Open-Source Software for Appraisal and Processing of Email at Scale

CHRISTOPHER (CAL) LEE

Abstract: The volume, diversity and internal complexity of email make it a challenge to appraise and process. This presentation will report of open-source software to curation of email, with a particular emphasis on selection/appraisal, review for sensitivities and response to open records requests. The Review, Appraisal and Triage of Mail (RATOM) project (2019-2020), funded by the Andrew W. Mellon Foundation, is a collaboration between the University of North Carolina and the State Archives of North Carolina. This presentation will highlight three main software offerings developed over the past year of the project: First, libratom efficiently extracts elements from mbox, OST and PST files and writes the results to a simple database structure for future processing. Second, a set of Jupyter notebooks allow users to learn about libratom's functionality without requiring any local software installation. Third, a browser-based graphic user interface allows archivists to review, search and tag email for topics, record status and sensitivities. We'll conclude with application of machine learning to email and relationships to core archival functions.

About the author:

Christopher (Cal) Lee is Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches courses and workshops in archives and records management; understanding information technology for managing digital collections; and digital forensics. His primary research focus is the long-term curation of digital collections. He is particularly interested in the professionalization of this work and the diffusion of existing tools and methods into professional practice. 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. He has served as Principal Investigator of the Digital Acquisition Learning Laboratory (DALL), BitCurator, BitCurator Access, BitCurator NLP, BitCuratorEdu, and Review, Appraisal and Triage of Mail (RATOM) projects. He has been Co-PI on OSSArcFlow, as well as several projects focused on digital curation education: DigCCurr, DigCCurr 2, Closing the Digital Curation Gap (CDCG), Educating Stewards of Public Information in the 21st Century (ESOPI-21), and Educating Stewards of Public Information Infrastructure (ESOPI2). Cal was also Senior Personnel on the DataNet Federation Consortium funded by the National Science Foundation. He is a Fellow of the Society of American Archivists, and he serves as editor of American Archivist.