THE NATIONAL ARCHIVES

Using AI for Digital Records Selection in Government

Guidance for records managers based on an evaluation of current marketplace solutions

October 2021

Contents

Abstract

Digital transformation in government has brought an increase in the scale and variety of public records along with a reduced emphasis on filing and organising data. Traditional processes designed for paper records cannot handle the volume, diversity, complexity and distributed nature of Departmental digital records. This report describes work done at TNA to explore the potential of Artificial Intelligence (AI) tools to assist with this challenge.

Five AI vendors applied their tools to classify a dataset supplied by TNA. The tools and platforms evaluated were Adlib Elevate, Amazon Web Services, Microsoft Azure, InSight by Iron Mountain and Records365 by RecordPoint. Promising results were obtained overall with no tool or approach consistently outperforming the others across all tasks.

The project found that while AI cannot replace the expertise of Records Managers, commercially available AI tools and pipelines can be successfully applied to aid the task of records selection in semi structured and unstructured collections. The products are evolving rapidly and this a good time for departments to engage with suppliers to benefit from current capability and influence the direction of development.

Key learnings for the application of these tools in government departments are

- Investment in careful preparation of training data and commitment to a process of testing and refining models yields significantly better results than may be achieved using any of the tools 'out of the box'.
- Factors including fit of the feature set with the department's specific requirements; compatibility with the department's technology environment; and cost should be considered alongside raw reported performance when choosing a tool because more sophisticated tools do not deliver better performance against all requirements.
- Records managers will require technical training and access to data science expertise if they are to deploy these tools successfully.

Audience

This paper is written for UK government officials who are familiar with the appraisal, selection, review and transfer of records to The National Archives for permanent preservation. These processes are described in TNA's guidance on *Digital Records Transfer*.

1. Introduction

1.1 Background

It is the duty of Public Record Bodies under the Public Records Act (PRA) to select records of enduring value for permanent preservation at The National Archives (TNA). Traditional processes designed for paper records cannot handle the volume, diversity and distributed nature of digital data. The *Better Information for Better Government* project (BI4BG) estimated in 2018 that UK Government data totalled more than 16 billion emails and 3 billion documents, amounting to 5PB of data. Without solutions to reduce

manual effort and address the scale of digital records, there is a risk that valuable records will be lost as departments struggle to meet their PRA obligations. At the same time, departments will bear the cost and risk of storing unstructured records of low value for extended periods of time. Hence there is an urgent need for automated solutions.

This report describes TNA's evaluation of a range of machine learning (ML) solutions for records selection and builds on our previous survey of *Rules based eDiscovery products for Technology Assisted Review*, which covered the wider appraisal, selection, and sensitivity review of born-digital material.

Section 1.2 below introduces some core concepts in applying AI approaches to this type of problem. *Sections 1.3 – 1.5* give an overview of TNA's investigation, the data used and the commercial tools that were trialled.

Section 2 provides guidance for departments wishing to adopt AI tools to assist with records selection, focusing on practical advice drawn from the findings of this project.

In *Appendix A* we describe the steps involved in creating an AI tool, tailored to the needs of a specific set of records (such as a Departmental digital records store).

Appendices B & C give more information about the products under investigation and a detailed breakdown of product performance against TNA's evaluation criteria. This is supplemented by the individual *Supplier Reports* published alongside this document.

1.2 Introduction to Artificial Intelligence approaches for Records Selection

In this report the term 'Artificial intelligence (AI)' is used to refer to systems that can make intelligent automated decisions. Machine Learning (ML) refers to a specific technical approach to implementing AI systems. The premise of the machine learning approach is that an AI system could 'learn by example' to undertake a task that cannot be precisely codified.

This report examines whether ML approaches can assist with selection of records for transfer to TNA. This type of categorisation exercise is termed a 'classification' task, whereby records are classified as either 'selected' or 'or not selected'.

The machine learning approach to developing a classifier is to create a 'model', based on learning from 'training data' which has already been classified, usually via manual effort by an expert. Model-building proceeds via a process of identifying patterns or features that are characteristic of the 'selected' records in the training data and uncharacteristic of the 'non-selected' records. This is not a precise process and it is common to build and test several different models to determine which gives the most accurate results. When evaluating models, accuracy must be defined based on business need. For example, for this task, a classifier that correctly identifies records of value while also including some ephemeral records would be considered to out-perform one which reduces the selection of ephemera, but with the loss of valuable records. For other applications, the converse might be true.

The best model or combination of models can then be used to build a 'classifier'. This is a software tool which can be applied at scale to divide records into 'selected' and 'not selected' categories. The most useful tools will report a measure of their confidence alongside the results of the classification. This can be useful in determining how much

checking or human intervention is required before a tool's predictions become records selection decisions. See *Appendix A* for a practical overview of the work involved in developing an AI model using ML for classification of digital records as selected or not selected.

1.3 The 'AI for Selection' project

In 2020, TNA conducted an investigation into the applicability of commercially available AI and ML tools to the task of digital records selection. The purpose of this project was to:

- Explore the effectiveness of AI based commercial products to aid the identification of digital records suitable for permanent preservation at The National Archives.
- Understand the current state of commercial ML products in this space.
- Understand how commercial ML products could be practically applied to aid selection.

