Natural Language Processing for Archival Description of Electronic Records: Potential and Priorities

ALEXANDRA CHASSANOFF, CHRISTOPHER (CAL) LEE, SUNITHA MISRA, and KAM WOODS

Abstract: Archival description and access are significantly driven by the ability to identify specific contextual entities (e.g. people, places, organizations, events). The identification and presentation of access points based on those entities has traditionally involved significant human effort. Applying natural language processing (NLP) methods to automatically identify and extract such information has the potential not only to free up the time of archivists so they can focus more on processing tasks that require their judgment, but also to expose access points that may never have been generated by an archivist.

The BitCurator Access project, funded by the Andrew W. Mellon Foundation, has developed BCA Webtools, which allow users to dynamically navigate the filesystems of disk images, as well as searching through the content of many common files types contained within the images. In this poster, we will present our vision for laying NLP and information visualization on top of the existing BitCurator environment and BitCurator Access Webtools. This would be beneficial both for archivists (to carry out appraisal, description, reference services, discovery and review for sensitivity) and end users. We will pose a series of questions to the audience about how these approaches could be applied in their local institutions and which types of access points should receive the highest priority.

About the Authors:

Alexandra Chassanoff is a newly minted CLIR Postdoctoral Fellow working in Software Curation at MIT Libraries beginning in September. She is currently the Project Manager at BitCurator Access. She will receive her doctorate this fall from the School of Information and Library Science at UNC Chapel Hill, and received her Master's in Information Science from UNC in 2009.

Christopher (Cal) Lee is Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches archival administration; records management; digital curation; understanding information technology for managing digital collections; and digital forensics. He is a lead organizer and instructor for the DigCCurr Professional Institute, and he teaches professional workshops on digital forensics methods and principles. Cal’s primary area of research is digital curation. Cal developed “A Framework for Contextual Information in Digital Collections,” and edited and provided several chapters to I, Digital: Personal Collections in the Digital Era. Cal is Principal Investigator (PI) of BitCurator Access and was PI of BitCurator. He was also PI of the Digital Acquisition Learning Laboratory (DALL) project, is Senior Personnel on the DataNet Federation Consortium, and has served as Co-PI on several digital curation education projects.