Data Science Transdisciplinary Area of Excellence

Data Salon 2025 — 2026

The Pitfalls and Possibilities of Handwritten Text Recognition Development: Ottoman Turkish as Case Study

Kent Schull, Department of History

Friday, Jan. 30, 2026, 12:15-1:15 PM
AD-148, with lunch served
Zoom Meeting ID: 957 6354 3977 Passcode: 379441

Abstract

Handwritten Text Recognition (HTR) is a rapidly expanding field that intersects digital humanities, data science, and large language models to harness artificial intelligence and machine learning to read and translate handwritten texts. While Optical Character Recognition (OCR) is highly developed at this point enabling AI to translate printed texts in numerous languages, HTR is still in its infancy and vastly more difficult, because unlike printed text with a limited number of standardized fonts and textual arrangement, handwriting is as unique to each person as a fingerprint. Additional major obstacles include creating the training data and dealing with hallucinations produced by Large Language Models, especially when working with dead languages in non-Latin scripts. Currently, the Center for Middle East and North Africa Studies (CMENAS)'s Ottoman Demographic, Social, and Family History Research Group at Binghamton University in partnership with FamilySearch International is making dramatic strides in developing HTR for Ottoman Turkish. This presentation engages the pitfalls and potentialities of HTR development, particularly as it relates to Ottoman Turkish, which was the high bureaucratic language of the Ottoman Empire for almost six-hundred years written in a modified Persian-Arabic script. We have developed techniques that are overcoming some of the challenges described above, particularly as they pertain to creating data training sets and LLM hallucinations. Additionally, our methodologies show great promise for other Middle East languages, such as Arabic, Ladino, and Persian to name only a few. Through interdisciplinary cooperation that combines humanities, linguistic, and computer science expertise here on campus, we can make major strides in HTR development.

About the speakers: Kent F. Schull is a two-time Fulbright scholar to Turkey. His research and teaching interests include the social and cultural history of the Ottoman Empire and modern Middle East, criminal justice, Middle East Diaspora Studies, Israeli and Palestinian History, missiology and forced migration in the MENA region.

About the Data Salon: Data Salon is a dynamic venue designed to foster the exchange of ideas and the formation of new collaborations. Each gathering includes a brief talk to inspire discussion, but the emphasis lies on the social dimension — creating an open and welcoming space where scholars, researchers, and practitioners can engage in dialogue, discover shared interests, and explore opportunities for collaboration. More than a lecture series, Data Salon is a catalyst for community-building and cross-disciplinary connection.