Te Rua Mahara o te Kāwanatanga
ARCHIVES
NEW ZEALAND

Te Kāwanatanga
o Aotearoa
New Zealand Government

Archives in the cloud:

Exploring machine learning to transform Archives New Zealand's digital services for agencies

Final v3.1 March 2024

Te Rua Mahara o te Kāwanatanga
archives
NEW ZEALAND

© Crown Copyright, 2024

Recommended citation: Archives New Zealand. (2024). *Archives in the cloud: Exploring machine learning to transform Archives New Zealand's digital services for agencies*.

This work is licensed under the Creative Commons Attribution 4.0 licence. In essence, you are free to copy, distribute and adapt the work as long as you attribute the work to the Department of Internal Affairs (and abide by the other licence terms – see the plain English licence terms at creativecommons.org/licenses/by/4.0). Please note that neither the DIA logo nor the New Zealand Government logo may not be used in any way which infringes any provision of the Flags, Emblems, and Names Protection Act 1981 – attribution to DIA should be in written form and not by reproduction of the DIA logo or New Zealand Government logo.

Contents

Poumanaaki Chief Archivist’s Foreword

E ngā mana, e ngā reo, e ngā karangatanga maha o ngā hau e whā, nei rā te mihi ki a koutou katoa.

To the esteemed, the orators, the many from the four corners of the world, greetings to you all.

Every day, public servants create records of long-term significance that will eventually find their way into Archives New Zealand’s custody, where we ensure they’re maintained and accessible for future generations. This archive – the memory of government – is a vital cornerstone of our democracy. Without record keeping, we cannot have accountability or good governance, just memory and opinion. If the information and taonga that fill our repositories is accessible, searchable, and properly preserved, we have the potential to make Aotearoa New Zealand a fairer, safer and more equitable place.

The rapidly moving technology environment that current generations are experiencing provides amazing opportunities, but also confronts us with complex challenges. How do we identify which records have enduring business, cultural and historical significance? Which ones need to be transferred to Archives’ custody for safe keeping? And how do we make sure that information of importance to people and communities in New Zealand can be found now and in the future?

After decades of operating in either a hybrid or fully digital mode, the challenge is the huge quantity of structured and unstructured government data that has been created. Unlike paper records, we can create endless digital records with ease. Everything from databases to documents can be produced, reproduced, amended, and saved with just a few clicks. For us to find the data of value, we need to appraise the whole, and either have records to be destroyed~~,~~ or transfer them to Archives – a process that would likely take hundreds of years to do manually. In a world of millions or billions of records, we have now long-since moved beyond what is possible to sort and sift with the human eye and hand.

Between February and July 2022, Archives led a proof of concept to see how best it could transfer and identify high value agency data, including information that is of interest to Māori. This involved analysing hyperscale cloud capabilities and machine learning tools – seeing how a cloud service provider can support our need to deliver. This was a practical way for us to address the three key compliance shortfall issues we discovered in the then most recent Government Recordkeeping Report to Parliament. We learned that these tools have huge potential, but to harness it, we need to make sure the right relationships, processes and systems settings are in place. We need to partner with Māori and collaborate with agencies to ensure our approaches are trusted and work for everyone. Our aim is to make compliance easy.

This report outlines the process and findings from our proof of concept. We're looking forward to building on our learnings as we explore what's next.

It's time to embrace these amazing opportunities to address big and complex challenges, including the opportunity to change our ways of thinking about archival practices. We can now test new tools to address these data issues and be part of the solution to system-wide challenges.

Kei ngā iwi o te ao nei, ehara taku toa te toa takitahi, engari he toa takitini. Nō reira, tēnā koutou, tēnā koutou, tēnā koutou katoa.

To the many people of the world, it is not through my strength alone but the strength of the many that we will succeed. Greetings thrice over to all.

Anahera Morehu, Poumanaaki Chief Archivist

Executive summary

From February to July 2022, Archives New Zealand led work on a proof of concept which showed that machine learning tools have the potential to auto-classify digital public records *and surface information of interest to Māori.*

Public offices like government agencies and Crown entities create huge amounts of digital information and data in their day-to-day work. Some of this information has long-term significance to Aotearoa and will eventually be transferred to Archives New Zealand's custody and form part of the lasting memory of the government. The roles of Archives and the Chief Archivist are set out in the Public Records Act 2005.

Our current systems for sorting, maintaining, and ensuring the accessibility of this information were designed with paper records in mind. There are now huge stores of digital information and data held by public offices, from databases with millions of emails to legacy systems and shared drives full of content. This information needs to be appraised by agencies in line with policy documents called disposal authorities to determine how long information will be kept and what will happen to it – usually either destruction or transfer to Archives. It is no longer possible for people to sort through all of this information manually. Without looking for new approaches to appraisal, disposal, and searching for information within our archives it is inevitable that there will be gaps in the memory of government.

