Deloitte.

ADLIB®

The National Archives

AI for Digital Selection

Deloitte | Analytics and Cognitive

Contents

1 Project background 3

1.1 Deloitte key contacts 3

2 Industry point of view (File Analytics) 4

2.1 Explosion of data 4
2.2 Introducing File Analytics 4
2.3 Intelligence driven Records Management 5

3 Adlib Elevate Overview 6

3.1 Overview 6
3.2 Architecture overview 6
3.3 Methodology: Content Elevation Process 7
3.4 Data classification techniques 8

4 Evaluation/ Modular Testing 9

4.1 Introduction 9
4.2 Test category: Data Collection 9
4.3 Test category: Pre-processing 13
4.4 Test category: Modelling 16
4.5 Test category: Deployment / GUI 17

5 Appendix A: Data from Modelling Activity 19

6 Appendix C: User interface 26

7 Appendix D: Model deployment 34

8 Appendix F: References 39

1 Project background

The National Archives (TNA) has to preserve and manage over 11 million historical and governmental records. A key part of their role is the accessioning of records from government departments. The growing volumes of digital records now make this a significant challenge for the TNA to determine what should be appraised and selected for permanent preservation as a historical record.

The National Archives are looking to research the capabilities required to automatically identify future records which they must retain and manage in line with their Records Management policies and mandate. To support this research they are running a number of Proof of Concepts (PoCs) with Artificial Intelligence (AI) software vendors to evaluate how they can automate the identification and classification of Records.

Adlib has been selected as one of the AI software vendors for evaluation. To support this assessment Deloitte's Analytics and Cognitive consulting team have partnered with Adlib to run and deliver the proof of concept.

The Proof of Concept on Adlib Elevate has been run to demonstrate the key requirements of the TNA and their defined outcomes for the testing. The testing has included:

- Modular testing – demonstrating Adlib's ability to meet specific requirements and demonstrate key functionality of the tool i.e. metadata extraction
- Outcome testing – the ability of the tool to automatically identify and classify a sample set of records to determine what should be retained by the TNA
- Usability testing – understand how the tool works including the setup, configuration and execution

This document captures a view on the industry and how AI is improving Records Management, an overview of Adlib and captures the features and outcomes of the testing as part of the Adlib Proof of Concept.

1.1 Deloitte key contacts

Hani Girgis
Deloitte Partner
hgirgis@deloitte.co.uk

Hani is Deloitte's Enterprise Content Management Practice Lead within Analytics & Information Management Services. He is an accomplished Information Management evangelist and a leading Subject Matter Expert in the field of Portals, Document and Records Management and Business Process Management, having presented in numerous international conferences.

Leon Duffield
Deloitte Director
lduffield@deloitte.co.uk

Leon leads the Records Management practice within Deloitte UK. He has over 12 years of experience in managing and delivering records management programmes. This experience includes managing all phases of business and technology transformation projects; from requirement gathering to implementation/transition into service activities.

2 Industry point of view (File Analytics)

2.1 Explosion of data

Technological and social advances have produced a flood of new digital applications and devices, as a result the amount of digital data available is growing faster than ever[6]. Approximately 90% of the world's data has been created in the last two years[5], and by 2022, it is forecast that 93% of all data in the digital world will be unstructured[3]. This explosion of data and the insights that can be revealed can be highly valuable to organisations from a strategic standpoint.

Across Public and Private sectors, many organisations have large collections of both structured and unstructured data which are sat idle. This is exposing organisations to challenges in storage, management, and adherence to regulatory and legal requirements.

While organisations have made progress on structured data, few organisations have been able to explore and address their unstructured data. One of the reasons why organisations are finding it difficult to manage their unstructured data is its sheer volume. It is also in non-traditional data types such as text messages, documents, email, video and audio files, and still images. These data types remain largely unmanaged, either because they don't live in a relational database or because until relatively recently, the tools and techniques needed to leverage them efficiently did not exist

2.2 Introducing File Analytics

Comprehensively capturing or extracting business insight from unstructured data is a challenge facing many organisations today. With an unprecedented number of sources of information, along with more demanding compliance mandates than ever before, it is imperative that unstructured data is captured and managed effectively.

Organisations are now seeking intelligent data extraction capabilities to assist them in gaining control of their unstructured data, whether it is contained in claims forms, emails, CAD diagrams, contracts, clinical trial documentation or design approval paperwork. Many organisations are now turning to File Analytics platforms to ingest, discover and provide insight from this previously untapped set of data sources.

While File Analytics technologies have been in existence for nearly a decade they are now maturing and advancing with Machine Learning capabilities which underpin and facilitate Artificial Intelligence (AI). These platforms are now enabling organisations to extract automatically, identify, classify, extract data and take action on this previously unmanaged data set.

To achieve this the File Analytics platforms provide a data extraction tool with Optical Character Recognition (OCR) and report mining capabilities that can automatically identify and extract meaningful information for decision-making. Many organisations in both the Public and Private sector are now using these tools to support the following use cases:

1. **Records Management** – auto-classification of historical records that can then be managed effectively through their lifecycle

2. **Information compliance** – intelligently search and discover unstructured data that is non-compliant to legal and regulatory requirements

3. **Cloud migration** – removing the redundant, obsolete and trivial data is an important first step when migrating from traditional file shares to platforms such as Office365

4. **Automation** – unstructured data sits at the heart of processes, and the ability to extract data from documents can feed RPA and analytics engines

2.3 Intelligence driven Records Management

Now more than ever, the reliability, security, control and transparency of information is critical to meeting regulatory and legal requirements. Typically the records, legal and IT stakeholders have failed to keep pace with the dramatic change in the information environment.

According to IBM roughly 2.5 billion GBs of data are created per day, and many organisations now hold so much digital data that they have no effective way of determining what they possess, where it is stored or who has access to it. This is exposing them to:-

- Increased storage costs
- Increased regulatory compliance risk
- Increased risk of data leakage
- Increased legal costs

Records Management and Information compliance continues to be a significant challenge across the industry and poor information management is now impacting many organisations ability to prepare for era of automation and artificial intelligence.

File Analytics platforms are now emerging as key tools to address legacy Records Management challenges and reduce the manual nature often associated with identifying, classifying and managing records. The platforms are being leveraged to support:

- **Automated classification** – application of categories, labels, tags, or metadata to content
- **Clustering of records** - establish relationships between documents and records management
- **Data clean-up –** identify redundant, obsolete and trivial data that should be disposed
- **Defensible disposition –** dispose of time expired records based on retention schedule
- **Respond to e-discovery** - identify records based on key words
- **Sensitive data identification and redaction –** in response to GDPR PII can be identified across the estate and data either retrieved, redacted or destroyed

For the TNA these tools/platforms provide an opportunity to support the accuracy and automation of the record accession process from government departments. A key consideration for the TNA is how they accept permanent digital records in the future to avoid a tsunami of data that needs to be reviewed/classified. Options to be thought through as this proof of concept matures:

1) TNA builds a hosted File Analytics platform and offers a service where government departments can subscribe to, manage their rules for permanent archive and automatically accession the records from chosen repositories

2) Government departments invest in File Analytics platforms and are used as a pre-requisite with evidence prior to providing permanent digital records to the TNA

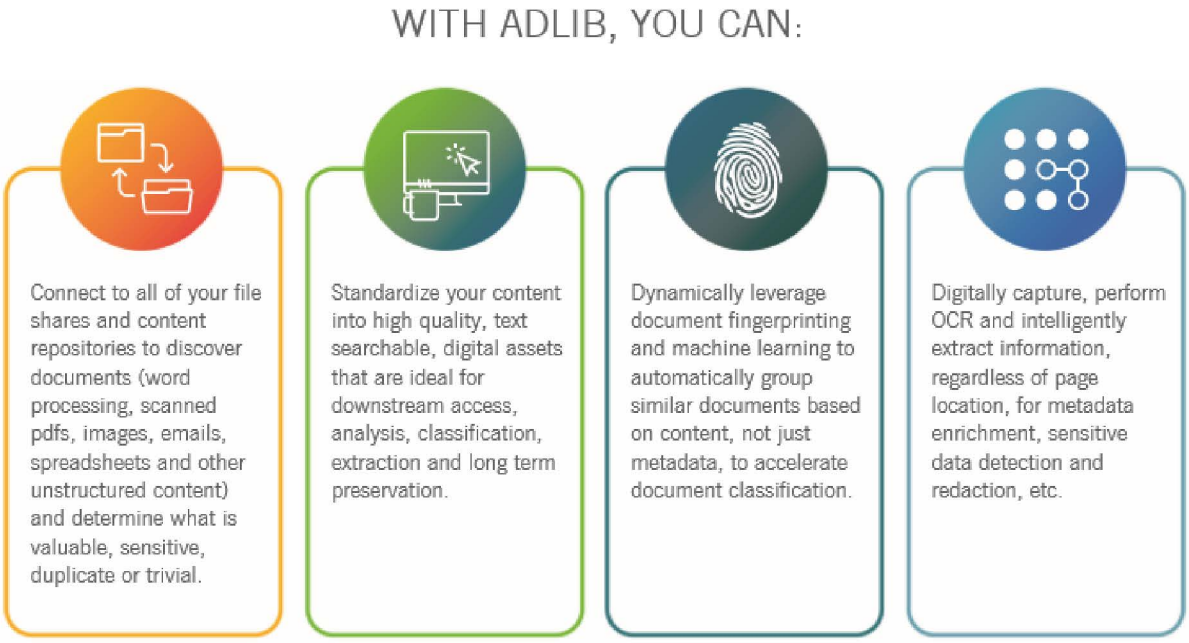
3 Adlib Elevate Overview

3.1 Overview

Adlib Elevate is a File Analytics platform to automate discovery, extraction and classification of vital data from complex documents to streamline data-intensive processes and accelerate process automation. The platform:

- Enables intelligent harvesting of data from a large volume of libraries and file types
- Provides a scalable and extensible platform for analysis, standardisation and classification of data
- Exposes the results of the analysis with supporting actions to address data
- Exports the data in standardised formats to strategic data repositories for long term preservation

Adlib Elevate provides a cloud based platform – designed for horizontal scaling – to support analysis of large volumes of documents. The platform has a modular architecture, enabling the easy integration of new components or applications as required.



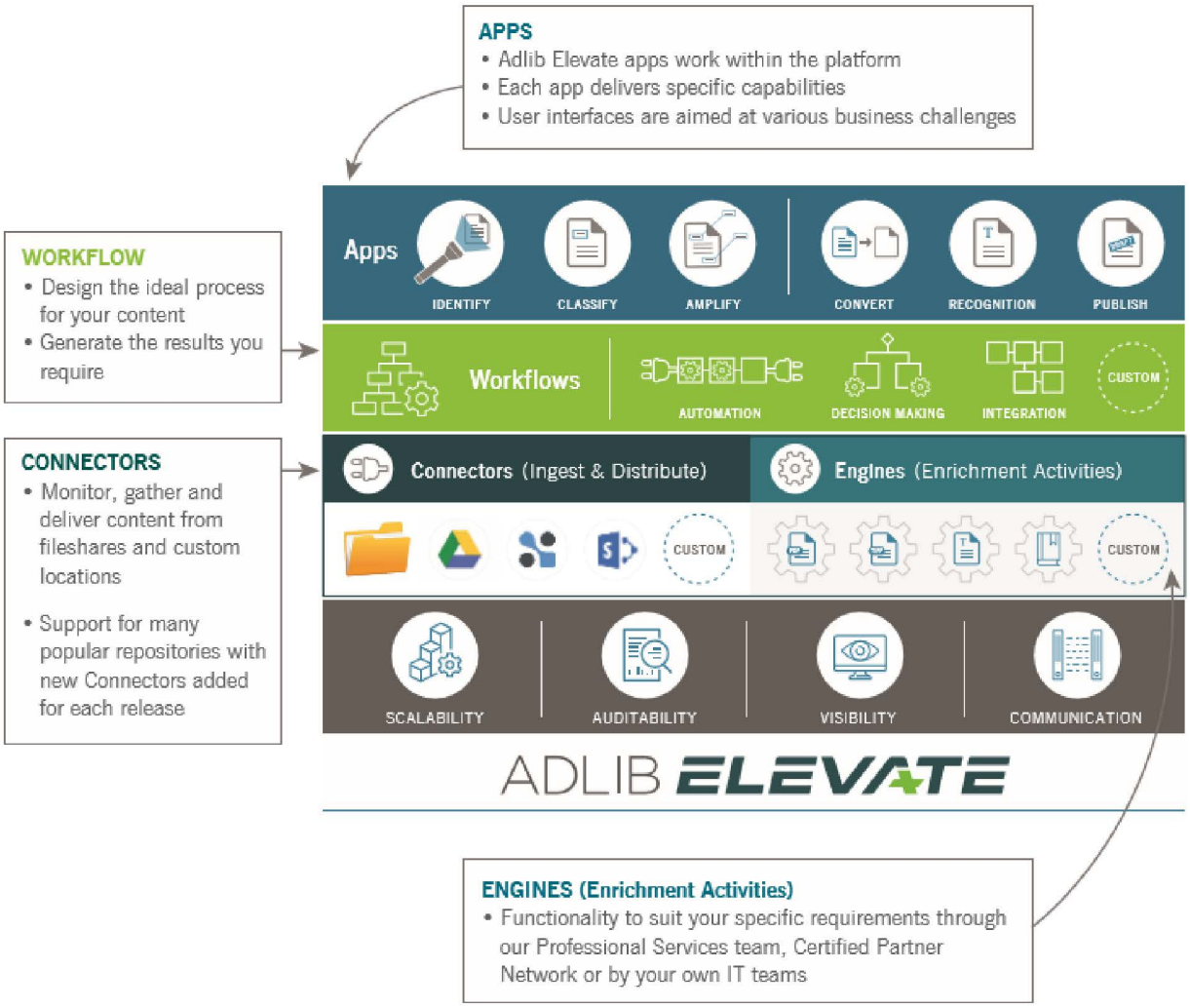
3.2 Architecture overview

Adlib Elevate is a scalable platform, with powerful workflow-driven applications and a large number connectors to enable organisations to discover and standardise their unstructured data.

