Software Lab:

Unified Search Engine for Distributed Project Files

Description

The Problem

Professionals in the AEC-O industry must navigate multiple isolated systems—local drives, cloud storage (e.g. Box), network drives (NAS/SFTP), and email—to locate project documents. Consider a construction manager trying to locate meeting minutes where the smart-building concept was discussed: they must search their laptop's Downloads folder, then switch to Box, then check their email attachments—each system with different search capabilities and interfaces. This fragmented workflow leads to wasted time, missed information, and duplicated effort.

The Solution

A cross-platform desktop application (macOS, Windows, Linux) that creates a unified, searchable index across multiple file sources. The application connects to various data sources, extracts text from documents, and provides a powerful search interface with both a visual file explorer and a search-first paradigm.

Core Features:

- Multi-source connectors: Local filesystem + Box API
- Text extraction from PDFs, Word, and Excel documents using Extractous
- Full-text search with fuzzy matching using Tantivy search engine
- Metadata filtering (date, file type, source)
- Hybrid interface: Visual file explorer combined with search-first paradigm

Optional/Bonus Features:

- Email connector (IMAP/Exchange)
- NAS/SFTP connectors for network drives
- Semantic search with vector embeddings
- Mobile companion apps (iOS/Android)

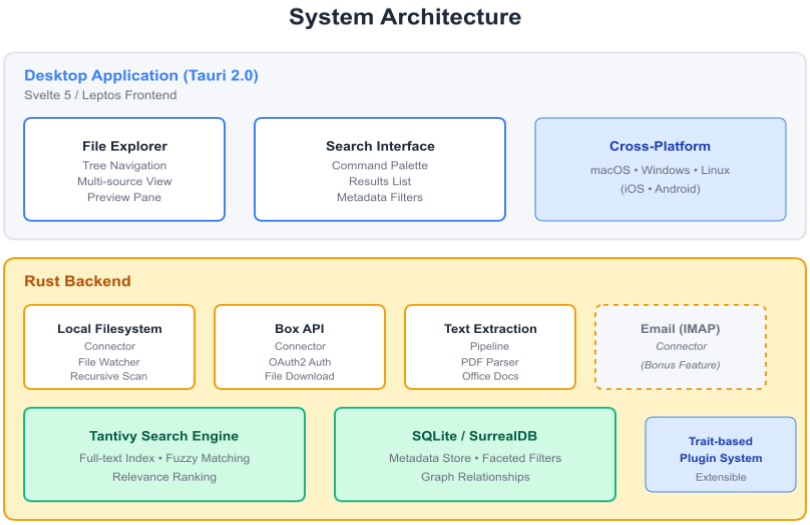

Figure 1: System Architecture

Technology Stack

- **Rust** backend for performance and safety
- **Tauri 2.0** framework for cross-platform desktop application
- **Svelte 5 or Leptos** for the frontend UI
- **Tantivy** for full-text search indexing
- **Extractous** for text extraction from documents
- **SQLite or SurrealDB** for metadata storage

Inspiration: The project draws inspiration from SpaceDrive's Virtual Distributed Filesystem concept, which provides a unified view across multiple storage locations.

Figure 2: User Interface Concept

Task

The project is divided into the following work packages:

1. **Project Setup & Architecture Design:** Set up Tauri 2.0 project structure, define data models, design connector interfaces
2. **Local Filesystem Connector:** Implement directory scanning, file watching for changes, metadata extraction
3. **Text Extraction Pipeline:** Integrate Extractous for PDF, Word, Excel text extraction
4. **Search Engine Integration:** Set up Tantivy indexing, implement full-text search with fuzzy matching
5. **Box API Connector:** OAuth authentication, file listing, content retrieval, change detection
6. **User Interface:** File explorer view, search interface, preview pane, filter controls
7. **Integration & Testing:** End-to-end testing, performance optimization, documentation

Bonus tasks (optional): Email connector (IMAP/Exchange), NAS/SFTP connector, semantic search with embeddings, mobile apps.

Supervisor

Sylvain Hellin, Chair of Computing in Civil and Building Engineering (CCBE),
sylvain.hellin@tum.de

References

1. Tauri Framework: https://tauri.app
2. Tantivy Search Engine: https://github.com/quickwit-oss/tantivy
3. Box API: https://developer.box.com
4. Extractous: https://github.com/yobix-ai/extractous
5. SpaceDrive: https://spacedrive.com
6. SurrealDB: https://surrealdb.com