Australian Government | GovAI

AI-powered Data Preparation

AI powered pipeline that cleans and structures Clik content to support accurate, scalable use in digital tools.

Department of Veterans' Affairs

Australian Government | GovAI

Executive Summary

Problem Faced

The Clik website contains thousands of pages of policy and legislative content essential to DVA's work. However, this content is difficult to search and interpret due to inconsistent formatting, duplication, and outdated material. Staff often spend significant time manually locating relevant information, which can slow down claims processing and increase the risk of inconsistent advice.

Existing tools are not designed to manage this level of scale or complexity. Manual review is unsustainable as demand for timely and accurate information grows. A more intelligent and scalable approach was needed to prepare and organise content to support reliable use in AI systems and improve service delivery.

Solution Overview

DVA developed a custom AI-powered data preparation pipeline to clean, structure and organise Clik website content for use in AI systems. This preparation is essential for enabling tools like chatbots to deliver accurate and trustworthy responses. Without well-prepared data, even advanced models can produce unreliable results.

Key features:

- Automated collection and removal of duplicated content
- AI-assisted extraction of key information blocks
- Conversion to a simplified format, reducing file size by ~75%
- Regular updates to keep content current and accurate

Benefits and Impact

The AI-powered pipeline has improved the quality, efficiency and reliability of preparing Clik content for digital use:

- **Cleaner, more relevant content**: Ensures only accurate, up-to-date material is retained, improving the reliability of AI tools.
- **Smaller, faster indexes**: Simplified formats reduced dataset size by ~75%, improving performance and reducing costs.
- **Faster updates**: Only changed pages are reprocessed, keeping content current with minimal effort.
- **Better outcomes for veterans**: Higher quality and more accessible policy information helps ensure faster, more consistent, and accurate support.

Target Audience and Stakeholders

Primary users are DVA technical teams who use the prepared content to build AI tools such as Clik Chat. These tools support frontline staff by making complex policy information easier to access and interpret.

Key stakeholders include:

- DVA claims processing teams, who rely on accurate content to support consistent service delivery.
- DVA ICT and data teams, who manage infrastructure and integration.
- GovAI, who advised on the solution and supported broader APS reuse.

Subject matter experts were engaged to validate extracted content and ensure it remained accurate and fit for purpose.

Risks and Mitigation Overview

To support responsible use of AI, the following risks were addressed:

- **Accuracy:** Mitigated through deduplication, targeted extraction, and regular updates to ensure only current and relevant content is used.
- **Bias and fairness:** AI extraction methods were reviewed to preserve the intent and structure of official policy content.
- **Accessibility:** Simplified formats improve readability and support future accessibility features.
- **Privacy:** No personal data was used.
- **Operational stability:** Indexing timeouts and service limits were managed through retry mechanisms, content simplification, and off-peak scheduling.

Use Case Status

Pilot

Use case timeline

- **March 2025** – Initial planning
- **April 2025** – Core pipeline development
- **May 2025** – Indexing, testing, and internal demonstrations
- **June 2025** – Refinement, documentation, and preparation for reuse

Additional Information

The pipeline is built on Azure infrastructure and integrates with DVA's existing systems. It uses virtual desktops for development, blob storage for content management, and a modular backend for processing and indexing. The system supports both public and synthetic data sources, with future extensions planned for OCR and handwriting recognition to support claims-related use cases.

Lessons Learned

- **Data quality is foundational**: Clean, well-structured content is essential for reliable AI performance.
- **Automation adds long-term value**: Early investment in deduplication and incremental updates improve maintainability.
- **Collaboration is critical**: Close engagement between technical teams and policy experts ensured the extracted content remained accurate and usable.
- **Design for reuse from the start**: Modularity of pipeline design made it easier to share across government and support broader APS capability uplif

Contact information

Responsible Entity Name

Department of Veterans' Affairs

Open for Collaboration?

Yes

Area of Entity

Digital Strategy & Planning Branch
Technology, Digital and Data Division

Use Case Contact

AI@dva.gov.au

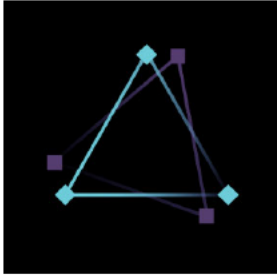
Use Case Website/s

N/A

Use Case Owner

Alicja Mosbauer
Chief Data Officer
Alicja.Mosbauer@dva.gov.au

Screenshot/s



Detailed Overview

Version Control

Version	Date	Author	Description of Changes
1.0	3 Feb 2025	GovAI	Version 1 created
1.1	17 Mar 2025	GovAI	Modified based on feedback

Index

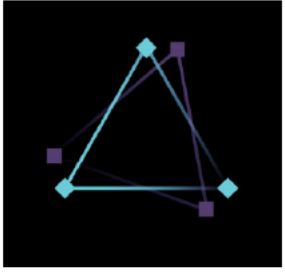
Note: For details about category items in the detailed overview, see *APS AI Use Case Repository Guidance- Guidance for Use Case Owners and Editors*.

