Document Understanding Proof of Concept – Archives New Zealand

Prepared by: Craig Hind - Public Sector Solution Architect
Version: 1.1
Date: 20 June 2022

Table of Contents

1. Problem

New Zealand government agencies generate large volumes of digital information and records. They have obligations relating to how they manage this information under the Public Records Act 2005. A key part of their overall information management process is the appraisal process which enables organisations to determine how they manage information and records over their life, how long they retain them, and when they can dispose of them. There are several disposal options one of which is transfer to Archives NZ.

With the volume of digital information being generated by agencies the following challenges arise:

1. How to determine the appropriate Disposal Authority to apply to information and records.
2. How to populate required and/or useful metadata about the information or record.

2. Proof of Concept Objectives

The 4 key business outcomes Archives NZ are seeking in the proof of concept are:

1. Correct policy for the retention and disposal of all files determined automatically
2. Information that may be of interest to Māori is identified
3. Security classification (PSR) that are automatically applied to the provided data is accurate
4. Manual intervention and input into business processes (DA) minimised

Source: Data Lakes PoC_Scope_v1.1.xlsx

3. Concepts / Terms

Artificial intelligence and machine learning (AI/ML) : Artificial Intelligence is generally defined as “creating intelligent systems that can simulate human intelligence”. Machine learning is a subset of AI and can be defined as “enabling machines to learn from past data or experiences without being explicitly programmed” .

Natural language processing : Natural Language Processing (NLP) is a way for computers to analyse, understand, and derive meaning from textual information in a smart and useful way. By utilizing NLP, you can extract important phrases, sentiment, syntax, key entities such as brand, date, location, person, etc., and the language of the text.

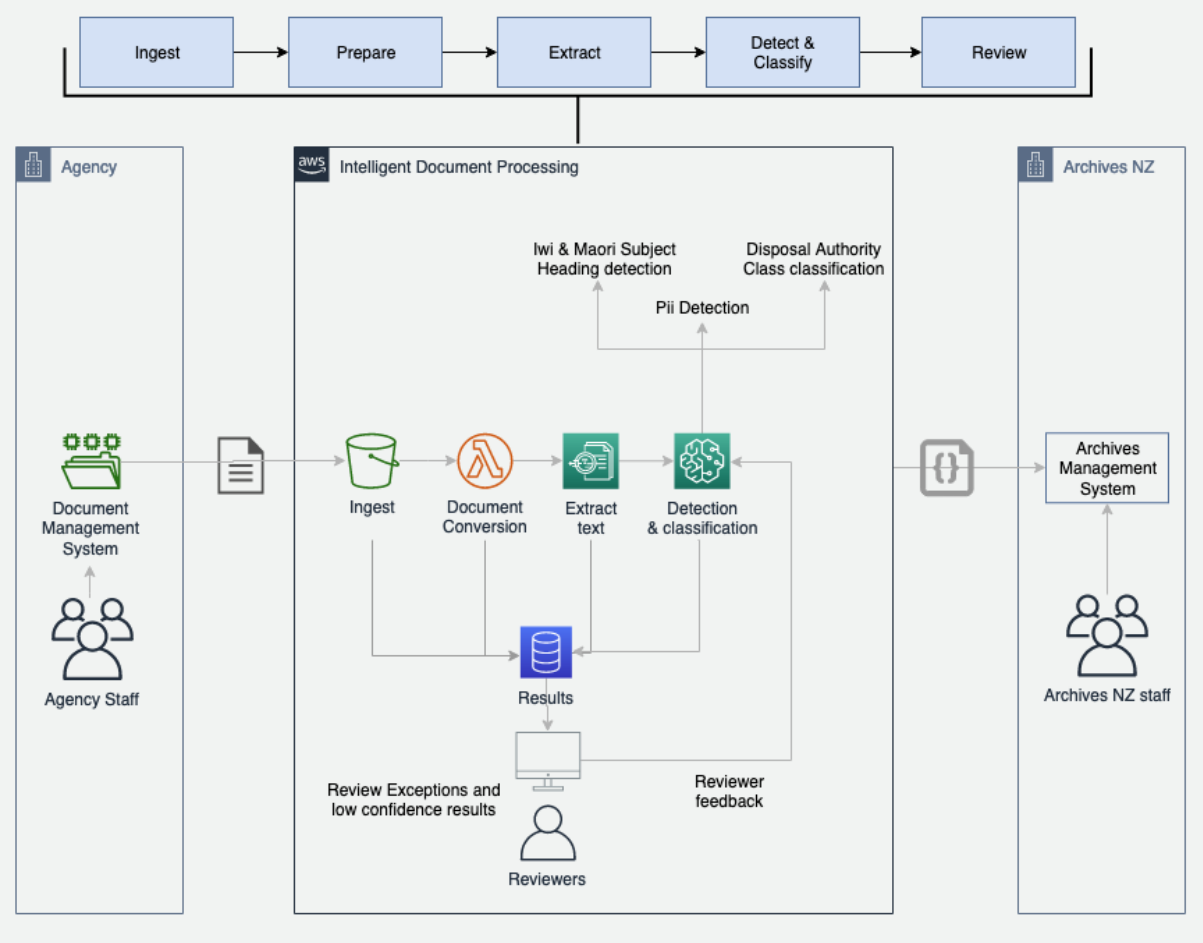
Entity: An entity is a textual reference to the unique name of a real-world object such as people, places, and commercial items, and to precise references to measures such as dates

and quantities. For example, in the text "John moved to 1313 Mockingbird Lane in 2012," "John" might be recognized as a PERSON, "1313 Mockingbird Lane" might be recognized as a LOCATION, and "2012" might be recognized as a DATE

Classification: Is the process of categorising text into groups of words, analysing that text and assigning labels.

4. Solution Design

The Document Understanding solution can be thought of as a processing pipeline. The diagram provides a high-level view of the document understanding pipeline.



The key aspects of the processing pipeline are:

1. **Ingest** – Documents will be ingested into the solution which will trigger processing. Documents will be stored within the solution post processing. Documents can be uploaded individually, or in bulk.