The project was designed to run in two phases. The first phase aimed to gain an overview of the types of products that were available and identify products for further investigation. In the second phase we engaged with suppliers (either product vendors or independent suppliers) who trialled the chosen products using The National Archives' own corporate data to help us understand how they worked and assess how well they could identify records for selection.

When briefing suppliers, we emphasised that the investigation was not a competition. The aim was not purely to evaluate the performance of the products but also to demonstrate the range of functionality available, to develop our understanding of the current state of the art and to explore the trade-offs involved in choosing a product for digital records selection.

1.4 Data used for proof-of-concept classifier development

The products were evaluated using data taken from TNA's corporate records management system. This comprised of 110,882 files and 12,462 folders, a total of 44.1GB in various formats, predominantly text-based records including emails, PDFs and Microsoft Office formats. This dataset was chosen as broadly representative of TNA's corporate records and will be referred to here as the 'representative data'. The documents came from TNA's EDRMS and were already organised into folders and subfolders according to TNA department and function or topic, in line with local Knowledge and Information Management (KIM) guidance. The subfolders were labelled with their retention schedule, which stated how long the records would be retained and identified those which had been selected for permanent preservation.

This document set came from one small department (TNA) and was highly curated. We know that this is not the case for all departmental 'digital heaps'. Satisfactory performance against this dataset does not imply that equal performance would be achieved against the records of a larger, more complex department. Departments will need to create their own 'representative datasets' in order to train and test models and achieve satisfactory levels of accuracy. See *Section 2* for a discussion of approaches to choosing representative data and training data.

1.5 Approaches used by the tools under investigation

The products selected covered a variety of approaches from off-the-shelf records management systems to cloud-based bespoke pipeline tools. *Appendix B* details the tools and the rationale behind their inclusion in the project.

The range of approaches used by the products made direct comparison difficult. The results presented here indicate the tools' accuracy but additional functionalities were evaluated to assess the broader value of the products to Records Managers in government. All suppliers were given a fixed period of time to create their prototype and/or carry out their tests but started at different stages as some had fully functioning, off-the-shelf/proprietary systems, while others were constructing systems from cloud hosted components. This meant that the off-the-shelf products had many of the requested features already available, while the consultancies working with cloud products ignored some of those features in order to achieve the classification goal in the time available.

Appendix C gives a full product comparison against our evaluation criteria.

2. What do Government Departments need to know to use AI for digital selection?

From our findings in evaluating the products and additional learning during the *AI for Selection* project, we can offer some advice to departments considering the use of AI to select digital records for preservation (or similar tasks).

2.1 Areas to consider when deciding the approach to AI for Selection

When working with suppliers departments should have realistic expectations of the software and expertise available in the market and should not underestimate the resource and expertise which will be required from the department. Departments will need to communicate with suppliers regularly and provide records knowledge and domain expertise to achieve the best results. The following areas should be considered when deciding on an approach.

Choosing training data

Training a supervised learning algorithm to carry out records selection essentially means trying to imprint the context of an appraisal policy via the training data. The process is one of showing the algorithm pre-classified data from which it can 'learn by example' and work out the rules for itself, followed by 'correcting' its answers so that it can refine those rules. The quality of the training data is critical to receiving good results. If a department has an existing set of labelled records (both selected and not selected), this could be used as a basis for training data. The data should be a representative and diverse set of records to reduce bias. If representative labelled records are not available for training the model, then they will need to be created. This is resource-intensive and needs a Records Manager (or team familiar with the records). Documenting the approaches taken in creating training data will be important in showing that it is representative of a department's appraisal policy.

When selecting training data Records Managers should be aware of the prevalence of

duplicate records and ensure there is enough variety in the training set. Duplication of training data can bias classifier results. For example if there are 100 records in the training set marked as 'selected for transfer', 10 of which are a copy of the same record then the model could be biased towards the selection of records similar to the duplicate. In these cases, it is often desirable to de-duplicate the training data. However, the presence of duplicates in archived records can provide valuable contextual information for future users and duplicates should not be routinely be removed as part of the records selection process. TNA accepts transfer of duplicate records.

Another important consideration is the volume of training data to be used. In the AI for Selection project, suppliers used up to 80% of the labelled training data to train the model and up to 20% to test (from a 44.1 GB dataset, 110,882 files / 12,462 folders). The model used the 80% of data to train (this used metadata and content of files) and then for the 20% test data the model was not given the labels, the labels were then used to see how accurate the model is. More information about data used can be found in *Appendix B*. The aim would be to use training data that is representative of a larger unlabelled dataset so then the model can be applied on a larger scale to determine selection in that dataset. The AI for Selection Project used data which was already labelled and in our EDRMS, though was under represented in some record categories, so was not suitable to test against our larger corpus of data.

The size of the training data required depends on the type of solution to be applied. Some tools require training data to be labelled manually within the software, generally working with smaller volumes of labelled data, while others allow larger volumes of labelled data to be loaded.

What level of accuracy is good enough?