The platform allows you to conceptualise, standardise, analyse, categorise and visualise unstructured data through the following features:

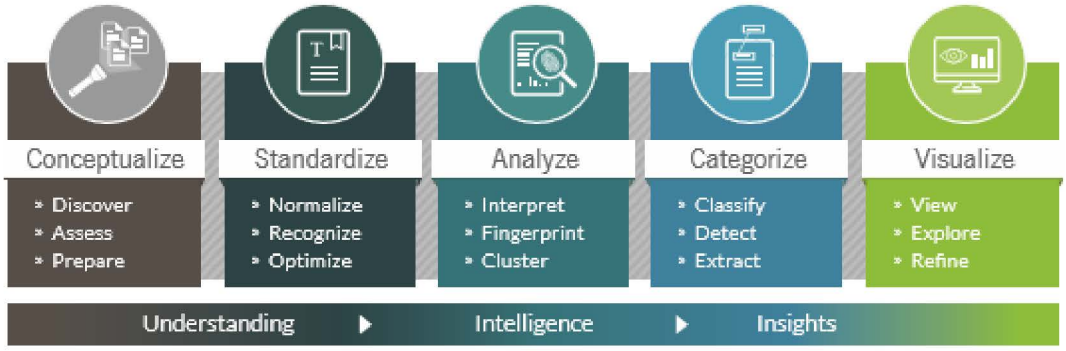
- Advanced rendering technology to standardise data
- Access and conversion 400+ file types
- Cloud based deployment

- Integration into large number of platforms (ECMs, PLMs, and ARs)
- Compression to reduce storage and file transfer costs
- Automation of archiving and identification follow long-term digital preservation guidelines



3.3 Methodology: Content Elevation Process

Adlib's Content Elevation Process (CEP) is designed to provide greater data insight on an organisations unstructured data and follows 5 key steps as detailed below:



1. Conceptualise: Discover file metadata for a library of content and apply file analytics on the set of data to identify ROT (redundant, obsolete, trivial) data within the content for potential remediation

2. Standardise: Capture documents from virtually any source, converting over 400 file types to a high-definition format with advanced OCR in preparation for effective classification and other document-centric processes

3. Analyse: Organise the textual information and document content flow into a unique file fingerprint. Automatically cluster documents based on fingerprint similarity to expedite the document classification process

4. Categorise: Classify documents based on relevant taxonomy and intelligently extract additional attribute information to strengthen the overall quality of the digital asset

5. Visualise: Provide a transparent relevant view of informed inventory across unstructured data sets to aid insights and comprehension of all content asset

3.4 Data classification techniques

Adlib Elevate's Progressive Classification™ capability takes two approaches to effectively classify documents and files:

1) Unsupervised machine learning for document clustering

2) Semi-supervised rule building which leverages a defined training set

Using advanced document conversion, clustering, and rules-based classifiers, Adlib's Progressive Classification solution allows enterprises to:

- Automatically process large volumes of unstructured data—creating actionable insights and revealing hidden value and risk.
- Take documents from multiple lines of business, remove the ROT, and ensure that valuable assets (contracts, agreements, reports, and regulatory submissions etc.) are grouped, enhanced, and easily searchable.
- Eliminate the cost, risk, and inconsistency of manually processing rapidly growing volumes and varieties of documents and data.
- Leverage a flexible and configurable system that learns and adapts as your business needs evolve.
- Group or cluster similar documents to enable targeted extraction of key data elements that can be added as metadata, validated against other systems, etc.
- Identify sensitive or pertinent data to manage compliance risk effectively (e.g. identify PII)
- Mitigate risk by identifying classified documents and applying appropriate security measures or redacting confidential information.

4 Evaluation/ Modular Testing

4.1 Introduction

To support the Proof of Concept on Adlib Elevate, the TNA provided set of evaluation and modular testing topics to be demonstrated. In addition to this two different data were provided by the TNA to support the evaluation.

The first set (referred to as DataSet1) was data extracted from the TNA EDRMS system and contained documents in many files formats. Dataset1 was already classified into disposal schedules. The second set of data referred to as DataSet2) was a random collection of files from a network shared drive and was not classified into disposal schedules.

Using DataSet1 as a training dataset, models were built using the Adlib Elevate and the files in DataSet1 were processed against the model to identify how well the model would predict the disposal schedule against the actual disposal schedule assigned by TNA. Once the processing of DataSet1 was complete, the files in DataSet2 were also run through the model to see if any would be assigned a disposal schedule.
The stats and findings on both sets of data as well as the outputs from the modular testing are set out below.

The following sections in this chapter run through each of the test categories provided by TNA and detail an overview of how Adlib functionality works to meet the requirement. Where relevant we have provided evidence based on the TNA data sets provided.

4.2 Test category: Data Collection

Overview
Adlib Elevate is able to process documents from multiple source locations. These can be, but are not limited to Network Folders, SharePoint, email (POP3 and SMTP), Exchange, Documentum, OpenText, Google, Box, Salesforce, BluePrism and Automation Anywhere. Additionally, due to the open nature of the Adlib platform, new connectors can be created to cater for source locations not serviced by the out-of-the-box provided connectors.

Users can set up new location of files, to be processed and initiate the discovery process to crawl the location. Upon completion the system has created metadata on each file including:

- file location,
- file creation date,
- file last referenced date,
- file size,
- file format and
- file document name.

Additionally, graphical representations of the location(s) contents are provided, including information such as files created by calendar year, whether the files are word processing or spreadsheets or image files, and more. This process also allows for the removal of duplicate and empty files (reducing ROT – Redundant, Obsolete and Trivial data).

Evidence
The files in Dataset1 were copied from the supplied encrypted hard drive to the Adlib Elevate hosted environment. A total of 110,284 files were successfully copied.

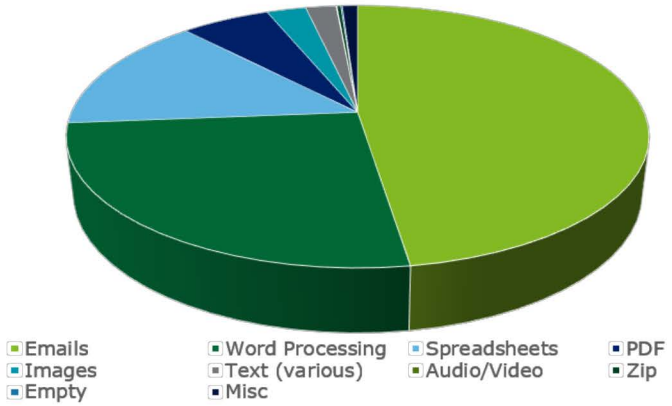
The files were initially analysed for file type. Adlib does not rely only on the file extension for this analysis, but looks at the header of each file to ensure that the contents of the file actually match the file extension.

The results of the initial Document Type analysis were as follows:

File Type	# of files
Adobe Acrobat (PDF)	6,499
Adobe Flash Video	4
Adobe InDesign	4
Adobe Photoshop Image (PSD)	10
AutoDesk AutoCAD	4
AVI Video	4
Corrupt File	4
Empty File	70
GIF Image (GIF)	35
GZ Archive	2
HTML Text Only	1,268
JPEG Exif	1,632
Log file	1
Microsoft Access Database (MDB)	10
Microsoft Bitmap (BMP)	39
Microsoft Document Image (MDI)	1
Microsoft EML	308
Microsoft Excel	15,838
Microsoft MSG	52,180
Microsoft OneNote	3
Microsoft PowerPoint	783
Microsoft Project	39
Microsoft Publisher	11
Microsoft RTF	3,077
Microsoft Visio	87
Microsoft Web Archive (MHT)	152
Microsoft Word	25,707
Microsoft XPS	2
MP3	6
MPEG-4 Video	29
NULL	42
Open Document Format	6
PCX Image	1
PNG Image (PNG)	310
PostScript (PS)	7
Scalable Vector Graphic	1
SGML Text Only	16
Shortcut/Link	8
Text (ASCII)	758
Text (Multibyte)	2

Text (UTF16)	3
Text (UTF8)	7
TIFF Image	848
Wang IWP	1
WAVE Audio	2
Windows EXE/DLL (EXE,DLL)	2
Windows Multimedia (WMA/WMV)	16
WordPerfect 5 - X4	4
XML document file	161
ZIP Archive	280
Grand Total	**110,284**

As can be seen from the (simplified) chart below, Email files were the largest grouping on DataSet1, followed by Microsoft Word documents, then Microsoft Excel files, PDF documents and various formats of Image files.



After extracting files from ZIP files and attachments from emails the following final statistics for files in DataSet1 were as follows:

110,284	Files brought in from dataSet1 disk
11,922	Duplicates
8,514	Additional duplicates found in email attachments and zip files
70	Empty Documents
22	Additional empty documents within zip and email attachments
180,968	Total Documents including all attachments
138,297	**Total Unique Documents**

The files in Dataset2 were copied from the supplied encrypted hard drive to the Adlib Elevate hosted environment. A total of **104,694 files** were successfully copied. Text was the largest document type with a count of **38,906**, followed by Document with **26,139** documents and lastly Image with **17,529** documents The overall analysis of DataSet2 is summarised below:

104,694	Files
108,190	Files After Expanding ZIPS and MSG Files
77.90%	Video Files by file size
8.64 GB	Largest File : D:\Websites\Webteam\Video-editing-PC-backup\maude-raw-footage.mpeg
55,429	(51.2%) Files within the Olympics Folder Structure that have a final folder level \0\. These are all PDF's of Olympic documents over the years with corresponding files for OCR, OCR Stats, RTF of the File and XML of stats.
3,727	Documents remaining after removing the Olympic files above that are most suited to be processed with the AI model
70,816	Files Older than 7 years
21,737	Duplicate Files
1,574	Empty Files
1,776	Files with duplicates in Set 1 (Most are Redundant, Obsolete or Trivial files - 1574 are empty files)
28	Distinct files with a duplicated file in Set 1
1	e-Transfer Form

4.3 Test category: Pre-processing

Overview

Identify:

The following table shows just some of the 400+ file formats supported by Adlib Elevate:

File Type	File Extensions
Adobe Illustrator	AI
ANSI Text	TXT
AutoCAD	DXF
Email Message	EML, ELMX
Graphics Interchange Format	GFA, GIF, GIFF
HTML	HTM, CSS, HTML
Joint Photographic Experts Group	JPEG, JPG, JPE, JIF
Microsoft Excel for Mac	XLS, XLSX
Microsoft Excel for Office 365	XLS, XLSX
Microsoft Excel for Windows	XLS, CLSX, XLT, XLTX, XLSB
Microsoft Outlook	MSG
Microsoft PowerPoint for Mac	PPT, PPTX
Microsoft PowerPoint for Office 365	PPT, PPTX
Microsoft PowerPoint for Windows	PPT, PPTX, PPS, POT
Microsoft Visio	VSD, VSDX
Microsoft Word for DOS	DOC
Microsoft Word for MAC	DOC, DOCX
Microsoft Word for Windows	DOC, DOCX
Portable Network Graphic	PNG
Rich Text Format	RTF
Tagged Image Format File	TIF, TIFF
Unicode Text	TXT
Windows Meta File	WMF
WordPerfect for Windows	WPD, WP5

De-duplication: Adlib identifies any duplicate files across the input Dataset. This is done using a Hashing technique and does not rely on file names. This means duplicates are accurately identified and only one version of the file will be included in subsequent processing. Duplicates are automatically excluded.

Standardisation: Adlib is able to support such a wide range of file types and provide consistent processing (such as classification, Extraction and Analytics on them) because Adlib uses a technique call "Standardisation". The standardisation process takes place as each file is ingested into the Adlib system. Each file is processed by Adlib using the native application if available or if the native application is not available a generic engine is used. A PDF version of each file is created and is used as the work file for all post processing actions such as classification and extraction. Adlib are experts on the PDF format and can process files at the bit level for extreme accuracy of results. It is this standardisation process which is key to powering the Text Classification and Analytics engine as well as providing input to the Machine Learning capability, both during

the training phase as well as during business as usual processing where documents are classified by the AI capability.

OCR: If the original file, such as a Word document also contains images, these images are passed to the Adlib OCR engine where any text is extracted. If the original document is an image file such as a TIFF file, it is also run through the Adlib OCR processing engine and a text version of file is created.

Classification: After initial discovery and standardisation, files are ready to be run through a series of classification workflows, each enable users to define, organise and extract the data within each document. Examples of these workflow processes include:

- Clustering Workflow – Used to group like documents based on content similarity
- Fingerprint Workflow – Used to compare/contrast/extract key information about each file's metadata, providing a unique identifier to the files
- Manually Classify Workflow – Used to manually assign a document type to a specific type of file

These workflows feed into building logic to ultimately predict document types, and automatically classify document sets given the assigned type.

Metadata extraction: Adlib accomplishes this by converting all content into one standard format, using Advanced Rendering-powered classification, enhanced Optical Character Recognition (OCR) techniques, image comparison technology and PDF interpretation to examine, interpret and pull metadata out and enter it back into the system where it becomes searchable, classifiable and can be safely delivered to a repository for long term storage.

