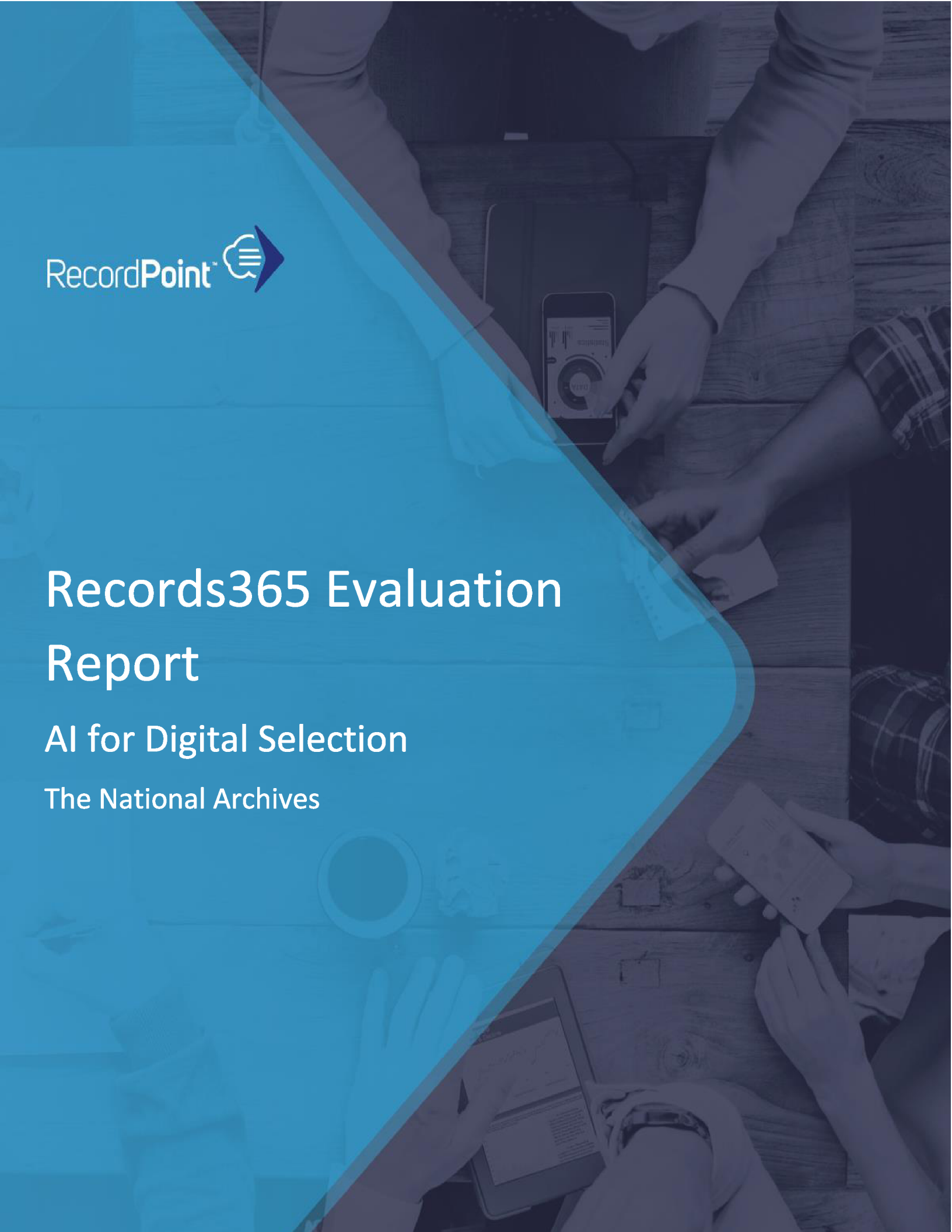


RecordPoint

Records365 Evaluation Report

AI for Digital Selection

The National Archives

Notice

Contact for all enquiries

Anthony Woodward

CTO and Co-Founder, RecordPoint

anthony.woodward@recordpoint.com

+1 425 245 6235

CONTENTS

Table of Figures

1. Executive Summary

Purpose

In order to help TNA to better understand solutions that can increase TNA's depth of capability in leveraging artificial intelligence (AI) tools to appraise and select data for permanent preservation, RecordPoint has been invited to be part of the AI for Digital Selection project with its Records365 platform.

This report describes the current state of the AI market as well as how RecordPoint's Records365 Record Management platform provides advanced governance capabilities that leverage machine learning (ML) to appraise the value and risk of high volumes of information with turnkey compliance.

Approach

Records365 is a cloud-based software-as-a-service platform that can connect to multiple content sources to enable organisations to apply federated governance across all their information, regardless of where it lives.

To help customers addressing the challenges they are facing as part of their digital transformation journey, RecordPoint is committed to bring customers continuous innovation by delivering solutions that are:

- Easy to use, so they can apply governance regardless of where the content lives,
- Intelligent, so that they provide the level of automation required to manage the massive amounts of structured, unstructured and semi-structured content that organisations have,
- Trusted, so that organisations know that their information is securely managed and have the required policy and controls in place.

As part of the project, TNA has provided RecordPoint with samples of labelled and unlabelled data that we have used to demonstrate the Records365's machine learning capabilities and increase TNA's understanding on how to leverage AI using the following approach:

Load Retention Schedule: Using the retention schedules spreadsheet provided, we loaded each disposal class and retention schedule into the Records365 global File Plan.

Create Rules for Labelled Dataset: In order to automatically assign a disposal class and retention schedule in Records365 to the labelled data, a set of declarative rules were created in the Records365 rules tree that mapped each document to a specific disposal class using its metadata.

Import Labelled Dataset: Since the data was provided on a hard drive, for the scope of this project we have decided to load the labelled dataset from a windows file share using the Records365 FileConnect connector. Once the connector was enabled and the documents were added to the file share, FileConnect looked for redundant/obsolete/trivial (ROT) documents. The FileConnect ROTBot performed deduplication, enriched the document with additional metadata and automatically submitted them to Records365. Once processed by the Intelligent Processing Engine, each document was classified according to the rules previously created.

Train Model on Labelled Dataset: The Records365 Classification Intelligence capabilities have been designed to be used by compliance and record management teams without requiring the involvement of a data scientist. The model was trained by simply selecting the different disposal classes on the file plan with enough data samples. The rest of the processing was automatically handled by Records365 without requiring user intervention.

Apply ML to Unlabelled Dataset: Once the model was trained, we proceeded at submitting to Records365 the unlabelled dataset using the same Windows file share and FileConnect Connector previously mentioned. Once again as the content was added to the file share, the FileConnect ROTBot performed deduplication and named entity extraction to enrich the context to the document to be used for e-Discovery purposes. Once received by Records365 the Intelligent Policy Engine applied the Machine Learning model to each of the unlabelled documents to suggest a relevant category. After that, the Records Management team is still fully in control on making a final decision and can review the suggestions made by accepting or correcting it. This feedback loop is then used to improve the model over time.

Future Research and Development

In addition to the Intelligent capabilities available in Records365 today, RecordPoint is making big additional investments in the AI space. We understand that organisations still struggle to control their information and make meaningful business decisions due to the out-of-control number of content sources that they are dealing with on a day-to-day basis which contain structured, semi-structured and unstructured content.

Some of the capabilities that customers can expect to see in Records365 in the future are:

- Context Enrichment
- Multi-Model Appraisal
- Unsupervised Learning
- Searchable Knowledge Graph
- Multi-Dimensional Appraisal
- Language Models
- AI-driven Content Analytics
- Intelligent Connectors
- AI based Risk & Value Scoring

We believe that machine learning capabilities will be at the core of helping organisations to reduce their current risk and to make better decisions faster and to do so, those capabilities need to be explainable and easy to use by regular users.

2. Key Observations Summary

As the outcome of the experiments undertaken during this project the following key results and findings were determined:

Permanent Preservation: Based on the training set provided by TNA the Records365 Classification Intelligence has identified a total of 3180 records as candidates for permanent preservation by TNA.

Duplicates: Out of the unlabeled dataset sample Records365 identified a total 4843 documents with at least one duplicate. 2252 out of that were unique documents and 2591 were duplicate versions that should be treated as redundant and disposed in the content source.

Model Details: As part of this project Records365 has selected a model that demonstrates an overall training accuracy of 74.5%, and an overall test accuracy of 71.8%. These results are related to the individual performance of the different disposal classes/categories as well as the number of samples available for training in each category. Some poorer performing categories like 25 (Research and Academic Liaison Administration) and 6 (Business Planning & Performance - HIGH Corporate LEVEL) had fewer samples on which to train.

File Analysis: As part of the file analysis performed a sample dataset* by the FileConnect ROTBot we have extracted the following observations:

- The most mentioned organisations across the entire data set were "The National Archives", "TNA", "MessageLabs" and "FOI".
- The most mentioned geopolitical entities across the entire data set were "Richmond", "UK", "London", "England" and "Wales".

Redacted under FOI exemption 40(2)

- 70% of the content was smaller than 100KB and 20% was between 100KB and 500KB. Just a small percentage of the content was bigger than 10MB.
- Most of the content hasn't been modified recently. 49% of the content was last modified 5 - 9 years ago, 18% of the content was last modified 9 – 15 years ago and 29% more than 15 years ago.

The results and findings of this project are further explained in detail later in this report in the section Results & Finding.

* As part of this project RecordPoint choose to just perform file analysis on a smaller set of the total data.

3. State of the Artificial Intelligence Market in the Information Management/Compliance space

The use of artificial intelligence in the information management space has lagged behind the state of the art for better part of two decades [1] , and the market is only now starting to catch up.

Problem Space

Most of the born-digital content within organisations is now unstructured or semi-structured with little available metadata or other contextual information which most current commercial software systems rely upon for identifying the value and risk of an individual record.

There is a long-held desire to conceptualise records as a continuum [2] . A record is not a single document in isolation; it is a stream of data that may be constantly changing through time and which may be embedded in a dynamic set of contexts (personal, organisational, cultural, security related, etc.). Recognising and being able to utilise this is essential to being able to fully understand the meaning and value of a record. Innovations in data science look set to start delivering upon these goals of including all the relevant context when assessing a record.

The most prevalent records and information management solutions currently available employ technologies that date back to the 1990s. These systems are extremely manual - users need to input documents, identify metadata and classification by hand. Platforms that automatically harvest and classify records and their metadata have been rare.