Responsible Organisation Category

Select the Classification of the Functions of Government – Australia (COFOG-A) 3-digit category that best identifies the functional area associated with your AI use case.

☒ **01 - General Public Services**	014 - Planning and statistical services
☐ **02 - Defence**	Choose an item.
☐ **03 - Public Order and Safety**	Choose an item.
☐ **04 - Economic Affairs**	Choose an item.
☐ **05 - Environmental Protection**	Choose an item.
☐ **06 - Housing and Community Amenities**	Choose an item.
☐ **07 - Health**	Choose an item.
☐ **08 - Recreation, Culture, and Religion**	Choose an item.
☐ **09 - Education**	Choose an item.
☐ **10 - Social Protection**	Choose an item.
☐ **11 - Transport**	Choose an item.

Scope of the Use Case



Use the dropdown menus below to identify the scope of your use case.

Geographical focus Choose the region for implementation from the dropdown list	National
Primary type of government interaction Choose the type of government interaction from the dropdown list	Government-to-government (G2G)
Cross-features - Sector Indicate if the use case describes a solution that can be used across sectors or in cross-sector scenarios (Yes/No).	Yes
Cross-features – Jurisdiction Indicate if the use case describes a solution that can be used across State/Federal borders or in cross-border scenarios (Yes/No)	Yes

Ethical Considerations

Accuracy, Fairness, Accessibility, Bias and Discrimination	The pipeline ensures accuracy through automated deduplication, targeted extraction, and regular updates to retain only current and relevant content. Fairness is defined in consultation with policy experts to preserve the intent and structure of official materials. Accessibility is supported by simplified formats that improve readability and enable future accessibility features. Bias is mitigated through subject matter expert validation and review of AI extraction methods to ensure representative and non-discriminatory outputs. No personal data is used.
Privacy	No personal data is used. The system processes only public and synthetic content. Privacy risks are mitigated by design. No identifiable information is collected, stored, or processed. Azure infrastructure with

	secure blob storage and modular processing ensures data integrity and protection.
Rights of Users	Internal users are informed through documentation and subject matter expert validation. While not yet public-facing, future reuse planning includes transparency measures and responsible AI principles. Users are notified of AI use through internal briefings, and traceability of content supports auditability and review.

Value of the Use Case

Identify the public value that the solution provides or is expected to provide. Select from the multi-select options.

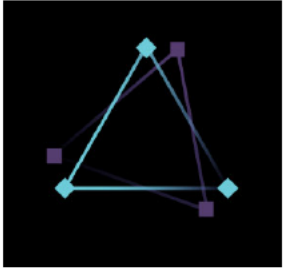
Improved public service This category refers to solutions that enhance the services provided to end users, whether they are citizens or businesses.	☐ Personalised services ☐ Public (citizen)-centred services ☒ Increased quality of public information and services ☒ More responsive, efficient and cost-effective public services ☐ New services or channels
Improved administrative efficiency This category refers to solutions that increase efficiency, effectiveness, and quality while reducing costs within administrative processes, systems, and services.	☐ Cost reduction ☐ Responsiveness of government operation ☒ Improved management of public resources ☐ Increased quality of processes and systems ☐ Better collaboration and better communication ☐ Reduced risk of corruption and abuse of the law by public servants ☐ Greater fairness, honesty and equality enabled
Open government capabilities This category refers to solutions that enhance the level of openness, transparency, engagement, and communication within public organisations.	☐ Increased transparency of public sector operations ☐ Increased public participation in government actions and policymaking

	☐ Improved public control of and influence on government actions and policies

AI Process Type

Select the types of tasks within government operations that the AI solution is performing or expected to perform

Supporting Decision Making- Tasks that support formal or informal agency decision-making on benefits or rights.	☐ Taking decisions on benefits ☐ Managing copyright and intellectual property rights
Analysis, monitoring and regulatory research – Tasks that collect or analyse information that shapes agency policymaking.	☒ Information analysis processes ☐ Monitoring policy implementation ☐ Innovating public policy ☒ Prediction and planning
Enforcement - Tasks that identify or prioritise targets of agency enforcement action.	☐ Smart recognition processes ☐ Management of auditing and logging ☐ Predictive enforcement processes ☐ Supporting inspection processes ☐ Improving cybersecurity ☐ Registration and data notarisation processes ☐ Certification and validation processes
Internal management - Tasks that support agency management of resources, including employee management, procurement, and maintenance of technology systems.	☐ Internal primary processes ☒ Internal support processes ☐ Internal management processes ☐ Procurement management ☐ Financial management and support
Public services and engagement - Tasks that support the direct provision of services to the public or facilitate communication with the public for regulatory or other purposes.	☐ Engagement management ☐ Data-sharing management ☐ Governance and voting ☐ Payments and international transactions ☐ Supporting disintermediation ☐ Authentication of self-sovereign digital ID services ☐ Service integration ☐ Service personalisation ☐ Tracking of goods and assets along the supply chain



AI Technologies Utilised

Select the types of AI technologies proposed / utilised to deliver the use case.

