Autonomous Data Cataloging Agent

Field Session Project Proposal – Summer 2026

Company Background

Infinity Technology Systems is a software and managed IT company on a mission to bring modern technology to one of the most underserved industries in the world - construction and trades. While other sectors have been transformed by software, this industry still runs on spreadsheets, paper forms, and phone calls. That's the gap we're closing.

We build purpose-built software suites, AI-powered workflows, and full infrastructure solutions designed around how these businesses actually operate. The problems are real, the technology is cutting-edge, and the opportunity to make a meaningful impact is enormous.

We're growing fast and building in real time. The projects we bring to Field Session are genuine contributions to that roadmap, not classroom exercises. Students who work with us are building things that get used.

Project Background

ITS is building an autonomous data cataloging agent purpose-built for the construction industry. Large construction businesses manage a massive volume of documents across servers, SharePoint libraries, and local drives. With no consistent structure, no tagging, and no unified way to search across it all, that data is effectively invisible. To ensure the agent is built around the real operational demands of these environments, Infinity Commercial and Industrial is serving in an advisory capacity throughout the project, bringing deep firsthand experience with exactly these challenges.

The solution is an autonomous AI agent that crawls a company's entire file infrastructure on a schedule, reads documents across multiple formats, and maintains a structured, searchable catalog without any manual input. What sets this apart from a simple indexing tool is that the agent genuinely reasons about each file, classifying it by type, extracting key metadata, assigning tags, and flagging duplicates based on content rather than filename or folder location. The agent decides how to handle each document, what to extract, and how to categorize it through a language model with tool-use capability, where those tools include filesystem traversal, document parsing, content classification, and database writes. Building that reasoning loop and getting it to perform reliably on a real enterprise file system is the core engineering challenge.

The catalog is built to feed directly into downstream AI applications, documents will be chunked and embedded for semantic search, labeled for potential use in model training, and tagged to be findable by content, date, project, and division. The student team will also deliver a web dashboard to browse, search, and audit everything the agent has cataloged. This is real infrastructure that gets used. Every AI tool ITS builds after this one runs on top of what this agent produces. For students interested in agentic systems, data engineering, or enterprise software, this is a project with real stakes and a tangible output.

Recommended Stack

We've thought through the architecture and have some initial direction to share, but we want the team to own the technical decisions. For the agent framework, we've been looking at LangGraph, AutoGen, and the Anthropic Agent SDK all of which seem well-suited to the kind of reasoning loop this system needs. For document parsing, PyMuPDF and python-docx cover the file formats we typically work with. The catalog data needs to live somewhere structured and queryable, and we've been considering PostgreSQL and Azure SQL. Python feels like the natural language for this kind of work, and a React dashboard makes sense for the front-end browsing interface.

That said, these are starting points, not requirements. If the team has a different take on the stack or sees a better way to approach the problem, we genuinely want to hear it. The goal is the best solution, not a specific set of tools.

Desired Skillset

- Experience with Python backend development, or a strong willingness to learn
- Interest in agentic AI systems and LLM tool-use
- Familiarity with databases and structured data
- Interest in data engineering and document processing pipelines
- Comfortable working in an enterprise software context with real production stakes
- Adaptable and self-directed — this project requires independent research and the ability to pick up new tools on the fly

Preferred Team Size

We recommend a team of 3 to 5 students to ensure meaningful collaboration and a well-rounded skill set across the project.

Location

All work and collaboration for this project can be done remotely. The ITS team will provide consistent support and regular feedback throughout the engagement via virtual meetings and online communication tools. Infinity Commercial and Industrial will also be available in their advisory capacity for industry context and guidance as needed.