2. **Prepare data** – Pre-processing will occur identify the kind of document and perform any specific pre-processing required for each document. Within the PoC this will be primarily involve document conversion so it can be passed into the extraction stage

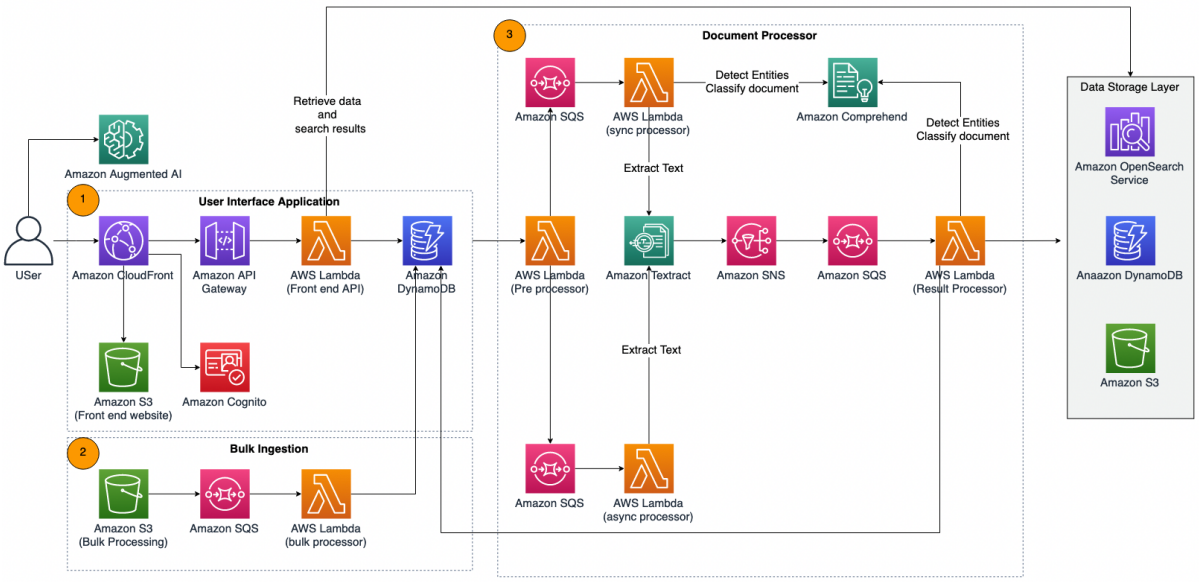
3. **Extract text** – The text from a document will be extracted prior to performing any entity recognition and classification steps. The resulting text will be passed through to the meta data extraction phase

4. **Detect and classify** – The extracted text will be passed through a number of steps to perform entity recognition, including out of the box entity types and selected custom

entity types. Custom classification will also be performed to meet specific requirements.

5. **Review** – A web user interface will allow PoC users to search for processed documents and view the meta data and classifications associated with those.

The diagram below provides a technical overview of the solution.



The PoC solution is based on an existing open source solution that is designed and built by AWS. The solution utilises a number of AWS services to provide the document understanding capability and the user interface. This includes AWS AI services to support text extraction and finding meaning and insight from text. It uses purpose-built storage including object storage for content, No SQL based storage for meta data and processing job details, and data stores designed for providing search capability.

Key aspects of this solution are:

1. **Web front end**:

The solution provides a web interface that enables users to view individual documents that have been processed by the solution. It enables users to view the Entities, tables, and keywords related to the document. For the PoC this provides features to interact with processed documents. The web application utilises several AWS services that support the development and hosting of modern serverless based web applications. These include:

1. Lambda
2. CloudFront
3. API Gateway
4. Cognito
5. Dynamo DB

The User interface provides some of the review functionality but the PoC also utilised the **Amazon Augmented AI** service to provide a workflow engine and interface for users to review and remediate predictions that did not meet the confidence threshold.

2. **Bulk Ingestion**

The solution provides a feature that supports bulk ingestion of files. Files are loaded into an S3 bucket which then triggers an event that loads each file into the solution for processing. The *Post Proof of concept* section discusses file ingestion in more detail as additional methods are likely to be required depending on the chosen operating model for the solution.

3. Document Processing

The core of the solution is the document processing engine. This provides the following capability:

Pre-processing:

Pre-processing includes tasks required to prepare a document for processing. Within the PoC solution this will determine the file type and where necessary convert the file to another format for text extraction. Once pre-processing is complete the job is added to a queue for processing. The pre-processing stage will use AWS Lambda functions to perform any business logic required. This provides an extensible design that enables additional pre-processing steps to be added as needed.

Text extraction: Before the solution natural language processing to identify specific entities or terms, or perform classification of a document the text needs to be extracted from it. **Amazon Textract** is a document analysis service that detects and extracts printed text, handwriting, structured data (such as fields of interest and their values) and tables from images and scans of documents. Amazon Textract's machine learning models have been trained on millions of documents so that virtually any document type you upload is automatically recognized and processed for text extraction. When information is extracted from documents, the service returns a confidence score for each element it identifies so that you can make informed decisions about how you want to use the results.

Language understanding: Amazon Comprehend is used to identify the language of the text, extract key phrases, places, people, brands, or events, understand sentiment about products or services, and identify the main topics from a library of documents. The source of this text could be web pages, social media feeds, emails, or articles. You can also feed Amazon Comprehend a set of text documents, and it will identify topics (or group of words) that best represent the information in the collection.

The extracted text will be passed to Amazon Comprehend which will identify the following standard items:

- **Entities** – References to the names of people, places, items, and locations contained in a document.
- **Key phrases** – Phrases that appear in a document. For example, a document about a basketball game might return the names of the teams, the name of the venue, and the final score.

- **Personally Identifiable Information (PII)** – Personal data that can identify an individual, such as an address, bank account number, or phone number.

In addition, custom entity recognisers and classifiers will be incorporated into the solution to :

- Identify Iwi and Māori Subject Headings
- Determine the correct Disposal Authority class classification.

Note that for the Proof of Concept pre-packaged AWS AI Services were used. These provided text extraction and natural language processing capabilities. Based on the learnings form the PoC in respect to entity extraction and document classifications we would likely develop custom machine learning models for these tasks. See the following sections for more detail.

5. Proof of Concept Outcomes

This section provides an analysis of the PoC's achievement of the key business outcomes taken from Data Lake PoC Scope 1.1 spreadsheet provided by Archives NZ

The table below provides a summary of the key metrics and outcomes

Documents		
	Total Provided in MPI dataset	1391
	Converted and processed	1277
Disposal Authority Class classification		
	Accuracy	88%
	Processed Documents	1277
	Time taken to classify	10 mins
Iwi Entity and Māori Subject Heading recognition		
	Iwi Accuracy	63%
	Māori subject headings	83%
	Processed Documents	1277
	Time taken	13 min
PII Detection		
	Documents processed	1277
	PII detected	~89000
	Time taken	12 min

1. Correct policy for the retention and disposal of all files determined automatically

The PoC produced a model that was able to determine a disposal authority class for a piece of content. The AWS Comprehend service was used to produce a natural language

processing model. The service abstracts the complexity of identifying the appropriate machine learning algorithm to use and the associated training and validation processes for the model. A set of training data is provided to AWS comprehend and the service produces a model that supports document classification. This service was chosen for the PoC as it simplifies the training and deployment process.

Several iterations of the model were produced during the PoC. The data included below is from the model with the best model training scores.

