AI x Engineering - Transforming Organisational Capability for Sustainable Impact

Group Plan

Group Name: Maccas

Yilin Chen

Shuo Du

Mohan Guo

Shiying Shen

Tongxuan Zhao

Andrew Li

Supervisor: Dr. Fabian Held

Faculty of Science
The University of Sydney
Australia

7 September 2025

Word Count: 1897

Table of Contents

1. Introduction and Problem Statement

1.1 Introduction

In the modern digital-driven economy, a company's data has become a critical asset of competitive advantage (Mohan et al., 2025). However, for many organizations with a long history, the effective management and utilization of vast amounts of accumulated unstructured legacy data remain a pressing challenge. These historical data assets often exist in isolated and unstructured forms, such as paper archives, handwritten notes, and scanned documents, which severely impact the employee’s efficiency and collaboration, even forcing them to perform tasks they have already done (*How Poor Knowledge Management Is Harming Your Business*, 2024).

Tonkin Engineering, a leading engineering consulting firm with over seven decades of history, is facing this problem. Although its archives contain a wealth of intellectual assets, the lack of a structured management system has left this information siloed, leading to inefficiencies and severely impacting the company's ability to learn and innovate. This project aims to explore how artificial intelligence can address this complex problem, and figure out how to allow the engineers using natural language to quickly retrieve and summarize relevant data (Chen, 2024).

1.2 Stakeholder Perspectives

Framing Tonkin’s data management problem through the perspectives of key stakeholders shows how different groups experience distinct challenges.

For the industry partner and company employees, the problem lies in the fragmented and unstructured nature of Tokin’s archives. More than seventy years of project reports, drawings and inspection data remain scattered and difficult to retrieve. This results in low efficiency in daily operations and increases the risk of knowledge loss when staff leave the organization. At the same time, new employees encounter difficulties during onboarding and training, as they cannot easily access or learn from past decisions.

For clients, the problem shows lower trust and less clarity in project delivery. When past data and technical records cannot be found quickly, Tonkin struggles to give consistent and fact-based advice. This can lead to delays in reports, differences in quality between projects, and difficulty showing how current solutions build on proven experience.

For external collaborators, the lack of a unified system creates barriers in communication and teamwork. Scattered records mean that partners often cannot access the same information at the same time which leads to different understandings of the same issue. This raises the risk of repeated work, inconsistent standards, and project delays caused by confusion. Since the underlying knowledge base is not well organized or open, collaborators also find it difficult to keep their own practices aligned with Tonkin’s way of working.

The development team also faces challenges. We must deal with scatter data, unclear requirements, and high system complexity. These issues slow down building and testing the system, and make maintenance and updates more difficult.

By considering the needs of employees, industry partners, clients, external collaborators and development team together, it becomes clear that Tonkin's data management problem is both complex and multi-layered. The lack of a structured and unified system creates hurdles for both company employees and external partners.

1.3 Problem Statement

Tonkin Engineering's principal challenge is that data remains unstructured and siloed, making it difficult to access through existing systems. This severely degrades archival retrieval efficiency—for example, due to messy metadata and inconsistent naming—and undermines the value of legacy data.

We will design an enterprise-grade database that integrates a data repository system with AI models to unify heterogeneous historical records and deliver intelligent retrieval and analysis. The goal is to unlock the reusable value of legacy data, empowering cross-project learning, project delivery, and business innovation, and helping Tonkin Engineering sustain its competitive edge in engineering consulting.

Our Aim:

1. **Governed, Maintainable Repository:**
    - **Heterogeneous data backbone:** Ingest and store drawings, images, PDFs, and structured tables under a consistent information architecture.
    - **Sustainable Maintenance:** An easy-to-maintain system with streamlined operations and automated housekeeping to keep the repository reliable and cost-effective over time.
    - **Privacy & Safety:** Implement a secure repository with layered, role-based access controls; minimise internal data leakage.
2. **AI Empowerment:**
    - **Intelligent Assistance:** Provide assistive capabilities such as rapid understanding and summarization, project analysis, and visual presentation.
    - **Related content & suggestions:** Surface similar projects and related assets (photos, drawings, and documents) tied to the search results.
3. **Semantic Retrieval & Enhanced Fundamental Search:**
    - **Semantics-first discovery:** Interpret user intent and context to surface conceptually related items; handle synonyms, abbreviations, and domain terms.
    - **Smarter keyword search:** Improve matching and ranking with typo tolerance, phrase, field weighting, and recency/authority boosts.

- **Cross-format coverage:** Search across drawings, PDFs, images/CCD, emails, and structured records.

