Processing, Appraisal, and Iterative Selection of Email in Context

International Digital Curation Conference
February 4-7, 2019
Melbourne, Australia

Christopher (Cal) Lee
School of Information and Library Science
University of North Carolina at Chapel Hill

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

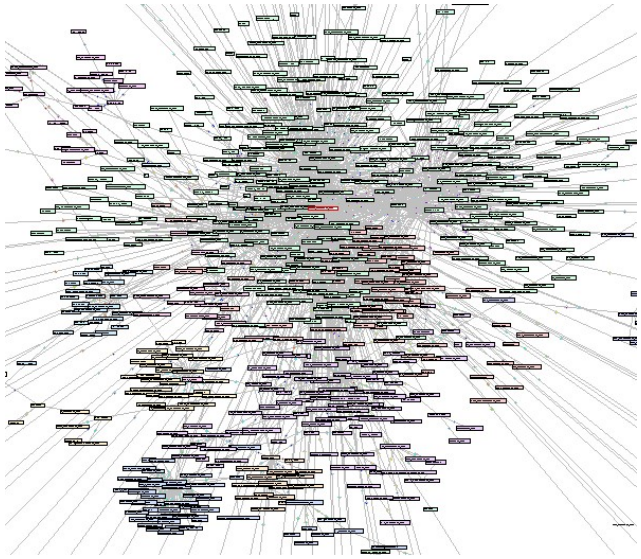
NC DEPARTMENT OF
NATURAL AND CULTURAL RESOURCES

Motivation – Selection/Appraisal

- Despite progress on various technologies to support data management and digital preservation, relatively little progress on software support for the core activities of selection and appraisal
- Selection/appraisal decisions are based on various patterns
- When patterns can be identified algorithmically, software can assist the process
- GLAMs frequently want to take actions that reflect contextual relationships
- Timeline representations and visualizations can also provide useful, high-level views of materials

Motivation - Email

- 48 years of email creation
- Hundreds of billions of messages generated every day
- Most has little long-term retention value, but some absolutely does
- Despite presence of numerous other modalities, email still deeply embedded in activities, serving as massive source of evidence and information
- Often found in collections and acquisitions with other types of materials