Current Solutions

Current automated records classification systems, where they exist, have mostly been based on the concept of an expert system [3]. In an expert system, a domain expert (e.g. a records manager) needs to hand-author a set of rules that can classify the records by examining their metadata. This is an expensive process that is difficult to scale. In addition to that, these expert systems are very brittle. If the data varies slightly from the formats expected, or there is a gap in the rules coverage, these systems will fail to classify the records.

Compounding this problem is the fact that records are now federated across an increasing variety of content sources, such as email or social media. The increasing variety of content sources makes it difficult, if not impossible, to impose a metadata schema or information architecture upon the entire records corpus of an organisation.

Rise of AI and Data Science

Only in the last few years have we started to see AI coming into use in the records management and compliance space with companies like RecordPoint adding natural language processing (NLP) and text-based classification to their appraisal capabilities. These techniques allow classification of records based on a statistical analysis of their text, rather than by using declarative rules that consider only record

metadata. A University of Glasgow researcher, Graham McDonald, has been looking at using NLP techniques to appraise records for sensitivity [4], for use in serving freedom of information (FOI) requests and declassifying secure documents. Only few vendors operating specifically in the information management space have strong or thoroughly integrated AI-based offerings in their product suites. Adding AI capabilities to information management software is just the beginning of the applications of AI to this space in the future.

Data science and artificial intelligence are broad domains that are currently undergoing an unprecedented spike in technological growth, so it is difficult to predict what new applications will be possible even next year. Barriers remain in terms of operationalising AI applications for a scalable, production, user-facing system, but the infrastructure and best practices employed are quickly maturing along with the science that backs them.

We will soon start seeing metadata enrichment capabilities becoming mainstream. This includes new techniques to harvest intrinsic metadata from record documents as well as the adoption of AI-based techniques like named entity recognition (NER) and sentiment analysis to seek new information from document text. In addition, a wider use of the current text-classification machine learning techniques can be employed to add custom dimensions to records based on statistical properties of the text.

We should also start to see clustering and other unsupervised machine learning processes becoming available to assist in the initial assessment of large bodies of unlabelled and unstructured records.

Systems like Records365 now have enough data available that we can start to graph the relationships between record documents, users, and content sources. This lets us establish the context and dynamic nature of records, the records continuum, as opposed to the old view of a record as a static document.

Perhaps the single biggest technological innovation to emerge from the NLP space in recent years is the pre-trained language model [5]. These models have enabled a raft of new applications that work on text. There are neural network-based technologies than in many ways supersede the statistical techniques that are currently finding their way into the marketplace (see Modelling below for more detail).

Language models are pre-trained on massive corpuses of text (for example, all of Wikipedia) and then fine-tuned to more specific applications. This transfer-learning approach allows us to create more powerful classifiers with substantially less data. They also offer many powerful new applications, including:

- de-corrupting damaged data;
- text summarisation;
- question answering; and
- text generation.

These technologies will be invaluable in understanding the content of records under management and leveraging them to provide improved compliance and a better experience for users.

4. Records365 Overview

Records365 is a SaaS (Software-as-a-Service) federated data management platform that unlocks compliance to all content sources by managing content no matter where it lives.

Records365 extends the native document management and collaboration capabilities of major enterprise content sources by providing advanced governance required to meet the complex compliance and regulatory requirements of modern organisations. It does that by providing a depth of capabilities such as:

- **Electronic and Physical Records** managed together using a single and centralised experience
- **Complex Retention Workflows** that ensure fully defensible disposal with capabilities such as multi-stage disposition and event-based triggers
- **Intelligent Policy Engine** that can appraise high volumes of structured, semi-structured and unstructured information either through rules using a declarative approach or leveraging the power of Machine Learning
- **Long Term Preservation** so content can be either kept permanently within the organisation or transferred to the relevant archives
- **Federated Search** that enables organisations to search content across many content sources directly within the Records365 interface
- **Legal Holds** that can be applied to content that might be under legal litigation ensuring it is kept accordingly until the legal hold is lifted
- **Full Version History and Audit Trails** of all records managed by Records365 for full compliance and defensible disposal
- **Security** that ensures only the right people are authorised to access certain records and perform certain actions
- **Content Analytics (ROTBot)** that identifies the value and risk of content living in the organisation enabling better business decisions faster
- **Reporting** to enable organisations to report and collaborate on various aspects of the records under management by Records365

In addition to this, Records365 provides a range of out-of-the-box connectors that customers can leverage, with new connectors being released on a regular basis. Some examples are:

- SharePoint Online
- SharePoint On-Premises
- OneDrive for Business
- File Shares
- Dropbox
- Box
- Exchange Online
- Teams

For more bespoke content sources, organisations also have the option to create custom connectors by using the open-source Records365 SDK.

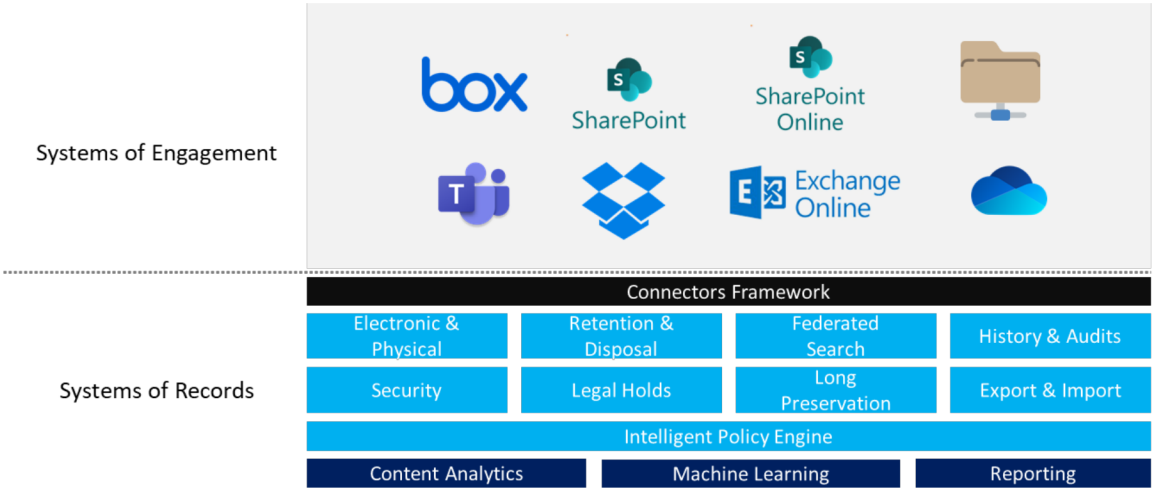


Figure 1. Records365 Context Landscape

Records365 is designed to seamlessly classify, track and retain records to ensure they are appropriately managed in accordance with corporate, regulatory and legislative rules while being completely transparent and not impacting the productivity of the end-users.

Records365 is certified to meet the VERS standard for addressing the long-term preservation of records and built to meet the requirements of ISO 15489 and ISO 16175, as well as other industry standards, such as IS40, Sarbanes-Oxley, 21 CFR Part 11, HIPAA, NAA, MoReq 2010 and ANSI.

As an Easy, Intelligent and Trusted solution Records365 provides the following benefits:

- **Adoption and Usability:** Records365 is designed to provide the level of automation required to manage the massive amounts of structured, unstructured and semi-structured content that organisations have nowadays with a seamless interface reducing the need to train end-users on specific governance tasks and allows them to focus on the important part - the content.
- **Centralised Control:** Impose corporate retention and storage policies and reuse information to enhance existing business processes, reduce management and governance overhead, and boost productivity, all from a central, online platform that provides a holistic view of the entire solution.
- **Transparent Access:** The way that content is submitted and managed as a record is completely automated and seamless to end-users, providing transparent organisational control over content without impacting business productivity.
- **In-Place Management**: Records365 is designed to manage all documents in-place, while providing governance capabilities through the Records365 user interface. No content is duplicated, ensuring the users benefit from all the native capabilities of the different content sources.

With Records365, organisations have a platform that can meet the immediate and longer-term compliance challenges posed to them regularly and provide the performance, scalability and flexibility to evolve alongside the information management practices as systems and processes change over time.

5. Narrative around areas of evaluation template

Data Collection

Records365 is designed to ingest and manage content from any content source. We have a range of out-of-the-box connectors which are designed to easily allow administrators to enable governance on any content source with minimal effort.

Our currently published connectors include SharePoint Online, SharePoint On-Premises, OneDrive for Business, Exchange, Box, Dropbox, Teams, as well as FileConnect, which enables organisations to manage content on file shares. New connectors are released on a regular basis.

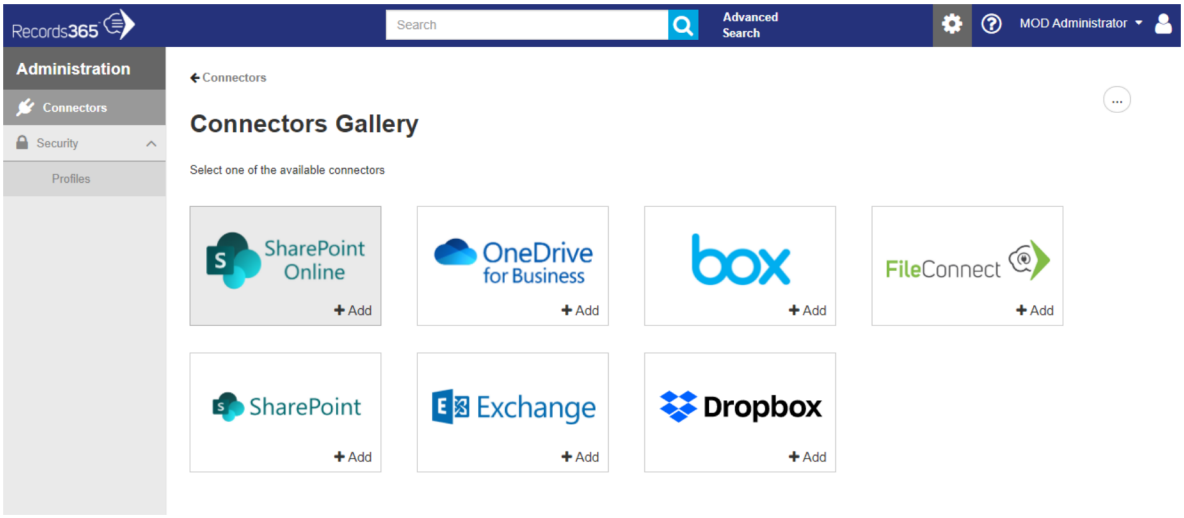


Figure 2. Adding a New Content Source Connector

We also publish an open-source connector SDK, available on GitHub, which allows organisations to create bespoke connectors for content sources that we don't yet support, or any content sources that are custom to an organisation.