Real life data is often not good enough for ML purposes without some intervention – usually data cleansing. The data is often skewed so that certain types of important files (i.e. files that should be selected) may be under-represented and missed by the training stage of the process. When assessing the accuracy of a product, it is important to consider

(i) whether it is acceptable for a few historically important files to be missed and therefore not selected in order to avoid selecting a great many unimportant files as well

or

(ii) whether the model must select *every* important file (and there may be very few files of this type) even if this could result in a great many unimportant files being selected too.

There is generally a trade-off between the two considerations. If it is important to ensure selection of 100% historically significant documents then you will probably include a high volume of 'unimportant' ones too. If you were worried about selecting a high volume of 'unimportant' records this increases the risk that you will miss some important

documents. TNA favours approaches which select more records of enduring value, even if this means that the selection is 'messier' or that greater volumes of 'unimportant' records are transferred.

Irrespective of approach, ML models are unlikely to achieve the best accuracy without human involvement. To improve accuracy, incorrect classifications can be corrected by records management teams and the model can be retrained. This step should be repeated several times until satisfactory classifications are achieved and will be an on-going commitment even when algorithmically assisted selection is used operationally.

For some of the products evaluated in the *AI for Selection* project at TNA, this re-training is only undertaken in the initial setup by the engineering teams (usually requiring extensive discussions with records managers or record specialists), while others provide options for records managers to undertake corrections of classifications and retrain the model inside their software on a continuing basis.

It is necessary to monitor the accuracy of the model, reviewing its performance periodically. The frequency of retraining will depend on changes to a department's view of what a 'selected' record is, such as a new appraisal policy. This could also be affected by a major event changing the importance of certain topics, for example records related to 'coronavirus' will have increased in importance over the last few years. If the criteria used for record selection change, the model will not perform as originally intended. It may require retraining with new training data or even the creation of a new model to avoid bias from the outdated selection decisions that are embedded into it. Similarly, if the nature of the records to be classified changes, new model(s) may be required.

How many models to train and test?

The number of models depends on the variety and types of records in the collection. A model developed for organised and structured data held in a document management system will not give the same accuracy with unstructured data held on a shared drive. Similarly, different models may be required for different file types such as text, media and images.

Understanding and explaining your results

How best to explain and interpret results is still a challenge for the Machine Learning community. The metrics provided by systems can be interpreted by data scientists to assess a product's effectiveness but are often less accessible to a wider audience. Records Managers and departments should be able to exercise a degree of control over how products assign selection decisions, leveraging technology to make their job possible, rather than relying on it to perform the selection task for them.

Transparency is vital and more work is needed by suppliers of these products to be able to explain algorithmic decisions. Government departments need to be able to show how their use of these tools aligns with records selection policies. Where departments adopt ML approaches, the processes followed and rationale behind the choice of training data

could be published alongside appraisal policies to help future users of the records understand potential biases. Additional methods to aid transparency could also be undertaken, potentially the training set itself could be selected as a 'record' to show how machine learning models were trained.

Handling sensitive data

Depending on the sensitivity of records in a government department, products may require appropriate levels of certification and security. This may limit the choice of products if, for example, suppliers require records to be uploaded into their systems. For the AI for Selection project using TNA corporate data we established data sharing agreements with external suppliers to ensure appropriate data handling measures and security were in place. The time required to complete these steps appropriately should not be underestimated.

Are you ready to implement automation?

Finally, the department should consider whether it is ready to invest in an ML solution. Initial steps could be taken prior to investment (for example, to identify a potential training set and label records for selection) which would facilitate an ML approach later.

2.2 Select an approach

The AI for Selection project identified the two main types of approaches offered by suppliers:

(i) **Off-the-shelf** record management solution with AI functionalities OR a product for viewing and working with results after modelling done via consultancy

(ii) **Bespoke solution** built by external specialists or by an in-house team of data scientists and developers

The department may have a preference based on local policy and previous experience. Where this is not the case the choice of approach should consider:

- The level of control and autonomy: a bespoke solution gives greater control than an off-the-shelf product which gives greater control than a product run for the Department by a consultancy. Security constraints should also be a factor.
- The capacity and skill of a department's technical team: a bespoke solution will require greater technical skills to commission and use than an off-the-shelf system which will require greater technical skills than a product run by an external consultancy.
- Additional benefit: Records management features included in off-the-shelf products could be used to meet other business needs.

2.3 Cost

It is important to be aware of the costs which come with different types of products. Off-the-shelf solutions will require a license and renewal fees, whereas bespoke solutions will incur higher costs for highly technical staff to build ML pipelines and ongoing support of the solution.

Three of the solutions reviewed by the AI for Selection project were records management products which incorporated ML technology to aid selection. A cost-effective way to use them might be for a department to also use the product as its record management system, but the ultimate choice of a document management system should depend on more than its record selection functionality.