Text and data extraction: The Amplify Tool is an extension of Adlib's Elevate platform, used for data extraction, highlighting and redaction from document sets. Users can create rules via Regular Expressions to perform advanced text searches, text and content extraction from document sets. Extracted information can be output in XML, CSV or json format. Additionally, sensitive information within a document set can be highlighted and redacted if deemed necessary.

Text search: this engine allows for data to be extracted from completely unstructured documents, such as emails, contracts, and other forms of written text that do not conform to any standard pattern, format or schema etc., such as this snippet of a legal agreement between two parties:

Highlight and redact: this engine allows for text based data within documents to be highlighted or redacted. It can do either operation based on hard coded, exact coordinates or it can work similar to the TextSearch engine where any text that matches a certain regular expression pattern can be highlighted or redacted regardless of where that content appears in the output.

Redacted under FOI exemption 40(2)

Evidence

After analysing dataSet1, Adlib Elevate identified Authors, and Document Titles, a small example of which are outlined below:

Document Authors	# of files
Authorised User	12180
legislation.gov.uk	7453
The National Archives	3306
[redacted]	
Administrator	1256
[redacted]	
Public Record Office	1056
[redacted]	
[redacted]	
[redacted]	
[redacted]	
HMSO_Web_Server@theso.co.uk	745
[redacted]	
tsetup50	734
copyright	674
sisaw2	603
[redacted]	
[redacted]	

Document title extraction with a subgrouping of retention schedule:

Document Titles	Count of Title
e Transfer Form	**5361**
04 Appraisal Decisions (Close at end of year)	5337
03 Appraisal Administration (Close file + 20years)	21
23 Projects - Informal internal projects (Close file when Project ends)	2
07 Communication (Close at end of year)	1
Document References and Date	**2201**
23 Projects - Informal internal projects (Close file when Project ends)	2201
Reference code	**1083**
24 Projects - Full Projects (Close file when Project ends)	817
02 Advice & Guidance - High level (Superseded + 7 years)	62
32 Time Category 10	43

Document Titles	Count of Title
03 Appraisal Administration (Close file + 20years)	43
21 Policy - Key Final (Close file when superseded)	36
24b Projects - Full Projects Reviewed Permanent (Transfer 20 years after Project ends)	28
33 Time Category Permanent	27
05 Business planning & Performance - Departmental & low level corporate (Close at end of year)	27

4.4 Test category: Modelling

Overview

The AI models within Adlib Elevate are built using a component called Workbench. Workbench is the AI component within Elevate that enables users to take advantage of machine learning to automatically recognise and classify large volumes of documents.

While the Adlib Elevate system can be used by a data scientist or expert technical user (dedicated areas for these users such as the Workbench are provided), it has been designed to be as simple as possible so that Records Manager or other user can utilise the system without being overly technical.

Creating a model to be used for automatic classification in Elevate involves six main steps:

1. Creating a New Project
2. Creating Training & QA Data Sets
3. Building a Model
    a. Training Models
    b. QA (testing) Models
4. Testing Models
5. Reviewing Test Results
6. Deploy Final Model to Elevate

To demonstrate the power and simplicity of building, testing and deploying models with Adlib Elevate, example flow of screens necessary fully build, test and deploy an Adlib Elevate AI model have been included in the appendix.

Evidence

An AI model was created to support the classification and identification of priority documents requiring permanent retention. This AI model was built on the first data set provided by TNA and then applied to the second set of data.

The Adlib Elevate models take into account various attributes such a Vocabulary, Concepts and NLP in order to make a classification decision. During the training and testing of the initial model some experimentation was done to try to assess what would lead to the best model. One of the tests that produced good results was to subsample only Word documents and create a model to run against other Word Documents. This initially produced good results however this didn't apply well to other document types. Another experiment sampled 5,000 documents for the training set however it was found that some of the disposal schedules that had few documents in the training set produced very low accuracies.

Overall, the final model used was the one suggested by TNA using the 80% training document set. This model was used as is without further refining. The results produced an overall 77% accuracy. When evaluating the precision of the model against the test set it was noted that disposal schedule 05 was only at 57%. Upon further analysis it was noted that when the documents flagged as outliers were removed from this set, the precision jumped to 83%. This observation was then reviewed on the training model and it was found that disposal schedule 05 had issues there as well. We then looked at a second confusion matrix removing the documents in the model that didn't pass the threshold of 70%. This resulted in the number of documents in the model decreasing by 9.5% but the K Fold cross validation jumped from 77% originally to 83%. We didn't have time to rerun the test set against the optimized model however it is expected that it too would increase by the same 5%. It is also likely that more documents would meet the threshold value of 70% confidence.

It is difficult to know what caused the issues with disposal schedule 05 but it was noted that within repository Strategic Projects and under folders Procurement, Project Delivery and Contract Management \ Project Delivery \ Projects – Closed \ 20 Year Rule and Collection Strategy \ Government Engagement and in folder 2015 Autumn Records Transfer Report, the 44 documents here had a schedule of 05 however all other similarly named folders were assigned 24. This seems to be a misclassified small set of documents and isn't likely responsible for the variance but there could be other misclassifications here. Additionally it could simply be that the documents in this schedule vary significantly.

In order to obtain a better performing model, additional testing using the method of removing documents that did not meet threshold could be performed. Additionally, an in-depth working session with TNA would benefit. This would allow us to review the documents that didn't meet threshold to understand if these documents are truly classed as they should be. If it is truly correct, then we could add more similar documents to the model. If it is wrong it should be removed or reclassified. Further to this if we take another subset of documents that didn't fit well into a class, we could manually classify these and put these back into the model for retraining.

All files from DataSet2 were then run through the model created using DataSet1. Due to the difference in the data sets this is not an advisable operation. Inferences that allow a model to accurately predict a disposal schedule from one dataset are not universal truths and may not achieve the same outcome when applied to a second completely different dataset. Additionally when looking at the patterns in the data in DataSet2, there is a clear pattern of over 50% of the files belong to the Websites\Webteam\Projects\Olympics folder. This is a single project with a significantly varied dataset (photos from the Olympics over 100 years with associated OCR confidence data). This could likely be better classified faster and with less variance manually rather than using a model.

All data to support the above comments are shown in Appendix A below and are also supplied in the accompanying Excel sheet. This includes the raw data and confusion matrices as well as the K fold cross validation data and the histogram of files to keep.

4.5 Test category: Deployment / GUI

Overview

The Adlib Elevate solution is delivered as a Cloud based solution. The solution can be installed quickly and easily in any cloud hosted infrastructure which supports Microsoft Windows Servers with MS SQL databases. Environments such as Azure, AWS and Oracle Cloud can be utilised. Access to the Cloud Based solution is highly secure and can be controlled in several ways.

User access control: External Access to the Cloud Based solution can be controlled by limiting access to specific IP addresses via the Cloud Firewall. Only IP addresses, such as the IP for The National Archives internal network can be authorised and all other attempts to sign-in rejected. In this way External Access can controlled and restricted.

Additionally, once access has been granted by the Firewall, access by individuals or groups of users is controlled by authenticating users when signing-in to the solution via LDAP. Furthermore, once authenticated

at sign-in a user is assigned an Access Level which will control access to the solution, limiting access to functionality based upon roles.

Each Adlib user is assigned specific permissions, based on the type of function they perform as it relates to the application. The function they perform is known as a Role.

User interface: Adlib Elevate has three key user interfaces. It is possible to run each user interface in a separate tab if the user is a power user or administrator.

User Interface	Purpose
Identify + Classify	The Identify and Classify screens are where the majority of Data Analysts will work. In this interface new data sets can be discovered and submitted to the Elevate system for processing.
Vision Console	The Vision Console is mainly used by System Administrators and is where the Elevate system is configured.
Workbench	The Workbench is the interface for the Machine Learning and Artificial Intelligence aspects of the Adlib Elevate Solution. New models can be created in the Workbench. Model training is also performed from here. When models are ready for production they are exported from the "Workbench" and would then be consumed by activities in the Identify and Classify interface.

Evidence

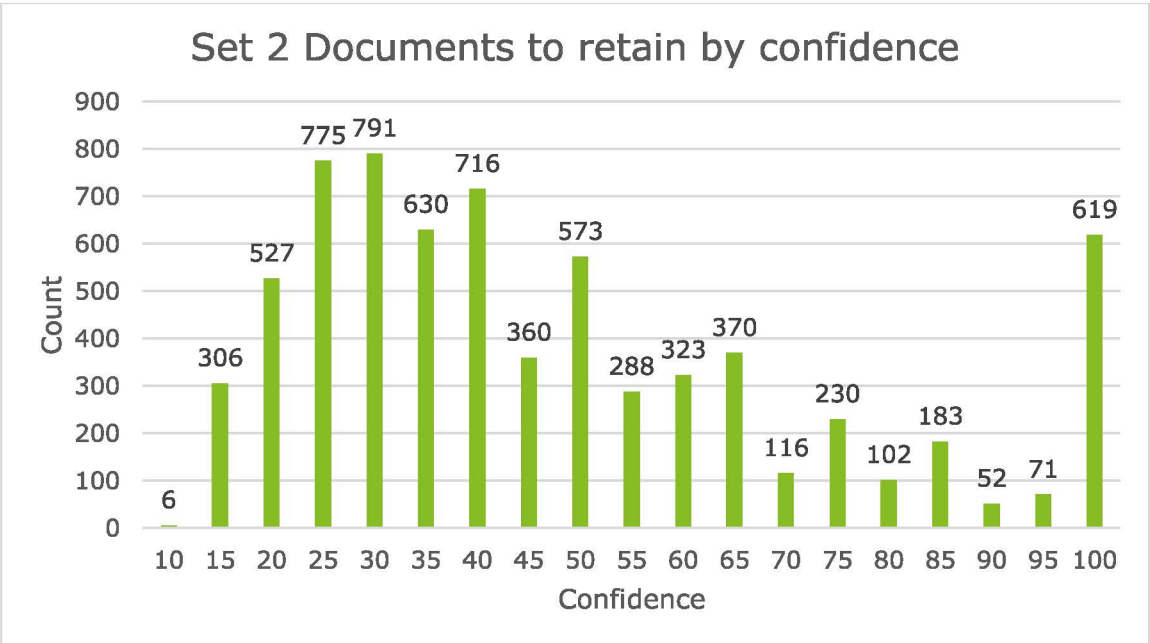
To support the evidence we have included a set of screenshots in the appendix on the key screens in Adlib Elevate.

5 Appendix A: Data from Modelling Activity

Set 2 Retention Chart:

Retention for Retention Schedules 4, 6, 15b, 17, 21 and 33

BIN	Percentage	Count
2	10	6
3	15	306
4	20	527
5	25	775
6	30	791
7	35	630
8	40	716
9	45	360
10	50	573
11	55	288
12	60	323
13	65	370
14	70	116
15	75	230
16	80	102
17	85	183
18	90	52
19	95	71
20	100	619



K Fold Cross Validation Using 5 folds:

Initial Results	Results after removing Outliers
Accuracy: 0.77 (+/-0.00)	Accuracy: 0.83 (+/-0.00)
Precision (weighted): 0.77 (+/-0.01)	Precision (weighted): 0.83 (+/-0.00)
Recall (weighted): 0.77 (+/-0.00)	Recall (weighted): 0.83 (+/-0.00)
F1 (weighted): 0.77 (+/-0.01)	F1 (weighted): 0.82 (+/-0.00)

Confusion Matrix for the test set not considering any thresholds or outliers:

		Predicted																				
		2	**3**	**4**	**5**	**6**	**7**	**10**	**11**	**16**	**20**	**21**	**23**	**24**	**24a**	**24b**	**25**	**27**	**28**	**32**	**33**	Total	Recall
2	**Actual**	2198	172	20	65	0	25	2	14	5	12	30	23	85	0	3	0	10	12	3	179	2858	0.77
3		151	837	18	93	0	2	0	4	2	0	11	6	18	0	0	0	6	4	2	19	1173	0.71
4		26	18	1551	8	0	0	0	2	0	3	14	7	20	0	0	0	2	0	12	26	1689	0.92
5		41	5	1	1560	0	15	5	26	1	4	4	29	91	0	7	0	12	11	14	59	1885	0.83
6		0	0	0	4	0	1	0	0	0	0	0	1	5	0	0	0	0	0	0	1	12	0.00
7		39	4	2	51	1	343	1	9	3	7	6	11	42	0	1	0	7	5	3	39	574	0.60
10		9	6	0	30	0	7	351	5	0	2	3	4	33	0	0	1	3	3	0	12	469	0.75
11		22	2	0	109	0	6	1	399	2	4	1	11	38	0	1	0	5	7	0	31	639	0.62
16		23	0	1	27	0	1	0	0	272	0	0	5	18	0	0	0	1	1	0	1	350	0.78
20		38	6	0	20	0	5	1	7	1	394	3	13	43	0	1	0	3	1	7	40	583	0.68
21		82	3	12	53	0	8	0	7	0	1	353	37	28	0	0	0	0	1	3	64	652	0.54
23		56	11	1	118	0	19	0	12	1	3	13	1243	63	0	2	0	8	5	0	55	1610	0.77
24		86	25	4	214	0	10	12	29	4	13	10	55	4560	0	10	0	37	17	15	104	5205	0.88
24a		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
24b		7	0	1	21	0	2	1	1	1	0	2	7	38	0	196	0	3	3	0	13	296	0.66
25		6	0	0	14	0	1	1	0	0	0	1	2	6	0	2	7	3	0	0	7	50	0.14
27		21	10	1	93	0	2	2	10	1	0	3	7	76	0	3	2	342	2	3	9	587	0.58
28		19	1	0	99	0	3	0	23	1	3	2	1	34	0	1	0	7	195	0	23	412	0.47
32		49	10	20	32	0	3	0	2	0	22	3	3	44	0	1	0	1	1	279	27	497	0.56
33		138	15	3	125	0	15	1	17	1	10	17	19	94	0	0	1	4	8	7	1786	2261	0.79
Total		3011	1125	1635	2736	1	468	378	567	295	478	476	1484	5336	0	228	11	454	276	348	2495	21802	
Precision		0.73	0.74	0.95	0.57	0.00	0.73	0.93	0.70	0.92	0.82	0.74	0.84	0.85	0.00	0.86	0.64	0.75	0.71	0.80	0.72		

Accuracy	**0.77**

Confusion Matrix for the test set after removing documents that did not meet threshold:

This reduces the dataset from 21,802 documents to 13,230 documents.