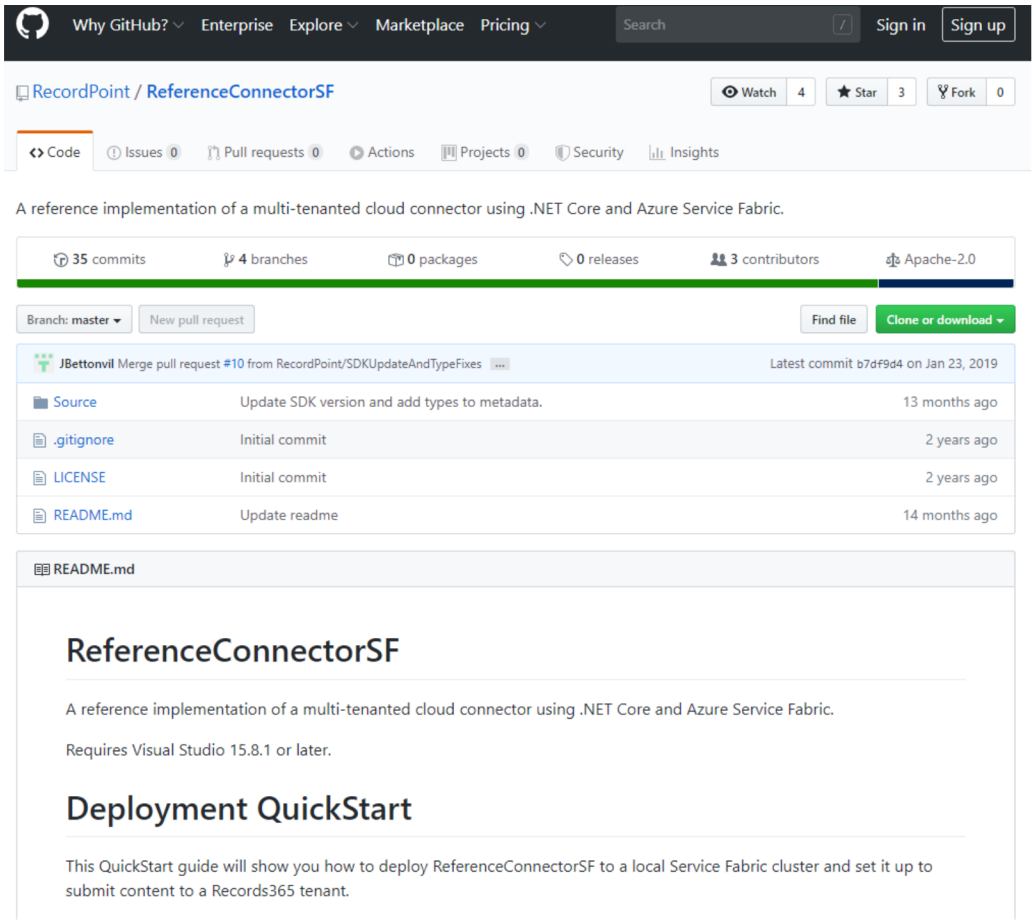


Figure 3. RecordPoint Connector SDK Example

Content is managed in place and ingested via our connector framework whenever it is created or updated on the content source. We capture all the metadata, as well as the binary and the audit trail of each document, so that we can manage its full lifecycle as a record.

Metadata captured includes basic information such as the file name, location, date the document was published, date the document was filed as a record, the size of the file, as well as any custom organisational metadata. Prior to submission the ROTBot enriches the metadata in several ways. It adds a hash of the document binary that is used for duplicate detection and to ensure the integrity of the record, and it also performs named entity recognition to find people, organisations, and geopolitical entities in the text of the documents that can be leveraged for e-Discovery purposes.

Once ingested, our Intelligent Policy Engine classifies the record against a disposal class. Each disposal class has an associated retention schedule which is also applied to the record.

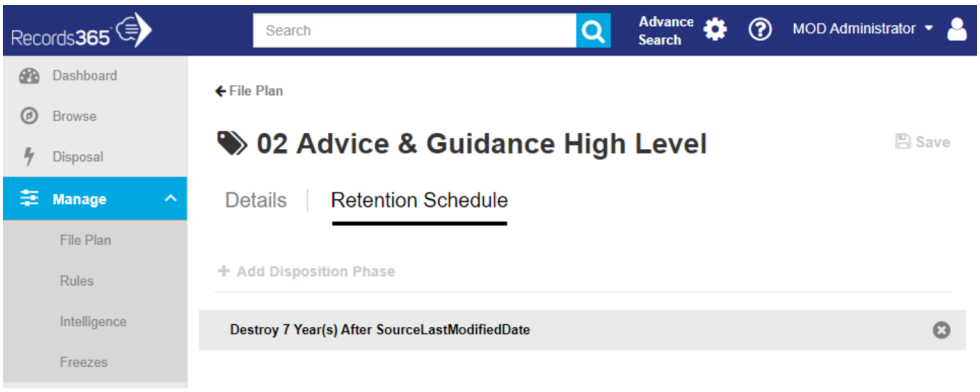


Figure 4. A Disposal Class with Its Associated Retention Schedule

The Intelligent Policy Engine utilises both declarative metadata-based classification, as well as machine learning classification to assign disposal schedules. Rules use a declarative approach and can be created to classify records based on any metadata field, whether captured from the content source or added by Records365 using metadata enrichment.

Incoming records first pass through the rules engine. If the record matches an appropriate rule it is classified using that rule. If a record does not match any rules, it will be classified by a statistical model based on the record's actual text.

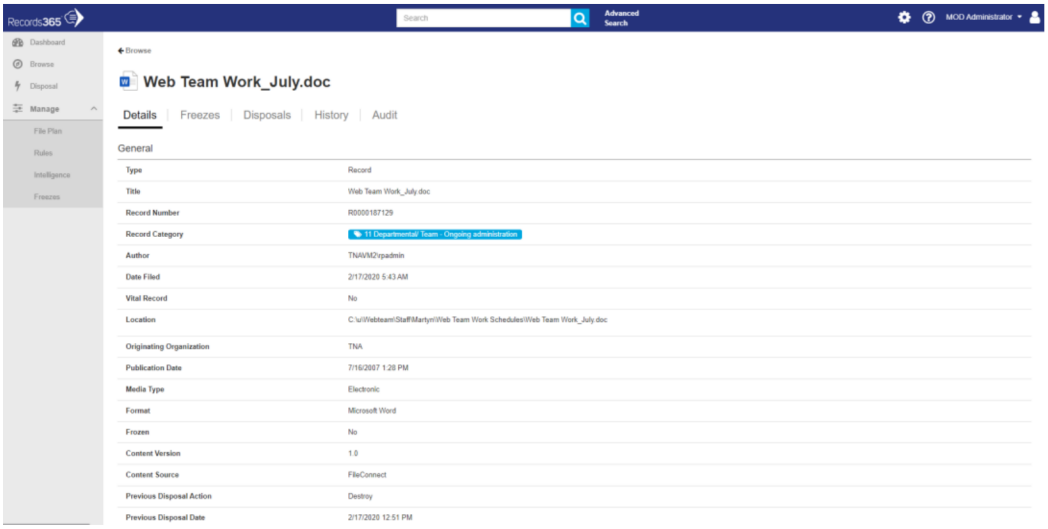


Figure 5. Details of a Record Under Management by Records365

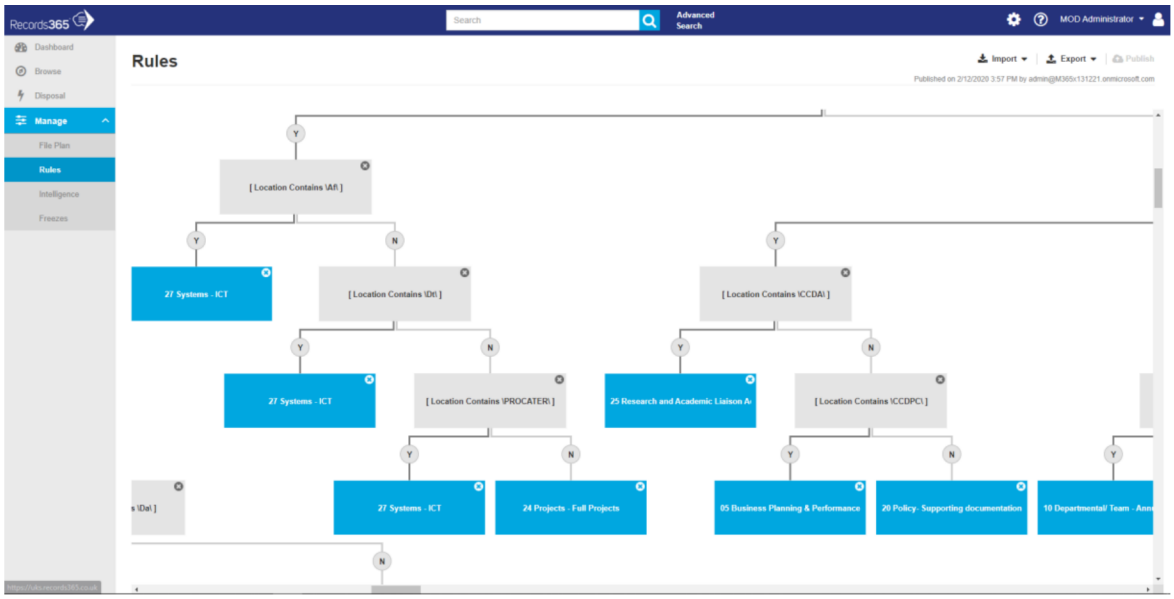


Figure 6. Rules Tree Used to Classify Records Declaratively

Pre-Processing

The pre-processing of record documents has several phases. First, we crack the documents to extract all the text from them. This process employs a variety of libraries and custom code to maximize our ability to find useful text in a document.

Records365 itself can manage any document regardless of format. For machine learning purposes, we extract the text from a subset of document formats, including using Optical Character Recognition (OCR) to extract the text from image-based formats, including image-based PDFs. The supported file types include several email-based formats, and we extract the text from attachments as well as the email message itself. See Appendix 1 – File Formats Supported for a list of file formats that we can utilise for machine learning.