One of the suppliers provided an off-the-shelf product to view results of modelling done by the supplier. If modelling is run by a consultancy, the initial cost should be considered alongside the anticipated frequency of re-training models or creating new ones, as this could become difficult to scale – though this approach may still be more cost-effective than employing a full time data scientist. Departments often already have teams with the requisite skills for this work, but the needs of records managers may have low visibility and supporting the use of AI for records selection may not be a priority for data-analytical teams. Early engagement and raising the profile of the selection task could assist records managers in gaining support from these teams.

Our experience with cloud platform suppliers indicates that the cost of the commercial cloud ML components of the pipeline can be significant. We would expect these to reduce over time as the products mature but for high volumes (millions) of records, departments may currently find the costs prohibitive. An alternative would be to build an in-house solution, taking advantage of the scalability of cloud, but using open-source libraries rather than commercial ones.

2.4 Considerations for implementing (or reusing) a bespoke solution

If a department's capabilities allow a bespoke process to be developed, then there are some technical questions that should be considered.

How much data will you be processing?

The dataset used in the *AI for Selection* project was relatively small (approximately 100k records. This could be processed, analysed and used to train ML models on laptops with basic configuration (16GB RAM, core i5 processor, no dedicated graphics card, SSD drive). To address the same task some of the products built bespoke applications using containerisation which could easily scale to very large datasets. Scaling up cloud-based solutions was not tested as part of the *AI for Selection* project. When choosing an approach, it is important to identify what is 'good enough' for the task, to avoid over-engineering. An initial proof of concept may be conducted on a small collection to evaluate an approach and scaled up later.

What platform are you comfortable with?

All major cloud providers (Azure, AWS, Google, IBM, etc.) offer multiple ways to build ML applications at all levels of complexity. When choosing a platform for AI for Selection, it is worth starting by exploring the platform(s) which are already in use by your department and where you may already have some expertise.

2.5 Developing Market

The *AI for Selection* project showed that ML has the potential to be an important part of the solution for dealing with the scale of digital records selection. Departments should be aware of the amount of work and expertise they will require to implement one of the approaches listed above. It has also shown that there are areas were improved options could be explored, such as understanding and being able to explain the results. ML approaches and tools are developing and great amounts of research are being undertaken in understanding results, as well as other areas such as reducing the amount of training data required. All of the products we saw were in early stages of development which means this is a great time to engage with suppliers and influence their future development so that they work for Records Managers and Archivists.

Conclusion

The advice presented here is drawn from a small project which evaluated a limited number of bespoke and off-the-shelf applications to classify records for selection for permanent preservation. The project enabled TNA to have a good understanding of the current market for AI tools in this area. The project achieved promising results and demonstrated that AI approaches can be applied to the records selection task.

The solutions under investigation solved the problem in different ways and offer different features. Our evaluation criteria allow the products and approaches to be compared *(Appendix C).* Departments wishing to implement AI approaches to assist selection can build upon this work to evaluate options before choosing a supplier.

It is clear from this project that AI cannot replace the expertise of Records Managers but can be a useful tool to help deal with the scale of digital records collections in Departments. Records Managers' knowledge of the records in their custody is essential for any of the products or approaches described here to work effectively. While the products require different levels of technical expertise in the team, it is equally clear that information management teams will require access to skills in data analysis and ML if they are to implement such tools successfully. This may be achieved by through building this capability within the team, engaging with data scientists within the Department or via external suppliers. Whichever approach is taken, Records Managers will benefit from training in the concepts of ML, the practicalities of creating data to train the system and the ongoing task of reviewing and refining models.

Appendix

Appendix A - Steps for AI classifier development

This section provides a brief introduction to the stages involved in building an AI classifier[1] tool capable of classifying records as *selected* or *not selected* for transfer to TNA (see Figure 1 below for an overview of the stages). Records Managers applying ML tools will require an awareness of how the product they are using addresses these stages. These steps may be carried out by the department or the supplier but time and effort invested at this point will determine the performance of the classifier.

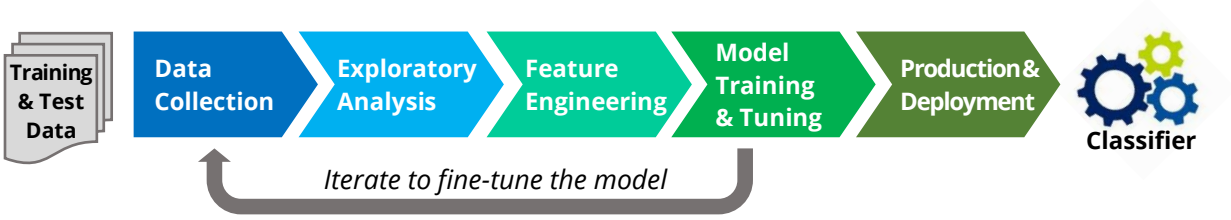

Figure 1. Overview of steps in developing a machine learning classifier.

1. Data collection

Data collection is comprised of (i) data acquisition and (ii) data pre-processing.

(i) *Data acquisition*: In this step, the data to be classified is collected and aggregated. Most ML tools require data to be available in a single location. This can be a challenge for government data which is often distributed across multiple storage locations including shared drives on a corporate network, cloud storage, stand-alone hard drives, in multiple databases or on different servers.