Archives wanted to see if machine learning tools and hyperscale cloud capabilities can help to sort this information and solve other information and archival challenges that have arisen in the digital era. Archives received funding from the Digital Government Partnership Innovation Fund to carry out a proof of concept (PoC) from February to July 2022. The PoC aimed to test if it was possible to use these tools to:

1. **Streamline the appraisal process**, specifically whether auto-classification could determine the appropriate disposal authority to apply to information and records.

2. **Identify material of importance to communities**, specifically whether available tools could identify and surface information of interest to Māori**.**

We worked with agencies (the Ministry of Justice and Ministry for Primary Industries), technology partners (Microsoft and AWS) and information management experts for the PoC. The aim was to test if machine learning and cloud computing tools could classify data in line with disposal authorities – the rules for keeping or disposing of information. We also tested whether these tools could surface information of interest to Māori.

We worked together in a nimble and iterative way to develop this PoC and through issues like where the data would live and how to keep it safe, and what key outcomes we wanted to test in this small-scale experiment.

Within the limited timeframe available, both Microsoft and AWS successfully developed solutions using their suites of tools that could auto-classify records and find Māori subject headings within records. With further training, the models would likely become more accurate, and further refinement and consultation could help ensure the relevance and accuracy of the Māori records identified.

The potential of these technologies is huge, and we want to continue developing processes and approaches to help to address the challenges we have and to grasp opportunities. For future work in this area, we will need to get the right resources in place, work alongside Māori, and ensure that the wider processes are fit for purpose and in line with the Algorithm Charter. We also need to think about the wider information context across government. For example, it is likely for any large-scale project to be successful we will need to rethink how we develop disposal authorities and ensure an all-of-government ontology is built and available.

Our proposed next step to build on this PoC is to continue work on approaches to auto-classification of digital records and information under general disposal authorities. We expect that auto-classification approaches will make a significant positive impact to information managers and agencies more broadly, as it will allow them to carry out their work more efficiently.

Part One: Background and context

About the role of Archives New Zealand

Archives New Zealand Te Rua Mahara o te Kāwanatanga (Archives) preserves and protects government records which have long-term value to Aotearoa New Zealand, so people can access them now and in the future. Archives also has a regulatory role to ensure public offices create and manage information in a way that supports transparency, accountability and the rights of New Zealanders. The role of Archives New Zealand and the Chief Archivist is set out in the Public Records Act 2005.

Records are created across government

Public offices like government agencies and Crown entities create records as part of their usual business practice, from finance and human resources to the specific business of the agency or organisation. Currently, most records are created in digital formats, including emails, databases, documents, spreadsheets, and PDFs.

In this report, we also refer to these records as “information” and “data”.

What happens to records created by public offices?

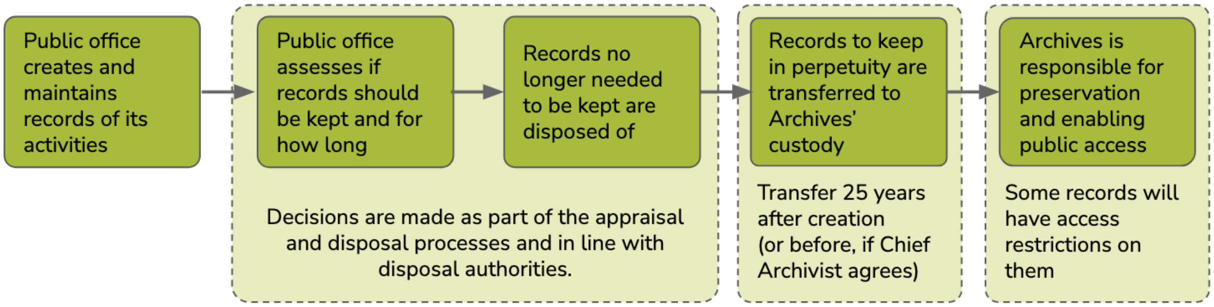
Decisions about keeping or getting rid of public records are made in line with disposal authorities – policy documents which outline how long certain kinds of records need to be kept, and what happens to them when they are no longer needed for operational purposes. These documents are agreed between the Chief Archivist and agencies and developed in consultation with communities. There are two kinds of disposal authority: general and organisation-specific. General disposal authorities (GDAs) cover non-core business information and records that are common across organisations like administration, corporate services, human resources, and finance. Organisation-specific disposal authorities identify the information and records classes that are specific to an organisation, and across government there are thousands of classes and subclasses of record.

The two main ways for public offices to dispose of information and records they no longer need are to transfer them to Archives New Zealand or, if they have no long-term value, destroy them (with approval from the Chief Archivist). Before doing so, they must follow a process to understand what can be destroyed or transferred and when – this is called appraisal.

Detailed information about the appraisal and disposal of public records is on "How to manage your information" of the Archives website.