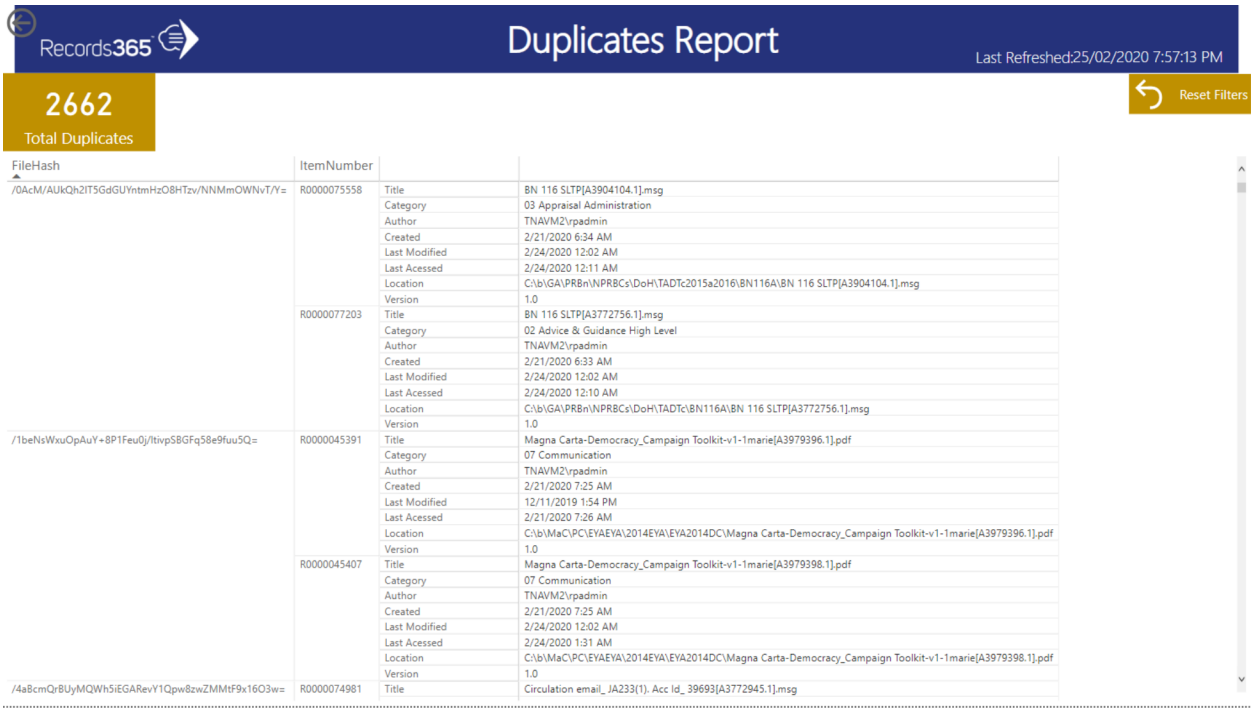
The current supported formats are based on the file formats that are most commonly in use by our customers. We always seek to expand our supported formats based on feedback from customers, and video and audio files are on our roadmap.

Once we have cracked the documents, we remove any duplicates with the same extracted text, then we transform the text to be all lowercase. We then locate bigrams (two-word combinations) in the text and select the most frequently occurring of these. Next, we tokenize the text, and remove all the punctuation and stop-words ('and', of', 'the' and other articles). We then drop a percentage of the very most common words across the set.

We are now able to determine a vocabulary for the document corpus - a set of the words we are going to use to compare all the text. Depending on the number and size of documents in the experiment, there may be thousands or tens of thousands of terms in the vocabulary.

After this we count the number of occurrences of every term in the vocabulary for every document, providing us with a term frequency vector (TF). We then divide these word counts by the number of times each term appears in all documents (inverse document frequency, or IDF). This gives us a scaled numerical value for each term, for each document, which gives greater importance to words that appear less frequently throughout the document corpus. These are called TF-IDF vectors. We will use these TF-IDF vectors as signals to train a machine learning model to classify text.

As well as removing duplicate documents when creating the vocabulary and vectors for model training, we also enable duplicate detection across all our content sources by calculating the hash of each record's binary as we ingest it. This hash is stored as metadata on the record, and organisations can report on the duplicates across their whole corpus of documents. See Figure 7 for an example duplicates report.



FileHash	ItemNumber		
/0AcM/AUkQh2IT5GdGUYntmHzO8HTzv/NNMmOWNvT/Y=	R0000075558	Title	BN 116 SLTP[A3904104.1].msg
		Category	03 Appraisal Administration
		Author	TNAVM2\rpadmin
		Created	2/21/2020 6:34 AM
		Last Modified	2/24/2020 12:02 AM
		Last Acessed	2/24/2020 12:11 AM
		Location	C:\b\GA\PRBn\NPRBCs\DoH\TADTc2015a2016\BN116A\BN 116 SLTP[A3904104.1].msg
		Version	1.0
	R0000077203	Title	BN 116 SLTP[A3772756.1].msg
		Category	02 Advice & Guidance High Level
		Author	TNAVM2\rpadmin
		Created	2/21/2020 6:33 AM
		Last Modified	2/24/2020 12:02 AM
		Last Acessed	2/24/2020 12:10 AM
		Location	C:\b\GA\PRBn\NPRBCs\DoH\TADTc\BN116A\BN 116 SLTP[A3772756.1].msg
		Version	1.0
/1beNsWxuOpAuY+8P1Feu0j/ItivpSBGFq58e9fuu5Q=	R0000045391	Title	Magna Carta-Democracy_Campaign Toolkit-v1-1marie[A3979396.1].pdf
		Category	07 Communication
		Author	TNAVM2\rpadmin
		Created	2/21/2020 7:25 AM
		Last Modified	12/11/2019 1:54 PM
		Last Acessed	2/21/2020 7:26 AM
		Location	C:\b\MaC\PC\EYAEYA\2014EYA\EYA2014DC\Magna Carta-Democracy_Campaign Toolkit-v1-1marie[A3979396.1].pdf
		Version	1.0
	R0000045407	Title	Magna Carta-Democracy_Campaign Toolkit-v1-1marie[A3979398.1].pdf
		Category	07 Communication
		Author	TNAVM2\rpadmin
		Created	2/21/2020 7:25 AM
		Last Modified	2/24/2020 12:02 AM
		Last Acessed	2/24/2020 1:31 AM
		Location	C:\b\MaC\PC\EYAEYA\2014EYA\EYA2014DC\Magna Carta-Democracy_Campaign Toolkit-v1-1marie[A3979398.1].pdf
		Version	1.0
/4aBcmQrBUyMQWh5iEGARevY1Qpw8zwZMMtF9x16O3w=	R0000074981	Title	Circulation email_ JA233(1). Acc Id_ 39693[A3772945.1].msg

Figure 7. Example Duplicates Report

Modelling

Records365 manages the detail of all modelling and training for users of our machine learning classification model. Rather than requiring users to have data science training, we have abstracted away both the operations side of this process and the experiment design. All that users need to do is select the disposal classes from the File plan that they want the model to cover. Our systems will then train, select and deploy a model without further user intervention.

Figure 8 shows the File Plan, with all the disposal classes selected. When two or more disposal classes are selected a user can clicks "Train Model", which will start the training process. The status of the training can then be monitored using our reporting tools (see more details around reporting in Tracking Progress).

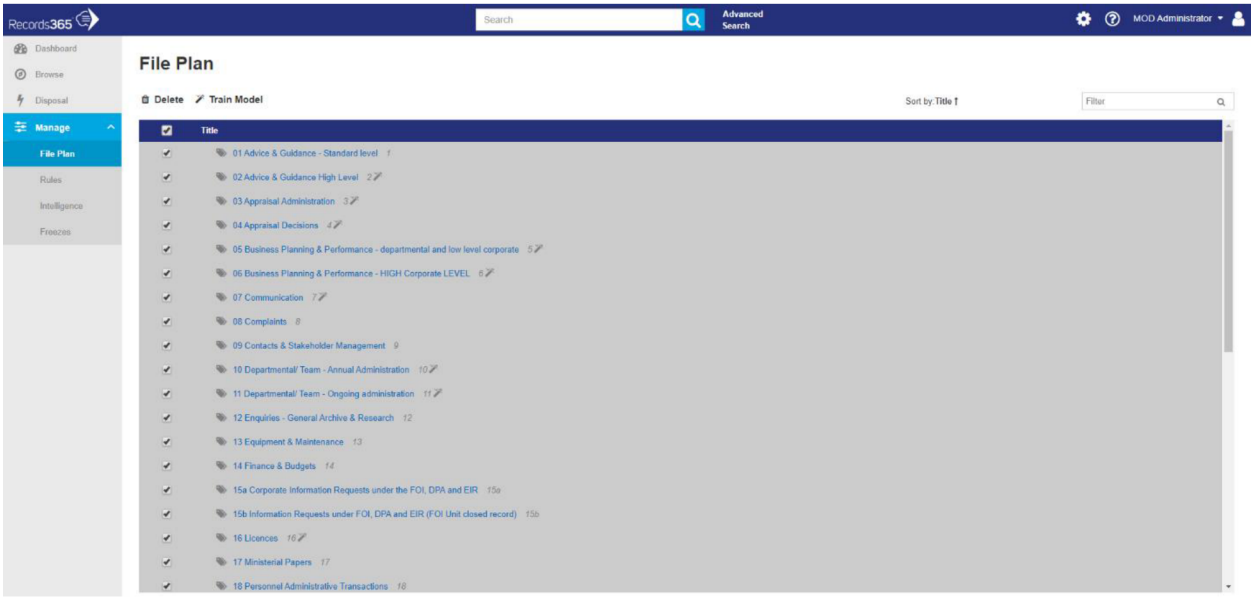


Figure 8. Disposal Classes selected for Model Training

Training Set Selection

Rather than attempting to model all the available data, Records365's models are adaptive and are designed to continuously improve by frequently retraining. We only use a subset of the data in order to keep the training set at a manageable size for frequent retraining. We have found that increasing the amount of data trained upon does not improve the accuracy of the model but serves to increase the cost and time of model training, making frequent training infeasible.

When Records365 trains a machine learning model it selects a subset of data from the body of classified records. The selection algorithm favours more recent data and has a staleness threshold to ensure that the model can adapt to changing data. It will also try to balance the number of records selected across all categories if there is enough data. This balance will help to create a model that works well across all the categories, not just a few. Real data is messy, but we employ several heuristic measures to train our models from the best available data.