For the AI for Selection project TNA provided aggregated representative data either on a hard drive or uploaded to the supplier's cloud.

(ii) *Data pre-processing*: The acquired data usually requires pre-processing to prepare and format it for use with the chosen tool. Pre-processing often includes data cleansing, for example, removing duplicate and corrupt files or converting dates to uniform formats.

2. Exploratory data analysis

The huge volume and diversity of data in use in government departments makes it difficult to offer general guidance on the type of model that should be applied. Further insights into the nature of the data under consideration are required. It might be worth using statistical techniques to visualise your data to help reveal inherent patterns and missing information e.g. under represented categories in training data.

[1] Refer to glossary of terms

Usually, understanding the final outputs provided by an AI tool depends on understanding the data it is trained on. Important characteristics include formats, size (both volume and number of files), content and metadata. The exploratory data analysis step enables data scientists (either in house or from product suppliers) to identify which features of the data could be used to support classification and, if needed, identify improvements. The risk of omitting this analysis is that significant resources could be invested into building a model which produces poor results.

Along with the complete data records, TNA also provided highly structured metadata as an additional spreadsheet. This contained important information about the file path, date last modified, retention schedules (e.g 04 Appraisal Decisions - Permanent Preservation), selection decision and other important information. Exploratory data analysis revealed that while records belonging to certain retention schedules were abundant, some important categories of retention schedules (including ones for permanent preservation) were under-represented in the training data. Suppliers took this into consideration when selecting modelling approaches.

3. Feature engineering

Features are the characteristics of data that can be used to create a classification model. For example, the file-path, words in the documents or the presence of named entities such as organisations or people. The feature engineering step is also an opportunity to augment the data with human knowledge.

During this stage, issues identified during the exploratory data analysis data are rectified. For example, we may need to add new features, either by manual labelling or by combining the data with external resources to make it easier for the modelling algorithm to identify differences between records.

In TNA's representative data, a metadata feature 'file-path' is provided. This is an important feature which provides us the information about the originating unit within TNA. This is often highly relevant to the selection decision for a record. However, when records are stored in common shared drives, the file-path alone may not provide enough information and may need to be combined with features such as the author of the document. Using this human knowledge, one may engineer two features (the file-path & author) by combining them.

4. Model training and tuning

Depending on the data to be classified, a variety of mathematical approaches can be applied to build a model. The initial choice is usually a simple approach. A portion of the labelled data is used for training the model. The remaining portion is set aside for testing the model. Multiple models may be trained with the same data to provide a basis for comparison of approaches and to determine which model(s) perform well for the data under consideration. Since machine learning algorithms (models) are developed to extract patterns from data, the most appropriate choice of algorithm will be dependent on the data being analysed. Once a model has been chosen, it is fine-tuned to get the best running model for classifying the data.

When training models, 80% of the representative data is typically used to train the model (called "training data") and the model's performance is then evaluated on the remaining 20% (called "test data") before applying the model to new, unknown data (for example to evaluate a supplier's solution as part of a tender or to use the model in a live environment to classify records for selection). The 80/20 ratio is only a guide and can

vary according to the specific problem). When working with an external supplier, a department may choose to hold onto some of their representative data so that they can test the model themselves, independently of the supplier.

One or more modelling algorithms are trained with the training data and then used to classify the test data. The results are analysed and compared to prior knowledge of the 'correct' classification to evaluate the performance of each model and assess its suitability for the task. This evaluation may be quantitative (for example, comparing the accuracy with which different models can classify the records into categories) or qualitative (for example, examining individual errors or considering which types of records were wrongly classified). The model and the features should be examined and adjusted iteratively until satisfactory performance is achieved.

5. Deliver to production and deployment

Once the model provides satisfactory results, it is developed into an application which can be used by Records Managers and others to classify unlabelled documents. Usable applications tend to feature a graphical user interface (GUI) designed for non-specialist users, unlike the command-line interfaces which may be encountered during the training and refinement stage. Ideally the application will support a mechanism which enables users to provide feedback on the results, which can then be used to further tune and improve the product's performance and generate reports for the user to review or act on.

Once an application is ready to be used a department will have to consider if it is ready for deployment. This will depend on the level of confidence in the tool's accuracy. Even when there is a high degree of confidence in a tool, a level of manual checking will still be required, though this may be focused on items which a tool has marked as having a low level of confidence in its decision.

Appendix B - The Products

This section briefly describes the five products tested by the suppliers and a sixth product, developed by TNA for use as a benchmarking tool. More details can be found in the suppliers' reports on the report (see *Appendix D* for links to supplier reports) along with a more detailed description of TNA's benchmarking tool.

An experienced consultant was appointed to carry out desk research to identify four products that would potentially meet our requirements for records selection. The consultant provided an in-depth review explaining the reasoning behind each choice along with reasons that other products were excluded. The recommended products covered a variety of approaches from off-the-shelf records management systems to cloud-based bespoke pipeline tools. We added a fifth product with which we had prior experience and which differed in approach from the four tools already chosen. The products and suppliers included in the project were Adlib Elevate by Adlib, Amazon Web Services (AWS) used by Kainos, Azure used by Adatis, InSight by Iron Mountain and Records365 by RecordPoint.

