

Discovery OCR & Party Information Extraction Tool

Team Members: Jennifer Lias | Peter Uy | Thomas Ogden | Nadia Hernandez | Lemeng Zhao | Daniel Concepcion | Joseph Lam | Addison Zhou | Jesus Villa | Jose Holguin | Tommy Works

Faculty Advisor: Dr. Jung Soo Lim

Santa Barbara Public Defender's Office

Liaisons: Alexander Voisan | Angella Stokke | Shawna Mateer | Jalena Jones | Luis Ramirez

Department of Computer Science

College of Engineering, Computer Science, and Technology

California State University, Los Angeles



Background & Problem

Background

Within the Santa Barbara Public Defender's Office, Legal Office Professionals manually extract police report party information from terabytes-worth of discovery documents and input it into their Case Management System (CMS).

Problem

Extracting report party information to prepare for CMS is a manual, multi-hour process that is tedious and error-prone.

Objectives

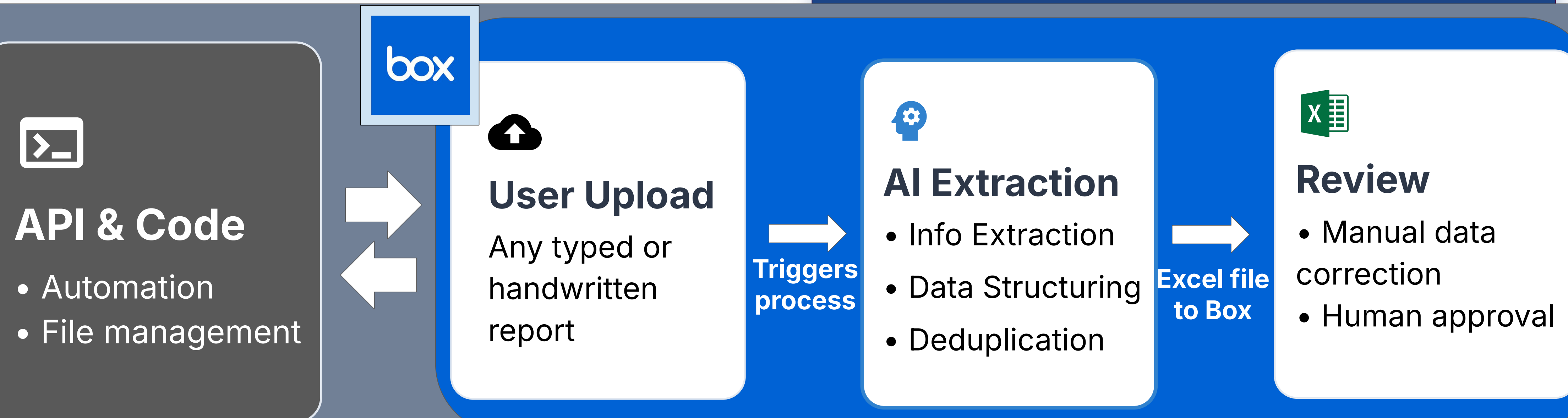
Main: Develop an extraction tool that integrates AI and Box features to reduce manual processing effort by **at least 50%**.

Also important:

Security: Files must remain in Box.com.

Accuracy: Extract **90%** of relevant data correctly

Workflow



FEATURES

Final Structured Export File

Validated data in JSON file export-ready for outside Content Management System

Mar 5 at 4:00 PM	LPD Initial 2_1772755200.xlsx	Complete
Mar 5 at 3:57 PM	LPD Initial 2_1772755048.xlsx	Stalled
Feb 5 at 10:49 PM	C-SBSO Supplemental Report (1)_176...	Complete
Feb 5 at 10:46 PM	C-SBSO Supplemental Report (1)_176...	In Progress

Figure 2: File Auditing & Tracking features

Figure 1: Sample Excel Extraction Showing Party Types

Party Type	Name	Role	Date of Birth
Person	Uy, Peter	arrestee	05-09-1976
Person	Hernandez, Nadia	other	07-03-2001
Person	Lam, Joseph	other	01-01-1984
Officer	Lias, Jennifer	Officer	
Officer	Zhao, Lemeng	Officer	