If a records manager rejects the suggested category for a new record, that record is deemed to be 'interesting' and is added to the training set for the next experiment run. This creates a feedback loop, iteratively improving the model, and including the records manager as an authoritative signal.

Model Training

Once the training set has been selected, we pass it to a training experiment, which will create a machine learning model that the platform can leverage to predict the categories of incoming records.

First, we pre-process all the data and render it down to TF-IDF vectors, as discussed earlier. Then we split 20% off the data to use for a validation set. The remaining 80% is used to train and tune a variety of models using the K-Nearest Neighbours, Random Forest and Support Vector Machine algorithms. We select the best model based on accuracy (the proportion of correct classifications) across all categories. We employ K-folds cross-validation (with ten folds) to estimate the accuracy, and to reduce the likelihood of under- or over-fitting the model.

Once we have selected the best model, we train it again using the full 80% test set. We then determine the model's overall test accuracy by using it to classify the records in the 20% validation set. Finally, we train the model using 100% of the available data and publish this. This final model is then used to classify incoming records until a new model supersedes it.

Model Outputs

Once the model is trained, any records that are ingested by Records365 which do not get classified using the rules engine are classified by machine learning. Records that aren't classified through the rule's engine are considered uncategorised. They therefore don't have a classification or associated retention schedule.

These uncategorised records are sent through the machine learning pipeline and are assigned a suggested category. This suggested category can then be accepted by a records manager, or rescheduled to another classification, providing different signals to the model in the future.

Records that have been assigned a suggested category with a permanent preservation outcome are therefore selected for permanent preservation.

Interpretation

As Records365 approaches machine learning from the records and document management perspective, each model has been trained against specific disposal classes. Each model has an identifier, allowing the version to be tracked, and the disposal classes used to train it are also stored with the model.

By correlating the records that have been classified with the model that was used to classify them, and further the disposal classes that were used to create the model, it is possible to obtain an understanding of why a specific disposal class, and its associated retention schedule, have been applied to a record.

Technology Assisted Review

Records365, including our machine learning capabilities, is designed to be used by records managers. In a risk averse industry, it is important to ensure that users have control over the outcomes of a system.

Our intelligence dashboard, shown in Figure 9, allows users to review the suggested categories that have been supplied by machine learning, and action these either by accepting the suggested category, or by rescheduling to a different disposal class. Only after these records have been reviewed are the disposal classes assigned to the record and therefore their retention schedules started.

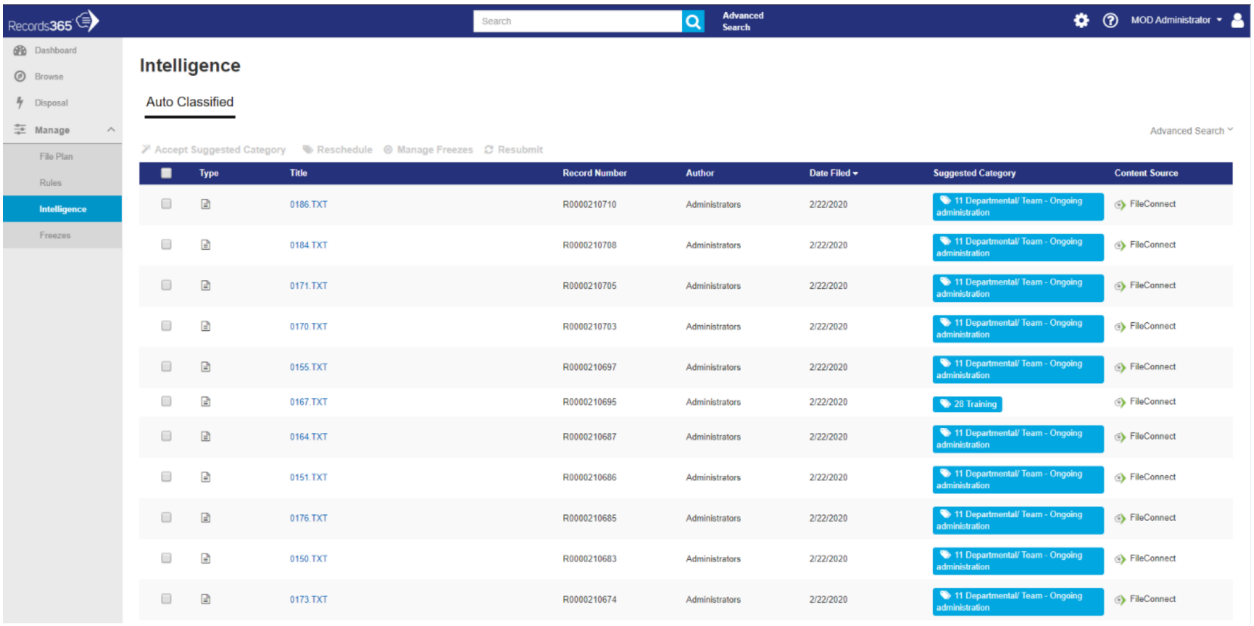


Figure 9. Intelligence Dashboard

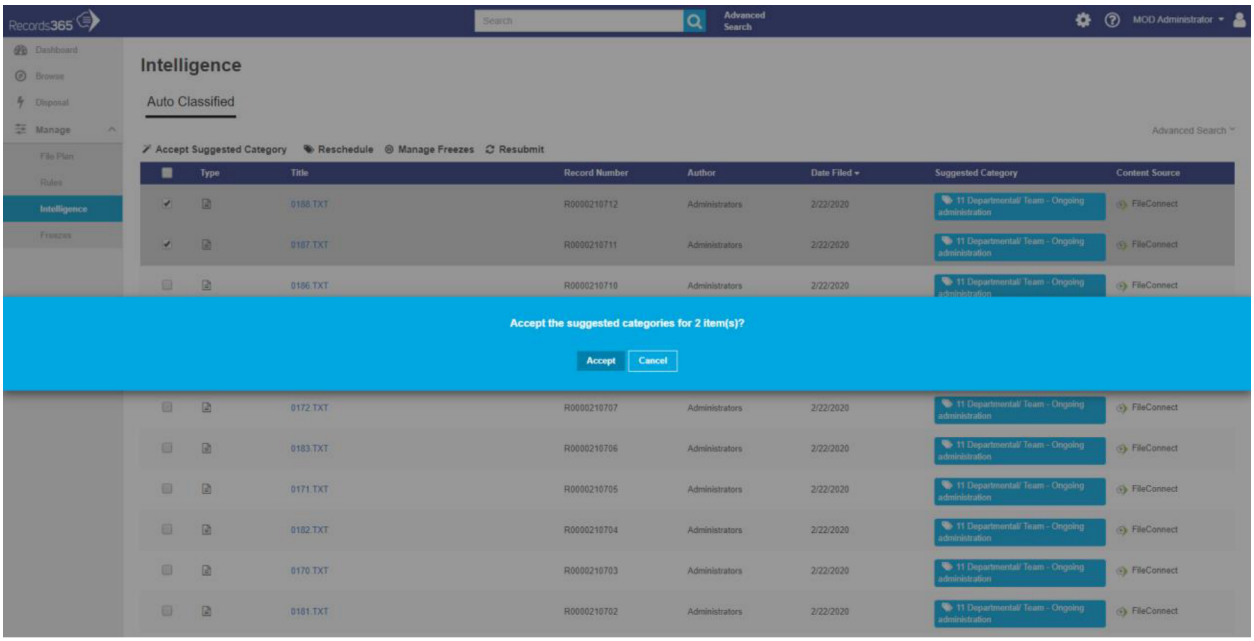


Figure 10. Accepting the Suggested Category

When a records manager reschedules a record to a different disposal class, we consider this action to be "interesting", so we ensure that when the model is retrained the rescheduled record is included in the training set, adding to the feedback loop of model training.

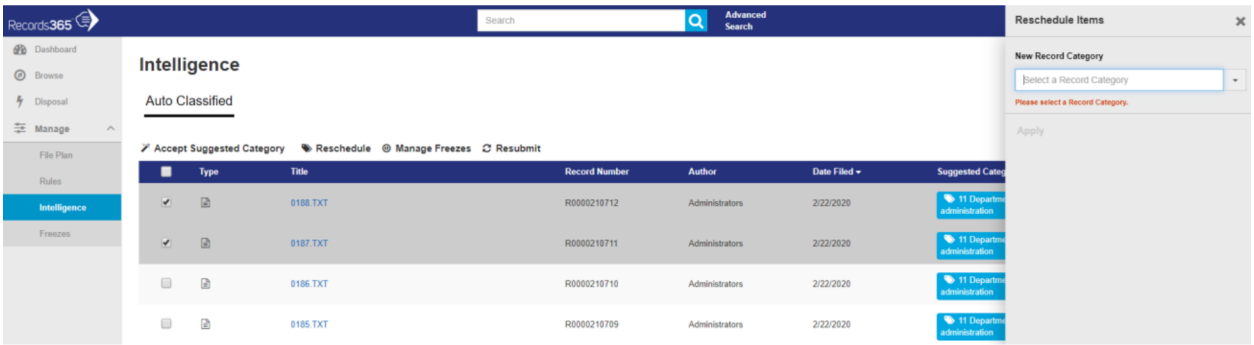


Figure 11. Rescheduling Records to correct Disposal Class

Deployment/GUI

User Access Control

Records365 leverages enterprise standard Azure Active Directory to provide identity and authentication within our service. This allows organisations to use their existing corporate identities and rely upon the security and compliance offered from Azure Active Directory.

Within Records365, users and groups from within Azure Active Directory can be assigned fine-grained access to different aspects and actions within Records365 ensure that only the right people can access certain content and perform certain actions.