The matrix below shows the test against version 5 of the model using a test sample size of 318 documents. (Note 956 were used to train the model) .For each “actual” disposal authority class , as provided in the MPI meta data, it shows the count of the “predicted” class. For example, disposal authority class 1.1.2 had a total of 112 samples in the test set. Of those the model predicted the correct class (1.1.2) on 106 of those, but it incorrectly classified 6 samples (one as 1.2.1, another as 1.2.5, two as 1.3.1, and two as 4.1.1)

		PREDICTED CLASS								
	Count	Column Lab								
	Row Labels	1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	Grand Total
ACTUAL CLASS	1.1.2	106		1	1	2		2		112
	1.1.4	3	25							28
	1.2.1	2		3						5
	1.2.5	4			15	1				20
	1.3.1	11			1	100				112
	1.3.2	3					3			6
	4.1.1	1						26		27
	7.1.2	1							5	6
	8.1.1	2								2
	Grand Total	**133**	**25**	**4**	**17**	**103**	**3**	**28**	**5**	**318**

The matrix below is the same as the one above but rather than showing counts it shows percentages. For example, disposal authority class 1.1.2 had a total of 112 samples in the test set. Of those the model predicted the correct class (1.1.2) on 94.64% of those, but it incorrectly classified 6 samples representing 5.36%.

		PREDICTED CLASS								
	%	Column Lab								
	Row Labels	1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	Grand Total
ACTUAL CLASS	1.1.2	94.64%	0.00%	0.89%	0.89%	1.79%	0.00%	1.79%	0.00%	100.00%
	1.1.4	10.71%	89.29%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
	1.2.1	40.00%	0.00%	60.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
	1.2.5	20.00%	0.00%	0.00%	75.00%	5.00%	0.00%	0.00%	0.00%	100.00%
	1.3.1	9.82%	0.00%	0.00%	0.89%	89.29%	0.00%	0.00%	0.00%	100.00%
	1.3.2	50.00%	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%	0.00%	100.00%
	4.1.1	3.70%	0.00%	0.00%	0.00%	0.00%	0.00%	96.30%	0.00%	100.00%
	7.1.2	16.67%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	83.33%	100.00%
	8.1.1	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
	Grand Total	**41.82%**	**7.86%**	**1.26%**	**5.35%**	**32.39%**	**0.94%**	**8.81%**	**1.57%**	**100.00%**

The table below contains the key model metrics which are common to classification type models, for version 5 of the model. These metrics were created when the model was trained.

Metric	Score	Description
Accuracy	0.88	The number of disposal authority classes predicated exactly right when training the model. 0.88 is 88%
Precision	0.84	When the model predicts a disposal authority class it gets it correct 84% of the time. e.g. If the model predicts 1.1.2 then 84% of the time it will be correct as opposed to a false positive
Recall	0.78	When the model predicts a disposal authority class it will be a true positive 78% of the time. The remaining will be false negatives. e.g. If the model predicts that a document is not 1.1.2 but it is actually a 1.1.2 (False negative)
F1 score	0.80	Derived from both precision and recall F1 gives a balanced view that takes into account both Precision and recall metrics.

The data shows the model is performing well in some classes (1.1.2, 4.1.1) but not in others (1.3.2, 1.2.1). Additional work will be required to tune the model further to increase the accuracy level. Additional training data, especially for disposal classes with smaller sample sizes (*See proof of concept learnings section below for more detail.*) . Additional data on the model training process can be found in *Appendix 3 disposal class training data*

Recommendations for production solution:

- Engage an AWS partner with experience in developing classification machine learning models who can help increase the overall performance to the desired levels. A partner will bring the skills and experience to create, train, and fine tune the model. They will also be able to provide more insight into the model from a transparency perspective, and work with Archives on the ongoing operational aspects such as model monitoring, feedback loops, and retraining.
- Investigate moving to Amazon Sagemaker for the development, training and testing of the model. Amazon Sagemaker is a suite of services that support the full development lifecycle of machine learning models. It provides more flexibility and control over the fine tuning of machine learning models than what is available within the AWS Comprehend service used for the PoC.

2. Information that may be of interest to Māori is identified

The PoC produced a named entity recognition model that successfully identified Iwi and Māori Subject entities (keywords) within the content using the AWS Comprehend natural language processing service.

A custom model was trained with a derived data set as no pre-labelled data was provided in the samples. No Māori subject headings were specified so the top terms from the National

libraries Ngā Upoko Tukutuku [1] site were used to train the model. Some additional fisheries related terms were added given the source of the sample data set was MPI. See appendix 2 for the full list of Māori subject headings. An Iwi list was derived from the Stats NZ Iwi Classification set[2].

The performance of the model during training is shown below. This indicates the named entity recognition model is good at identifying Iwi and Māori Subjects based on the sample of training data it used for model validation.

Entity	F1 Score	Precision	Recall	Number of training mentions
Entire Model	99.32	99.09	99.54	N/A
IWI	99.49	98.99	100	2327
MAORISUBJECT	99.74	99.49	100	3208

1. *Precision answers the question "What proportion of positive identifications were actually correct?" e.g., out of 100 samples a model may predict that 70 contained an Iwi. Of those only 60 may have been correct (true positive) with the remaining 10 being false positives.*

2. *Recall answers the question "What proportion of actual positives were identified correctly?" e.g., Out of 100 samples a model may identify an Iwi in 70 samples but missed 10 additional instances. (False negatives)*

There was no pre-labelled data provided in the sample data sets therefore no automated validation against the "actual" labels was performed to determine accuracy of the model against the PoC data sets.

A manual analysis of the outputs against the Archives NZ provided data set (57 documents) showed a false positive rate of 36% for Iwi and 16% for Māori Subject heading. It should be noted that the top 10 terms account for 66% of the false positives. Note that the MPI data set provided very few results (51 documents containing Iwi and 21 containing Māori subject headings out of a sample set of 1277).

The concept of Māori subject headings will need to be defined in more detail to produce more valuable results. The model would need to be trained on a wider group of Māori subject headings instead of the 20-30 used in the PoC. Archives NZ would also need to define how that data may be used. An example is the term Tikanga which is one of the top 20 Māori Subject Headings however one document alone contained over 500 mentions of Tikanga. The question is are some terms too generic to be used to help identify information that may be of interest to Māori? An option that was not pursued in the PoC would to determine content themes within the document. For example if a document had a number

[1] https://natlib.govt.nz/librarians/nga-upoko-tukutuku/top-terms-index
[2] http://aria.stats.govt.nz/aria/?_ga=2.172651427.1491418232.1655595527-236560222.1648424807#ClassificationView:uri=http://stats.govt.nz/cms/ClassificationVersion/QjykuWT5BTJrGdVs

of references to waka, Kai moana, and Hī ika would Archives classify that as “fishing” related and try to align it to a subject area or ontology.

Recommendations for production solution:

- Engage an AWS partner with experience in developing named entity recognition models who can help increase the overall performance to the desired levels.
- Investigate the accuracy of a code based approach for Iwi extraction that does not use AI/ML.
- Develop a clearer definition of how content may be “of interest to Māori” to better inform what technology solution can be used to meet the need. This may align to the AoG ontology work (See below for details)

3. Security that are automatically applied to the provided data is accurate

This was not achieved within the PoC as the data sets provided for the PoC did not provide any security classification meta data. This meant it was not possible to train an AI/ML model to determine the security classification. Determining a security classification with the use of AI/ML technology is achievable. However this would require an appropriately sized , and representative sample data set to train and validate a model. It will be challenging to get a representative sample that includes Restricted and above classifications due to the nature of the content within those classifications. If the model cannot be trained to identify restricted and above documents then the value of having the solution classify the security classification may be diminished as it will not be able to predict if a document should have a higher security classification.