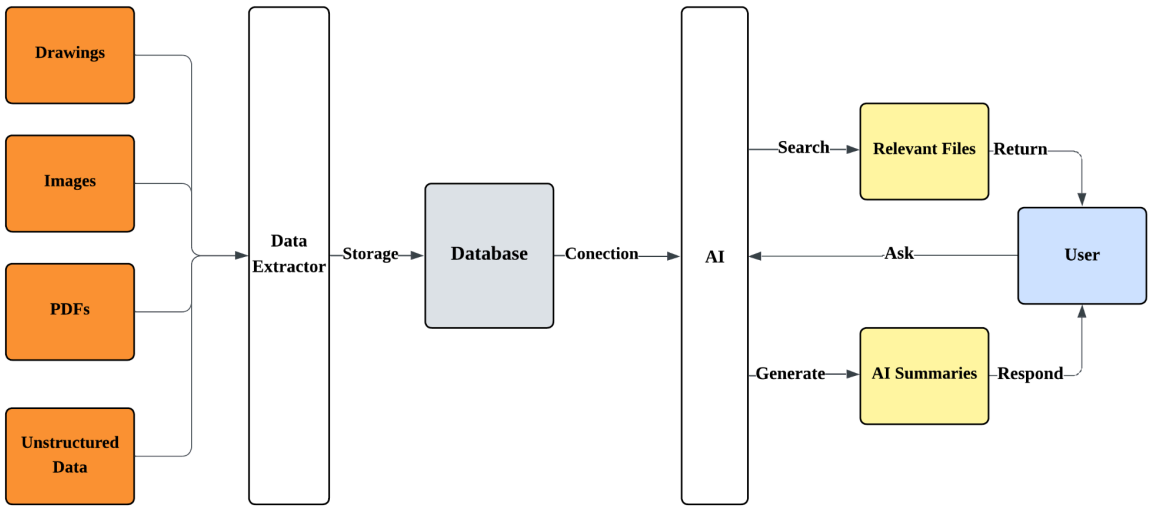

Figure 1. Project Architecture Example

2. Research Questions and Plan

2.1 Research Questions

The central aim of our project is to explore methods that can help Tonkin unlock the value of their numerous legacy data by transforming it into a structured and searchable knowledge system. To achieve this, we have identified three research questions that address the major unknowns in this challenge:

1. **Extraction**: What efficient and robust AI methods are available for extracting information from diverse legacy data formats, including typed text, handwritten notes, and engineering drawings?
2. **Storage and Connection**: What approaches are most effective for storing and connecting extracted information in ways that capture relationships between project details and professional decisions?
3. **Search and Question Answering**: What AI methods are able to search and answer questions efficiently and accurately from the information that has been retrieved and stored?

Together, these research questions define the scope of our project by addressing the full process from data extraction to knowledge application, ensuring that both technical and organizational perspectives are considered carefully.

2.2 Integrated Approach

To address our problem statement, it is difficult to rely on a single disciplinary perspective. Instead, our project requires an integrative approach that combines insights from each team member. We begin by pooling our existing expertise to establish a comprehensive understanding of the problem. While the challenge is broad and complex, parts of it align with our background knowledge, such as data science and machine learning. This process allows us to identify what we already know and, equally importantly, what we still need to investigate. As (Repko & Szostak, 2021) notes, interdisciplinary integration begins with identifying the relevant disciplines and drawing together our insights to define both strengths and gaps in knowledge, which in turn informs the formulation of interdisciplinary research questions.

For areas where no team member has prior experience, we plan to adopt an agile development mindset. Agile emphasises iterative progress, starting with high-level goals and refining methods as understanding develops (Beck et al., 2001). This approach is particularly suited to our project, where many details will only become clear as we engage with the problem and review evidence. By combining disciplinary knowledge with an iterative, adaptive process, we aim to build a plan that is both grounded in expertise and flexible enough to respond to new insights as they emerge.

2.3 Research Activities and Methods

To address our three research questions, our group has identified a set of research activities that will allow us to gather evidence, evaluate potential methods, and formulate well-grounded recommendations. These activities combine literature review, stakeholder engagement, and comparative evaluation.

2.3.1 Literature Review

We will begin by conducting a targeted literature review to identify methods that address each of the three research questions. For **extraction**, this will involve reviewing academic work on natural language processing for typed text, such as Transformer-based models (Vaswani et al., 2017), as well as computer vision approaches to interpreting complex engineering drawings (Jamieson et al., 2024). For **storage and connection**, we will examine research on knowledge graph construction (Zhao et al., 2023) and the use of vector databases and vector embeddings to represent high-dimensional relationships in engineering data (Kukreja et al., 2023). For **search and question answering**, we will analyse literature on semantic search, case-based reasoning, and retrieval-augmented generation (Lewis et al., 2021), which has been shown to enhance access to knowledge-intensive domains.

