

REVIEW, APPRAISAL, AND TRIAGE OF MAIL:

A Tutorial

Christopher A. Lee

University of North Carolina

USA

callee@ils.unc.edu

Kam Woods

University of North Carolina

USA

kamwoods@email.unc.edu

Abstract - This full-day tutorial will focus on open-source software (OSS) tools and methods for libraries, archives and museums (LAMs) to identify email in born-digital collections, review email sources for sensitive or restricted materials, and perform appraisal and triage tasks to identify and annotate records. We will specifically explore products of the Review, Appraisal and Triage of Mail (RATOM) project's use of machine learning to separate records from non-records, along with natural language processing methods to identify entities of interest within those records. In addition to gaining hands-on experience using the tools, participants will also learn about the rationale for their development, how they relate to other available software, and how processing of email can fit into larger digital curation workflows. We will conclude with a discussion of implications for participants in their own institutions.

Keywords - email curation, appraisal, natural language processing, machine learning, timelines

Conference Topics - Exploring New Horizons; The Cutting Edge: Technical Infrastructure and Implementation

I. BACKGROUND

Despite progress on various technologies to support both digital preservation and archival description, we have still seen relatively little progress on software support for the core library, archives and museum (LAM) activities of selection and appraisal.

The BitCurator environment allows LAMs to engage in acquisition of data from media, as well as carrying out characterization and triage of the data, exposing numerous data points that can inform selection and appraisal decisions, including file types, file sizes, timestamps, original directory

structures, potentially sensitive features. However, these various forms of output are generally static. Users have expressed interest in additional ways to iteratively make judgements. Software developed by the BitCurator Access and BitCurator NLP projects - including topic modelling and named entity extraction - can further facilitate such decision-making. By incorporating timeline representations and visualizations, LAM professionals also can get useful, high-level views of the materials that they are investigating.

Email is often part of collections and acquisitions that include other types of materials, and LAMs frequently want to take actions that reflect those contextual relationships. The Transforming Online Mail with Embedded Semantics (TOMES) project has developed software to identify email accounts of public officials with enduring value in order to capture, preserve and provide access to important government records.

The Review, Appraisal and Triage of Mail (RATOM) project (2019-2020), funded by the Andrew W. Mellon Foundation, is a partnership between the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the Division of Archives and Records (DAR) of the North Carolina Department of Natural and Cultural Resources (NC DNCR). RATOM is developing capabilities for interactive selection and appraisal in the BitCurator environment in ways that can be directly connected to the output of TOMES to support iterative processing, in which information discovered at various points in the processing workflow can support further selection, redaction or description actions. Among other activities, this involves mapping of timestamp, entity, sensitive features and other elements across the tools.

II. TUTORIAL OVERVIEW

This full-day tutorial will focus on open-source software (OSS) tools and methods for libraries, archives and museums (LAMs) to identify email in born-digital collections, review email sources for sensitive or restricted materials, and perform appraisal and triage tasks to identify and annotate records. We will specifically explore products of the Review, Appraisal and Triage of Mail (RATOM) project's use of machine learning to separate records from non-records, along with natural language processing methods to identify entities of interest within those records. In addition to gaining hands-on experience using the tools, participants will also learn about the rationale for their development, how they relate to other available software, and how processing of email can fit into larger digital curation workflows. We will conclude with a discussion of implications for participants in their own institutions.

III. INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for curation of digital collections that are composed – in whole or in part – of email. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how to incorporate various open-source tools and services into institutional workflows (including through combinations with other proprietary software).

IV. PREPARATION AND PREREQUISITE KNOWLEDGE

Participants are not required to have prerequisite knowledge in any specific machine learning or NLP tools or methods. However, some previous knowledge of email structure and some command line experience will both be helpful. The tutorial will make use of a combination of hosted and local software. Participants will be expected to download and install software in advance.