Another requirement was the ability for the solution to determine if a piece of content, or it’s meta data, should be marked as ‘restricted’ access. The primary determinate for this would be the documents security classification. This would be achievable through programmatic means (business rules in code) based on the document’s security classification. For example : If the security classification is ‘Restricted’ then set the “restricted” meta data attributes to the appropriate values.

Recommendations for production solution:

1. Review the requirement and move to an approach where we are extracting security classification from documents should it exist, rather than trying to determine the appropriate security classification to apply to the docment.

4. Manual intervention and input into business processes (DA) minimised

The PoC did not prove, or disprove, that a reduction in Human intervention would be achievable as this depends on a range or factors outside the PoC’s scope.
However, there are many proven examples of how technology , and AI/ML specifically , can perform tasks which have previously been human centric.

Tasks such as disposal authority class classification are an example of this. The model developed for the PoC could be used to automatically determine the disposal authority class for a given piece of content, therefore removing a task previously performed by an agency or Archives NZ staff member. The solution could process thousands of documents an hour where as a human would not be able to achieve the level of throughput. With the PoC we used AI/ML to perform 3 tasks. The table below summarises the time taken to process the 1277 documents in the MPI dataset.

	Time taken	Documents per second
Disposal Authority Class classification	10 minutes	~2
PII Detection	12 minutes	~1.7
Iwi and Māori subject heading detection	13 minutes	~1.6

There will still be a level of human input required to handle exceptions, to perform periodic sampling of the model outputs, and to process examples where the model outputs do not meet an agreed threshold. For example, if the confidence level of a classification is less than 95% then the document is put into an exceptions process that requires a human to review the output and modify if necessary.

5. Other outcomes

Personally identifiable Information (PII) Detection

Another PoC goal was the identification of potential PII within the content being processed. AWS Comprehend, a natural language processing service, provides a feature that supports the detection of common PII including names, dates, addresses, email, bank account, Driver Identity (i.e. drivers license number), and credit card numbers.

AWS Comprehend extracted 89,000 potential pieces of PII from the 1277 MPI documents that were passed through the PII detection service. The table below provides a break down by PII type. Note that because this is a global service there are some country specific types it supports.

Row Labels	Count
ADDRESS	6097
AGE	3
AWS_SECRET_KEY	3
BANK_ACCOUNT_NUMBER	105
CA_SOCIAL_INSURANCE_NUMBER	4
CREDIT_DEBIT_NUMBER	4
DATE_TIME	49259
DRIVER_ID	7
EMAIL	55
IN_NREGA	6
IP_ADDRESS	22
LICENSE_PLATE	4
MAC_ADDRESS	6
NAME	32298
PASSWORD	3
PHONE	1184
UK_NATIONAL_HEALTH_SERVICE_NUMBER	8
UK_NATIONAL_INSURANCE_NUMBER	1
UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER	1
URL	233
US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER	2
USERNAME	47
(blank)	
Grand Total	**89352**

Key observations from the analysis of the PII results include:

1. PII detection for bank account and credit card types included false positive identifications when assessing text extracted from spreadsheets. The key cause being that when data is extracted from a row in a spreadsheet into plain text the result can look like credit card or bank number.
2. Name and Date types had very high detection rates but the usefulness of the output may be limited. For example the identification of a date could be a birth date, which is PII, however the majority are likely to not be birth dates.
3. There was no specific test data for NZ forms of PII such as driver license, passport, National Health Index (NHI), or IRD number so it was not possible to determine if these would be detected and what type they would be detected as.

Recommendations for production solution:

- Prepare specific PII data set including NZ specific PII types (Driver's license, passport, NHI,IRD etc) and retest AWS Comprehend PII model to determine ability to accurately identify NZ specific kinds of PII data.
- Optionally investigate if a custom model may perform better at identifying New Zealand specific PII than the general AWS Comprehend PII detection feature
- Identify what are the key kinds of PII that Archives NZ / agencies are looking for and investigate if additional context is required to determine if the PII detection needs to be followed up on as part of a human based workflow. For example if a lower confidence credit card number detection occurs as the source file type was a spreadsheet then should that be raised for review?

Algorithmic transparency

The use of Artificial Intelligence and machine learning (AI/ML) in Government is increasing as is the need to have appropriate policies and processes in place that govern it's use. One aspect is transparency in terms of understanding how the model works. There are two key terms used:

1. **Interpretability** — If a business wants high model transparency and wants to understand exactly why and how the model is generating predictions, they need to observe the inner mechanics of the AI/ML method. This leads to interpreting the model's weights and features to determine the given output. This is interpretability.

2. **Explainability** — Explainability is how to take an ML model and explain the behaviour in human terms. With complex models you cannot fully understand how and why the inner mechanics impact the prediction. However, through model agnostic methods you can discover meaning between input data attributions and model outputs, which enables you to explain the nature and behaviour of the AI/ML model.

The PoC uses AWS AI services, specifically Amazon Textract for text extraction , and Amazon Comprehend for entity recognition (PII, Iwi, Māori subject headings) and document classification (disposal authority class). These services abstract away the complexity of machine learning model development, however provide less transparency in terms of their inner workings as that is proprietary to a large extent. We can gain a level of explainability but not a deep understanding. We do not know what specific algorithm was selected, or what model parameters were used.

If the level of explainability or interprebiltity required is higher than what can be derived from the available information about packaged AI services such as AWS Comprehend then developing models from scratch would provide more options for model explainability and interpretability. Within the AWS machine learning ecosystem the Amazon SageMaker Clarify service can be used as part of the end to end model development process to get greater visibility into models so you can identify and limit bias and explain predictions.

Recommendations for production solution:

- Apply an appropriate risk based assessment of how AI/ML is being used within the solution and use that to determine the level of transparency required. The level of transparency required will inform the technology choices we make for capabilities like entity recognition and document classification.
- If custom models were developed to achieve higher levels of accuracy we would utilise the Amazon Sagemaker product family for this which provides features to explain the models as we would be in control of the full model development and training process.

All of Government Ontology

The PoC did not explore the application of AI/ML and ontologies due to time constraints. There is a lot of opportunity to leverage AI/ML to automatically classify documents to the proposed AoG ontology.

Recommendations for production solution:

- This is an area that should be focused on early in parallel to the development and adoption of an AoG Ontology so that the technology is ready and proven when organisations begin to adopt the ontology.

6. Proof of Concept learnings

This section details the key learnings which will need to be taken into consideration as a path to production is planned.