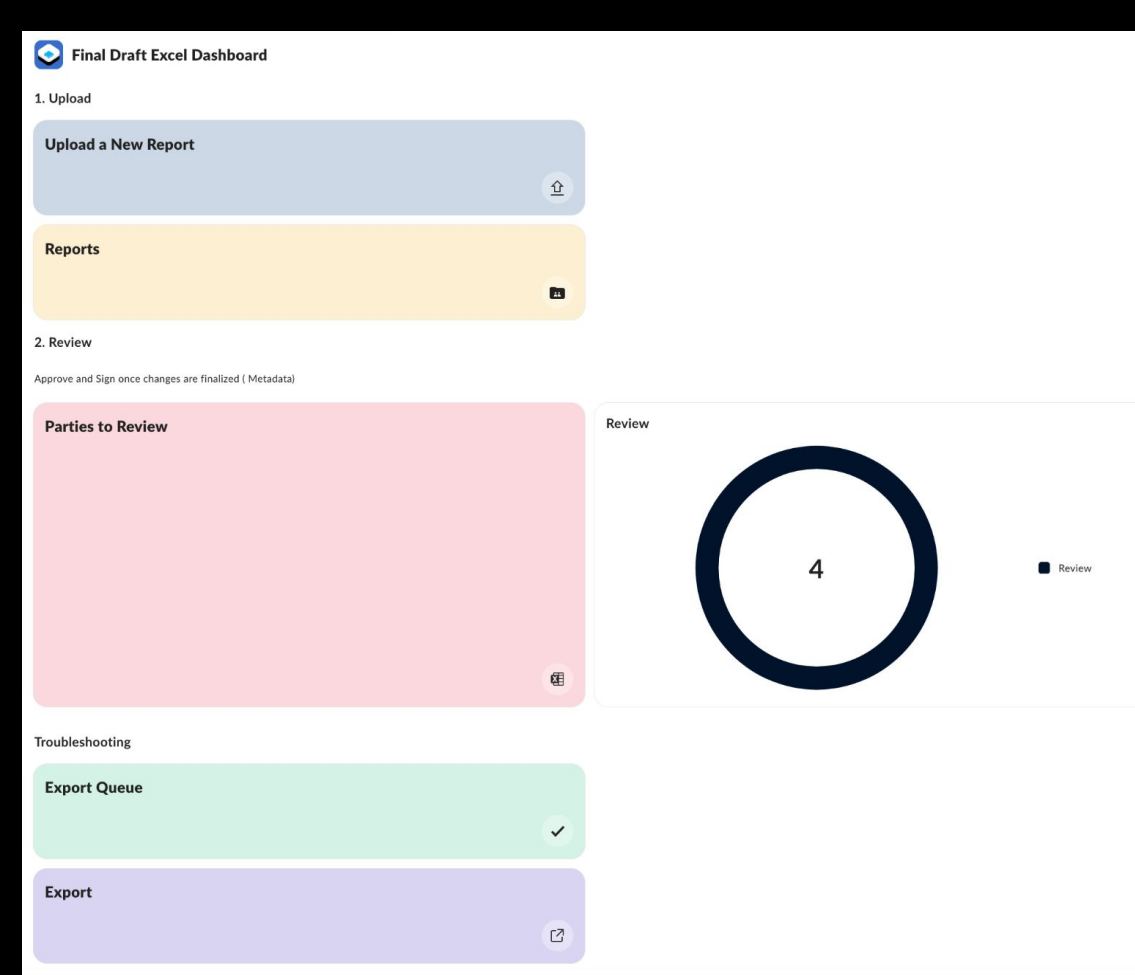


Figure 3: User Dashboard interface

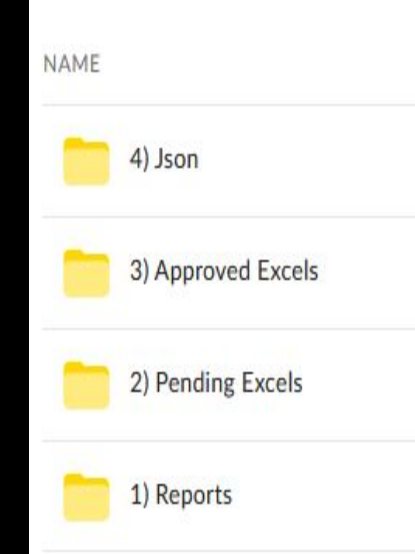


Figure 5: Simple Box folder structure that connects the whole workflow together.