		Predicted																					
		2	3	4	5	6	7	10	11	16	20	21	23	24	24a	24b	25	27	28	32	33	Total	Recall
2	Actual	1536	117	1	1	0	0	0	3	0	0	3	1	8	0	0	0	2	3	0	19	1694	0.91
3		100	673	4	7	0	0	0	2	0	0	0	3	0	0	0	0	0	0	1	3	793	0.85
4		2	4	1431	0	0	0	0	1	0	0	4	1	2	0	0	0	0	0	2	2	1449	0.99
5		2	0	0	1188	0	1	2	2	0	1	0	5	4	0	1	0	0	5	2	1	1214	0.98
6		0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	2	0.00
7		1	0	0	4	0	222	1	1	0	0	1	1	0	0	0	0	1	0	0	6	238	0.93
10		1	0	0	3	0	0	317	0	0	0	0	0	4	0	0	1	1	1	0	0	328	0.97
11		0	0	0	2	0	0	0	275	1	0	0	2	9	0	0	0	2	4	0	1	296	0.93
16		9	0	0	1	0	0	0	0	209	0	0	0	2	0	0	0	0	0	0	0	221	0.95
20		2	0	0	1	0	0	0	4	0	245	0	3	3	0	0	0	0	0	0	2	260	0.94
21		9	0	5	3	0	0	0	0	0	0	166	5	0	0	0	0	0	0	0	4	192	0.86
23		6	1	0	0	0	0	0	0	0	0	1	924	6	0	0	0	2	1	0	1	942	0.98
24		11	3	0	11	0	2	3	5	0	0	1	7	3796	0	1	0	9	3	4	8	3864	0.98
24a		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
24b		0	0	0	2	0	2	0	0	0	0	0	1	7	0	126	0	0	0	0	0	138	0.91
25		0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	7	0.86
27		3	2	0	4	0	0	1	2	0	0	3	2	11	0	0	1	201	0	2	1	233	0.86
28		0	1	0	1	0	0	0	11	0	0	0	0	3	0	0	0	1	145	0	2	164	0.88
32		0	3	2	3	0	0	0	0	0	2	0	1	2	0	0	0	0	0	150	2	165	0.91
33		17	3	1	4	0	1	0	2	1	1	1	2	16	0	0	0	0	0	3	978	1030	0.95
Total		1699	807	1444	1235	0	228	324	308	211	249	180	959	3875	0	128	8	219	162	164	1030	13230	
Precision		0.90	0.83	0.99	0.96	0.00	0.97	0.98	0.89	0.99	0.98	0.92	0.96	0.98	0.00	0.98	0.75	0.92	0.90	0.91	0.95		
		Accuracy	0.95																				

Confusion Matrix for the test set after removing results flagged as outliers:

Overall, 1360 files were flagged as outliers

		Predicted																					
		2	3	4	5	6	7	10	11	16	20	21	23	24	24a	24b	25	27	28	32	33	Outlier	Total	Recall
2	Actual	2185	167	19	29	0	25	2	14	5	12	29	23	81	0	3	0	10	12	3	178	61	2858	0.76
3		148	833	16	15	0	2	0	4	2	0	11	6	18	0	0	0	6	4	2	19	87	1173	0.71
4		26	17	1508	3	0	0	0	2	0	3	13	7	20	0	0	0	2	0	12	25	51	1689	0.89
5		41	5	1	1380	0	15	5	26	1	4	4	29	90	0	7	0	12	11	14	57	183	1885	0.73
6		0	0	0	2	0	1	0	0	0	0	0	1	5	0	0	0	0	0	0	1	2	12	0.00
7		39	4	2	16	1	341	1	9	3	7	6	11	41	0	1	0	7	5	3	39	38	574	0.59
10		8	6	0	11	0	7	351	4	0	2	3	3	32	0	0	1	3	3	0	12	23	469	0.75
11		22	2	0	12	0	6	1	394	2	4	1	10	37	0	1	0	5	7	0	28	107	639	0.62
16		23	0	1	5	0	1	0	0	267	0	0	5	17	0	0	0	1	1	0	1	28	350	0.76
20		38	6	0	9	0	5	1	7	1	372	3	13	42	0	1	0	3	1	7	40	34	583	0.64
21		82	3	12	9	0	8	0	7	0	1	344	37	27	0	0	0	0	1	3	62	56	652	0.53
23		55	11	1	18	0	19	0	12	1	3	13	1235	63	0	2	0	8	5	0	54	110	1610	0.77
24		82	25	4	53	0	10	12	29	4	13	10	54	4469	0	10	0	37	17	15	103	258	5205	0.86
24a		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
24b		7	0	1	10	0	2	1	1	1	0	2	7	38	0	196	0	3	3	0	13	11	296	0.66
25		6	0	0	2	0	1	1	0	0	0	1	2	6	0	2	7	3	0	0	7	12	50	0.14
27		21	10	1	28	0	2	2	10	1	0	3	7	75	0	3	2	340	2	3	9	68	587	0.58
28		19	1	0	10	0	3	0	23	1	3	2	1	34	0	1	0	7	194	0	23	90	412	0.47
32		48	10	20	20	0	3	0	2	0	18	3	3	44	0	1	0	1	1	271	27	25	497	0.55
33		136	15	3	37	0	15	1	17	1	10	16	19	94	0	0	1	4	8	7	1761	116	2261	0.78
Total		2986	1115	1589	1669	1	466	378	561	290	452	464	1473	5233	0	228	11	452	275	340	2459	1360	21802	
Precision		0.73	0.75	0.95	0.83	0.00	0.73	0.93	0.70	0.92	0.82	0.74	0.84	0.85	0.00	0.86	0.64	0.75	0.71	0.80	0.72			
Accuracy Adjusted		0.80	(the 1360 outliers removed from the calculation)																					

Confusion Matrix for the training set not considering any thresholds or outliers:

		Predicted																				
		2	3	4	5	6	7	10	11	16	20	21	23	24	24a	24b	25	27	28	32	33	Total	Recall
2	Actual	10941	47	0	142	0	4	0	3	20	0	0	3	72	0	1	0	2	5	0	6	11246	0.97
3		519	3816	0	324	0	1	0	1	0	0	1	10	13	0	0	0	6	1	0	0	4692	0.81
4		7	10	6703	24	0	1	0	2	0	2	3	1	4	0	0	0	0	2	0	0	6759	0.99
5		9	15	0	7357	0	1	0	29	1	0	1	4	63	0	0	0	10	10	3	4	7507	0.98
6		1	0	0	10	33	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	47	0.70
7		13	6	1	140	0	2088	0	4	1	0	1	4	24	0	0	0	5	8	0	3	2298	0.91
10		3	3	0	92	0	2	1708	2	0	0	0	2	10	0	0	0	1	1	0	1	1825	0.94
11		4	5	1	495	0	5	0	2022	0	0	1	4	27	0	0	0	5	9	0	0	2578	0.78
16		19	2	0	41	0	1	0	0	1276	0	0	1	56	0	0	0	0	1	0	0	1397	0.91
20		2	4	1	78	0	1	1	3	0	2220	0	0	8	0	0	0	3	1	0	1	2323	0.96
21		10	2	13	143	0	1	0	1	3	0	2427	0	11	0	0	0	1	0	0	1	2613	0.93
23		19	19	2	417	0	9	1	21	2	5	10	5798	26	0	0	0	6	6	0	4	6345	0.91
24		57	26	1	647	0	9	9	29	61	6	6	31	19874	0	1	0	16	14	0	6	20793	0.96
24a		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
24b		19	1	0	35	0	1	0	6	4	0	2	0	21	0	1087	0	1	0	0	4	1181	0.92
25		2	0	0	27	0	0	0	0	0	0	0	1	13	0	1	180	0	2	0	0	226	0.80
27		8	13	0	261	0	2	1	21	0	2	2	12	50	0	1	0	1980	4	0	1	2358	0.84
28		10	2	1	384	0	7	3	44	2	1	1	2	40	0	0	0	7	1109	0	1	1614	0.69
32		1	7	2	63	1	1	0	1	1	5	0	3	5	0	0	0	3	2	1895	3	1993	0.95
33		24	4	6	308	0	9	2	8	2	2	4	6	51	0	0	0	7	9	3	8600	9045	0.95
Total		11668	3982	6731	10988	34	2143	1725	2197	1373	2243	2459	5882	20369	0	1091	180	2055	1184	1901	8635	86840	
Precision		0.94	0.96	1.00	0.67	0.97	0.97	0.99	0.92	0.93	0.99	0.99	0.99	0.98	0.00	1.00	1.00	0.96	0.94	1.00	1.00		
		Accuracy	0.93																				

Confusion Matrix for the test set after removing documents that did not meet threshold:

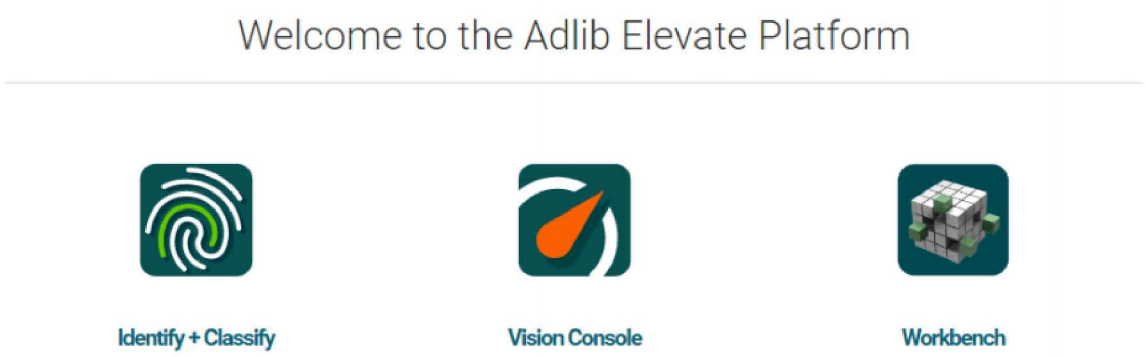
This reduces the dataset from 86,840 documents to 78,592 documents.

		Predicted																				
		2	3	4	5	6	7	10	11	16	20	21	23	24	24a	24b	25	27	28	32	33	Total	Recall
2	Actual	10277	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	10282	1.00
3		0	3665	0	4	0	0	0	1	0	0	0	2	1	0	0	0	0	0	0	0	3673	1.00
4		1	2	6680	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	6685	1.00
5		0	0	0	6515	0	0	0	2	0	0	0	0	2	0	0	0	0	1	0	0	6520	1.00
6		0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	1.00
7		0	0	0	0	0	2051	0	0	0	0	0	1	1	0	0	0	0	0	0	0	2053	1.00
10		0	0	0	2	0	0	1691	0	0	0	0	0	1	0	0	0	0	0	0	1	1695	1.00
11		0	0	0	2	0	0	0	1938	0	0	0	0	2	0	0	0	0	1	0	0	1943	1.00
16		0	0	0	0	0	0	0	0	1132	0	0	0	0	0	0	0	0	0	0	0	1132	1.00
20		0	0	0	0	0	0	0	1	0	2200	0	0	0	0	0	0	0	0	0	0	2201	1.00
21		0	0	0	0	0	0	0	0	0	0	2403	0	0	0	0	0	0	0	0	0	2403	1.00
23		4	1	0	0	0	0	0	0	0	0	0	5725	2	0	0	0	0	1	0	0	5733	1.00
24		1	1	0	4	0	0	0	3	0	0	0	1	19506	0	1	0	2	0	0	0	19519	1.00
24a		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00
24b		1	0	0	0	0	0	0	0	0	0	0	0	0	0	1083	0	0	0	0	0	1084	1.00
25		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	180	0	0	0	0	180	1.00
27		0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1936	0	0	0	1937	1.00
28		0	0	0	1	0	0	0	11	0	0	0	0	1	0	0	0	0	1046	0	0	1059	0.99
32		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1887	0	1887	1.00
33		0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	8570	8574	1.00
Total		10284	3671	6680	6530	32	2051	1691	1958	1132	2201	2403	5730	19518	0	1084	180	1938	1050	1887	8572	78592	
Precision		1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00		
		Accuracy	1.00																				

6 Appendix C: User interface

As a cloud based solution, the Adlib User Interface is delivered through a Web Browser such as Chrome or MS Edge. It has been designed to be accessible to all skill levels. Access to elements of the Adlib solution can be controlled by role, giving access only to the areas of the solution applicable to the user's job role.