Data Preparation

A significant amount of time was spent preparing the sample data sets for two key parts of the PoC. 1) Converting documents to text and 2) Preparing training data for the AI/ML models.

Document conversion - The AWS service used for the entity recognition and classification tasks needs plain text as the input, therefore text extraction from the source document is a critical step. The MPI sample set included a variety of file formats including older Microsoft Office formats (.doc, .xls, .ppt) and more obscure formats (See appendix 1). The source files had to be converted to formats that the text extraction service, AWS Textract, could process. A separate programme was written to convert older formats of Office documents and Word-Perfect documents to PDF. The more obscure formats could not be converted given the PoC timeframes. The solution will need to incorporate a document conversion function for a core set of common file formats. It will require a feature to allow unknown , or unsupported file types, to be raised as exceptions and bought to the attention of users.

Also given that the transfer process may be dealing with documents 5, 10, 20 years old consideration needs to be given to how we can support converting documents created in applications that are now obsolete. See *Post Proof of Concept Considerations* section below for additional detail on this.

Training Data – Entity recognition - The preparation of training data can take significant amounts of time. The custom entity recognition model that identified Iwi and Māori Subject headings required a set of training data had to be generated. This involved writing a programme to extract text from a set of sample documents, and loop though the results (~400,000 lines) to identify lines that contained an Iwi or one of the Māori subject headings. This data was then used to prepare training data sets for the Entity Recognition model. Archives NZ will need to consider this process when identifying the Entities that the solution

should be capable of extracting from content. Each additional entity type will require additional time to prepare training data sets , and go through the model training and testing process.

Training Data – Disposal Class classifier - The custom classifier that was produced to determine the disposal authority class required the use of the MPI provided data set to train the model. The initial sample size was 1391 documents, of which we could use 1277. 114 documents could not be converted due to file type, file corruption, password protected files, and unexplained conversion errors. The 1277 usable documents were split into a training set , to train and test the model, and a validation set which was used to produce the results that Archives NZ validated as part of their test process.

The MPI data set included several classes with smaller sample sets including 8.1.1, 8.1.3, 1.2.1 . (See table below for full breakdown) The result of this is fewer samples of those classes in the training set even when we ensured a representative sample was included. For example, for class 8.1.1 the training set included 5 of the 7 samples. Of those 5 training samples 4 of them were used to train the model and 1 was used to test the model. The remain 2 were in the validation set. This demonstrates the importance of having good quality data sets for model training.

Row Labels	Count of Object Name
DA613	**545**
1.2.1	20
1.3.1	495
1.3.2	30
GDA6	**154**
4.1.1	114
7.1.2	30
8.1.1	7
8.1.3	3
GDA6	**692**
1.1.2	491
1.1.4	122
1.2.5	79
(blank)	
(blank)	
Grand Total	**1391**

AI/ML service selection - AWS provides a range of AI and Machine learning capability to support a wide range of AI/ML use cases. The PoC used AI services which abstract away the underlying complexity of algorithm selection, model tuning, and model training. These AI services make it easier to incorporate AI capability into applications and reduce the time required to do so. For example to determine the disposal class authority we provided the

AWS Comprehend natural language processing service with a set of training data and specified that we wanted it to determine the disposal authority class. The AWS Comprehend service then went through the process of algorithm selection, tuning and training. This produced good results in terms of model accuracy however because the AI service abstracts the complexity away we have less options to fine tune the model to improve accuracy levels.

To improve the accuracy of the model we will pivot from using the AWS Comprehend service to developing a model using Amazon Sagemaker. Amazon Sagemaker is an AWS service that support the end to end development lifecycle of a machine learning model. It provides more flexibility for organisations over all aspects of model development. The move to developing a bespoke model will require specialist machine learning and data science skills.

7. Post Proof of Concept Considerations

This section outlines items that will need to be considered after the Proof of Concept when planning a path to production.

Operating model considerations

Archives NZ will need to identify an appropriate operating model for the solution. The PoC is a custom built solution as opposed to a commercial packaged solution. There will need to be a lead agency responsible for the development, operation , and ongoing evolution of the solution.

The solution could be offered “as a service”. In this approach Archives NZ would design, implement, and operate the solution and offer it as a consumable software as a service solution. Agencies can then choose to consume this service to solve a business need within their agencies. Archives NZ can also use it for their own purposes such as validating the data transferred by agencies. An appropriate commercial model would be designed cover the cost of the solution which may involve the solution having a billing feature that attributes costs to subscribed agencies and generates invoices. This model will require Archives NZ / DIA to have the required people to develop and operate what is essentially a Software as a Service product. The skills required would include front end developers, back end developers, and Infrastructure engineers for the application components such as the web front end and the processing engine. It would include data scientists for the development, tuning, and ongoing evolution of the machine learning models. Additional skills would include product management, product support / helpdesk , and billing management. Archives could partner with a 3rd party software company to outsource most of the requirements such as software development, data science, and support while retaining core responsibilities such as Product management, and billing management.

Alternatively each agency could have their own instances within their own environment with the associated implementation costs and on-going operational costs. An open source

software type approach could be used in this model. Archives NZ could be the lead agency and develop the solution themselves if they have the required skills, or engage a 3rd party software company to build out the solution on their behalf. Consideration needs to be given to the size of the open source community that could be built up around such as solution. The burden of maintenance and ongoing evolution of the solution may fall on Archives NZ if other agencies do not contribute in a meaningful way.

While these items clearly sit with Archives NZ the decision on operating model will potentially impact the design and architecture of the solution. For example if the solution is going to be offered "as a service" then the solution will need to cater for multiple tenants, and have a cost attribution feature, and possibly a billing feature.

Solution functionality considerations

The Proof of Concept had a narrow focus on proving the ability of AWS services to meet specific PoC requirements. The path to production will need to take into account several additional solution capabilities that have not been addressed during the PoC.

