 pm: Break

6:00pm: Dinner (Old Maryland Grill located off of the lobby of The Hotel)