The journey of a government record: from creation to archive

The diagram below sets out at a high level the journey of a record from creation to its disposal or transfer to Archives.


Public office creates and maintains records of its activities
Public office assesses if records should be kept and for how long
Records no longer needed to be kept are disposed of
Decisions are made as part of the appraisal and disposal processes and in line with disposal authorities.
Records to keep in perpetuity are transferred to Archives' custody
Transfer 25 years after creation (or before, if Chief Archivist agrees)
Archives is responsible for preservation and enabling public access
Some records will have access restrictions on them

Government has gone digital – that means *a lot* of data

Digital records have been produced as a matter of course for over two decades and today most public records are created digitally.

Digital records have several characteristics that make them different from their physical counterparts. Most notably, digital technologies allow much larger quantities of information and data, including copies, to be created and stored. In addition, digital records need regular maintenance to ensure they are accessible over time, unlike records stored on paper. Some storage and software formats, like floppy disks or Word Perfect, are now out of date and no longer easily accessible using contemporary computer operating systems. This means information in legacy systems or created using older software can be hard to access if not maintained over time.

The challenges we're facing

While digital records have been created by public offices for decades, the practice of appraisal, transfer and disposal of digital records is in early stages of maturity. At the time of writing, only a few digital transfers to Archives have taken place. There are huge stores of digital information across public offices, including collections of millions of emails, huge shared drives, and legacy systems that are no longer accessible with current software or able to be migrated to new servers – and this is only growing.

Many digital records are approaching 25 years since their creation. This means the amount of information which needs to be appraised and transferred to Archives or disposed of in the coming years, is going to increase significantly. Much of this (with the exception of some very valuable records) is duplicate information or of little long-term value, and most has not been tagged, classified, or identified effectively. At present content within records that is of interest to Māori cannot be easily identified, which is something we need to address in order to ensure we are meeting our Te Tiriti o Waitangi responsibilities. Overall, this huge volume of digital information held by agencies is already too large for effective and timely processing by human beings, and the problem is only growing.

The data storage capacity required to store the digital records held across agencies is significant, and will continue to increase. The already-high cost of this storage is compounded by the fact that each of the approximately 500 core public offices have individual storage arrangements.

The old paradigms for storing, appraising, and transferring information and data are not fit-for-purpose in a digital world. These models were designed in a paper-based era for a paper-based era, making it difficult to implement the processes using automation or other digital capabilities now available to us. These systems and processes desperately need updating for the digital age, to prevent us from losing vital records and risking a gap in the memory of government. Otherwise, we risk being unable to meet our obligations to uphold Te Tiriti and fulfil our role as information stewards.

The big opportunity

There is currently a big opportunity to take a new approach to information management and archival processes. We can address the problems and add immeasurable further value by using digital tools to unlock insights that empower future decision-making and exploration of our history. We need a different method for appraising and transferring records that reimagines the way information is stored, classified, and transferred to Archives.

In embracing a digital-centred approach, we have the opportunity to make the most of powerful hyperscale cloud storage and advanced machine learning capabilities. This includes looking to machine learning tools and automated classification to process high volumes of data and identify valuable information from within huge pools of unsorted data. It also includes working with tangata whenua to ensure Māori can access information that is of interest to them and building capabilities to manage information at an all-of-government scale while maintaining privacy, security, and cultural requirements.

Moving towards a digital-centred approach to archival processes will have positive long-term impacts for Archives, public offices, and current and future generations of New Zealanders. It will be a big journey. But any big journey starts with a series of small steps. We will need to start modestly, test ideas, and test options for addressing some of these big issues in a way that is inclusive, equitable, and makes the most of the new tools available. This report reflects on one of these vital first steps – a proof of concept to test new methods for classifying and surfacing key information from huge volumes of digital records.

Part Two: a proof of concept

Solidifying our idea and applying for funding

Seeing an opportunity to collaborate on an innovative cloud-based approach to solving real world archival challenges, Archives applied for, and was granted, funding from the Digital Government Partnership Innovation Fund. We proposed a proof of concept (PoC) to test whether hyper scale cloud storage solutions and machine learning tools can be used to auto-classify information and surface material of interest. The project ran from February to July 2022.

We wanted to test was whether using hyperscale cloud capability and machine learning tools could improve how we classify and manage information, in order to:

- **provide agencies with tools to make service improvements and enable better government decision-making,** for example through enabling access to insights afforded through more effective data analysis
- **lower costs** through more efficient data storage and information access across government and within Archives
- **meet our Te Tiriti obligations** by identifying information of interest to Māori and in doing so improving access.
- **enable agency regulatory compliance** by enabling appraisal at scale and reducing risk that information will become unusable or unfindable.

If successful, pre-appraisal auto-classification could be used to help agencies to filter and manage the vast quantities of unsorted information they hold, and help to ensure people can find information that is of relevance to them. However, this PoC is only a first exploratory step. Any policy decisions or redesigns of our archival management systems would be a separate process.