This review will allow us to understand the technical landscape across these three areas, highlight the strengths and weaknesses of different approaches, and establish a foundation for evaluating their suitability to Tonkin's context.

2.3.2 Stakeholder Engagement

Engaging with Tonkin's stakeholders will be essential for aligning potential methods with real organizational needs. We plan to capture the perspectives of engineers, project managers, and executives to understand their needs. For example, engineers may prioritize quick retrieval of technical details, while executives may emphasize decision-making based on past experience. These perspectives will help us evaluate whether candidate methods are not only technically feasible but also trusted and practical in real work.

2.3.3 Comparative Evaluation

We will then articulate insights from the literature review and stakeholder perspectives into a comparative framework. Each potential method for extraction, storage and connection, and search will be assessed against criteria such as accuracy, scalability, and organizational fit. This structured evaluation will allow us to balance trade-offs and provide evidence-based recommendations for Tonkin.

2.4 Limitations

We recognize several limitations in our approach. First, access to Tonkin's legacy data may be restricted due to confidentiality or sensitivity, limiting opportunities to test methods on real material. Second, technical methods themselves have limitations. Optical character recognition often struggles with unclear or illegible handwritten notes, and computer vision may face performance degradation on complex engineering drawings. Third, the timeframe of the project restricts us to a high-level investigation rather than a full implementation. Finally, interdisciplinary collaboration poses challenges, as different disciplinary perspectives may use different terminology and assumptions, requiring additional effort to build shared understanding within the team.

2.5 Team Allocation

We divided the team into three main groups based on the Research Questions: **Data Extraction**, **Data Storage and Connection**, and **AI Implementation**, and assigned members according to their skills.

Name	**Skills**	**Allocation**
Mohan Guo	Coding, Model Training, Data Analysis	Data Extraction

Shiying Shen	Python, Java, C, SQL, R, Weka, Machine Learning, Data Structures and Algorithms, OOP, Data Analysis	Data Extraction
Sihao LI	Coding, Machine Learning, EDA, Visualizations	Data Storage and Connection
Tongxuan Zhao	Python, SQL, R, Data Analysis, Data Structure and Algorithms	Data Storage and Connection
Shuo Du	Data Analysis, Research & Literature Review, Teamwork in Interdisciplinary Contexts	AI Implementation
Yilin Chen	Python Programming, Machine Learning, Research Methods	AI Implementation

Additionally, we have a responsibility for **Report Writing**, which requires the participation of all team members.

3. Anticipated Benefits

Our proposed system will create substantial value for Tonkin and its stakeholders by consolidating decades of files in multiple formats into a unified, searchable database. This will enable employees to efficiently access engineering knowledge and historical records, reducing duplication and improving decision-making. An AI-powered search and conversational interface will further enhance accessibility, allowing users without technical expertise to obtain relevant insights with ease.

For the industry partner, the solution represents a long-term investment in knowledge management and sustainable practice. It will improve productivity by reducing the time needed to locate historical information, while supporting training and knowledge transfer for new employees. The development team will benefit from clearer data organization and standardized workflows, which simplify testing, debugging, and maintenance, reducing future technical debt. Clients and external collaborators will also gain faster, more reliable access to project knowledge, resulting in better service delivery and outcomes.

Meanwhile, the project shows how legacy engineering data can be transformed into a dynamic, reusable resource, driving broader innovation. It offers a model for sustainable data management, aligns with digital transformation initiatives across the engineering sector (Sodola, 2025), and may contribute to future research in information retrieval, semantic search, and applied AI.

4. Timeline

Our project timeline begins in Week 5 with a partner engagement session to confirm Tonkin's needs and define the practical scope of our work. This early interaction ensures that our research direction remains anchored to real organizational priorities.

From Week 6 through Week 10, we run three overlapping strands of research. The Literature Review starts in Week 6 and continues until the end of Week 10, providing the foundation for understanding existing AI methods for extraction, storage, and search. From the first day of Week 7, we begin Literature Review Synthesis, consolidating individual findings into shared insights. Starting from Week 8, we also develop the Comparative Evaluation Framework, running in parallel with the other tasks until the final day of Week 10. This structure ensures that by the end of Week 10, we have both comprehensive knowledge and a tested evaluative framework that considers technical feasibility, scalability, and organizational fit.

In Week 11, our focus shifts to integration and rehearsal, combining literature findings, comparative evaluation, and sample data testing into a coherent prototype. Week 12 is dedicated to delivering the group presentation to Tonkin. Finally, in Week 13, we will submit the final report, incorporate presentation feedback and ensure evidence-based recommendations.