Once a user has been authenticated using Active Directory, a choice of User Interfaces is offered (dependent upon permissions).

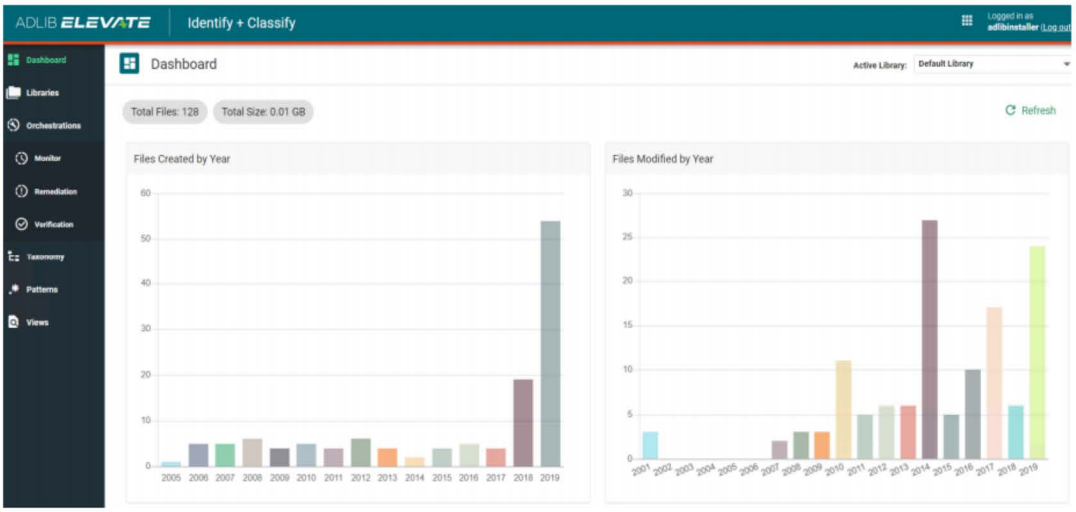


The different user interfaces serve different purposes and can be selected as required. It is possible to run each user interface in a separate tab if the user is a power user or administrator.

Identify & Classify User Interface

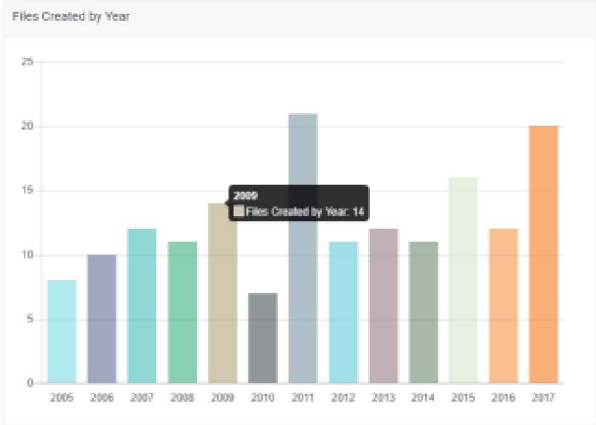
The Dashboard

The Identify Dashboard provides a breakdown of all documents within a specified library, with useful statistics on document details (e.g. number of files, breakdown of file format, file size etc.)

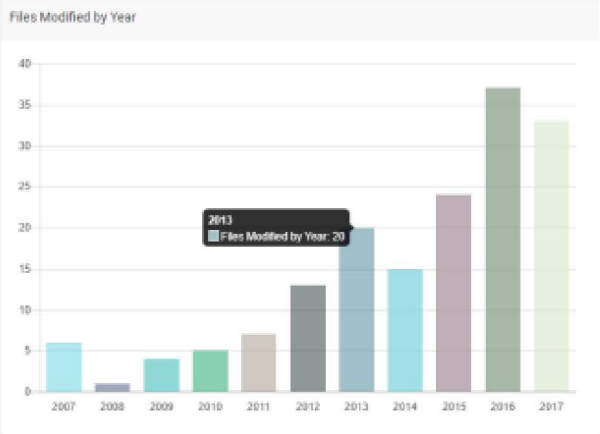


To Get a Visual Representation of Documents:

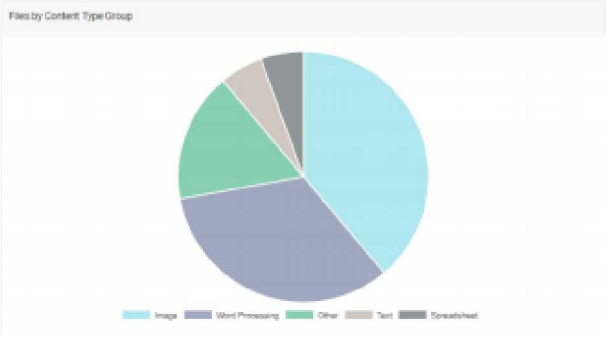
Files Created by Year – Groups a library's files by the year each document was published



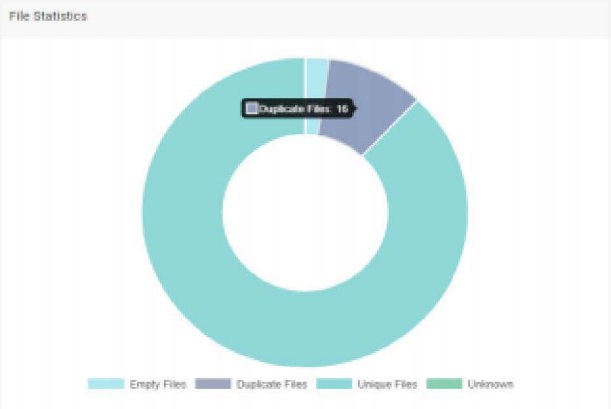
Files Modified by Year - Groups a library's files by the year each document was last modified.



Files by Content Type Group - Groups a library's files by the type of file (e.g. image file, text file, etc.)



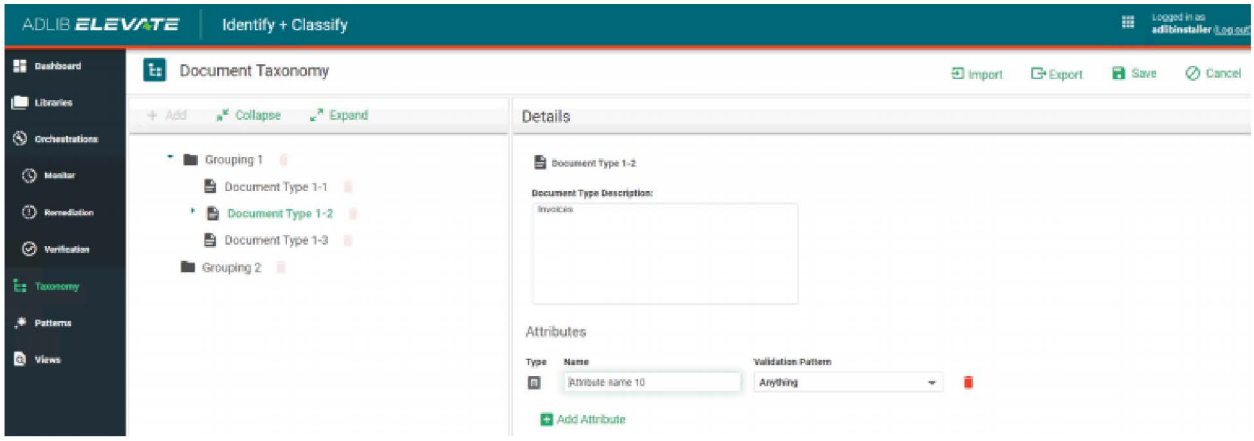
File Statistics - Groups a library's files by unique content (identifying duplicate files)



Taxonomy Page

Within Adlib, creating a taxonomy is the act of naming and defining documents that share the same characteristics. Document Taxonomies (also called document types) are used to help categorise/subcategorise etc. documents, by using specific properties that belong to certain document types (e.g. contract agreement). The Taxonomy page allows users to create a tree-structured classification of documents, specific to their lines of business. Users can easily create parent/children relationships and define document types with assigned document properties.

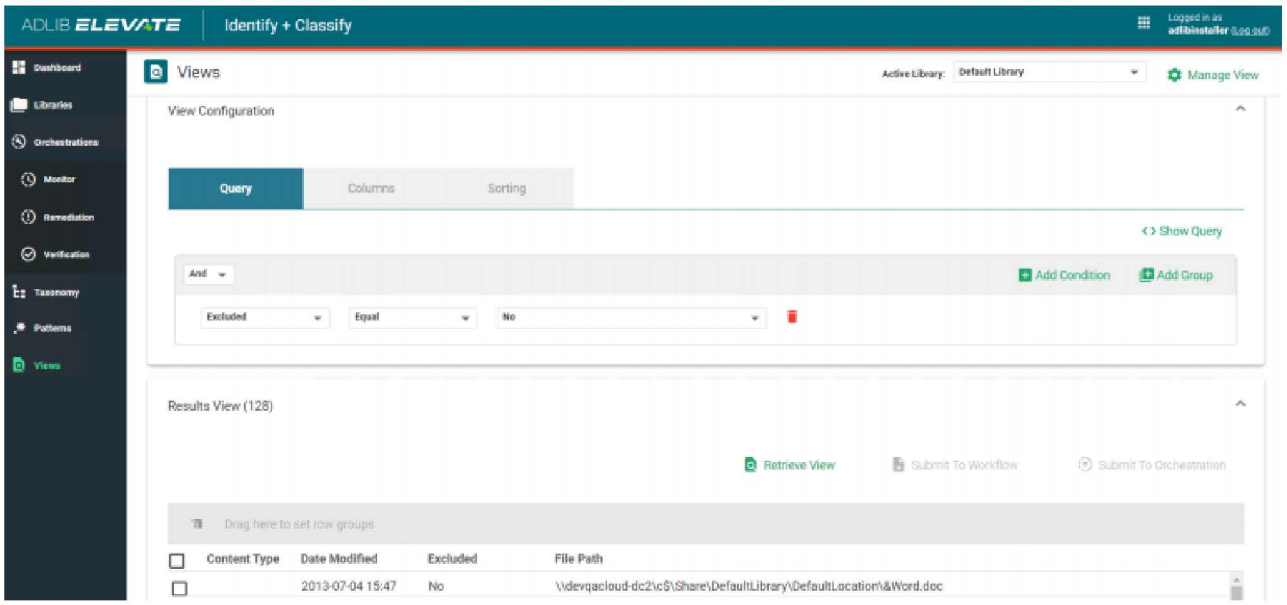
To help organise multiple taxonomies, Taxonomy groupings can be created. The groupings allow for related document types to be segmented.



Views Page

The Views page enables users to build powerful filters that target a specific set/type of files, for eventual processing in Elevate. This tool is of great value to organisations with a multitude of silos, departments and unknown files.

Locations are created - and libraries discovered - within Identify. The discovered files display within the Views page. The user creates and executes filters based on the type of files being targeted (e.g. filter on multimedia files, Word docs). Filtered items can then be submitted to a workflow within Elevate for further processing.



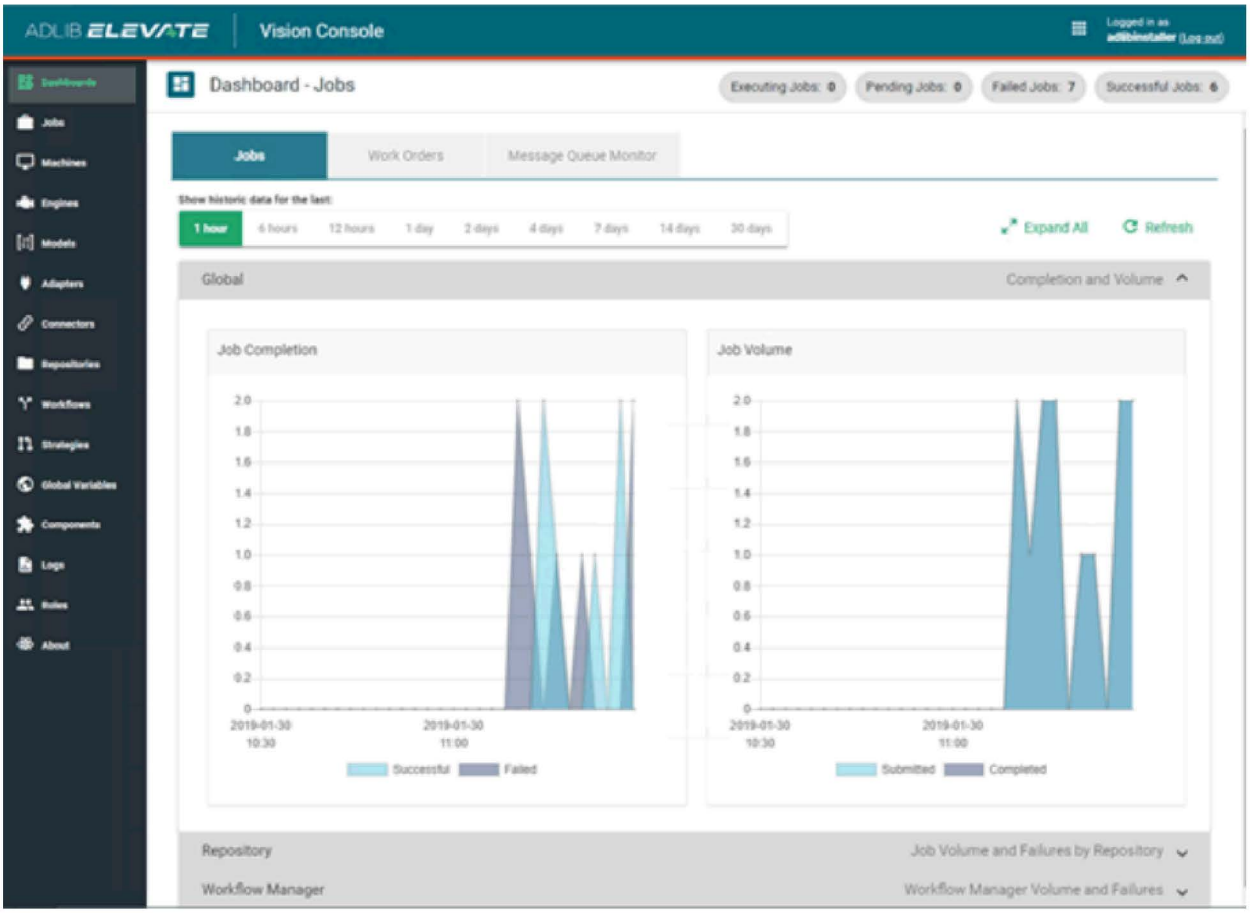
Vision Console

The Dashboard

The Dashboards page provides a series of reports that document the application's processing activity. Administrative users can utilise the dashboards to get a high-level view of the system's activity, and quickly identify potential issues.