The funding we received allowed Archives to bring on dedicated project resources to plan and implement the PoC in collaboration with agencies and technology partners – vital for a technically complex project like this. Without this additional funding, Archives would not have had the resources within its budgets to carry out a PoC like this.

What is a proof of concept?

A proof of concept (PoC) is a term used in software development and other industries to describe an exercise to test the real-world potential of a design concept or business idea. The PoC helps determine that an idea is feasible and likely to work in practice before putting production-level resources behind it.

After a PoC is completed and the idea is verified, organisations tend to put together a prototype – a working mock-up of the full system that would implement the idea from the PoC so it can be tested and tweaked. Once the system is ready to go, a pilot can be launched. Pilots tend to involve rolling out a fully functional system in a real-world environment to test with a limited group of real users, before fully implementing the system or scaling it more widely.

Bringing together collaborators

We took an iterative, collaborative approach to developing and executing the PoC, recognising that the project would not only need to meet the needs of Archives, but also solve problems for agencies. We also wanted to collaborate with cloud providers to understand how some of the advanced machine learning tools and hyperscale capabilities available in their suites of services could be used to solve complex information challenges.

For the project, we brought together:

- **project support**, including project management and coordination roles
- **technology partners,** which for this PoC were Amazon Web Services (AWS) and Microsoft
- **agency partners,** which for this PoC were Ministry of Primary Industries (MPI), and the Ministry of Justice (MoJ)
- **the Cloud Programme team** at the Department of Internal Affairs Te Tari Taiwhenua (DIA)
- **Archives New Zealand expertise,** including in digital preservation, information system design and policy development, archives appraisal and ontologies

- **external auto-classification and ontology expertise** to help with the planning stages of the PoC.

A project team, which included practitioners from Archives and partner agencies worked with the cloud providers on two separate test environments. A governance group provided strategic direction and guidance throughout the project.

Collaborating with agencies

To ensure the PoC could help achieve outcomes that were useful for the wider government information management system, it was important that Archives collaborated with public offices throughout the process.

We worked with two government agencies that generously gave their time, expertise and other resources to the project:

Ministry for Primary Industries (MPI) contributed the data set for the PoC. MPI officials also participated at both the project team and governance level. Like most public offices, MPI has multiple decades of digital records that need appraisal and, where appropriate, transfer to Archives. This is a complex, time consuming task and MPI officials saw the potential for digital tools to assist in the appraisal process. The complexity of the information management landscape at MPI is compounded by a wide range of current and legacy information management systems, reflecting that it has a wide range of roles, multiple Ministers, a history as three separate organisations, and many legacy information systems. MPI has developed ontologies covering some of its core business functions, which made them well-placed to contribute to this PoC.

Ministry of Justice (MoJ) officials participated as part of the project team and governance group. MoJ recently completed a similar proof of concept to auto-classify information across their shared drives, testing a similar concept within their own specific and complex operating environment. Sharing findings and comparing PoC processes and strategies with MoJ officials was invaluable, and helped ensure we work strategically and learn from peer agencies.

Collaborating with cloud providers

For this PoC, our technology partners were Microsoft and Amazon Web Services (AWS), which both developed a technical solution using their suites of tools and services. Both providers have the hyperscale capabilities we were looking to test and are recognised government suppliers of cloud services, with some agencies already using Microsoft and AWS cloud services to generate, store, and organise information. For future projects, we are open to working with other government cloud providers to test approaches and tools.

Both technology partners have an interest in working with government agencies to develop solutions for complex real-world data problems they might be facing and are investing in New Zealand-based hyperscale data storage facilities. Participating in these PoCs gave the cloud providers an opportunity to apply their data science and machine learning tools and trial new capabilities, and the agencies an opportunity to co-develop information and data management solutions that could be applied and scaled.

The technology partners provided their time and capabilities for free, with Microsoft bringing in their partner Insight Enterprises to build the solution on their behalf. This offer to provide time and capabilities at no cost enabled us to have some of the best engineering minds on the job without committing additional financial resources. However, there is no commitment for Archives or other agencies to work with these companies on future PoCs or pilot projects.

Proof of concept process and outcomes

The PoC was carried out in four stages:

1. **Planning**, which included scoping the key project objectives, engaging with agency and technology partners to build a shared understanding of the project goals, and putting in place commercial agreements around the project. We also bought in some specialist auto-classification expertise at this stage, along with ontology and metadata experts, to help us think through key issues and help set the project up for success.

2. **Preparing the test environments,** which involved setting up the spaces where the technology partners would test their software and tools using the MPI data (in this case, the data stayed within an MPI system). This meant ensuring the data was going to be safe, organising the required clearances and approvals, and preparing the environments in which the solutions were developed.

