



Australian Government



MAGDA++

AI-assisted Decision Making with a Structured Data Commons

CSIRO





Executive Summary

Problem Faced

A major barrier in government and industry decision-making is the ability to answer complex questions when relevant knowledge is scattered across structured and unstructured siloed documents, and the tacit expertise residing in individual employees' heads. Meanwhile, it is also challenging to automatically capture provenance, the underlying data sources, and the reasoning steps for continuous knowledge accumulation and auditability. In CSIRO, we built a system to make tacit, tested knowledge explicit and usable in the question-answering process. We have demonstrated the capability of the system in the food industry by assisting cross-domain protein value chain optimisation.

Solution Overview

- Built on MAGDA's open-source core to support advanced data analytics. MAGDA++ leverages foundation models to fuse structured /semi-structured data and specific AI/ML models on such data for predictability.
- Interactive, explainable data analysis to capture the decision-making workflow to infer tacit knowledge for solution recommendation.
- Designed as an agentic platform to link multiple science disciplines from genomics to food chemistry to agronomy to increase predictability and reduce uncertainty in business value chain.
- Multi-Layer AI Integration: Supports both in-browser LLMs (on-device privacy-preserving interaction) and powerful, domain-specific AI agents running on secure local servers.

Benefits and Impact

We use MAGDA++ on protein value chain research with an aim of increasing the predictability of plant-based protein innovations. In this case, relevant knowledge is distributed across Word/Excel documents/PDFs, databases, and expert minds. Our approach leverages both global and Australian data to support sustainable farming for climate adaptation and improve product quality tailored to consumer needs. It leads to the accumulation of tacit knowledge of domain experts and reduces trial and error cost in R&D.

The same data common system can be expanded to support decision making in other government agencies.



Target Audience and Stakeholders

- CSIRO Agriculture, Health and Environment Science researchers
- Protein ingredient and food manufacturers
- Government Science & Research Agencies: Seeking AI-enabled platforms to accelerate discovery while maintaining data sovereignty.
- Current and previous clients:
 - data.gov.au (DTA)
 - data.nsw (NSW DCS)
 - digital twin (NSW DCS)
 - PMC MAGDA trials for COVID-19 data collection (PMC)
 - Private data collection (Depart of Agriculture & DEE)

Risks and Mitigation Overview

Risk 1: Data Privacy Concerns. Risk of sensitive data exposure if AI processing is outsourced to third-party cloud services.
Mitigation Strategy: AI runs locally in-browser or on controlled internal infrastructure.

Risk 2: Performance Trade-offs. In-browser AI may not handle large-scale analysis alone.
Mitigation Strategy: Architected to use lightweight browser agents for initial discovery and server-based models for complex or compute-intensive tasks.

Risk 3: AI biases affecting decision-making.
Mitigation Strategy: Metadata enrichment for shared data understanding and support of model prediction explanation