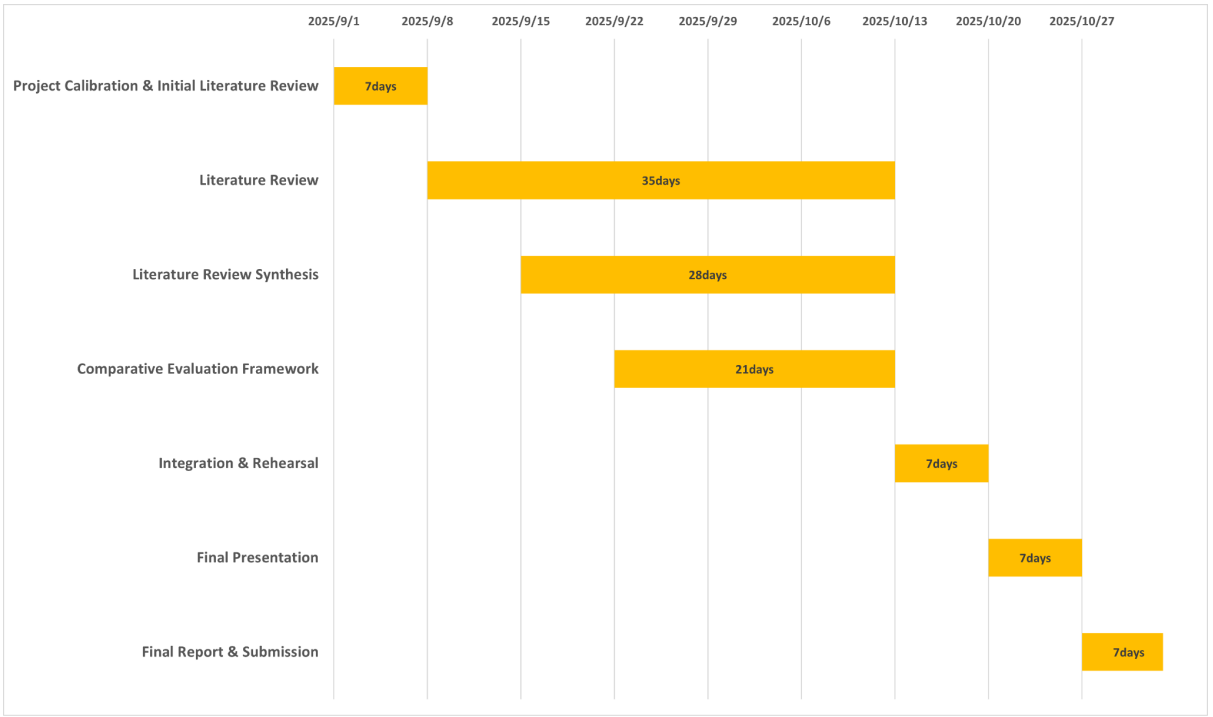

Figure 2. Project Timeline

5. AI Statement

We acknowledge the use of **ChatGPT** (version GPT-5, OpenAI, https://chatgpt.com) for assistance in this work.

Specifically, we used AI to summarise our initial ideas, allocate tasks according to our individual strengths, polish our wording, perform grammar checks, and assess whether the issues stakeholders might face are reasonable.

References

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., & Thomas, D. (2001). *Manifesto for agile software development*. Agile Manifesto. https://agilemanifesto.org/

Chen, E. (2024). Empowering artificial intelligence for knowledge management augmentation. *Issues in Information Systems*, *25*(4). https://doi.org/10.48009/4_iis_2024_132

How poor knowledge management is harming your business. (2024, April 26). Www.rezolve.ai. https://www.rezolve.ai/blog/how-poor-knowledge-management-is-harming-your-business

Jamieson, L., Carlos Francisco Moreno-García, & Eyad Elyan. (2024). A review of deep learning methods for digitisation of complex documents and engineering diagrams. *Artificial Intelligence Review*, *57*(6). https://doi.org/10.1007/s10462-024-10779-2

Kukreja, S., Kumar, T., Vishal Bharate, Purohit, A., Dasgupta, A., & Guha, D. (2023). Vector databases and vector embeddings-review. *International Workshop on Artificial*

Intelligence and Image Processing. https://doi.org/10.1109/iwaiip58158.2023.10462847

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*. https://doi.org/10.48550/arXiv.2005.11401

Mohan, S. K., Bharathy, G., & Jalan, A. (2025). Enterprise data valuation—a targeted literature review. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12705

Repko, A. F., & Szostak, R. (2021). *Interdisciplinary research: Process and theory.* (4th ed.). Sage Publications.

Sodola, O. O. (2025). Digital transformation in engineering project management.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Zhao, Z., Luo, X., Chen, M., & Ma, L. (2023). A survey of knowledge graph construction using machine learning. *Computer Modeling in Engineering & Sciences*, *139*(1), 225–257. https://doi.org/10.32604/cmes.2023.031513