3. **Preparing the data**, which began with MPI identifying a data set that would be useful and appropriate to test, and doing some pre-curation to ensure the data was secure and fit for purpose. It then involved working with the technology partners

and MPI to safely export the data into both test environments. Officials at MPI had carried out manual appraisal of the records against the GDA 6 and 7 and MPI disposal authorities. The pre-appraised files meant that technology partners had accurate training data to use for the machine learning models.

4. **Solution building and testing,** which saw each technology partner develop an approach to addressing the specific challenges and outcomes and producing a trial to test their approach. Both technology partners developed their test environment within a relatively constrained timeframe of a couple of weeks. The technology partners used the data set provided by MPI to test their approach, the results of which were then shared back with the project team in a demo, and detailed further in a project report.

About the proof of concept data set

The PoC was conducted using a test data set provided by the Ministry for Primary Industries (MPI). The data set was a series of about 1400 records created in the late 1990s by the Ministry of Agriculture and Forestry in relation to its preparation for Y2K. The data was in a number of digital formats including Word Perfect, and older Word versions.

Refining the challenges to address with the proof of concept

During the planning process, a key task was to take the wider set of challenges and narrow them down to a targeted set of actionable areas to test in this PoC.

While the challenges of assessing, managing, and surfacing insights from vast pools of government information are broad and complex, we selected two vital tasks for the technology partners to test in the PoC:

1. **streamlining the appraisal process**, specifically whether auto-classification could determine the appropriate disposal authority to apply to information and records. It is important that the privacy and security of information is maintained, so we also wanted to test whether automated processes would recognise and maintain security protocols.

2. **identifying material of importance to communities**, specifically whether available tools could identify and surface information of interest to Māori. This will help Archives to ensure it meets its Te Tiriti obligations. More broadly, a key need for

ongoing information management and access is the ability to surface material with cultural importance.

Focusing on this narrower set of issues allowed us to ensure a clearly defined scope for the proof of concept, allowing us to test whether auto-classification is likely to be a good fit for our needs.

The outcomes we wanted to achieve

Taking the two challenges as a starting point, we identified a range of key requirements for the technology partners to consider when developing their solutions in the PoC.

The key outcomes we were seeking from this proof of concept were:

1. **the correct policy for the retention and disposal of files is determined automatically.** The appropriate disposal authority and specific document class number are identified for each file, enabling the retention time and disposal method to be automatically identified.

2. **information that is of interest to Māori is identified.** For this PoC, the Ngā Upoko Tukutuku were used, which were developed by the Māori Subject Headings Project. The tool provides a structured path to a Māori world view within library and archival cataloguing and description and includes a wide range of terms.

3. **the security and privacy categories applied to the provided records in the data set are accurate.** Any security classification applied to the records (for example written in document footers) was identified and records that contained personal information were identified.

The other key parameters we set for the technology partners were that manual intervention and input into these processes should be minimised and algorithmic transparency should be emphasised.

About the solutions developed by the technology partners

Following the planning and scoping process, the project team worked with each technology partner as they built a solution to reach the desired outcomes. Both AWS and Insight (on behalf of Microsoft) developed solutions that drew on their respective company's existing tools and capabilities. This included storage and file ingest systems, machine learning tools to carry out text extraction and analytics, and front end user interfaces.

Each technology partner then trained and tested their approach with the MPI data. The aim was to test whether the concepts – auto-classification and surfacing of information – were technically possible and identifying areas for further refinement.

Both technology partners built solutions that could classify and identify information

Each technology partner trained machine learning models that were able to identify the correct disposal authority class with a high level of accuracy, and in both cases noted that the models became more accurate with more training. For example, Insight's Microsoft solution delivered 35-45% accuracy with training on 100 documents, which was raised to 75-85% when trained on 1000 documents. The final model that AWS demonstrated was the fifth iteration, which delivered 88% accuracy. It is expected that with additional training that the accuracy of both models would improve.

Both technology partners took slightly different approaches to identifying security classifications and private (personally identifiable) information. This reflects that the files did not consistently have security classifications included either in their metadata or in the documents themselves (for example written in the footer). Microsoft noted that it was able to identify security classifications of files when they had been entered as metadata in the Microsoft SharePoint document management system. AWS was able to identify about 89,000 instances of personally identifiable information within documents (for example addresses, bank account numbers or phone numbers) from a set of 1277 documents.

