ACCELERATING TRANSITION OF DIGITAL RECORDS TO THE NATIONAL ARCHIVES USING ARTIFICIAL INTELLIGENCE

"The *AI for Digital Selection* project aims to learn more about existing AI tools that could be used to carry out the appraisal and selection of the 'digital heap' of documents, emails, datasets and other types of information held across government."

The National Archives

More details will be available in the forthcoming report authored by The National Archives entitled: "Using AI for Digital Selection in Government: an evaluation of marketplace solutions using machine learning to select digital records for permanent preservation."

CHALLENGE

The UK government needed to find an efficient way of handling the first batch of documents from the dawn of the digital age. These were due to be placed in The National Archives under the 20-year rule for preservation of records of historical value.

SOLUTION

Iron Mountain leveraged the latest developments in Artificial Intelligence (AI) and Machine Learning (ML) to train their system to recognise candidate records for permanent preservation, detect duplicates for disposition, extract entities and provide file analysis.

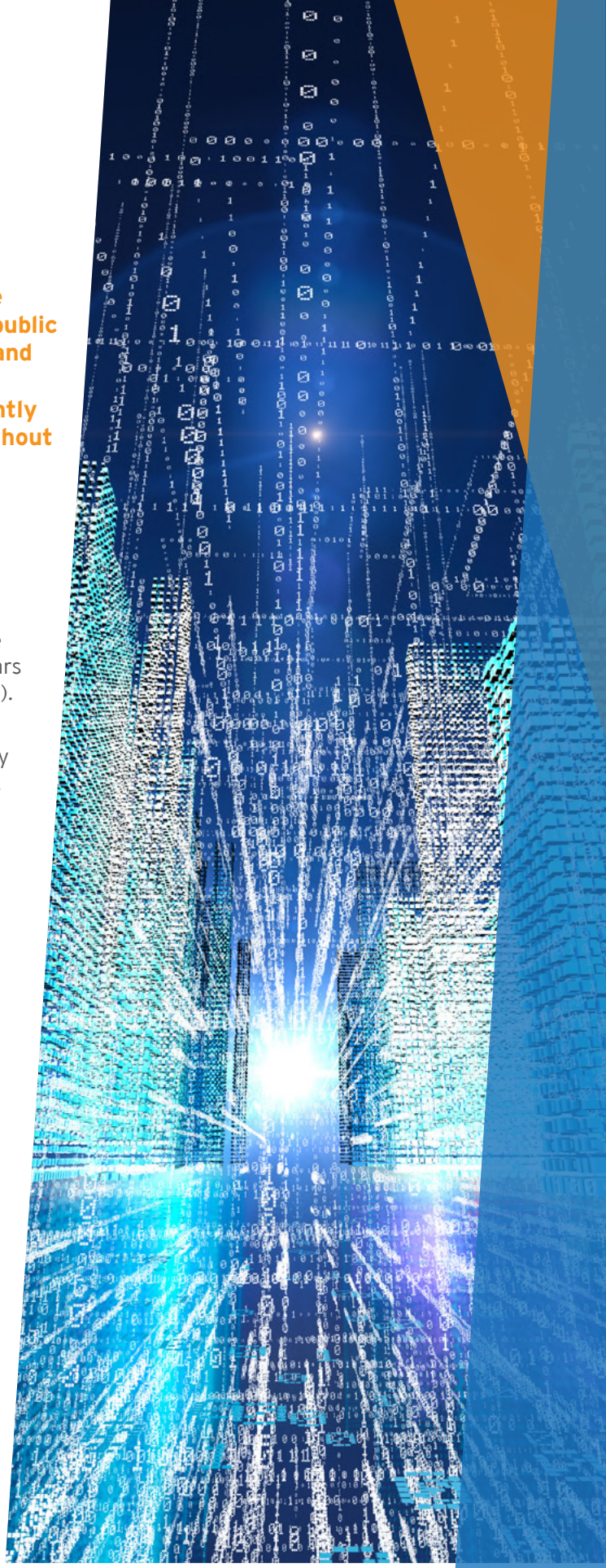
RESULTS

AI and ML capabilities will enable government departments to confidently transfer records far faster and more efficiently than working through the same volume of materials without recourse to these tools.

THE CHALLENGE FACING GOVERNMENT

"The first batch of UK government documents from the dawn of the digital age are about to be released to the public under the 20-year rule. Successfully making use of AI and ML capabilities will enable government departments to confidently transfer records far faster and more efficiently than working through the same volumes of material without these tools."