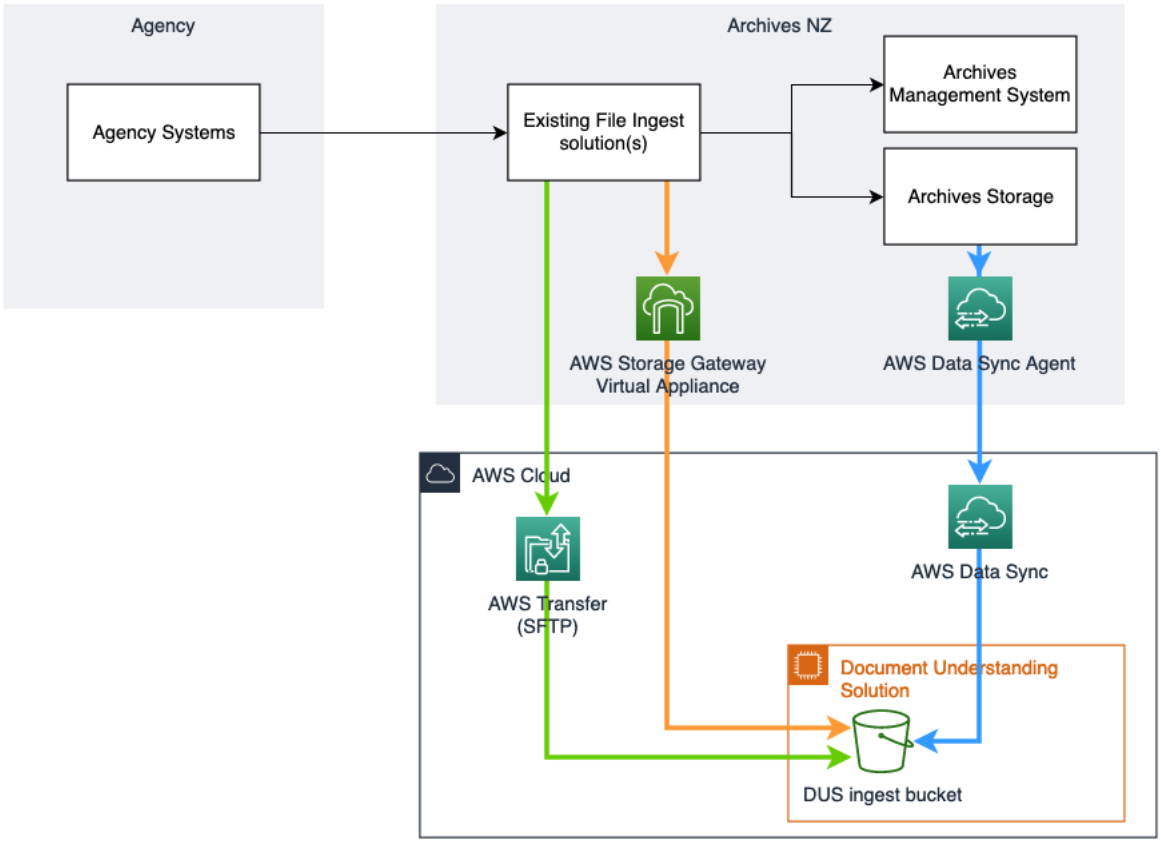
Metadata extraction –. The PoC solution did not include extracting meta-data passed in from agencies with the content. For example, filename, file type, creation data, keywords etc. It is assumed that if meta data is passed in with the documents, then Archives has an existing way of extracting that and associating it with the document. If Archives NZ wants the document understanding solution to also extract the passed in meta-data and incorporate that with the system identified meta data (e.g. Iwi, hapu, Māori subject headings) then this will need to be incorporated into the overall design.

Human intervention – Archives NZ will need a mechanism for reviewing the outputs of the solution and determining if human intervention is required. For example, if the retention and disposal authority classification confidence level is lower than an agreed percentage then it requires human review. It will also need a way to perform administration tyope functions and to review system metrics and reports. The PoC demonstrated the use of the Amazon Augmented AI service as an example of the interface for reviewing classification results. Consideration will need to be given to the tasks that people will need to perform in the system and how the user interface part of the solution can be best support those.

Metrics - The solution will need to provide Archives NZ and/or agencies with insight and visibility into the classification processes and outputs. Insights will provide Archives NZ with key business metrics so they can see what the solution is doing. Examples of key metrics may be number of documents processed, Documents processed by source agency, model accuracy / confidence levels, errors encountered. We will need to work with Archives NZ to identify the most appropriate metrics that need to be monitored.

File Ingestion – The PoC solution provides 2 methods of ingesting files: a web-based interface for single file upload and a bulk upload to S3. In a production context the primary

ingestion approach would be bulk ingestion. When moving to production consideration needs to be given to how the files will flow from the source agency to the solution, should the solution be hosted by Archives. The diagram below shows a high-level view of 3 possible options that assume Archives will be hosting the solution.



- AWS Transfer (Secure FTP) - Files can be ingested into the AWS solution via secure FTP. AWS Transfer is a managed SFTP service that can be provisioned to allow archives to send files to the solution via FTP. This could leverage existing managed file transfer solutions that Archives NZ already use.

- AWS Storage Gateway – This service allows organisations to create a file system gateway on premise. On Premise servers or systems can then access the gateway using SMB or NFS protocols. AWS Storage Gateway then coordinates the upload of the documents to AWS.

- AWS Data Sync allows organisations to easily transfer content stored in on-premise files systems to AWS. The agent can monitor a source file system and when files are added or changed it will replicate those to AWS.

Document type support – Files may have a range of different file types. Additional work will be required in the pre-processing stage to adequately handle the variety of different file types. Common file formats will need to be incorporated into the solution from day 1

however consideration will need to be given to how additional file formats are incorporated into the system. Also given the time periods that Archives deals with consideration will need to be given to how we support applications from 10-25 years ago.

Integration with existing systems –If the solution is intended to be hosted by Archives NZ, as opposed to the agencies themselves, then consideration needs to be given to how the extracted meta data gets into other Archives NZ systems. The Document Understanding Solution will produce outputs from the classification's steps. The solution could integrate with Archives NZ systems by calling existing API's or by preparing a data file that could be ingested by an existing system. If the solution is to be hosted by individual agencies then the current approach to transferring data to Archives NZ will continue to be used therefore no additional integration should be required.

8. Pricing

This section outlines a pricing model for the solution outlined in the *Solution Design* section above . It provides details on the various cost components and key considerations for those. It should be noted that AWS has a "pay for what you use" pricing philosophy. Therefore if you use less then you pay less. The estimates below provide an indicative view of projected AWS run costs given the set of estimated volumes / usage patterns.

Web Application

The web application provides the user interface for users to perform administration and review type functions. This part of the solution is mainly built using a serverless architecture along with a small search cluster. The pricing estimate is based on the following assumptions:

Web site requests per month:	1 million
Data served by web application to end users per month	1000 gb
Active users per month :	1000
Database reads/writes per month:	1 million / 1 million

Approximate monthly cost:	
Approximate cost 12 months:	

Note that a portion of the monthly cost , approx. $XXX NZD , is fixed cost for a search cluster which provides the document search function for the web user interface. The remaining costs are variable based on the web site usage.

Document Processing engine

The document processing engine contains the components that process documents and perform the text extraction, standard entity recognition, PII detection, custom entity extraction, and document classification steps. The cost of the document processing is proportional to the amount of documents processed and the size of those documents.

The key contributors to the costs for document processing are custom models for entity recognition (Iwi, Māori subject heading) and classification (Disposal class authority) .

No statistics were available so the following estimates of volume have been used to calculate the price.