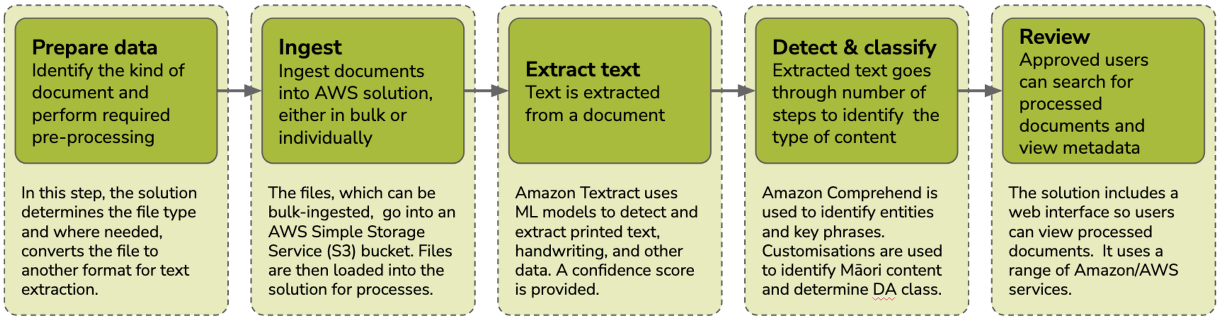
AWS noted that while it was able to identify Māori subject headings and iwi names in documents, further work would be needed to ensure that the results are both accurate and relevant to Māori. For example, words like "tikanga" may be used hundreds of times, and the particular relevance may vary depending on context. Microsoft was able to recognise and classify documents that contained selected language anywhere in the document text, including the names of iwi. However, further work would be needed if documents were to be assessed in terms of their relevance or importance to Māori (beyond containing a certain word).

Further development of these approaches

Both technology partners provided recommendations on potential next steps towards piloting and operationalising the tasks they tested during the PoC. This reflects that there was limited time during the PoCs to automate processes, customise software or train the models.

Solution overview: Amazon Web Services (AWS)

The AWS team developed a solution that it described as a processing pipeline with five key components. The image below (produced by Archives) outlines these components at a high level. AWS produced more detailed diagrams of their architecture as part of their project report.



Prepare data
Identify the kind of document and perform required pre-processing

In this step, the solution determines the file type and where needed, converts the file to another format for text extraction.

Ingest
Ingest documents into AWS solution, either in bulk or individually

The files, which can be bulk-ingested, go into an AWS Simple Storage Service (S3) bucket. Files are then loaded into the solution for processes.

Extract text
Text is extracted from a document

Amazon Textract uses ML models to detect and extract printed text, handwriting, and other data. A confidence score is provided.

Detect & classify
Extracted text goes through number of steps to identify the type of content

Amazon Comprehend is used to identify entities and key phrases. Customisations are used to identify Māori content and determine DA class.

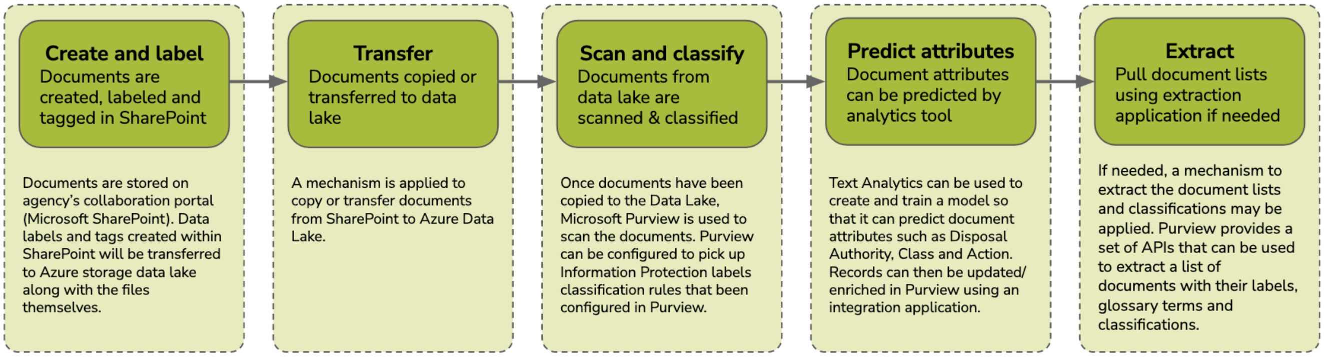
Review
Approved users can search for processed documents and view metadata

The solution includes a web interface so users can view processed documents. It uses a range of Amazon/AWS services.

The processing pipeline was built on an existing open source solution designed and built by AWS. It uses a number of AWS services to enable understanding of document content and to provide the user interface.

Solution overview: Microsoft

The Microsoft team developed a solution for identifying and classifying documents from a data lake, using both rule-based classifications and machine learning auto-classification to classify documents and surface insights. The image below (produced by Archives) outlines the solutions components at a high level. Insight (for Microsoft) produced more detailed diagrams of their architecture as part of their project report.



Create and label
Documents are created, labeled and tagged in SharePoint

Documents are stored on agency's collaboration portal (Microsoft SharePoint). Data labels and tags created within SharePoint will be transferred to Azure storage data lake along with the files themselves.

Transfer
Documents copied or transferred to data lake

A mechanism is applied to copy or transfer documents from SharePoint to Azure Data Lake.

Scan and classify
Documents from data lake are scanned & classified

Once documents have been copied to the Data Lake, Microsoft Purview is used to scan the documents. Purview can be configured to pick up Information Protection labels classification rules that been configured in Purview.