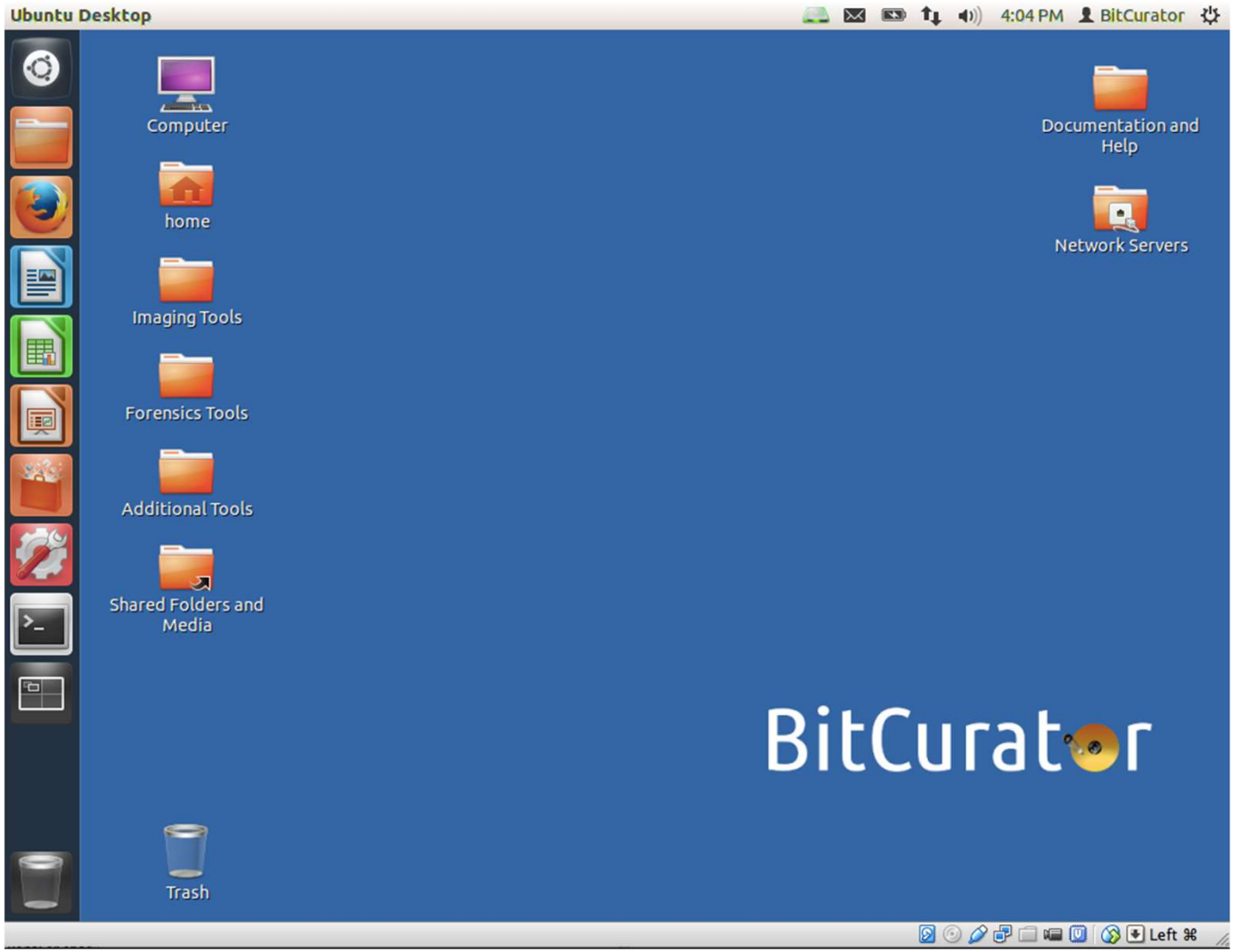


http://hci.stanford.edu/~jheer/projects/enron/v1/

Background – BitCurator (2011-2014)

- BitCurator environment allows GLAMs to:
  - acquire data from media
  - characterize and triage data
  - expose numerous data points that can inform selection and appraisal decisions, including file types, file sizes, timestamps, original directory structures, potentially sensitive features
- Output is generally static
- Users have expressed interest in additional ways to iteratively make judgements