Use Case Status

Implemented

Use case timeline

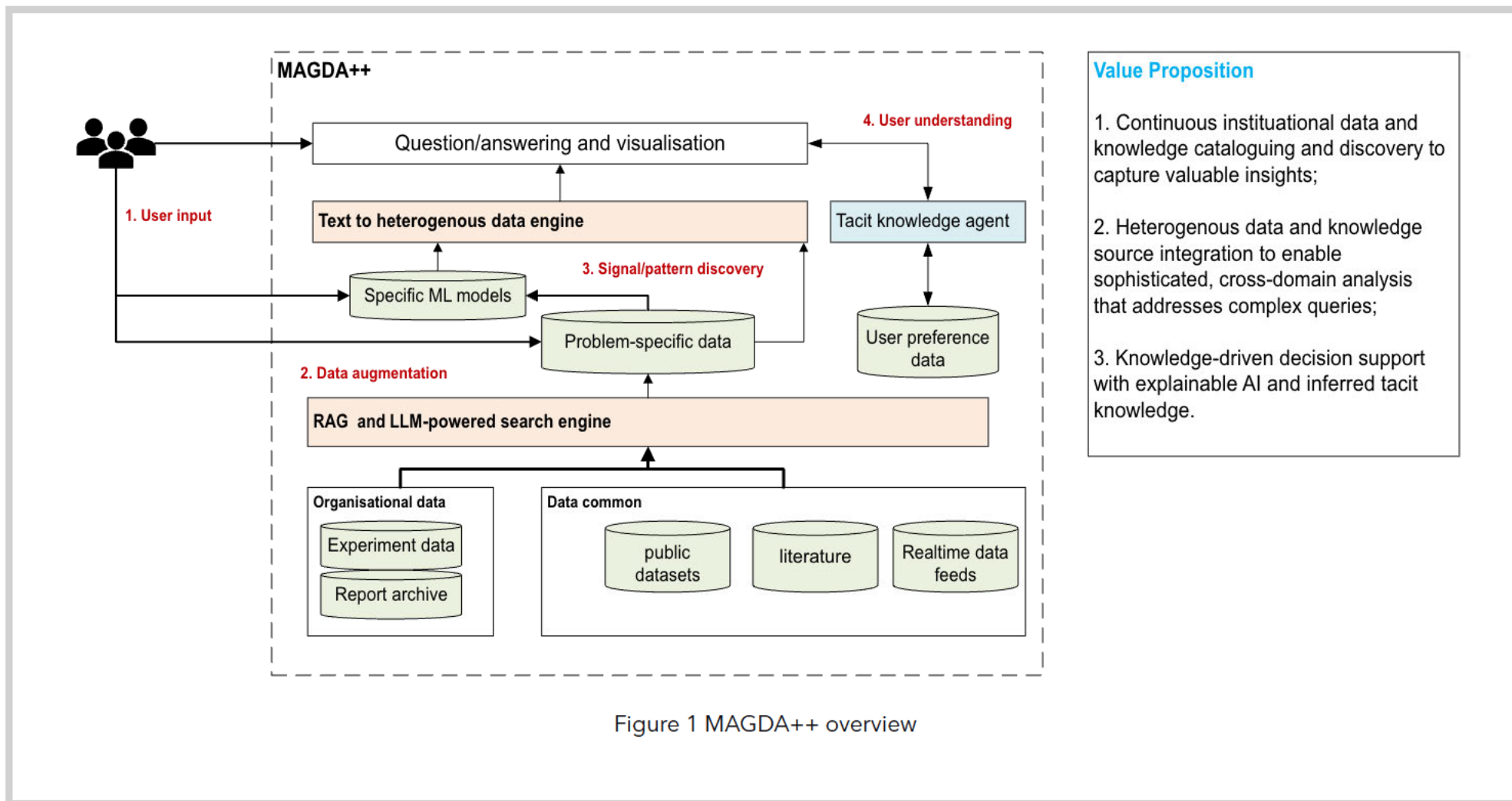
July 2023: Project started
Dec 2023: Initial data integration completed; tabular data retrieval-augmented generation (RAG) released.
June 2024: Sequence data and 3D structure data module released. Specific ML models integrated. Knowledge graph for papers constructed.
Dec 2024: Graph data retrieval-augmented generation implemented.
2025 and onward: Complex query answering and additional tool integration implemented.



Additional Information		Lessons Learned		Contact information	
Our approach builds on the data catalogues that most government agencies already maintain. It can unlock substantial value from existing assets without resorting to massive new datasets. MAGDA++ is highly extensible with its underlying open-source data catalogue engine. Existing data can be ingested into the system easily. Its AI module adopts an Agentic architecture with customised user question routing engine and configurable large language models, which is also extensible.		Data is the key for meaningful predictions and responsible decision making in an organisation. Surfacing tacit and tested knowledge delivers outsized value. A solid data common enables AI for insight discovery and accumulating knowledge inside an organisation.		Responsible Entity Name	Open for Collaboration?
				CSIRO Agriculture and Food and CSIRO Data61	Interested in collaboration with other Commonwealth entities
				Area of Entity	Use Case Contact
				Software & Computational Systems	Chen Wang, chen.wang@csiro.au Jacky Jiang, jacky.jiang@csiro.au Jean-Philippe Ral, jean.ral@csiro.au Yipeng Zhang, yipeng.zhang@csiro.au
				Use Case Website/s	Use Case Owner
				https://www.csiro.au/en/news/All/Articles/2025/April/AI-for-plant-proteins	Liming Zhu, liming.zhu@data61.csiro.au