Although the project endeavoured to standardise evaluation it became clear that direct comparison across different products and approaches is difficult. Summary results are available in *Appendix C* with further detail in the suppliers' reports (see *Appendix D*).

Adlib Elevate by Adlib

Elevate is a cloud-based off-the-shelf record management product developed by Adlib. Although the platform is designed for ease of use by Records Managers, it also provides functionality for data scientists to engage with the model building process. It can be used to transform large collections of unstructured data into structured data to carry out business processes.

Amazon Web Services (AWS) used by Kainos

Kainos are a technology supplier who built a bespoke cloud hosted Machine Learning as a Service (MLaaS) solution using Amazon Web Services (AWS). Their solution trains a model using representative data but model selection and parameter tuning are opaque and not configurable. The AWS platform offers open ML algorithms but the supplier was specifically asked to test Amazon Comprehend, a Natural Language Processing product which was chosen for evaluation. Records and their labels are loaded into the system via an API (Application Programming Interface) which requires some programming skill. A full pipeline could be built programmatically using the platform.

Azure used by Adatis

Adatis are a technology supplier who built a bespoke cloud hosted MLaaS platform using Microsoft Azure. The solution uses existing algorithms (Azure Cognitive Search, Azure Cognitive Services, Azure Databricks) and includes the ability to build an ML pipeline. A prototype GUI was also created to demonstrate how a Records Manager could execute workflows on the platform.

InSight by Iron Mountain

Iron Mountain developed InSight as a content-services-platform focussed on records analytics, with ML capability provided by Google Cloud's ML and Artificial Intelligence service. The ML training is a managed service performed by their own data scientists, rather than the user. The interface was therefore focused around organising and searching records.

Records365 by RecordPoint

Records365 is a cloud-based off–the-shelf Software-as-a-Service (SaaS) platform, developed by RecordPoint. It is designed to enable Records Managers to use ML without the input of a data scientist. Training a model is achieved by manually classifying the sampled records through the GUI. Following training, test records are imported and the ML classifier assigns predicted labels. These are not final until they have been approved or corrected by the Records Manager. The model can be iteratively re-trained as the user works through this process.

Benchmarking classification tool by The National Archives

This was a lightweight AI-based classification tool, developed in-house at TNA using open-source software libraries, to provide a baseline against which the results of the five commercial products could be compared.

Appendix C – Product comparison

The five products under consideration applied different approaches to the selection task. For this reason, it is difficult to directly compare the results obtained. We have developed an initial evaluation form that identifies factors we consider important in evaluating results..

This section discusses in more detail the functionalities of the various products, structured by the categories used in the evaluation form. These categories broadly align with the phases in the ML process described in *Appendix A* above.above.

The individual supplier reports have been published and provide further information (see *Appendix D* for links to supplier reports).

1. Data Collection

Products Adlib Elevate and RecordPoint's Record365 retrieved files directly from cloud-based storage services (such as Google Drive, Microsoft SharePoint, Dropbox, Exchange, etc.). InSight by Iron Mountain accessed files stored on cloud servers (Google, Azure).

Adlib Elevate and Records365 provided custom connectors that can be reused. Both bespoke solutions could add connectors, given time and resource. While Adatis favoured Azure blob storage, Kainos used Amazon s3 to store the dataset.

Data Collection		**ADLIB Elevate by Adlib**	**AWS (used by Kainos)**	**Azure (used by Adatis)**	**InSight by Iron Mountain**	**Records 365 by Record Point**	**TNA Baseline**
Source of data for import	**Cloud based (Dropbox / Sharepoint / Google Drive / etc)**	**Y**	**N**	**N**	**N**	**Y**	**N**
	Cloud server based (Azure, Google, etc)	**Y**	**Y**	**Y**	**Y**	**Y**	**N**
	Data from hard drive	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**

Table 1: Products' data collection ability

2. Data Pre-processing

We decided to concentrate on the text-based and email file formats as they are the key file formats predominantly used in government departments. All suppliers managed to pre-process these file formats. Even though the contents of emails are text-based, they were not included by some suppliers due to a lack of time and processing complexity.

The next issue that arose was the existence of duplicates in the representative data. Duplicates are common in departmental records and it is not unusual for multiple versions of a record to be transferred to TNA. As it was intended to be 'representative' data, the dataset provided to suppliers contained duplicates too. However, when used in a training set, the presence of duplicates can bias the model. For example, too many copies of an unimportant (i.e. labelled as not selected) document would mislead the product to classify all the records originating from the same source as 'not selected'. The

product would then make an incorrect prediction for other important records that should be selected. Mathematical approaches to detect duplicates (cosine similarity, hashing) were applied by the suppliers.

Our experience of running TNA's own benchmarking tool suggested that removing duplicates for the purpose of creating a model could be difficult because of lack of policy on which instance should be retained. Eliminating duplicates would require a manual review which is time-consuming. One of the learnings from this project is the need for additional staff time to remove duplicates for future versions of the tool or to reduce bias by ensuring sufficient variety in the training set. Some products ignored duplication, while some detected and removed duplicates as part of their pre-processing.

