Adam in the Archive: Discovering Named Entities in the Cold War Papers

ALEXIS ANTRACOLI, WILL CLEMENTS, and CLIFF WULFMAN

The Cold War Papers named entity extraction project explores the use of named-entity recognition to enhance discoverability in digitized archival collections. The project team began with several hypotheses: (1) using named-entity recognition on uncorrected OCR would produce useful information that is not already well-represented in the finding aids; (2) that we could develop MPLP-friendly automated and/or semi-automated processes to do the work; and (3) that connecting named entities with finding-aid components would reveal previously hidden connections between people, places, and organizations.

The project focuses on over 350,000 digitized page images created as part of an NHPRC-funded grant project from 2013 to 2015. Several months of research focused on the George Kennan papers have produced a set of useful results via automated and semi-automated processes, including linked-data knowledge graphs connecting text spans, names, and entities. The research has also revealed several challenges, including the difficulty of working with dirty OCR and the limits of existing tools for working with archival materials. This presentation will provide a summary of research efforts to date, including methodologies and results, and will present a roadmap for future work on the project and potential applications beyond the current project.

About the authors:

Alexis Antracoli is Assistant University Librarian for Special Collections Technical Services. She leads the Archival Description and Processing team and supervises the Curatorial Services team. She also acquires manuscripts in American history to 1900. Previously she worked at Drexel University Libraries and the University of Michigan Bentley Historical Library. She has published on web archiving, inclusive description, and the archiving of born-digital audio visual content. She also teaches the Archives and Manuscripts course at the Rutgers University School of Communication and Information. She is especially interested in applying user experience research and user-center design to archival discovery systems, developing and applying inclusive description practices, web archiving, and computational approaches to archives. She holds an M.S.I. in Archives and Records Management from the University of Michigan, a Ph.D. in American History from Brandeis University, and a B.A. in History from Boston College.

Will Clements is currently Public Policy Papers Archivist at the Seeley Mudd Manuscript Library at Princeton University. In this role he processes personal papers and organizational records in the Public Policy Papers, which comprises a number of collections documenting United States public policy, diplomatic history, and international development. Will's current work at Mudd also includes co-managing Princeton Special Collections' ArchivesSpace instance. Previously, Will was Digital Projects Archivist at Washington State University's Center for Digital Scholarship and Curation, where he managed descriptive metadata of digital cultural heritage materials on behalf of a number of Native American tribes and nations in the Northwest.

Cliff Wulfman holds a Ph.D. in English from Yale University and a M.S. in Computer Science from the University of Pennsylvania. His current areas of interest include knowledge-based discovery systems and the semantic web. At Princeton, he serves as Coordinator of Periodicals Digitization.