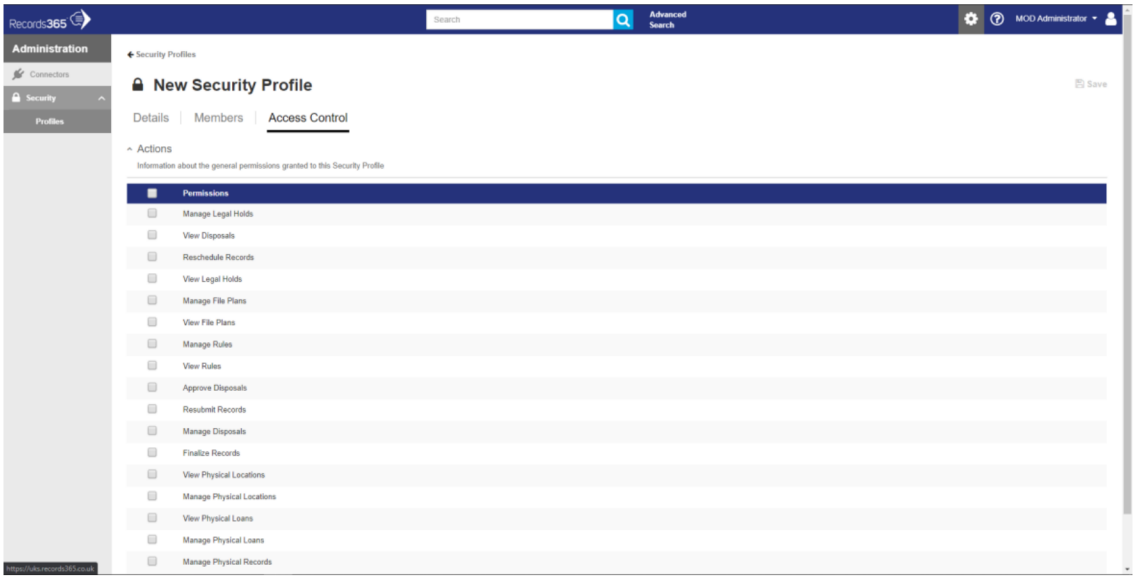


Figure 12. Security Profiles for User Access Control

Tracking Progress

Once the model training has been started, progress can be monitored using Records365's reporting tools. On Figure 13 you can see an example report, showing all the models that have been trained in the system, including their status and their testing accuracy. These dashboards are available in the administration area of Records365 as well as through the PowerBI reports shown in figures 13 and 14.

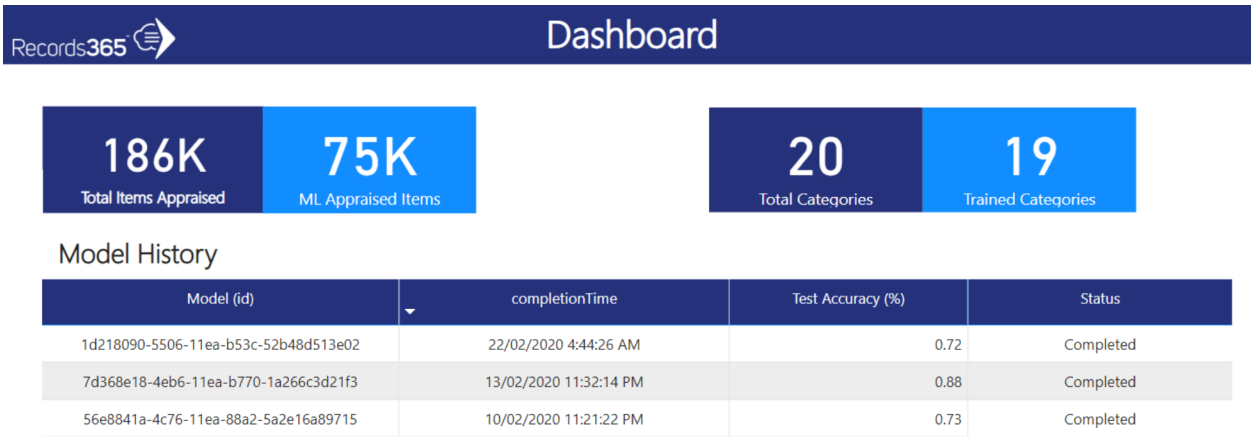


Figure 13. Machine Learning Dashboard

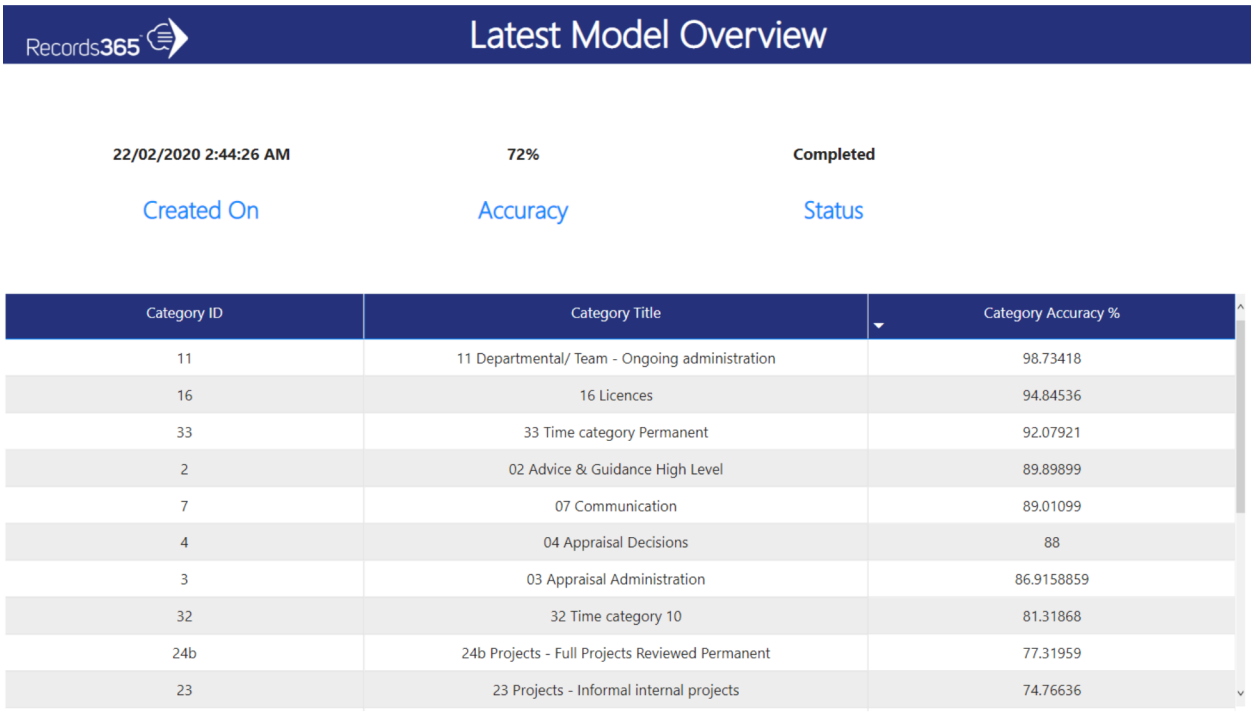


Figure 14. Machine Learning Model Overview

Collaboration

Though training set selection and model training are handled by Records365, with no user input, reviewing the machine learning classified records is handled by users, providing control to the compliance and records management teams to make a final assessment. This review can be done collaboratively by any Records365 user. Based on existing customers using this capability, it is common for records from different departments to be handled by different users.

Skill Levels

Records365 with Intelligent Classification is designed as a fully featured records management solution, making managing records across all content sources simple and easy. We have infused machine learning into our product so that organisations can benefit from machine learning without any expert users required.

Once a user has selected the disposal classes and chosen to train a model, there is no other user input necessary. Records365 itself will select the training data set, train and deploy a model, and then for incoming records that do not match a rule, the platform will use the model to automatically classify the records.

6. Results and Findings

Evaluation Spreadsheet

The evaluation spreadsheet can be found in the attached document "Five Tools Comparison Reporting – RecordPoint.xlxs"

Unlabelled Sample

The detail of all the unlabelled documents that have been appraised by machine learning can be found in the attached document "TNA – Records appraised by ML.csv".

As you can see in Figure 13, not all of the unlabelled items could be appraised by the machine learning model. This is because we only request a ML appraisal for records that have not been classified by the metadata rules. Also, records that do not have a supported document type cannot be appraised by the ML model. Finally, some records might have been skipped because they were corrupted, encrypted, or simply too large (the Records365 platform supports documents up to 500MB in size).

Duplicates

The details of the duplicates found during the project can be found in the attached document "TNA - Duplicate Report.csv". This document shows each record which has a duplicate as assessed by hashing the binary. There are 4843 records found that have at least one duplicate. 2252 out of that were unique documents and 2591 were duplicate versions. From some cursory analysis, many duplicates have very similar filenames, some with "copy" or "draft" appended.

It can also be seen that though the majority of duplicates have the same retention schedule, there are a significant number that have different retention schedules. Having the same data used in different categories will have an effect on model training. During this project we chose to import and report on all duplicates rather than choosing which duplicate to remove but removing duplicates would lead to a cleaner data set.

Model Details

Records365 has selected a support vector machine algorithm from the models considered. This model demonstrates an overall train accuracy of 74.5%, and an overall test accuracy of 71.8%. The cross validated testing and training metrics are shown in Figure 15. We select the model and its parameters based on the cross validated test accuracy, which for this model is 69.9%.

Metric	Value
xvalTestAccuracy	0.699
xvalTestF1	0.664
xvalTestPrecision	0.775
xvalTestRecall	0.648
xvalTrainAccuracy	0.728
xvalTrainF1	0.689
xvalTrainPrecision	0.785
xvalTrainRecall	0.674

Figure 15. Mean accuracy across 10 cross validation folds

The F1 score of about 66.4% suggests we exercise some additional caution in our trust of this model, which suffers from and imbalance in the class distribution, as shown in Figure 16.