This page is displayed upon initial login. At the top right is a brief job processing summary. Immediately following are three tabs:

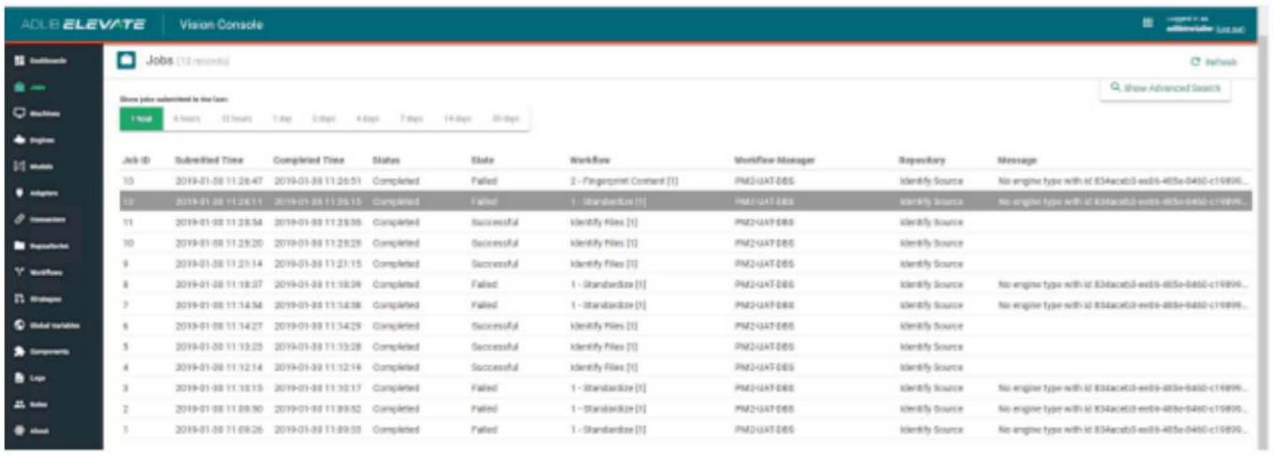
Jobs – shows a series of graphs that slice job statistics in a wide variety of ways, including job volume, processing statistics per work order, per machine, by repository and much more. Note that the headings of each area of statistics can be expanded and collapsed as required for ease of viewing.



Jobs

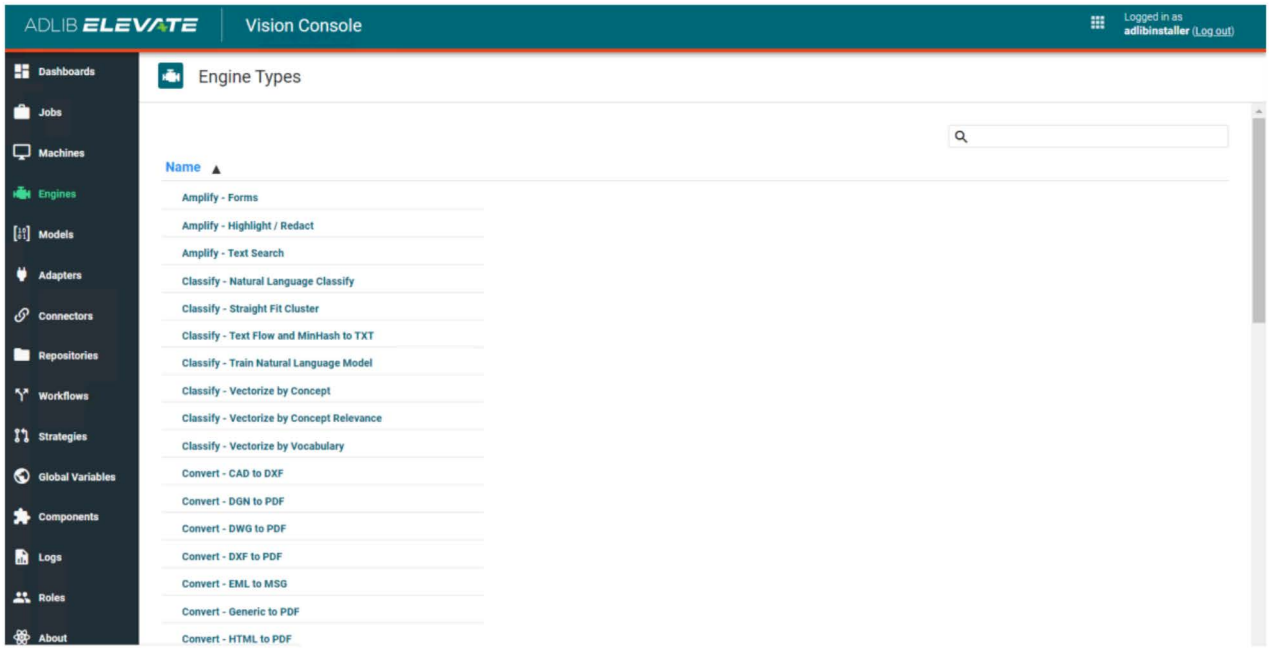
At the heart of the application is the Jobs page. Jobs are created by Connectors and Elevate applications. A Job involves one or more documents that are submitted to the system with metadata.

The Jobs page is one of the most useful pages within the UI. Here, users can view the workflow processing logs. These include not just the input and output payloads to each workflow executed in the system, but also the input and output payloads that move between the various workflow activities within a workflow.



Engines

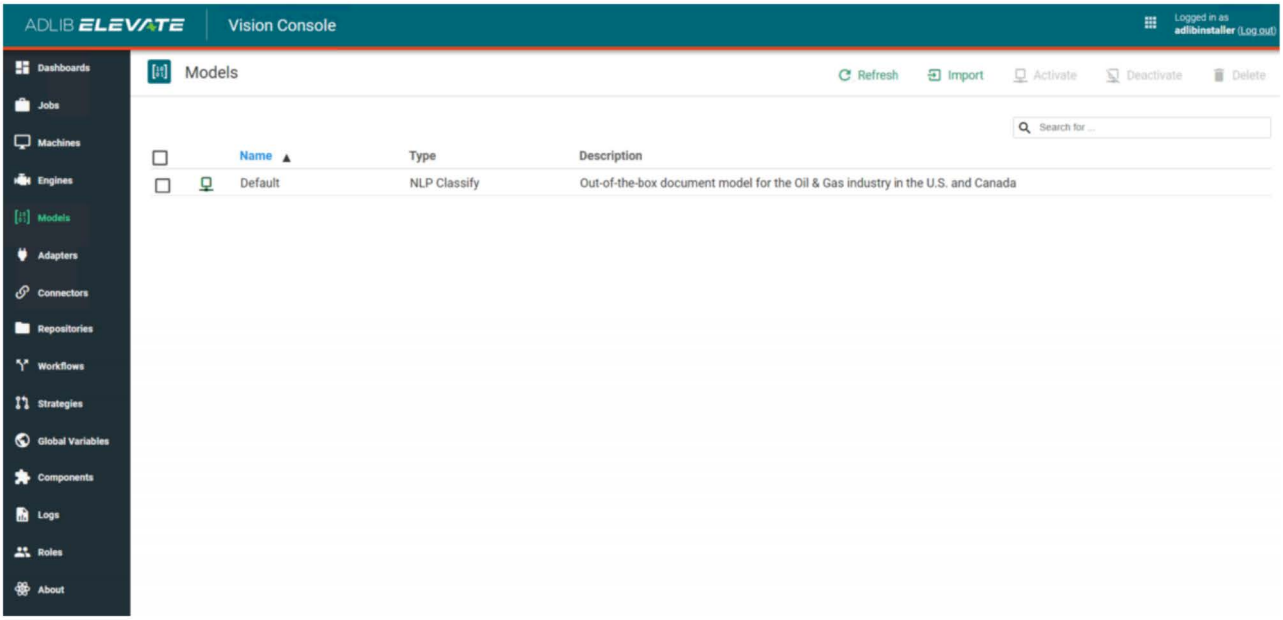
Engines are used to distribute units of process-intensive work throughout an Elevate farm (network). The Engines page lists all available Engine Types. The engines tab shows all engines registered in the system.



Models

Leveraging AI and Machine Learning technology within Elevate (via the Workbench App) requires the use of models. A Model is an algorithm or equation that has been built and trained by data scientists to automatically determine the content of a group of documents.

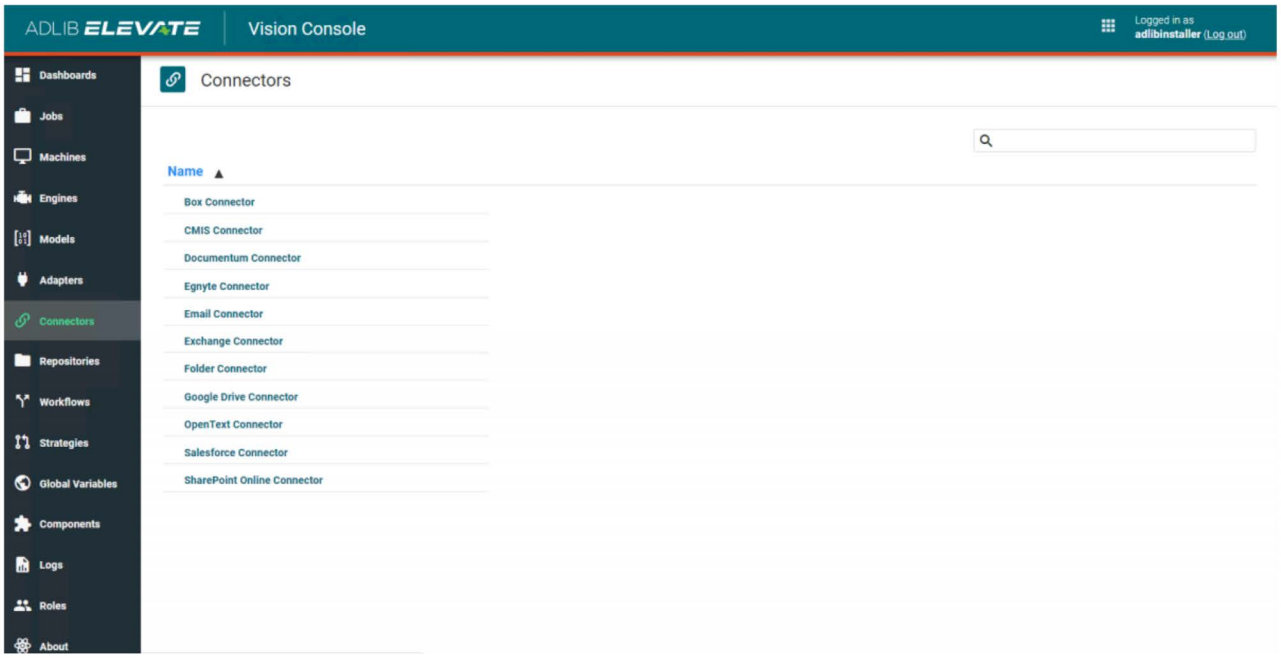
The Models page is used to manage AI models created, and to easily leverage them with user workflows.



Connectors

Connectors link a repository (e.g. Documentum) to the Elevate platform. This allows documents to flow from that repository to the platform for processing and back.