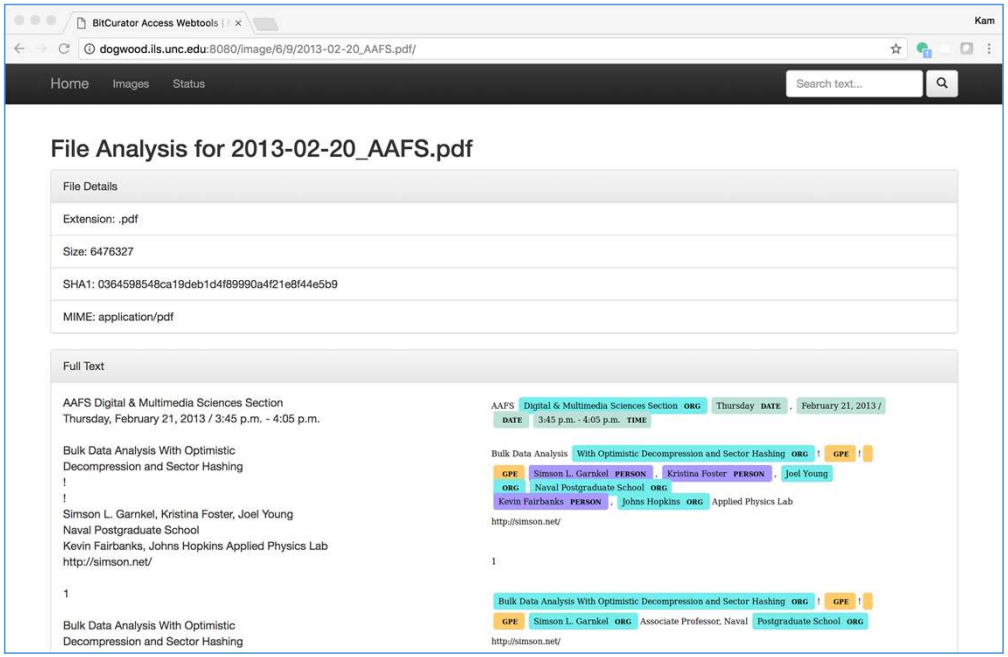
BitCurator

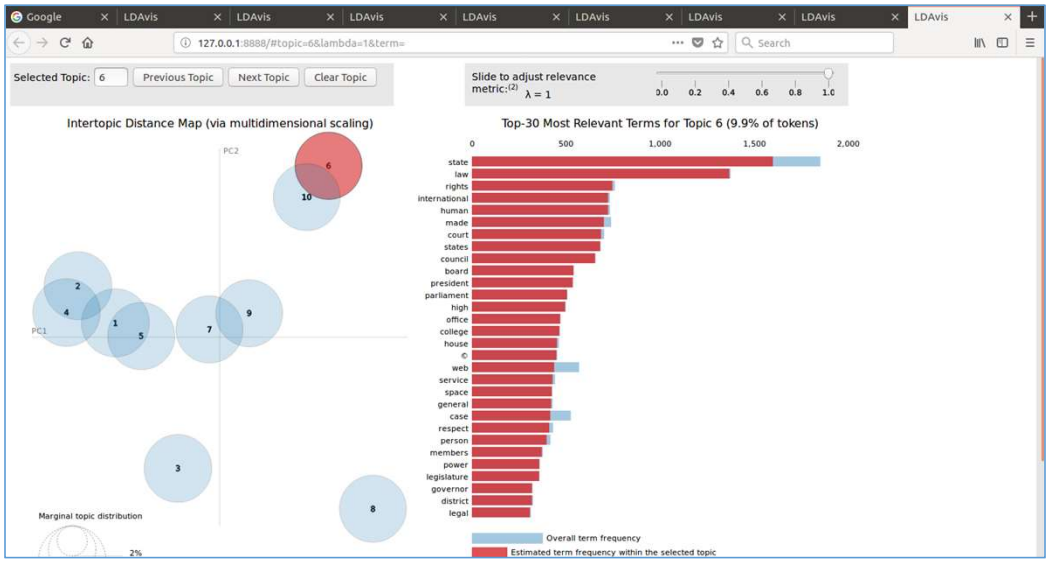


http://bitcurator.github.io/

Background – BitCurator Access and BitCurator NLP (2014-2018)

- Developed and repurposed software (topic modelling and named entity extraction) that can facilitate appraisal/selection





Background – TOMES (2015-2018)

- Transforming Online Mail with Embedded Semantics (TOMES) project developed software to identify email accounts of public officials with enduring value to capture, preserve and provide access to important government records
- Outcomes:
  - cross platform .pst to EAXS XML parser
  - processing a set of test email accounts based on Capstone rules
  - publishing NLP dictionary to flag named entities unique to state and local government

Review, Appraisal and Triage of Mail (RATOM)

- Funded by Andrew W. Mellon Foundation (2019-2020)
- Developing and repurposing software (including NLP and machine learning) for selection/appraisal in BitCurator environment with hooks and enhancements to TOMES output
- Support iterative processing - information discovered at various points in the processing workflow can support further selection, redaction or description actions
- Mapping of timestamp, entity, sensitive features and other elements across the tools

Ray Tomlinson

https://upload.wikimedia.org/wikipedia/commons/0/01/Ray_Tomlinson_%28cropped%29.jpg

RATOM Project Team at UNC

- Christopher (Cal) Lee, Principal Investigator
- Kam Woods, Co-PI and Technical Lead
- Antoine de Torcy, Software Developer
- Anusha Suresh, Project Manager

RATOM Project Team at State Archives of NC

- Camille Tyndall Watson, Co-Principal Investigator
- Jamie Patrick-Burns, Investigator
- Nitin Arora, Software Developer

Thank you

http://ratom.web.unc.edu/