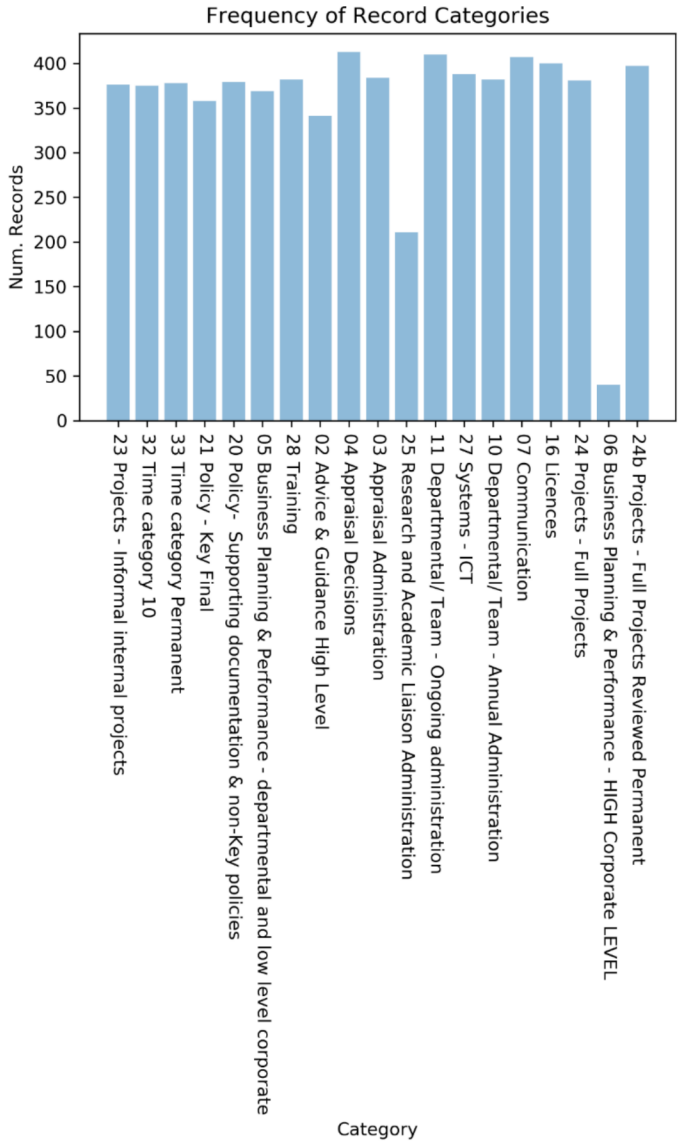


Figure 16. Category distribution

Our experiments were designed with the assumption that the class distribution would be balanced and in future we will likely use the F1 score instead of accuracy as the model selection criterion in situations where this assumption is not met.

Error! Reference source not found.7 and 18 show the test and train confusion matrices. Broadly speaking, the higher the number on the top-left to bottom-right diagonal the better the model's ability to correctly predict that class and we expect to see the highest numbers along this line. High numbers off the diagonal indicate false positives or true negatives and also indicate places where the model confuses two categories.

Predicted \ Actual	23	32	33	21	20	5	28	2	4	3	25	11	27	10	7	16	24	6	24b
23	80	0	9	0	0	0	0	1	0	0	0	17	0	0	0	0	0	0	0
32	0	74	10	0	0	0	0	0	0	1	0	6	0	0	0	0	0	0	0
33	0	4	93	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
21	0	0	25	50	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
20	0	7	37	0	42	0	0	1	0	0	0	17	0	0	0	0	0	0	0
5	0	1	12	0	0	65	0	0	0	0	0	3	0	8	0	0	3	0	0
28	0	2	11	0	0	2	52	0	0	0	0	45	0	0	0	0	0	0	0
2	0	0	2	0	0	0	0	89	0	0	0	8	0	0	0	0	0	0	0
4	0	0	5	0	0	0	0	0	66	0	0	4	0	0	0	0	0	0	0
3	0	0	8	0	0	0	0	0	2	93	0	4	0	0	0	0	0	0	0
25	0	0	36	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0
11	0	0	0	0	0	0	1	0	0	0	0	78	0	0	0	0	0	0	0
27	0	0	10	0	0	1	0	0	0	0	0	14	61	0	0	0	0	0	0
10	0	0	16	0	0	5	0	0	0	0	0	19	0	62	0	0	0	0	0
7	0	0	2	0	0	2	0	0	0	0	0	6	0	0	81	0	0	0	0
16	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	92	0	0	0
24	0	2	34	0	0	7	0	0	0	0	0	5	0	0	0	0	62	0	0
6	0	0	18	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
24b	0	0	15	0	0	5	0	0	0	0	0	2	0	0	0	0	0	0	75

Figure 17. Test Confusion Matrix

Predicted \ Actual	23	32	33	21	20	5	28	2	4	3	25	11	27	10	7	16	24	6	24b
23	309	0	10	0	0	0	0	0	0	0	0	57	0	0	0	0	0	0	0
32	0	306	30	1	0	1	1	1	0	3	0	32	0	0	0	0	0	0	0
33	0	19	342	0	0	0	1	0	0	0	0	14	0	0	2	0	0	0	0
21	0	4	92	240	0	1	0	4	0	0	0	17	0	0	0	0	0	0	0
20	0	35	147	0	120	3	6	11	0	0	0	57	0	0	0	0	0	0	0
5	0	2	46	0	0	264	0	0	1	0	0	19	0	25	0	0	12	0	0
28	0	4	33	0	0	0	206	4	0	0	0	134	1	0	0	0	0	0	0
2	0	0	13	0	0	0	0	300	0	0	0	28	0	0	0	0	0	0	0
4	0	0	23	0	0	0	0	0	365	2	0	23	0	0	0	0	0	0	0
3	0	0	19	0	0	0	0	0	3	344	0	18	0	0	0	0	0	0	0
25	0	0	137	0	0	0	1	0	0	0	0	71	0	0	2	0	0	0	0
11	0	0	2	0	0	0	1	0	0	0	0	407	0	0	0	0	0	0	0
27	0	0	24	0	0	2	0	0	0	0	0	57	302	1	2	0	0	0	0
10	0	1	43	0	0	16	0	1	0	2	0	63	0	256	0	0	0	0	0
7	0	0	6	0	0	1	1	0	0	0	0	36	0	0	363	0	0	0	0
16	0	0	6	0	0	5	0	0	0	0	0	3	0	0	0	386	0	0	0
24	0	5	110	0	0	33	0	0	0	1	0	21	1	0	0	0	210	0	0
6	0	1	37	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
24b	0	1	48	0	0	2	0	1	0	0	0	20	0	0	0	0	0	0	325

Figure 18. Train Confusion Matrix

In these matrices we can see that the poorer performing categories, 25 (Research and Academic Liaison Administration) and 6 (Business Planning & Performance - HIGH Corporate LEVEL), had fewer samples on which to train, and that both classes are likely to be mistaken for category 33 (Time category Permanent). We would almost certainly see better results if we were to increase the number of labelled records in these categories and train a new model. In situations where there is much less data available for certain categories, it may be worth combining them, or, if possible, tagging those records with metadata that can be used to distinguish them by the rules.

There is a relatively high level of confusion where several of the different categories are mistaken for category 33. This may indicate that the documents labelled within category 33 may vary widely, and not be well defined by vocabulary. In the initial training run that we presented at the workshop with TNA, we

experimented training the model without this category and the resulting model had higher testing and training accuracy. When records in a category are not descriptive of what should be in a category ML will tend to perform poorly, this is the case of category 33.

Chosen for Permanent Preservation

The details of the unlabelled documents that our machine learning selected for permanent preservation can be found in the attached document "Selected for Permanent Preservation.csv". There are 3180 records that were selected for permanent preservation from the unlabelled dataset. Of these 31 were determined to be of the disposal schedule "21 Policy - Key Final", while the remainder received the disposal schedule "33 Time category Permanent".

We have designed the machine learning capabilities to over compensative so, that important records are not missed and leave the final decision on the hand of the records manager so, it is expected that this number of records is higher than the actual number of records that should be permanently preserved.

It can also be noted that there are a number of records selected that are of a very small size. For this set of unlabelled data, a better outcome could be achieved by utilising ROTBot capabilities that allow users to dispose of ROT records before being submitted to be managed as records.

File Analysis Results

In the figures below you can see some example reporting on the ROTBot's file analysis and named entity recognition capabilities applied on a sample of the dataset.

- The most mentioned organisations across the entire data set were "The National Archives", "TNA", "MessageLabs" and "FOI".

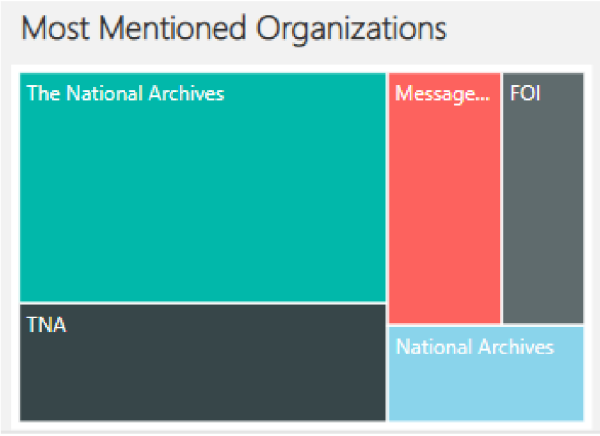


Figure 19. Top Five Mentioned Organisations

- The most mentioned geopolitical entities across the entire data set were "Richmond", "UK", "London", "England" and "Wales".

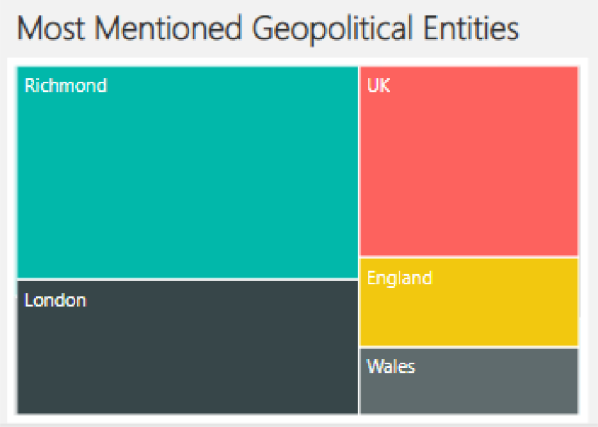


Figure 20. Top Five Mentioned Geopolitical Entities

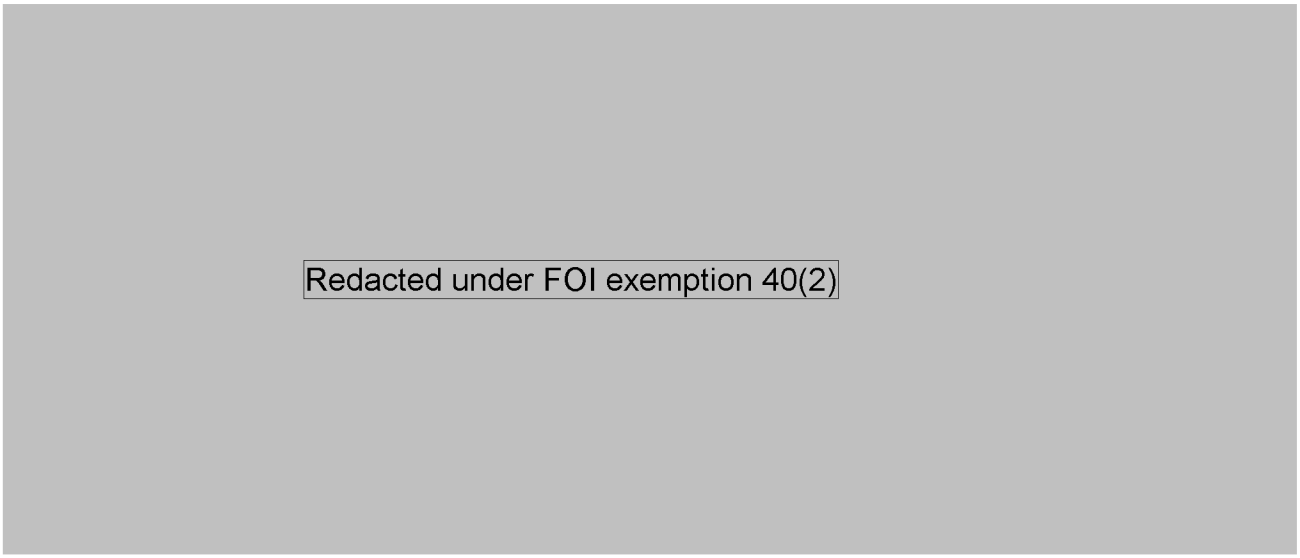


Figure 20. Top Five Mentioned People

- 70% of the content was smaller than 100KB and 20% was between 100KB and 500KB. Just a small percentage of the content was bigger than 10MB.