Ed Irving, Business Development Director, Central Government, Iron Mountain



All UK government departments are responsible for reviewing their records. Those identified as having historical value must be transferred to The National Archives (TNA) once they are 20 years old under the requirements of the Public Records Act 1958 (PRA).

A key step in this process is a Sensitivity Review. This ensures that records transferred do not contain any information that may cause harm to others or reputational risk to the UK government.

The dawn of the digital age

Most of the main UK government departments began to move to digital working around 2004; some departments and agencies have only ever existed in a digital era (although still create some paper). This means that many of the records subject to the PRA now exist only in a digital format. These include a wide range of information from structured datasets to emails, word documents and spreadsheets.

The scale and variety of this digital information means identifying and selecting records of value is near impossible using the traditional 'human effort alone' approach. Government departments are beginning to look at three key questions:

- **Can Artificial Intelligence (AI) or Machine Learning (ML) technologies accelerate and simplify the classification and review process of digital records?**
- **How effectively can this technology sort records of value and ephemeral data or duplicate files that are of no value?**
- **Is it feasible to ingest everything into a single system once selected?**

In support of the government challenge, TNA put together the *AI for Digital Selection* project. The project aims to learn more about existing AI tools that could be used to carry out the appraisal of and selection from the 'digital heap' of documents, emails, datasets and other types of information held across government.

DEVELOPING AI FOR DIGITAL SELECTION

"Given the importance and complexity of the challenge, The National Archives decided to run a pilot programme to ensure that any processing tools were capable of achieving the required levels of accuracy and control. It also needed a partner with a proven track record of safely and efficiently handling large, complex projects."

TNA carried out a review of available tools to identify between three and five for in-depth testing with a set of their own corporate records. The plan was to find how the tools fare in identifying records that should be selected for permanent preservation and those that should not. In addition, TNA wanted to help government departments in using AI for selection. This involved identifying where these techniques could be incorporated into the process or workflow of selecting digital records for transfer to The National Archives.

A well-established relationship

Iron Mountain already has a long and well-established relationship with TNA, as well as with some of the UK government's largest departments, having supported paper record transfer at scale for a number of years. The company also offers InSight™ services, which includes Intelligent Document Processing combining Iron Mountain content analytics, data management and information governance expertise with Machine Learning (ML) and Artificial Intelligence (AI) capabilities.

Iron Mountain was selected to participate in the *AI for Digital Selection* project, to understand the effectiveness of AI in the process of digital selection.

Managing multiple file formats

The Iron Mountain tools are capable of handling a wide selection of media formats, which was essential as there were more than 100 different file types to be processed in the proof-of-concept study including audio, video and text-based documents. Some of the most obscure formats were sidelined for the purposes of the trial but the actual number of files that fell outside of the study was small – fewer than 10%. The files included in the proof-of-concept were then loaded into the Iron Mountain content services platform for analysis.

WHAT DID THE SOLUTION INVOLVE?

Ascertaining what needed to transfer

As part of the project, TNA provided Iron Mountain with labelled and unlabelled data sets to demonstrate AI capabilities intrinsic to InSight™ in identifying records relevant to selection criteria.

Removal of sensitive data

Even with files that are being made available to the public, there needs to be a stage at which sensitive information not authorised for release can be removed. In a paper-based world, redacting information is a physical process which takes place as part of a human-led review – documents are marked up or edited by hand. Digital forensic technology tools take a different approach – they obfuscate words or sentences, scrambling content where appropriate.

Steps to success

For the trial, Iron Mountain first loaded the 17,000 test documents into Google cloud bucket storage. The documents were processed using Optical Character Recognition technology to make them fully-searchable. The InSight™ Intelligent Document Processing platform then classified them into 20 pre-defined categories using natural language processing (NLP), a software process that can decipher the content of a document and the contextual nuances of the language being used. This enabled the platform to accurately extract the information contained in the files and build associations across the sample set to ensure meaningful search was possible.

InSight™'s machine learning capabilities enabled the project team to train the model in an iterative process over the course of the project. In the end, the tool achieved an F1 score of above 85%. The following outcomes were also delivered:

- Duplicates identified for disposal
- Candidate records for permanent preservation identified
- Entities including organisations and people extracted
- File analysis including content summary, age summary etc.
- Average Precision and Recall scores

Successful proof of concept

InSight™ delivered the required level of functionality, document classification and duplicate detection. Iron Mountain are pleased to support The National Archive with this programme of national interest. The solution can now be used across government to accelerate the transition of documents from all government departments via straightforward framework procurement routes.

1300 476 688 I IRONMOUNTAIN.COM/AU
0800 723 255 I IRONMOUNTAIN.COM/NZ