State Archives & Records

Using auto-classification to classify unmanaged records

April 18, 2018

Last week the Digital Implementers Group enjoyed a presentation by one of the members of the Group on auto-classification.

Following the end of a service provider's contract, a government agency received the property service records that the provider had been creating and managing for ten years. These records consisted of over 400,000 electronic documents contained in 31,000 folders, some up to 14 levels deep. Many of the records did not have consistent titling or match the agency's own records classification scheme. Due to the impending transition to a new service provider, the records needed to be migrated and classified in a matter of months.

With the scope and timeframe of this migration project rendering manual classification out of the question, it was the perfect opportunity to trial auto-classification.

How the auto-classification system work?

The project team chose to leverage existing investment in TRIM and pilot the use of the auto-classification module. The rationale was that the TRIM auto-classification module was more affordable than procuring a new system as it only required upgrading an existing system.

The auto-classification solution that the team used involved three components. The first stage was an Optical Character Recognition (OCR) program which transformed image files into readable text. The file was then indexed by a content indexing server, and finally forwarded to the auto-classification module to be classified.

While the OCR component of the project was slower and resource-heavy, there was still a strong business case to be made as the OCR component made documents searchable that were not previously.

An agile, continuous process for refining terms

The accuracy of the auto-classification system relied on the definition of a set of terms. When a collection of terms were identified in a record, the system filed it to the corresponding classification.

Initially, the team allowed the auto-classification program to train itself to define the terms for each category. This approach was not successful as the module identified many unknown or garbage terms. A subject matter expert then input manually terms they would expect to see for each classification. This was the most resource-heavy part of the project and the most critical for its success. Refining the terms, feeding new documents through, observing the results and then refining the terms again was an agile, continuous process.

Outcomes

During the testing phase, 5 -7,000 documents were uploaded into the system and were being auto-classified in under two hours, however this rate will change as the team are going to implement a bulk uploader. The OCR component was the most time-consuming in the process, and initially created a bottleneck in processing.

Key learnings

'Better but not best'

One of the members of the group asked about the risks of classifying documents which were not in fact 'records'. Due to the time limitations of the project, the team was unable to triage the documents so proceeded with an 'over capture' approach and accepted that 'non-records' would be captured.

The outcomes of the auto-classification project were described as 'better but not best.' Accepting that the outcomes will always be imperfect was one of the biggest lessons of the project.

Terms are vital

The success of the auto-classification depended on the definition and weighing of the terms involved. For the category 'Cleaning', 95% of the records were auto-classified correctly. This was because many terms were specific to that category. Other categories did not work as well, usually due to the duplication of terms across classifications. The team learned that auto-classification systems do not work straight out of the box, and accurate classification only happens when there is good implementation and definitions of terms for use cases.

Importance of a strong business case

One member who had worked on the project explained that their auto-classification system worked best if you were dealing with a 'mess' of records. They found that there needed to be a strong business case for spending a large amount of resources on the labour intensive parts of the project and this would be hard to justify unless there was a large volume of unorganised records.

Educate stakeholders to manage user expectations

Stakeholders often wanted to know how well the auto-classification system would classify records (e.g. would it correctly classify 9 out of 10 documents?) Due to the variables and unknowns in how the system would work and what the records actually held, that question was not able to be answered.
It was important in the face of these unknowns to educate stakeholders on the

processes you are applying and to set expectations low. The team originally estimated the system would correctly classify 50% of records, although system testing is providing a higher success rate now.

What's next?

The project team can see several other uses for the system. One of the ideas was to integrate the auto-classification system with front end customer service procedures. For example, the system could automatically classify routine forms for business services as soon as they were saved in the system.

Looking to the future, members discussed whether auto-classification could eventually make records managers redundant. Some members thought it could have the opposite effect, as auto-classification could allow records professionals to focus more on aspects of their work such as standards, procedures and programming rather than manual disposal and migration.

Photo by Matthew Paulson
Author: Shoshana Booth