Screenshot/s





Screenshot/s

The screenshot shows the GovAI search interface. The search bar contains the query "what are the known enzymes that can degrade plastics?". The results are displayed in a tabular format with two columns: "Microbial host/Enzyme/gene" and "GenBank/ UniProt/NCBI". The table lists several entries, including *Isocornelia sakaiensis* 201-F6, *Caldimonas brevitalea*, *Pseudomonas pelagia* DSM 25163, *Halopseudomonas aestuarii* VGO14, *Pseudomonas mendocina* ATCC 53652, *Pseudomonas saulimassiliensis*, *Halopseudomonas bauzanensis*, *Pseudomonas alcaligenes*, and *Moraxella* sp. TA144. The interface also includes a chat window on the right with a "Send" button.

Microbial host/Enzyme/gene	GenBank/ UniProt/NCBI
<i>Isocornelia sakaiensis</i> 201-F6, ISPEase, ISF_6_483.1	AGADKBP617
<i>Caldimonas brevitalea</i> , PET12 (PbCutSbCut), AAW51_2473	AGAD03B00_98URK
	ADK73612
<i>Pseudomonas pelagia</i> DSM 25163, Ppelatip	KU695573
<i>Halopseudomonas aestuarii</i> VGO14, PE-H, B7088_11490	AGATH6AD45
<i>Pseudomonas mendocina</i> ATCC 53652, PmC	N20M54ZM016
<i>Pseudomonas saulimassiliensis</i> , PbCut	AGAD78MG8
<i>Halopseudomonas bauzanensis</i> , PbautCut	AGAD51MKR8
<i>Pseudomonas alcaligenes</i> , PbCut	SUD16364.1
<i>Moraxella</i> sp. TA144, Ip1, Mors1	P19833

Figure 2

The screenshot shows the GovAI search interface. The search bar contains the query "what are the active sites in ADADKBP617 that enable the plastic degradation function?". The results are displayed in a text format. The text describes the enzyme with UniProt accession number AGADKBP617, known as PETase, which is involved in the degradation of Polyethylene terephthalate (PET) plastic. The active sites that enable its plastic degradation function are primarily located at the following positions: Serine 160, Histidine 237, and Aspartate 206. The interface also includes a chat window on the right with a "Send" button.

The enzyme with UniProt accession number AGADKBP617, known as PETase, is involved in the degradation of Polyethylene terephthalate (PET) plastic. The active sites that enable its plastic degradation function are primarily located at the following positions:

- **Serine 160:** Acts as a nucleophile.
- **Histidine 237:** Part of the charge relay system.
- **Aspartate 206:** Also part of the charge relay system.

These active sites are crucial for the enzyme's ability to catalyze the breakdown of PET into its monomers.

To visualize these active sites in the 3D structure of the enzyme, you can explore the structure using the following code block:

Figure 3

Figure 2 Tabular and graph data retrieval-augmented generation (RAG)

Figure 3 interactive chat interface and tool integration for complex structured data



Screenshot/s

The screenshot displays the MAGDA++ Scientific Data Navigator interface. On the left, a search bar is visible with the text "Search for data". Below it, there are filters for "Any Organisation", "Any Location", "Any Format", and "Any Data". The search results are listed under "RESULTS (21)". The first result is "Continent-wide agronomic experiment data", which is a scientific dataset with a linked data rating of 4 stars and 1 heart, and is available in PDF, CSV, RDS, ZIP, and R formats. The second result is "A Principal Odor Map Unifies Diverse Tasks In Olfactory Perception", a demo dataset with a linked data rating of 4 stars and 1 heart, available in PDF, CSV, and PLAIN formats. The third result is "PAZy - The Plastics-Active Enzymes Database", a dataset with a linked data rating of 4 stars and 1 heart, available in CSV and XLSX formats. The fourth result is "Inhibition of ALDH2 by quercetin glucuronide suggests a new hypothesis to explain red wine headache", a dataset with a linked data rating of 4 stars and 1 heart, available in PDF format. The fifth result is "Research paper collection; fatty acid oxidation", which has no linked data rating and is available in N/A format.