For clarity, while it is of'ten helpful to remove duplicates from training data to reduce bias in the model, 'real' records should not routinely be de-duplicated as part of the selection and transfer process. TNA accepts transfer of duplicate material.

3. Exploratory Analysis

The detailed exploratory analysis made by each of the suppliers is provided in their reports (see *Appendix D* for links to supplier reports).

4. Feature Engineering

Basic record metadata features (for example, file location, date, format type, size) were used by all suppliers to assist classification. However, to classify a record as selected, it is also necessary to understand the content. From our experience with the benchmarking tool, we now know that some of the retention schedules were under-represented in the training data (i.e. there were not enough records of this type in the training data / the training data was not sufficiently representative). While some suppliers (including TNA's benchmarking tool) tried various ways to overcome this, others ignored the under-represented categories. Both approaches had advantages and disadvantages (see links to reports in *Appendix D*). The retention schedules have a direct relationship with whether or not the record should be selected for permanent preservation. There were about 20 retention schedules. Even though, the task was to classify records into selected/not selected, it was useful to categorise records into retention schedules while evaluating the performance of the products.

- Data augmentation or synthetic data[2] was created and used to improve the representation of the under representative categories. Though synthetic data is the most prevalent method of improving data quality, generation of synthetic data is time-consuming and requires a very keen domain knowledge of the type of the missing data. Synthetic data is also more suitable for numeric than textual data.

- Ignoring the under-represented classes is a valid approach if we do not expect additional records of this type in the real data. If this is not the case, ignoring a class when building the model will lead to this category of records not being identified as 'selected' by the tool and, as a result, they will be destroyed.

It is difficult to create an algorithm that is sufficiently generalised to be applied to datasets from any department. Models (algorithms) must be specific to each dataset to produce satisfactory results. In the ML world, it is the data that directs the choice of the

[2] See glossary of terms

algorithm.

Table 2 shows the utilisation of various capabilities by the products under evaluation.

Data Pre-processing, Exploratory data analysis & Feature Engineering		**ADLIB Elevate by Adlib**	**AWS (used by Kainos)**	**Azure (used by Adatis)**	**InSight by Iron Mountain**	**Records 365 by Record Point**	**TNA Prototype**
Formats managed	**Text documents** (pdf / word / text / excel)	Y	Y	Y	Y	Y	Y
Content extracted through OCR/NLP/etc	**Emails and attachments**	Y	Partial	Y	Partial	Y	Partial
	Images	Y	Y	Y	Y	Y	N
	Videos / audio	Partial	N	N	N	N	N
Detect and ignore duplicates		Y	Y	Partial	N	Y	N
Retrieve file metadata (file location, date, format, document name, size)		Y	Y	Y	N	Y	Y
Feature (metadata) extraction	**Internal / external users**	N	N	N	N	N	N
	Date inference (subject date, document date)	N	Partial	N	N	N	N
	Organisation / gov department	Partial	Partial	Y	N	Y	Partial
	Internal department	Partial	N	N	N	N	Partial
	People	Partial	Partial	Y	N	Y	Partial
	Document importance & Security Classifications (official sensitive / commercial / personal / none)	N	N	N	N	N	N
	Title	Y	Y	Y	N	Y	Y
	Routine / non routine record type (example weekly updates)	N	N	N	N	N	N

Table 2: data pre-processing and utilisation of various capabilities by the products

5. Model Training and Tuning

Choice of features. All suppliers built the first version of their classifier models using metadata features alone. Models were trained on 80% of the data and tested on the remaining 20%. Since metadata features do not consider the content however, both accuracy and precision suffered as the algorithms struggled to predict the selection status of the test data. In addition, the *file location* feature worked well for organised

data but not for unmanaged shared drives, where using file location as a feature was a poor predictor. It became clear that a hybrid model using both metadata and content could work well.

Classification task. Since the data supplied belonged to multiple retention schedules, some of the classifiers focused on predicting these schedules which specified the length of time a record should be held (including *selected for permanent preservation*). Others focused only on identifying whether records should be selected for permanent preservation or not. Some suppliers performed multiple experiments on samples of the whole collection to avoid training on the entire dataset. More details of products' results can be seen in the suppliers' reports (see links to reports in *Appendix D*).

Model customisation. Most tools offered no way to customise a "built-in" model, whether off-the-shelf document management systems like Adlib and Record365 or when using an AI black box API as Kainos did with AWS Comprehend. For the benchmarking tool, simple models were initially chosen and a customised model created iteratively by selecting the most suitable features and tweaking the algorithm. Adatis also built a customised model similar to the benchmarking tool. In our view, this approach to customisation would work well as record managers or developers could use the model without any deep knowledge of ML. It should be noted that oversimplifying the model can be a risk as it may not perform accurately.

Versioning and sharing. There are many existing tools available for versioning, sharing and exporting models. Versioning enables us to try training multiple models and to roll-back when performance degrades. It also helps us to understand the predictions made and the parameters required for building the model. A sharing facility allows the re-use of successful models with other similar collections with few changes. It is important that departments are involved in the training and tuning process. This should be fully documented (and that documentation considered as a public record).