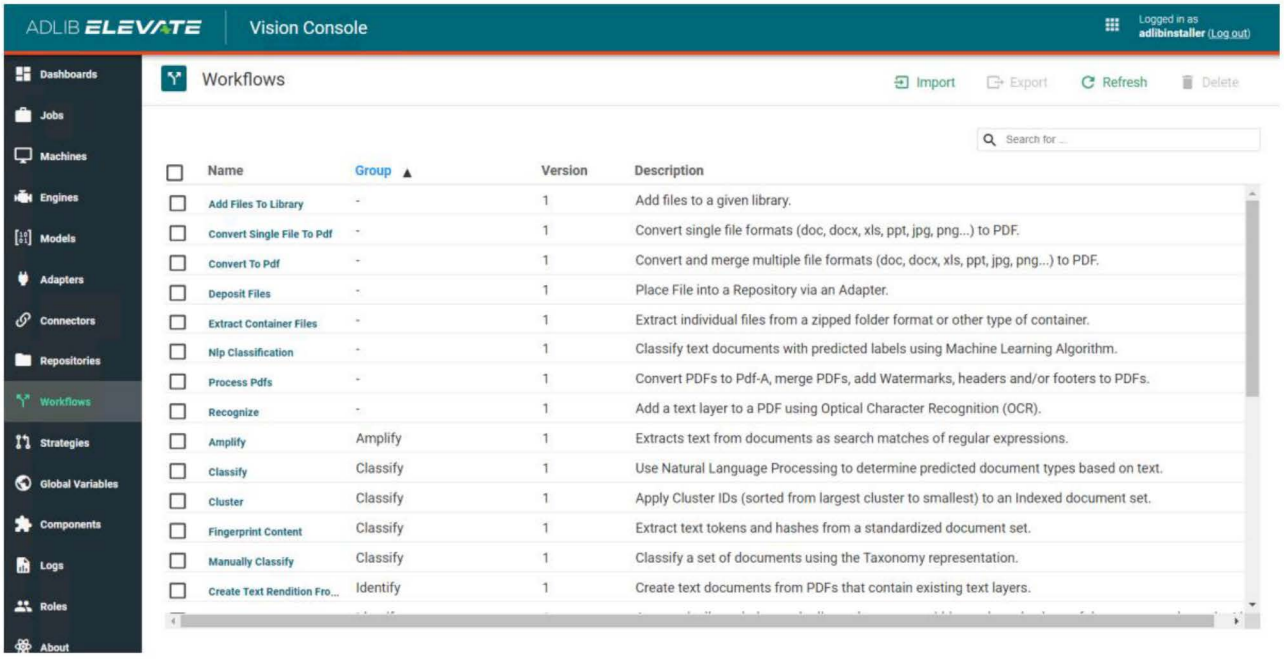
The Connectors page shows all connector available in the system. A connector is a component that serves as the start, and end, of job processing. Clicking on a connector will show you which machine it is installed on, as well as show initialisation rules and allow for them to be edited and/or added to.



Workflows

A workflow is a series of instructions, known as activities that the Workflow Manager uses to determine how a given job should be processed. Workflows are responsible for orchestrating the processing of a Job through various Engines, based on business requirements.

The Workflows page allows for Workflows to be configured. A simple workflow might consist of two engines, such as converting from Word to PDF, and then performing OCR on the converted PDF document.



A selection of workflows is immediately available for use upon installation.

Elevate Workflows

- **Add Files to Library** – Adds specified files to a given library
- **Convert Single File to PDF** – Converts single file formats to PDF
- **Convert to PDF** – Converts and merges multiple file formats to PDF
- **Deposit Files** – Places files into a Repository, using an Adapter
- **Extract Container Files** – Extracts individual files from a zipped folder or other type of container
- **NLP Classification** – Classifies text documents with predicted labels, using a Machine Learning algorithm
- **Process PDFs** – Converts PDFs to Pdf-A, merges PDFs, adds Watermarks, Headers and/or Footers to PDFs
- **Recognise PDFs** – Adds a text layer to a PDF using Optical Character Recognition (OCR)

Identify Workflows

- **Create Text Rendition from Recognised PDF** – Created text documents from PDFs that contain existing text layers.
- **Deduplicate** – Automatically excludes any duplicate documents within a selected subset of documents of the entire Library.
- **Get Library File and Put in Repository** – Retrieves files from a specific Identify Library and places it into a Repository
- **Get Library File PDF Rendition and Put in Repository** – Retrieves PDF renditions from a specific Identify Library and places it into a Repository
- **Identify Files** – Determines file types and file hashes and adds it as metadata to a file
- **Recognise PDF** – Adds a text layer to a PDF using Optical Character Recognition (OCR)
- **Standardise and Recognise** – Converts documents to PDF and then adds a text layer using Optical Character Recognition (OCR)
- **Standardise to PDF** – Converts documents to PDF format
- **Thumbnail PDF** – Creates PDF thumbnails for use in Clustering activity
- **Unpack Containers** – Recursively unpacks any archive files to create traceable Identify File Entries for each compressed file

- **Update File Metadata** – Bulk assigns/overwrites metadata values for all selected documents. Note: History of previous values is NOT kept

Amplify/Classify Workflows

- **Amplify** – Extracts text from documents as search matches of regular expressions
- **Classify** – Uses Natural Language Processing (NLP) to determine predicted document types based on text
- **Cluster** – Applies Cluster IDs (sorted from largest cluster to smallest) to an Indexed document set
- **Fingerprint Content** – Extracts text tokens and hashes from a standardised document set
- **Manually Classify** – Classifies a set of documents using the Taxonomy representation

Workbench Workflows

- **Manifest Generator** – Creates a summary/manifest of all documents within a subset for Workbench
- **Prepare Workbench Files** – Prepares files for use within Workbench
- **Test Model** – Compares model predicted labels against known labels
- **Train Model** – Runs a labelled document set against an algorithm to produce a machine-learning model

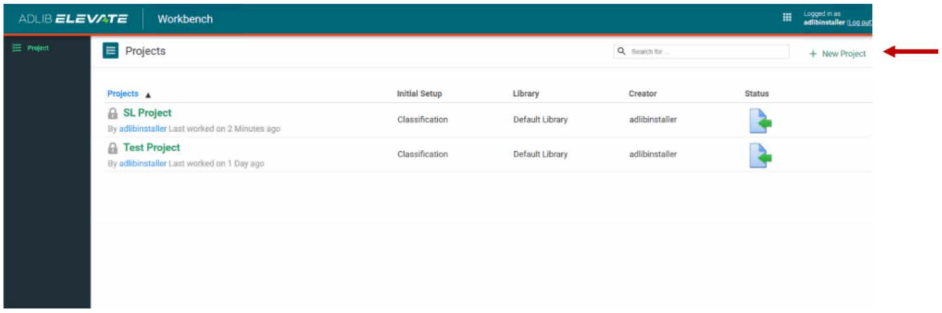
Orchestration Workflows

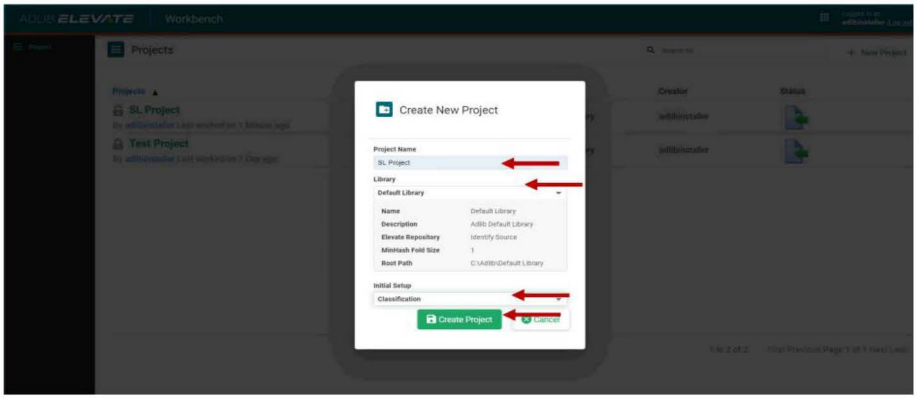
- **Start Orchestration** – Initiates an Orchestration process

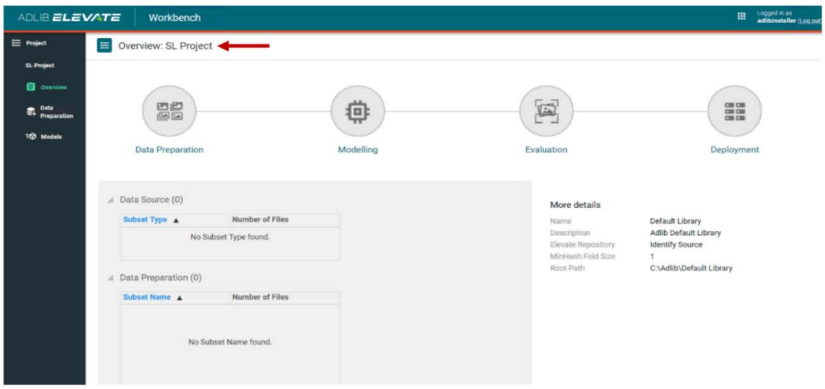
7 Appendix D: Model deployment

Creating a New Project

The first step in the process is to create a New Project within the Workbench tool.

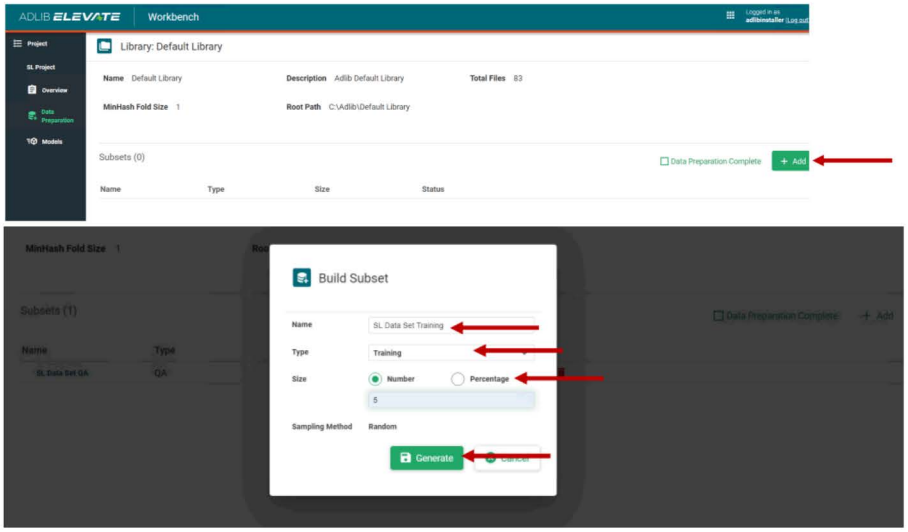


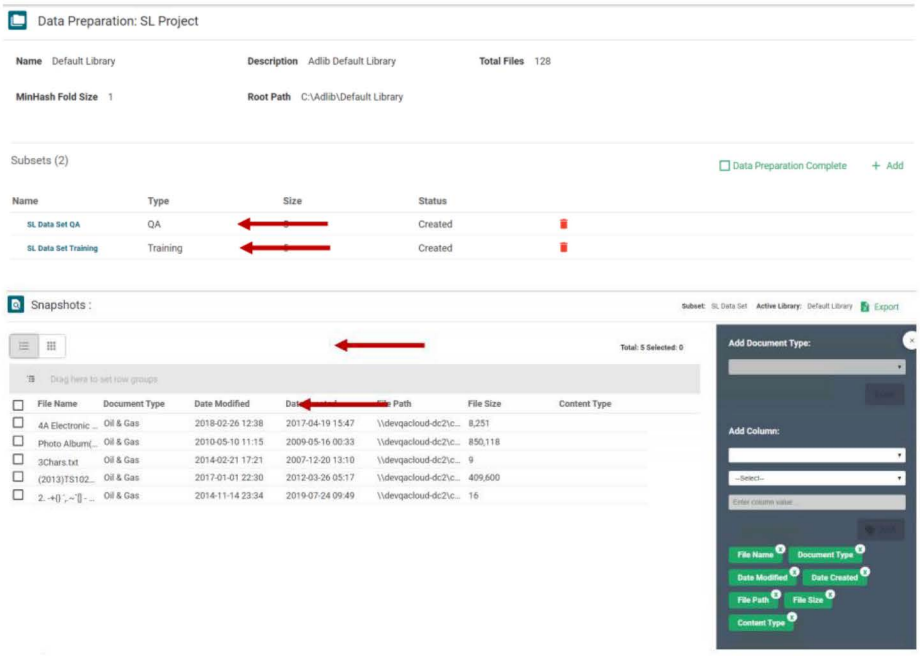


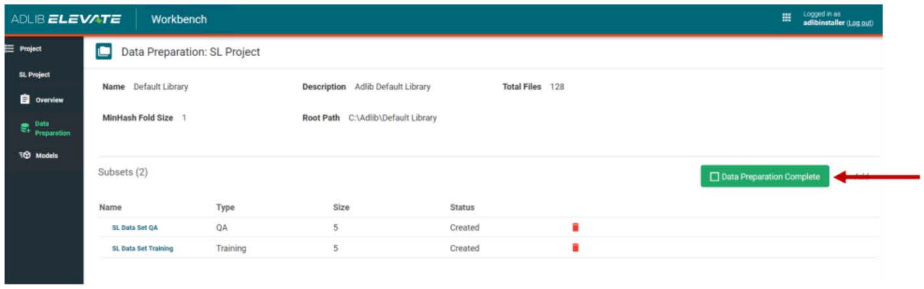


Creating Training & QA Data Sets

Once the project is created, a set of training/QA Data Sets should be defined to enable the model to be built and refined against.