On the right, a chat window titled "Chat to Magda" is open. It shows a user query: "Crop Prediction: How the chickpea protein percentage will change in Australia assuming the La Nina effect for the rest of this year. Please show the result on a map." The chat interface indicates that the system is "Determining the best assistant to serve your inquiry..." and has selected the "Crop Prediction Assistant". The assistant's response explains that to answer the query, it needs to consider the impact of La Nina on chickpea protein percentage, which is influenced by factors like rainfall, temperature, and soil moisture. It lists several features that are likely to be affected by La Nina, such as rainfall, temperature, and soil moisture. The assistant then provides a map of Australia showing the predicted changes in chickpea protein percentage across the country, with a table of data points. The table includes columns for SampleID, PredictedProteinPercentage, and Lat / Lon. The data points are as follows:

SampleID	PredictedProteinPercentage	Lat / Lon
1198	18.52	28.17516°S, 150.28571°E

The chat window also includes a "Send" button and an "Upload Files" button. At the bottom, there is a text input field for the user to type their question or drop files, and a "Send" button.

Figure 4 counterfactual question answering with integrated public data and specific prediction models



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

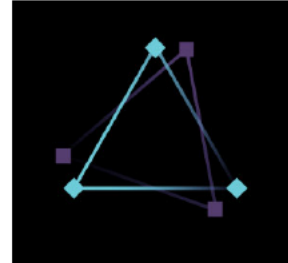
Responsible Organisation Category.....	7
Scope of the Use Case	8
Ethical Considerations	8
Value of the Use Case.....	8
AI Process Type.....	9
AI Technologies Utilised	10
Technical Elements.....	11

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	042 - Agriculture, forestry, fishing and hunting
☐ 05 - Environmental Protection	055 - Research and development in environmental protection
☐ 06 - Housing and Community Amenities	Choose an item.
☐ 07 - Health	077 - Research and development in health
☐ 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	Data quality control through metadata enrichment; prediction explanation to align with external constraints.
Privacy	Data are access controlled.
Rights of Users	We have not put transparency statements in the system yet, but users are informed about the AI nature of the system.

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	☒ Personalised services ☐ Public (citizen)-centred services
---	---



users, whether they are citizens or businesses.	☐ 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

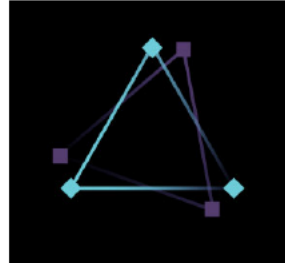


	☐ Supporting inspection processes ☐ Improving cybersecurity ☐ Registration and data notarisation processes ☐ Certification and validation processes
Internal management - Tasks that support agency management of resources, including employee management, procurement, and maintenance of technology systems.	☐ 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

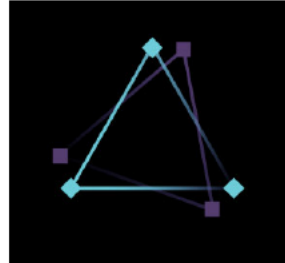
Select the types of AI technologies proposed / utilised to deliver the use case.

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	MAGDA++ can be deployed in private GCP platform or a native cluster with GPU resources.	
Model / Algorithm used	We use configurable GPT models and open weight models for different needs. The system also contains multiple specific ML models for curated data.	
Data Sources Select the types of data sources used and provide relevant details.	☒ Internal	☐ Third-party
	☒ Public	☐ Synthetic
	Details: MAGDA++ hosts a variety of data.	
Risk Assessment and Mitigation Details	Covered in executive summary.	
Security and Compliance Frameworks Select the security and compliance frameworks and measures implemented. Provide details or additional artifacts if relevant.	☐ Authority to Operate (ATO)	☐ Information Security Registered Assessors Program (IRAP)
	☐ System Security Plan (SSP)	☐ Penetration Testing
	Details: NA at this moment	
Assurance and Government Frameworks	NA	
Record maintenance	Design document and runtime logs are maintained in private cloud.	
Disengagement	NA	
Performance Metrics and Results	Specific ML models are evaluated using standard metrics such F1 score, roc, CCC etc. Interaction and decision support system are measured through human feedback.	