Reasoning or Knowledge Representation AI systems that store, structure, and process knowledge to make inferences, derive conclusions, or support decision-making.	☐ Knowledge Representation ☐ Automated Reasoning ☐ Commonsense Reasoning
Planning and Optimisation AI techniques that generate, refine, and optimise action sequences or resource allocation to achieve specific goals efficiently.	☐ Planning and Scheduling ☐ Searching ☐ Optimisation
Learning and Adaptation AI systems that identify patterns, extract insights, and improve performance over time based on data.	☐ Machine Learning ☐ Deep Learning ☐ Generative AI
Communication and Natural Language Processing AI systems that process, interpret, and generate human language for interaction, comprehension, and automation.	☒ Natural Language Processing (NLP) ☐ Text Generation ☒ Text Mining ☒ Machine Translation
Perception through the Senses AI systems that process and interpret sensory data, such as visual, auditory, or tactile inputs, to understand and respond to their environment.	☐ Computer Vision ☐ Audio Processing
Integration and Interaction with the Environment AI systems that interact with physical or digital environments, including autonomous agents, robotics, and interconnected systems.	☐ Multi-agent Systems ☐ Robotics and Automation ☐ Connected and Automated Vehicles (CAVs)

AI as a Service AI capabilities delivered through cloud-based platforms, offering tools, models, and infrastructure for AI-powered applications.	☐ AI Services (e.g., cognitive computing, machine learning frameworks, bots) ☒ Infrastructure as a Service (IaaS) ☒ Platform as a Service (PaaS) ☐ Software as a Service (SaaS)
Additional Comments or Explanation:	If you have selected any of the subcategories above, feel free to provide more detailed comments or a description of how these elements apply to your specific use case.

Technical Elements

Platform implementation	The pipeline is deployed on Microsoft Azure using a modular architecture. It leverages Azure Virtual Desktops for development, Azure Blob Storage for content ingestion and management, and Azure Functions for orchestrating processing tasks. Indexing and transformation are handled via containerised services, with scheduled jobs managed through Azure Logic Apps. Identity and access are controlled via Azure Active Directory with role-based access. The system is designed for scalability and maintainability, with incremental update logic to reduce compute costs. It integrates with DVA's internal systems and supports future extensions such as OCR and handwriting recognition.
Model / Algorithm used	The pipeline uses a hybrid approach combining rule-based logic with lightweight NLP components. Techniques include named entity recognition, structural parsing, and semantic deduplication using cosine similarity and TF-IDF scoring. These were selected for their transparency, deterministic behaviour, and ease of validation by policy SMEs. Open-source libraries such as spaCy and scikit-learn were evaluated and selectively integrated. The pipeline avoids

	black-box models to ensure traceability and auditability of content transformations.	
Data Sources Select the types of data sources used and provide relevant details.	☐ Internal ☒ Public	☐ Third-party ☒ Synthetic
	Details: Public: Legislation and regulatory material. Synthetic: Generated test data for validating pipeline logic and performance under edge cases.	
Risk Assessment and Mitigation Details	A structured risk assessment was conducted during the design and pilot phases. Key risks included content drift, indexing failures, and semantic misalignment. These were mitigated through incremental update logic, retry mechanisms, and SME validation workflows. Prompt ingestion was tuned to prioritise high-confidence content blocks. Cybersecurity risks were addressed through Azure-native controls including encryption at rest and in transit, network isolation, and audit logging. Data migration was staged with rollback support. The system is monitored for operational anomalies using Azure Monitor.	
Security and Compliance Frameworks Select the security and compliance frameworks and measures implemented. Provide details or additional artifacts if relevant.	☐ Authority to Operate (ATO) ☐ System Security Plan (SSP) ☐ Security Risk Management Plan (SRMP)	☐ Information Security Registered Assessors Program (IRAP) ☐ Penetration Testing
	Details: No Security and Compliance Frameworks have been implemented at this stage.	
Assurance and Government Frameworks	Not at this stage	
Record maintenance	The system includes regular updates and version control mechanisms. It supports traceability through modular processing and	

	indexing, and documentation was prepared during the refinement phase in June 2025.
Disengagement	The modular design and use of virtual desktops and blob storage means that components could be isolated or rolled back if needed.
Performance Metrics and Results	Outcomes include a ~75% reduction in file size, faster updates through incremental processing, and improved content quality.