Documents processed per month	100,000
Average pages per document	10
Pages processed per month (Documents * average pages)	1,000,000
Average Characters per page	2,500
Average document size	1 mb
Total Storage required per month	10TB

Approximate monthly cost:	TBC
Approximate cost 12 months:	TBC

Pricing Assumptions:

- Prices are estimates and not a quote or commitment
- All pricing is based on public pricing for the AWS Sydney region
- All pricing is in USD and has been converted to NZD using an exchange rate of $1.59
- Several AWS services include free tiers into their pricing. For transparency ,“free” tiers have been excluded from the pricing above, so the actual costs incurred may be lower once free tiers are applied.
- The pricing is based on the PoC solution architecture outlined in the Solution Design section with the exception that AWS Comprehend has been substituted for Amazon SageMaker. Amazon SageMaker is the AWS service that will enable more control over model training and tuning.
- Actual costs will depend on actual usage levels therefore costs will vary.
- Assumes machine learning “models” will be running 24 hours a day . Depending on the document ingestion pattern machines can be scaled down in periods of little or no inference which will reduce costs.
- No discounts have been applied.

The costs above are for business as usual operating costs of the AWS services used by the solution. In addition the project would need to determine costs associated with:

Solution development : Developing the solution to a point where it is production ready with the desired set of functionality. This would incorporate project activities such as requirements analysis, design, build, test , operational readiness, and project governance.

Operations : Once the solution goes live there will be operational costs over and above the AWS platform run costs. This would include application support, application and security monitoring, and application updates/deployments. There would also be ongoing

monitoring and refinement of the machine learning models by data scientists although this would not be a full time requirement.

Product Management: If a software as a service operating model is chosen there will be costs associated with managing the product. This would include product management, back log management, billing management, and sales/marketing management for example.

9. Appendices

Appendix 1 Data Preparation

The table below shows the last 3 characters of the “object” name. Shows 14 objects did not have a file extension (the ‘No’ group) and within the “Yes” group a range of obscure file formats such as ACA, AF#,DR1 and STA for example.

Row Labels	Count
no	**14**
V3	1
V4	1
000	5
028	2
100	1
199	2
799	1
y99	1
yes	**1377**
ACA	1
AF3	2
doc	923
DR1	1
gif	7
jpg	4
LDA	1
mpp	24
ppt	115
STA	1
wpd	58
xls	240
(blank)	
(blank)	
Grand Total	**1391**

Appendix 2 Māori Subject Headings

The following Māori subject headings were used to train the Māori Subject entity recognition model. Note that upper case and lower case versions were used as well as with and without macrons.

Ao wairua	Tangaroa	tikanga tuku iho

Atua	tapu	ture
Hauora	tiaki	tāngata
Mahi toi	Mātaitai reserve	tāngata whenua
Mātauranga	Tangata Kaitiaki	tōrangapū
Ohaoha	Mātaitai reserve	waka
Pakanga	Tangata Kaitiaki	whakapapa
Pāpāho	Mātaitai	whakapono
Pūtaiao	Mataitai	kai moana
Reo Māori	Taiāpure	rahui
Taiao	Taiapure	tahatai
Taonga	ao wairua	tangaroa
Tikanga	atua	tapu
Tikanga tuku iho	hauora	tiaki
Ture	mahi toi	mātaitai reserve
Tāngata	mātauranga	tangata kaitiaki
Tāngata whenua	ohaoha	mātaitai reserve
Tōrangapū	pakanga	tangata kaitiaki
Waka	pāpāho	mātaitai
Whakapapa	pūtaiao	mataitai
Whakapono	reo māori	taiāpure
kai moana	taiao	taiapure
Rahui	taonga	moana
tahatai	tikanga	

Appendix 3: Disposal Authority class training data

This section contains outputs from the model training process for the disposal class classification model. Each version includes 3 data points:

1. **Sample distribution across classes.-** Pivot table displays the distribution of samples across the classes.
2. **Actual vs prediction count matrix** - Displays the count of predictions for each "actual" disposal authority class in the training test data set. Should be read like this:

   When the "Actual" class was 1.1.2 the model predicted the class shown in the columns. So in the example below the model predicted the class was 1.1.2 26 times. It incorrectly predicted the class was 1.2.5 and 1.3.1 once each.

	PREDICTION									
	1.1.2	1.1.4	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	TOTAL	Sample %	Correct
1.1.2	26	0	1	1	0	0	0	28	36%	93%

3. **Actual vs Prediction percentage** – Displays the percentage of predictions for each "actual" disposal authority class in the training test data set.

   Should be read like this:

   When the "Actual" class was 1.1.2 the model predicted the class shown in the columns. So in the example below the model predicted the class was 1.1.2 92.86% of the time. It incorrectly predicted the class was 1.2.5 and 1.3.1 3.57% of the time. This gives us a view of the accuracy of the model predicting each disposal authority class.

	PREDICTION							
	1.1.2	1.1.4	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	Sample Count
1.1.2	92.86%	0.00%	3.57%	3.57%	0.00%	0.00%	0.00%	**28**

Version 1: Automatic split of 777 sample file into training and testing (60/40)

Key thing in this version is the unbalanced sample. The model test sample excluded 3 disposal authority classes.

Row Labels	Count	%
1.1.2	272	35.01%
1.1.4	70	9.01%
1.2.1	12	1.54%
1.2.5	46	5.92%
1.3.1	270	34.75%
1.3.2	18	2.32%
4.1.1	73	9.40%
7.1.2	11	1.42%
8.1.1	5	0.64%
(blank)		0.00%
Grand Total	**777**	**100.00%**

		Prediction									
		1.1.2	1.1.4	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	TOTAL	Sample %	Correct
Actual	1.1.2	26	0	1	1	0	0	0	28	36%	93%
	1.1.4	0	5	0	0	0	0	0	5	6%	100%
	1.2.5	0	0	4	0	0	0	0	4	5%	100%
	1.3.1	1	0	1	28	0	0	0	30	39%	93%
	1.3.2	2	0	0	1	1	0	0	4	5%	25%
	4.1.1	0	0	0	0	0	5	0	5	6%	100%
	7.1.2	1	0	0	0	0	0	0	1	1%	0%
									77	100%	
	NOTE:	Random train / test split did not incl;ude 1.2.1 and 8.1.1 and 8.1.3 in the test group									

		Prediction							
		1.1.2	1.1.4	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	Sample Count
Actual	1.1.2	92.86%	0	1	1	0	0	0	28
	1.1.4	0.00%	5	0	0	0	0	0	5
	1.2.5	0.00%	0	4	0	0	0	0	4
	1.3.1	3.33%	0	1	28	0	0	0	30
	1.3.2	50.00%	0	0	1	1	0	0	4
	4.1.1	0.00%	0	0	0	0	5	0	5
	7.1.2	100.00%	0	0	0	0	0	0	1
									77
	NOTE	Random train / test split did not incl;ude 1.2.1 and 8.1.1 and 8.1.3 in the test group							

Version 2: Manual split of 777 sample file into training (698) and testing (79) for more representative distribution of samples across class labels

This model retained the same training sample size off 777 files but used a 90/10 split for training data and the testing data.