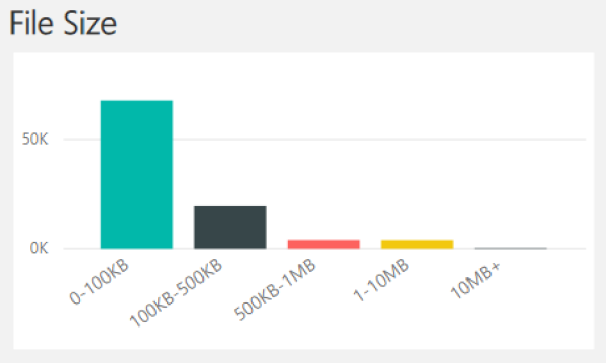


Figure 21. File Size Frequency

- Most of the content hasn't been modified recently with 49% of the content was last modified 5 - 9 years ago, 18% of the content was last modified 9 – 15 years ago and 29% more than 15 years ago.

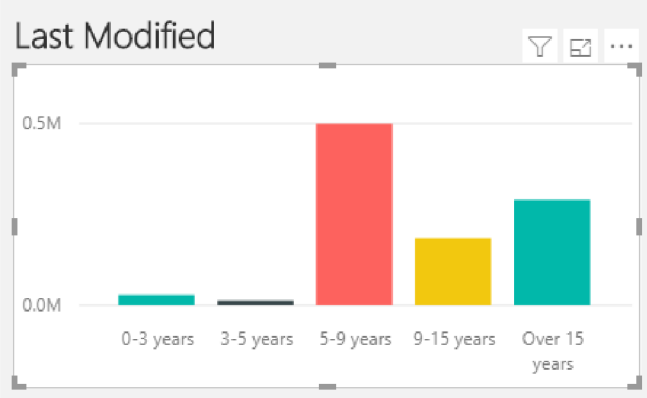


Figure 22. File Last Modified Date Proportion

7. Future Research & Development

Most of the innovations discussed in the State of the AI Market earlier in this document are on the current RecordPoint roadmap, which we have divided into the following items.

Context Enrichment

In the short term we are working to build a sophisticated, AI-driven data enrichment pipeline. This will bring natural language processing techniques such as image recognition, sentiment analysis, and additional named entity recognition (NER), as well as heuristic approaches to supplement document metadata into records as they are ingested. We will also use this pipeline to flag the presence of personally identifiable information (PII) and payment card industry (PCI) data in records, and to harvest metadata that are intrinsic to the document formats. All this additional metadata will be valuable for downstream appraisal tasks.

Multi-Model Appraisal

At present, Records365 concentrates on classifying records to a disposal schedule, but we have the capability to create machine learning models for any discrete set of categories. We will expand this to offer the ability to classify records in terms of value, risk, political sensitivity, sentiment, or other customizable sets of categories. These classifications will be stored as searchable properties on the records and will also be available for use by the rules system or for any other appraisal or workflow technologies.

Unsupervised Learning

Another application on our roadmap is to bring unsupervised learning techniques to record appraisal. Current market offerings around AI-driven appraisal use a supervised learning paradigm: they try to generalize a predictor based on some records that we have already classified. This approach works well when we know the appraisal categories up front and have example records for all of them, but if we are going in blind a different paradigm is required. Unsupervised learning tries to identify clusters of similar records in an unlabelled corpus of documents, which we can then compare by considering the most important terms used in each group. Once we are satisfied with the groupings, we can formalize these into categories and use these to train a supervised model that can be used to classify new data.

Searchable Knowledge Graph

RecordPoint is now at the point where we can graph the relationships between records as they move between content sources and authors. These graphs will allow us to model the ways in which users collaborate-on or share information between themselves. This will enable us to build some powerful forensic applications to detect scenarios such as IP theft, plagiarism and tampering, or simply to identify areas of value to users.

Multi-Dimensional Appraisal

There are ways to utilize both metadata and document text to appraise records. It is possible to integrate specialized machine learning models, which can be independently trained and managed, into a metadata-based expert system (RecordPoint's current rules engine), which can then be used to leverage the strengths of explicit logic as well as the statistical likelihoods that a machine learning model can offer.

Despite their limitations, expert systems do offer some benefits that natural language processing-based solutions do not. They are less opaque to human understanding, and they can provide certainty, where machine learning models offer only a statistical likelihood. Expert systems also enable us to base classifications on explicit requirements that may be imposed by regulatory agencies. For example, a document might belong in a certain category if it was signed by an individual during a specified period. This kind of classification is very difficult to achieve with only text classification techniques and so we are looking for ways to integrate structured data and even hand-coded rules with text classification to find the best of both worlds.

This approach would mitigate some of the brittleness of expert systems discussed earlier, so that we might still be able to use metadata to classify records even if the metadata is not complete, or the rules do not explicitly provide a pathway to the correct classification.

As well as a roadmap item, these concerns are the focus of Jason Franks' (RecordPoint's ML expert) academic research.

Language Models

Neural networks have recently enabled a lot of powerful new applications based on understanding human language and we will soon be looking to incorporate some of these features into our product suite. In addition to providing more accurate text classification than the TF-IDF techniques discussed earlier, we should be able to leverage these models to provide natural language text summaries for records.

Later it may be possible to repair documents that have been partly corrupted, and eventually to provide some limited capacity to ask questions about the content of a record that are more sophisticated than a keyword search.

AI-Driven Content Analytics (ROTBot)

Not all the information generated in an organisation has the same value and risk. However, it is difficult to correctly profile content and make the right decisions due to the vast amounts of information generated on a daily basis. To bring light to the types of information organisations have as well as how it is being used, RecordPoint will continue making further investments on its ROTBot. The ROTBot will use a combination of ML models augmented with heuristics coming from metadata to analyse vast amount of information to provide insights like Redundant, Obsolete, Trivial (ROT) content vs high value content. It will also detect duplicates and near duplicates and flag records that contain personally identifiable information (PII) and payment card information (PCI).

These insights will be actionable directly from the ROTBot to perform remedial actions to the content such as destroy ROT in place or put high value content under records management or archive low value / high value inactive content. More actions will be supported over time as Records365 expands its capabilities.

Intelligent Connectors

As processing power on smaller devices grows, we are looking to push some of our machine learning capabilities out of the platform to the edge. This means that our we will be building AI capabilities into Records365 connectors to external content sources, which will help us to pre-filter redundant/obsolete/trivial (ROT) records or content with no business value before ingesting them, saving on storage as well as pipeline processing. This will help organisations to ensure content is properly managed based on the risk and value associated to it.

In addition to that, it will also enable us to take smarter actions on the content source as the connectors observe events in real time, or in response to directions from Records365. At present we have prototyped some of this functionality on the file system connector and we are planning to offer a suite of AI capabilities across all content sources.

AI Based Risk & Value Scoring

As an enhancement to the risk and value profiling capabilities supported by Records365 we will be introducing a concept of scoring that provides a more transparent and granular approach to compare content within an organisation. This scoring will use a combination of ML and other heuristics to provide a fine-grained evaluation that will drive better insights and decisions.

8. References

[1] G. Rolan, G. Humphries, L. Jeffrey, E. Samaras, T. Antsoupova and K. Stuart, "More human that human? Artifical intelligece int he archive," *Archives and Manuscripts,* vol. 47, no. 2, pp. 179-203, 2019.

[2] S. McKemmish, "Placing records continuum theory and practice," *Archival Science,* vol. 1, no. 4, pp. 333-359, 2001.

[3] A. Gilliliand, "Designing Expert Systems for Archival Evaluation and Processing of Computer Mediated Communications: Frameworks and Methods," 2016.

[4] G. McDonald, C. Macdonald, I. Ounis and T. Gollins, "Towards a Classifier for Digital Sensitivity Review," *Proceedings of the 36th European Conference on Information Retrieval,* 2014.

[5] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

9. Appendix 1 - File Formats Supported

File Extension	Mime Type	Notes
.xml	text/xml application/xml	
.json	text/json	
.css	text/css	
.txt	text/plain	
.bat , .cmd , .btm	application/bat application/x-bat	
.csv	text/csv	
.lnk	application/x-ms-shortcut	
.mpp	application/x-project	
.php	text/x-php	
.url	text/x-url	
.xlsx, .xls, xlsm	application/vnd.ms-excel application/vnd.ms-excel.sheet.macroenabled.12 application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	
.pptx	application/vnd.openxmlformats-officedocument.presentationml.presentation	
.docx, .doc, .docm	application/msword application/vnd.ms-word.document.macroenabled.12 application/vnd.openxmlformats-officedocument.wordprocessingml.document	
.msg	application/vnd.ms-outlook	
.pdf	application/pdf	Some variants rely on OCR.
.rtf	application/rtf	
	application/octet-stream	
.eml	message/rfc822	
.jpg, .jpeg	image/jpg image/jpeg	Using OCR
.png	image/png	Using OCR
.gif	image/gif	Using OCR
.tif	image/tif	Using OCR
.bmp	image/bmp	Using OCR