Predict attributes
Document attributes can be predicted by analytics tool

Text Analytics can be used to create and train a model so that it can predict document attributes such as Disposal Authority, Class and Action. Records can then be updated/ enriched in Purview using an integration application.

Extract
Pull document lists using extraction application if needed

If needed, a mechanism to extract the document lists and classifications may be applied. Purview provides a set of APIs that can be used to extract a list of documents with their labels, glossary terms and classifications.

Microsoft's solution architecture was built on a series of existing tools and systems within the Microsoft ecosystem. The Microsoft solution imagines agencies using a Microsoft 365 or SharePoint environment to store and create their information. Files would then be copied or transferred into an Azure Storage data lake, before being managed and classified using Microsoft Purview and Azure Text Analytics tools.

From proof of concept to wider applications

The proof of concept showed that these tools can help with classification and surfacing information

The PoC showed there is significant potential of using advanced machine learning tools and hyperscale cloud capabilities to address the digital information management challenges and opportunities Archives and agencies are facing.

Both trials within the PoC demonstrated it is possible to carry out auto-classification and surface information with a high level of accuracy using machine learning and other digital tools. We think that auto-classification tools could help significantly relieve the appraisal burden for agencies and help information management specialists to work through the backlog of information that needs appraisal across agency systems.

> “Seeing the possibilities of these tools is really exciting, I can see huge potential to address some of the gnarly challenges we’re facing with needing to appraise at scale and enabling access in the future. Aligning this powerful technology with new processes and ways of working will allow Archives and agencies to work collaboratively on a pathway towards improved access to Aotearoa’s digital taonga, ensuring the memory of government is digitally accessible for generations to come.” – Anahera Morehu, Chief Archivist

The use of these tools alone will not be sufficient to address the challenges we’re facing. We also need to look at how we design future projects to auto-classify and surface information, and how this work fits into the wider government information and data ecosystem. We outline some of the considerations in each of these areas below.

Design considerations for future prototypes and pilots

When looking at future work that builds on this PoC, there are a number of design considerations that need to be made to ensure the approach is scalable, accurate, trusted and trustworthy.

We will need to look beyond the technical systems themselves and also examine the processes we build around them. The design considerations include:

- **Ensuring scalability.** For any auto-classification project to be successfully implemented, it will be essential that the project is granted the resources necessary to scale. Scalability is not possible with the current staff resources and access to technology, so additional dedicated funding and other resources will be needed to enable further project development or implementation roll-out.

- **Growing capability.** We executed the PoC with a reasonably constrained project budget, a small project team, and technical resources provided by technology partners for free. If we are to expand this work and implement auto-classification at any scale, we will need to build our internal capability, including ensuring we have data science skills in-house, and building leadership capability and knowledge. We will also need to support efforts to build capability across the government system and with Māori. A strategic plan around Archives’ capability building is likely a necessary first step. We will need to work closely with the Government Chief Digital Officer and Government Chief Data Steward.

- **Working with Māori.** Close collaboration and co-design with Māori will be an essential part of any future work to use machine learning tools to manage, archive, and categorise government digital information. This will involve making sure any auto-classification or machine learning approaches are implemented in consultation with Māori to ensure their accuracy and appropriateness. Archives will also need to be led by Māori in regard to which types of information are of interest when developing approaches to surface information within digital government records, and draw on work on Traditional Knowledge Labels to guide terms of access to specific collections.

- **Defining accuracy thresholds.** The two technical solutions were able to auto-classify documents with a reasonably high level of accuracy, and identified that with further training and refinement, this could increase further. In future, we will need to think about the appropriate level of accuracy we would be comfortable accepting for auto-classification of documents – especially if the outcome of auto-classification is used to automate disposal or retention decisions. Conversations will be needed about trust, the accuracy levels of auto-classification vs manual human classification, and what decisions we are comfortable automating vs where a human should remain in the loop.

- **Designing in the principles of the Algorithm Charter.** DIA (of which Archives is a part) is a signatory to the Algorithm Charter for Aotearoa New Zealand, which sets out six commitments for the use of medium and high risk algorithms. If this approach to classification and surfacing information is rolled out broadly, work will need to take place to ensure that the implementation of these tools aligns with the commitments to: transparency, partnership, people, data, privacy, ethics and human rights, and human oversight.

Wider information management system considerations

The challenges the PoC sets out to address are just one puzzle piece in a complex data and archives ecosystem – there are a number of dependencies that need to be coordinated if we are going to move the system forward. The issues that Archives and agencies face sit within a wider context that encompasses a regulatory framework for records management, Māori data governance approach, design of the government data system and the development of ontologies. We will need to collaborate with other information and data leads across government as we continue work in this area.

Below we describe three key areas where further work is required to ensure approaches to archives and information management can make the most of today's digital tools, and make the memory of government available now and into the future.