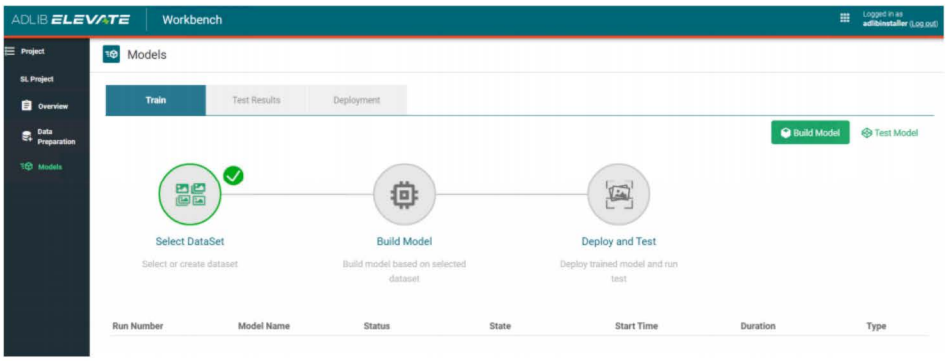


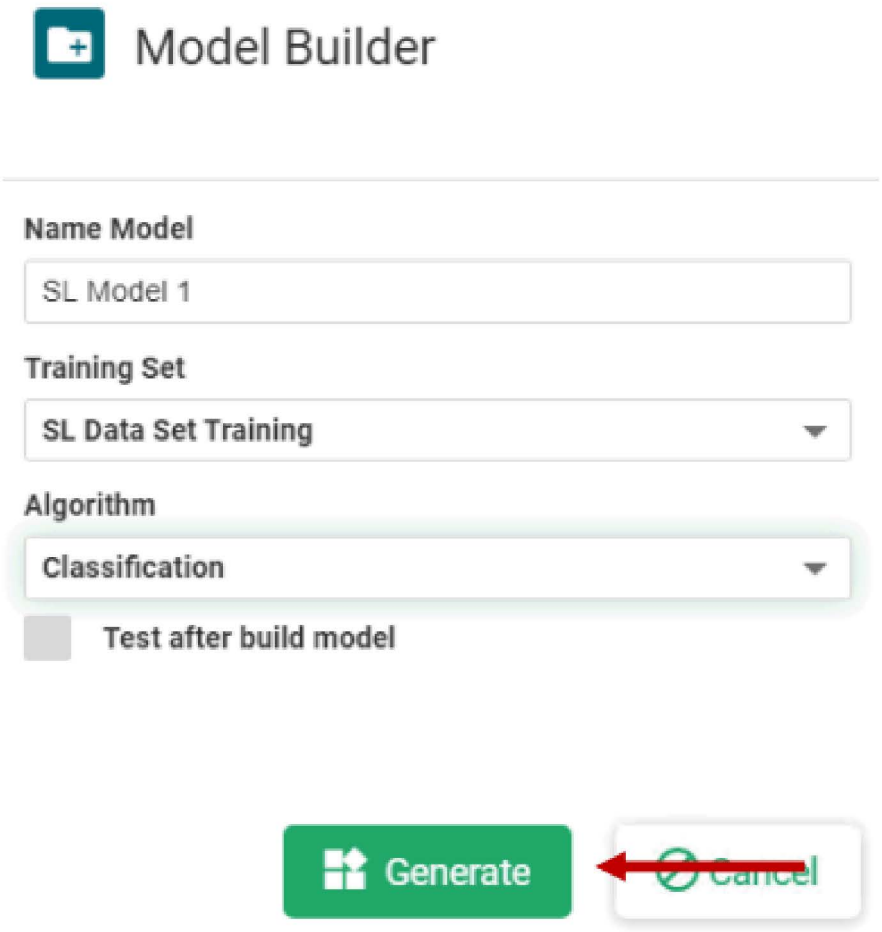


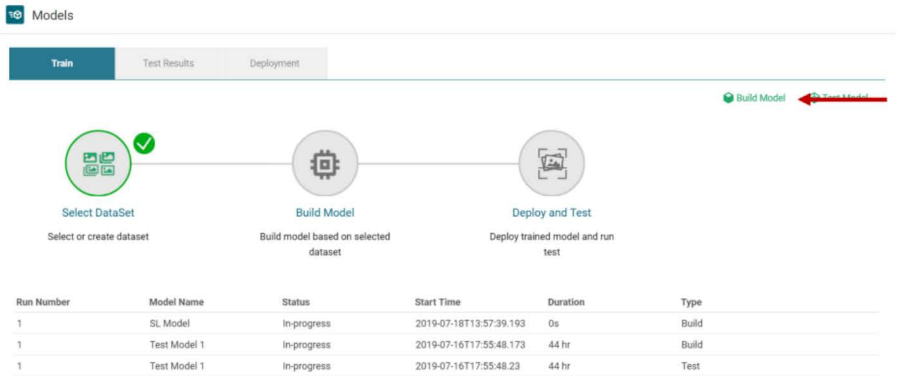


Building a Model

With Data Preparation complete and a new subset prepared, the next step is to build a model.

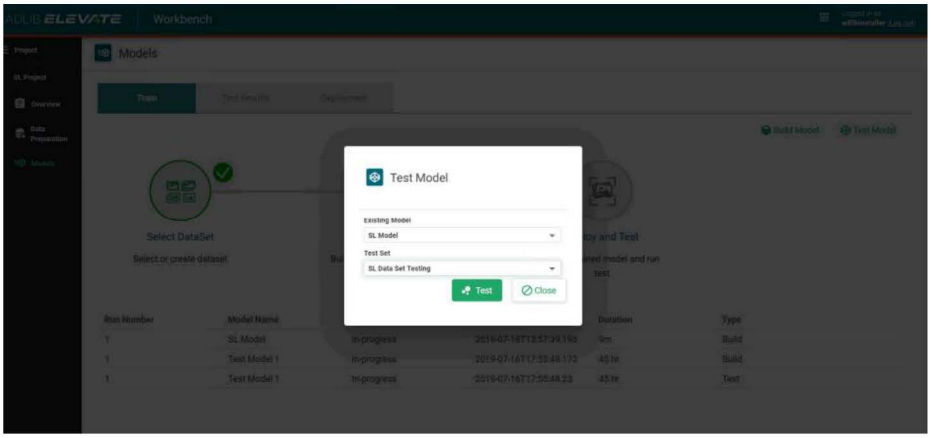

ADLIB ELEVATE
Workbench
Project
SL Project
Overview
Data Preparation
Models
Models
Train
Test Results
Deployment
Build Model
Test Model
Select DataSet
Select or create dataset
Build Model
Build model based on selected dataset
Deploy and Test
Deploy trained model and run test
Run Number
Model Name
Status
State
Start Time
Duration
Type

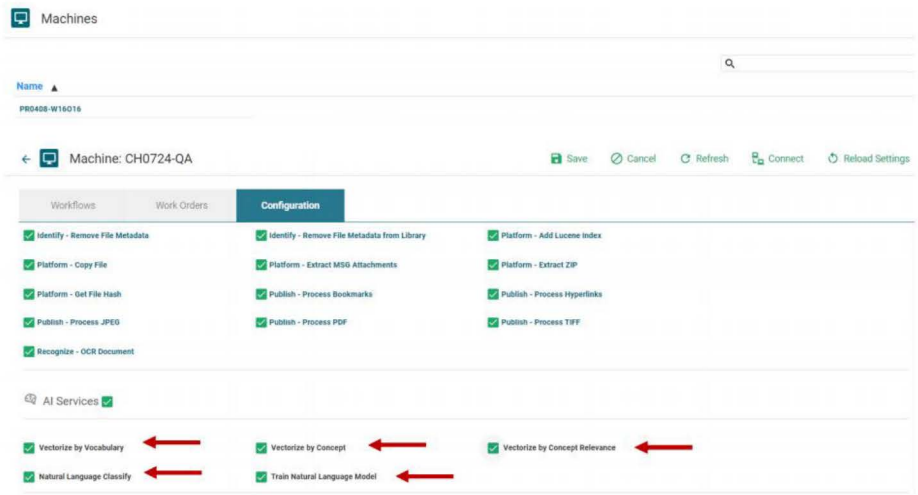

Model Builder
Name Model
SL Model 1
Training Set
SL Data Set Training
Algorithm
Classification
Test after build model
Generate
Cancel


Models
Train
Test Results
Deployment
Build Model
Select DataSet
Select or create dataset
Build Model
Build model based on selected dataset
Deploy and Test
Deploy trained model and run test
Run Number
Model Name
Status
Start Time
Duration
Type
1
SL Model
In-progress
2019-07-18T13:57:39.193
0s
Build
1
Test Model 1
In-progress
2019-07-16T17:55:48.173
44 hr
Build
1
Test Model 1
In-progress
2019-07-16T17:55:48.23
44 hr
Test

Testing a Model

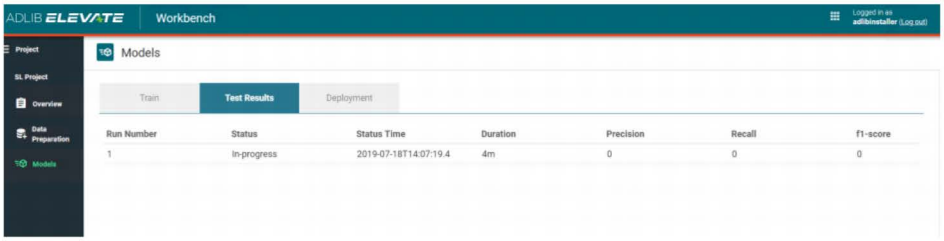
Testing a model is the process of running a subset against the model to see how accurate and precise the model is.





Reviewing Test Results

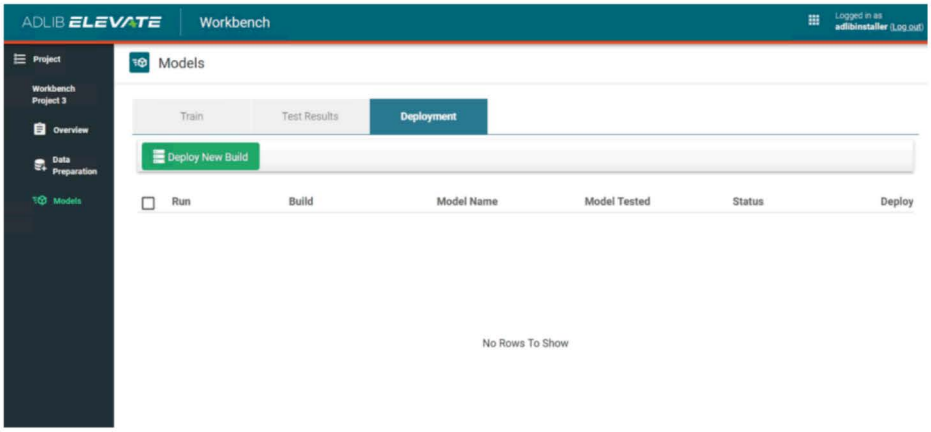
Once the model testing is complete, the results are ready for review.



The test results include a summary of the model test:

- Run Number – Test number
- Status – Progress of the test
- Status Time – The time is which the given status was assigned
- Duration – The length of test time
- Precision – The fraction of relevant instances among the retrieved instances
- F1 score – Measure of a test's accuracy
- Recall - The fraction of relevant instances that have been retrieved over the total amount of relevant
- instances

Once the optimal model has been chosen, the Model is ready to be deployed to Elevate.



The model can now be maintained from the Vision Console as outlined above and can also be consumed by jobs initiated from the Identify and Classify interface.

As you will see from the process outlined above, building models in Adlib Elevate is extremely simple and has been designed so that Records Manager or other users can create and maintain them. Adlib Elevate's Machine Learning and AI capabilities have been designed from the ground up to be able to be developed and run without needing a dedicated Data Scientist.

8 Appendix F: References

1 Ciklum. (2020). [online] Available at: https://www.ciklum.com/blog/big-data-and-the-challenge-of-unstructured-data/ [Accessed 29 Jan. 2020].

2 EY.com. (2020). [online] Available at: https://www.ey.com/Publication/vwLUAssets/ey-harnessing-insights-in-unstructured-data-through-nlp-and-ml/$FILE/ey-harnessing-insights-in-unstructured-data-through-nlp-and-ml.pdf [Accessed 29 Jan. 2020].

3 Forbes.com. (2020). What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone. [online] Available at: https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/#7ceef70615f6 [Accessed 29 Jan. 2020].

4 Ico.org.uk. (2020). [online] Available at: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf [Accessed 29 Jan. 2020].

5 Jacobson, R. (2020). IBM. [online] IBM Consumer Products Industry Blog. Available at: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/ [Accessed 30 Jan. 2020].

6 Klievink, B., Romijn, B., Cunningham, S. and de Bruijn, H. (2020). Big data in the public sector: Uncertainties and readiness. [online] Available at: https://link.springer.com/article/10.1007/s10796-016-9686-2 [Accessed 29 Jan. 2020].

7 Schneider, C. (2020). The biggest data challenges that you might not even know you have - Watson. [online] IBM. Available at: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/ [Accessed 29 Jan. 2020].

8 Woodward, A. (2020). Artificial Intelligence for Records Management | RecordPoint. [online] RecordPoint. Available at: https://www.recordpoint.com/artificial-intelligence-records-management/ [Accessed 29 Jan. 2020].

Deloitte.

This document is confidential and it is not to be copied or made available to any other party. Deloitte MCS Limited does not accept any liability for use of or reliance on the contents of this document by any person save by the intended recipient(s) to the extent agreed in a Deloitte MCS Limited engagement contract.

If this document contains details of an arrangement that could result in a tax or National Insurance saving, no such conditions of confidentiality apply to the details of that arrangement (for example, for the purpose of discussion with tax authorities).

Deloitte MCS Limited is registered in England and Wales with registered number 03311052 and its registered office at Hill House, 1 Little New Street, London, EC4A 3TR, United Kingdom.

Deloitte MCS Limited is a subsidiary of Deloitte LLP, which is the United Kingdom affiliate of Deloitte NSE LLP, a member firm of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"). DTTL and each of its member firms are legally separate and independent entities. DTTL and Deloitte NSE LLP do not provide services to clients. Please see www.deloitte.com/about to learn more about our global network of member firms.

© 2020 Deloitte MCS Limited. All rights reserved.