TOMES Software Overview

The TOMES software tool is the result of the Transforming Online Mail with Embedded Semantics (TOMES) project, which was funded by a 3-year grant from the National Historical Publications and Records Commission (NHPRC). Running from 2015 to 2018, the project was a partnership among the State Archives of North Carolina, Kansas State Historical Society, and the Utah State Archives. It aimed to identify email accounts of public officials whose emails contain records of enduring value and to capture, preserve, and provide access to those government records. Specifically, the grant sought a method to transfer email accounts out of hosted email platforms such as Office 365 and Gmail to convert and process them using natural language processing. This would streamline review and redaction of email prior to public access.

When email accounts are transferred to the State Archives, they are delivered as a large, unstructured body of data containing messages that include both public records and non-record material, as well as potentially sensitive or confidential information. Due to unstructured nature of the information, archival emails must be reviewed by an archivist, message-by-message, before being released to the public; without the assistance of software, the size of most email accounts makes message-by-message review time-consuming and unsustainable.

The TOMES software uses natural language processing (NLP) technology to “tag” specified information in an email account. NLP is an area of computer science that uses computers to analyze large amounts of natural language data—that is, written words, as opposed to computer code. There are multiple kinds of NLP tasks, including but not limited to tagging parts of speech; automatic translation to another language; named entity recognition (NER) to identify people, places, or organizations; topic recognition; automatic summarization; and text-to-speech. In the case of the TOMES software, we will be tagging:

- names
- locations
- organizations specific to state government
- sensitive personally identifiable information (PII), such as social security numbers or credit card numbers
- Information defined as confidential by law, such as personnel information or health records

These tags will allow archivists to more easily identify topics of interest, as well as sensitive and confidential information, when email accounts are included in public records requests, so they can more quickly restrict, redact, and make accessible email containing public records.

All tags are notated in an *entity dictionary*, which is a file that the software references while processing the messages to identify named entities. Other entities are automatically tagged by other NER software employed, including Stanford CoreNLP. The TOMES software also consists of a tool that allows other institutions to create their own entity dictionary, or for entities to be added or changed, so the output can be customized. Thus, agencies and institutions outside of the State Archives will be able to adapt the TOMES tool to their own records request needs, such as specialized confidentiality laws, and the State Archives will be able to adapt the entity dictionary to statutory or organizational changes within state government.