Working with Māori

Archives has a particular responsibility to ensure that iwi and Māori are able to access information of cultural importance and can exercise their mana motuhake, and working with iwi and Māori is, and will continue to be of importance to the way Archives does its work. It is our responsibility to ensure that data and information of importance to Māori is stewarded with care and respect.

It is important that Archives' future work on the auto-categorisation and surfacing of information aligns with these wider discussions and approaches to Māori data governance. Throughout Archives, government and beyond, big conversations are happening with Māori about a co-governance approach of data, how to appropriately store and categorise Māori data (recognising that the definition is still evolving), and how to ensure data and information is accessible and contributes to Māori self determination. The Data Iwi Leaders Group (DILG) is leading work to ensure full, free access and control over data about and for iwi, in order to empower iwi development. Through the DIA Mana Orite agreement, Archives will ensure that its approach to Māori data governance aligns with its responsibilities and obligations.

Ensuring disposal authorities are fit for the digital era

Disposal authorities, the rules that provide authorisation for the disposal of government information and records, are currently not fit for purpose or future-proofed for an increasingly digital public service. The two general disposal authorities (GDA 6 and GDA 7) issued by Archives NZ to guide agencies in their disposal or retention decisions for non-core business information were written in the context of paper-based information management systems and rely on manual assessment of each record by staff.

Currently, many aspects of the current disposal authorities are not fit-for-purpose. For example, some disposal authorities cannot be automated as their triggers assume human decision-makers and paper-based processes. There is a need to update disposal authorities in collaboration with a range of stakeholders to ensure they are suitable for a digital information ecosystem. There is also an opportunity for domain-wide disposal authorities to replace single agency authorities, or to embed disposal authorities into ontologies. These solutions would make it more streamlined to train machine learning models for auto-classification and reduce replication.

Building an all-of-government ontology

Archives has been exploring the possibility of developing an all-of-government ontology, to enable consistent management of information across government and improve access to information holdings. An ontology is a system for understanding and organising information within an organisation or discipline, and can provide a common vocabulary, concept definitions, and taxonomies that define how information is classified or organised. An all-of-government ontology would support data interoperability and consistency across agencies, and could serve as a bridge between legacy systems and future systems. It could also be an essential part of the shift to a digitally enabled approach, and facilitate the auto-categorisation of content or the automation of business processes, for example with machine learning tools.

What’s next: potential future actions

The PoC demonstrated the huge potential of machine learning tools to assist with the auto-classification of records and information at scale, and potentially the identification of information that is of interest to Māori. Alongside this capability, we think there is a broad range of other exciting possibilities for using advanced digital tools to transform the way we work, to support agencies and ensure we can deal with the significant volumes of digital data coming our way.

As outlined above, there are several system and design considerations needed to facilitate the wider implementation of these approaches. We also know that changing the way we do things, and working together to change the wider system, will take time and resources to develop. We need to plan for how to progress further, together.

Our proposed next step is to continue work on approaches to auto-classification of digital records and information under the GDAs. There are a few ways to approach this, and we will be working through the options. Our current thinking is that we will continue to work alongside agencies to develop a replicable approach to auto-classification that agencies can use on records in their current systems. These pre-appraised records can then be ring fenced as ready to transfer to Archives. We expect that auto-classification approaches will make a significant positive impact to information managers and agencies more broadly, as it will allow them to carry out their obligations more efficiently.

There will be a range of issues to work through before this approach can be rolled out to agencies more broadly, including working through whether disposal authorities need to be revised for the digital era before further work is progressed. At the same time, we are preparing for an influx of digital transfers in the years ahead, and there is a lot of foundational work to be completed in digital storage.

We know that to ensure we have a trusted archives system we need to move forward to make the most of the tools available to us, but we also need to get our processes and ways of working right.

Acknowledgements

Archives New Zealand would like to acknowledge everyone who participated in this PoC. We especially thank:

The governance group members and attendees: Angela Beaumont (then director of digital delivery, (Archives), Joanna Condon (Archives), Joanne Koreman (Archives), Anahera Morehu (Archives); Antonia Aloe, Brian Nevin, Tony Eyles (DIA); Andre Post (MPI); and Tony Breen (MoJ).

The project team: Shakil Akhter, Elisa Capecci, Joanne Koreman, Madeline Pierpoint, Spencer Tasker (DIA); Tanya Hurley, Sara Knight (MPI); Vanessa Cuthbert-King (MoJ); Marzieh Cameron, Jan Hutar, Carly Lenz, and Joshua Ng (Archives).

Technology Partners: Mark Gordon, Craig Hind, Matthew McCulloch (AWS); Nick Cater (Microsoft New Zealand); Bruce Anderson, Richard Malloch and JP Van Heerden (Insight).

Information and technical specialists: Namita Parshotam (DIA); Annabel Snow (Deloitte); Team IM.