On-going tuning. After initial deployment, tuning will need to continue and performance will need to be monitored. A Graphical User Interface (GUI) can help to make the process and models understandable (see *Table 4* below).

Interpreting performance. How can the product performance be interpreted by end users? Metrics such as *f1 score* and *confusion matrices* evaluate the model's performance but are not easily interpretable by decision makers (in our case, a records manager). The suppliers who used off-the-shelf products made use of the technology-assisted review features provided by the products to aid understanding. The other suppliers, who built their applications from scratch, decided to focus on building a platform to train their models (the ML part) but could not provide better GUI visualisations due to the lack of time. Adatis however, built a prototype demonstrating how a Records Manager could review and correct labels within their Azure pipeline.

The abilities of the products to create models and facilitate interpretation by decision makers are summarised in *Table 3* below.

Model Training and Tuning		ADLIB Elevate by Adlib	AWS (used by Kainos)	Azure (used by Adatis)	InSight by Iron Mountain	Records 365 by Record Point	TNA Prototype
Inputs	Metadata	N	N	Y	N	Y	Y
	Document contents	Y	Y	Y	Y	Y	Y
Outputs	Predict Selection	Y	Y	Y	Y	Y	Y
	Predict retention schedule	Y	Y	N	Y	Y	Y
Enable customisation of models	Customise models (e.g. choose ML algorithm)	N	N	Y	N	N	Y
	Specify which features to use	N	N	Y	N	N	Y
	Inject our own models and use them (e.g. via an open API)	N	N	N	N	N	N
Share / reuse / export	Versioning of models	N	Y	Y	N	Y	Y
	Share models (reuse deployed models / pipelines)	Y	Y	Y	N	N	Y
	Export models (export hyper parameters for archiving or later re-import)	Y	N	Y	N	Partial	Y
Interpretable / explainable results (e.g. show in decision tree why we made that prediction)		N	N	Partial	N	N	Partial
Support "technology assisted review"	Can the records manager override a tool's decision?	Y	N	Y	Y	Y	N
	Can the tool retrain its hyper-parameters from user feedback?	Y	N	N	Y	Y	N
	Deal with outliers (Partial: spot them for user correction, Yes: learn from user feedback & apply to all similar outliers)	Partial	N	N	N	N	N
	Provides aggregation / statistics / reporting to aid understanding of results	Y	Partial	N	N	Y	Y

Table 3: Ability of the products to create models and support decision makers

6. Graphical User Interface (GUI)

The two off-the-shelf products supported the full range of features in our requirements. The prototype by Adatis included the ability to confirm or correct the labels as well as search and filter on specific features. With more time this functionality could have been included in the Kainos pipeline. It was clear that the two bespoke systems which used cloud components could be designed to meet any requirements in terms of usability.

Table 4 (below) summarises the availability of features for user access control and progress tracking.

Deployment / GUI		**ADLIB Elevate by Adlib**	**AWS (used by Kainos)**	**Azure (used by Adatis)**	**InSight by Iron Mountain**	**Records 365 by Record Point**	**TNA Prototype**
User Access Control		**Y**	**Partial**	**Partial**	**Y**	**Y**	**N**
Users can track progress (through the whole process, through their review of classified documents)		**Y**	**N**	**N**	**N**	**Y**	**N**
Allow collaboration between users	**Continue work started by another user**	**Y**	**N**	**Y**	**Y**	**Y**	**N**
	Multiple people can work on the data	**Y**	**N**	**Partial**	**Y**	**Y**	**N**

Table 4: Ability of the products to provide assistance using GUI features

Some products offer the facility for visual interaction with the data. For example, Microsoft Azure provides a tool for labelling images; Records365 provides Power BI templates to help users explore the data with visualisation products: similar solutions could be implemented for any of the approaches we tested.

Appendix D. Links

1. Suppliers reports
    - Adlib Elevate by Adlib (in conjunction with Deloitte)
    - Amazon Web Services (AWS) used by Kainos
    - InSight by Iron Mountain
    - Records365 by RecordPoint
    - Azure used by Adatis
2. Report of First Phase Market Research
3. TNA Benchmarking Tool
4. TNA Guidance for Digital Records Transfer

5. Report - Rules based eDiscovery products for Technology Assisted Review
6. AI for Selection Evaluation Template

Glossary

Terms used in this report

Artificial Intelligence: General term for systems that can make intelligent automated decisions.

Machine Learning (ML): Specific technical approaches or functions used to implement AI systems.

Classification: Predictive categorisation of items based on pattern identification.

Model: A specific mathematical ML approach used to achieve the desired result of classification (i.e. classifying records as 'selected' or 'not selected' for permanent preservation).

Classifier: A tool used for classification of records into categories (in this case, 'selected' and 'not selected'). A classifier may be developed using one or more models to see which of the modelling techniques is best suited to the data under consideration.

Evaluation: The process of validating the classification model and understanding its performance.

Synthetic Data: 'Dummy' data with similar characteristics to training data for use when real data is not available or cannot be used.

MLaaS: Machine learning as a service