Overview of ECMS Projects

EZDesktop Application

The EZDesktop application is used by EPA employees to flag documents as official EPA records. The current iteration of the application is very simple, and allows employees to upload records without any additional metadata. This is a challenge for records management, since each record must be assigned to an official record schedule. Ideally, the employee uploading the document would provide a record schedule, allowing for proper retention and retrieval of the record.

The goal of this project is to update the EZDesktop application so that it requires users to provide a record at the time the document is uploaded. This is beneficial for records classification, but makes EZDesktop significantly less easy for users. Records classification is not always easy – there are over 500 different record schedules, each with their own specific requirements. There is a risk that employees will give records an incorrect schedule, or will be discouraged from submitting valid records because they dread the process of classifying the record. To ease this transition, EZDesktop will provide the user with suggested record classifications based on machine learning. It will also provide utilities for saving commonly used record schedules and searching record schedules by terms in the record schedule description, and using keywords identified by EPA records and language experts.

EZDesktop Outlook Integration

In addition to the EZDesktop application, we would like to develop an integration with Outlook for flagging emails as records. This will have some of the same functionality as the Desktop app, but there will be some differences since only email records can be flagged within Outlook. General emails are already retained under the Capstone project, so emails flagged within the plugin must require a more strict record schedule in order to qualify for entry into ECMS.

Understand and Classify Existing Records

ECMS currently holds roughly 30 million records. The majority of these records do not have an assigned record schedule. ECMS is unable to dispose of records which are not classified to a record schedule, resulting in the unnecessary retention of many records. Using machine learning, we may be able to roughly predict the appropriate record schedule for a document. By applying the model to the full corpus of documents, we can focus our attention on records which the model is confident we no longer need to retain, because they have already been held past the requirement of their predicted record schedule. If we become confident enough in the model, we may choose for a portion of these documents to be removed automatically, reducing the footprint of ECMS.

Using any automated method comes with risks. We will use cross validation to estimate the accuracy of our model. We will compare the estimated error rate of the model with the human error rate to understand how the risk of misclassification by the model compares with the risk of misclassification by EPA employees. This information can be used by senior management to decide whether the model should be used to automatically classify records.

Record Classification Model Analysis and Improvement

All of the projects above rely on our ability to predict the record schedule for a document based on its text and potentially some metadata about the origin of the document. The method currently used by the ECMS team leverages NLP modeling through DeepDetect. We will work to evaluate the model trained through DeepDetect and improve on it by testing other state of the art models. We will also take other steps to improve the model, by restricting the codes the model needs to learn based on requirements from NRMP, adding business logic to classify records where possible, and improving the processes for creating and managing training data.

PATT Upload

The PATT tool is a crucial component in EPAs digitization efforts. As paper records are digitized, they need to be uploaded into the ECMs system. We will need to provide a temporary file storage system in AWS S3 which allows files to be stored while they are still being validated by records management personnel. After records are validated, they will be moved to permanent storage in ECMS.

Enhanced Search Functionality

Within ECMS and across EPA, it is important to have effective methods to search documents. Open source keyword-based search is implemented in Elasticsearch. Elasticsearch can be further enhanced with vector/semantic based search by leveraging modern NLP models such as GPT and BERT. We will apply these methods to create search capabilities in ECMS, and leverage what we learn to create flexible ways to add search to other EPA systems.