Row Labels	Count	%
1.1.2	244	34.96%
1.1.4	63	9.03%
1.2.1	11	1.58%
1.2.5	41	5.87%
1.3.1	243	34.81%
1.3.2	16	2.29%
4.1.1	66	9.46%
7.1.2	10	1.43%
8.1.1	4	0.57%
(blank)		0.00%
Grand Total	**698**	**100.00%**

		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	TOTAL	Sample %	% Correct
Actual	1.1.2	24	1	0	1	1	0	0	0	0	27	35%	89%
	1.1.4	0	7	0	0	0	0	0	0	0	7	9%	100%
	1.2.1	0	0	1	0	0	0	0	0	0	1	1%	100%
	1.2.5	1	0	0	4	0	0	0	0	0	5	6%	80%
	1.3.1	2	0	0	1	24	0	0	0	0	27	35%	89%
	1.3.2	0	0	0	0	1	1	0	0	0	2	3%	50%
	4.1.1	1	0	0	0	0	0	6	0	0	7	9%	86%
	7.1.2	0	1	0	0	0	0	0	0	0	1	1%	0%
	8.1.1	0	0	0	1	0	0	0	0	0	1	1%	0%
											78	100%	
	NOTE:	Random train / test split did not include 8.1.3 in the test group.											

		PREDICTION									
		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	Sample Count
ACTUAL CLASS	1.1.2	88.89%	3.70%	0.00%	3.70%	3.70%	0.00%	0.00%	0.00%	0.00%	27
	1.1.4	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	7
	1.2.1	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
	1.2.5	20.00%	0.00%	0.00%	80.00%	0.00%	0.00%	0.00%	0.00%	0.00%	5
	1.3.1	7.41%	0.00%	0.00%	3.70%	88.89%	0.00%	0.00%	0.00%	0.00%	27
	1.3.2	0.00%	0.00%	0.00%	0.00%	50.00%	50.00%	0.00%	0.00%	0.00%	2
	4.1.1	14.29%	0.00%	0.00%	0.00%	0.00%	0.00%	85.71%	0.00%	0.00%	7
	7.1.2	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
	8.1.1	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
											78
	NOTE:	Random train / test split did not include 8.1.3 in the test group.									

Version 3: Larger training data set of 956 (approx. 75% of total 1277). Manual train/test split (90/10) with representative distribution of samples across class labels

Increased training set size to 956 files. Used stratified sampling to create the test file to ensure a representative set of samples were included.

Row Labels	Count	%
1.1.2	300	34.88%
1.1.4	77	8.95%
1.2.1	14	1.63%
1.2.5	53	6.16%
1.3.1	304	35.35%
1.3.2	18	2.09%
4.1.1	72	8.37%
7.1.2	16	1.86%
8.1.1	4	0.47%
8.1.3	2	0.23%
(blank)		0.00%
Grand Total	**860**	**100.00%**

		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	TOTAL	Sample %	% Correct
Actual	1.1.2	31	0	0	0	2	0	0	0	0	33	34%	94%
	1.1.4	0	9	0	0	0	0	0	0	0	9	9%	100%
	1.2.1	1	0	0	0	0	0	0	0	0	1	1%	0%
	1.2.5	0	0	0	5	1	0	0	0	0	6	6%	83%
	1.3.1	2	0	0	1	31	0	0	0	0	34	35%	91%
	1.3.2	0	0	0	0	1	1	0	0	0	2	2%	50%
	4.1.1	1	0	0	0	0	0	7	0	0	8	8%	88%
	7.1.2	0	0	0	0	0	0	0	2	0	2	2%	100%
	8.1.1	1	0	0	0	0	0	0	0	0	1	1%	0%
											96	100%	
	NOTE:	8.1.3 Remove from sample set as it only had 3 examples.											

		PREDICTION									
		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	Sample Count
ACTUAL CLASS	1.1.2	93.94%	0.00%	0.00%	0.00%	6.06%	0.00%	0.00%	0.00%	0.00%	33
	1.1.4	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9
	1.2.1	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
	1.2.5	0.00%	0.00%	0.00%	83.33%	16.67%	0.00%	0.00%	0.00%	0.00%	6
	1.3.1	5.88%	0.00%	0.00%	2.94%	91.18%	0.00%	0.00%	0.00%	0.00%	34
	1.3.2	0.00%	0.00%	0.00%	0.00%	50.00%	50.00%	0.00%	0.00%	0.00%	2
	4.1.1	12.50%	0.00%	0.00%	0.00%	0.00%	0.00%	87.50%	0.00%	0.00%	8
	7.1.2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	2
	8.1.1	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
											96
	NOTE:	8.1.3 Remove from sample set as it only had 3 examples.									

Version 4:

Version 4 encountered an error so has been omitted.

Version 5: Larger training data set of 956 (approx 75%). Manual train/test split (90/10) with representative distribution of samples across class labels. Removed 8.1.3 rows

Training set size to 956 files. Used stratified sampling to create the test file to ensure a representative set of samples were included. Removed the 8.1.3 samples from the training data.

Row Labels	Count	%
1.1.2	300	34.97%
1.1.4	77	8.97%
1.2.1	14	1.63%
1.2.5	53	6.18%
1.3.1	304	35.43%
1.3.2	18	2.10%
4.1.1	72	8.39%
7.1.2	16	1.86%
8.1.1	4	0.47%
(blank)		0.00%
Grand Total	858	100.00%

		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	TOTAL	%	% Correct
Actual	1.1.2	29	0	0	0	3	0	1	0	0	33	34%	88%
	1.1.4	1	8	0	0	0	0	0	0	0	9	9%	89%
	1.2.1	1	0	0	0	0	0	0	0	0	1	1%	0%
	1.2.5	0	0	0	6	0	0	0	0	0	6	6%	100%
	1.3.1	3	0	0	0	31	0	0	0	0	34	35%	91%
	1.3.2	1	0	0	0	0	1	0	0	0	2	2%	50%
	4.1.1	0	0	0	0	1	0	7	0	0	8	8%	88%
	7.1.2	0	0	0	0	0	0	0	2	0	2	2%	100%
	8.1.1	0	0	0	0	0	0	0	0	1	1	1%	100%
											96	100%	
	NOTE:	8.1.3 Remove from sample set as it only had 3 examples.											

		PREDICTION									
		1.1.2	1.1.4	1.2.1	1.2.5	1.3.1	1.3.2	4.1.1	7.1.2	8.1.1	Sample Count
ACTUAL CLASS	1.1.2	87.88%	0.00%	0.00%	0.00%	9.09%	0.00%	3.03%	0.00%	0.00%	33
	1.1.4	11.11%	88.89%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	9
	1.2.1	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
	1.2.5	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6
	1.3.1	8.82%	0.00%	0.00%	0.00%	91.18%	0.00%	0.00%	0.00%	0.00%	34
	1.3.2	50.00%	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%	0.00%	0.00%	2
	4.1.1	0.00%	0.00%	0.00%	0.00%	12.50%	0.00%	87.50%	0.00%	0.00%	8
	7.1.2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	2
	8.1.1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	1
											96
	NOTE:	8.1.3 Remove from sample set as it only had 3 examaples.									