- Box Relays
- Dashboards
- Code (API & .py libraries)

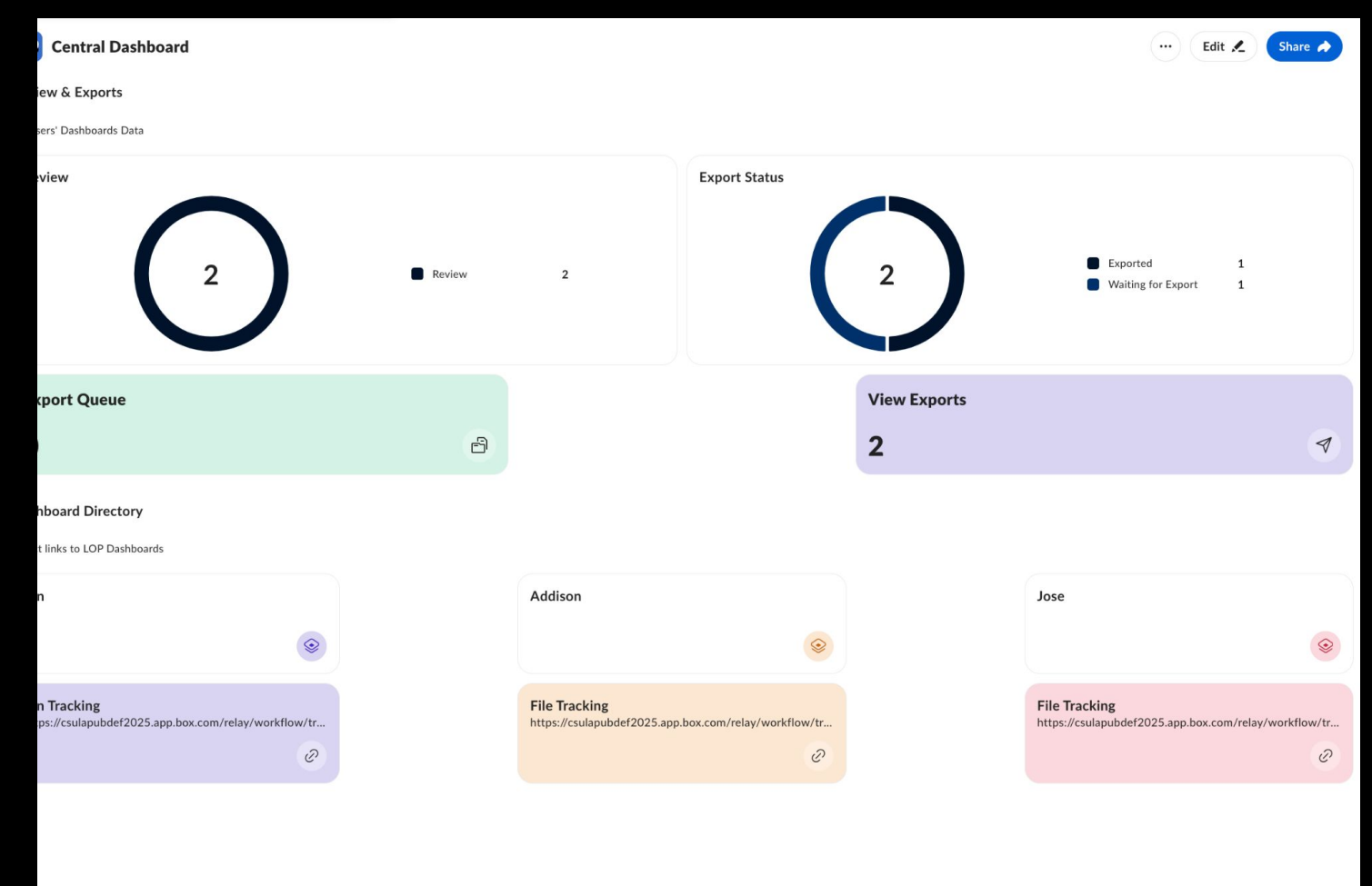


Figure 4: Administrator Dashboard interface

Accomplishments

A fully functional OCR and AI extraction system, halving manual data entry time!

Achieves at least **96% extraction accuracy** and **100% BATES identification accuracy**, while maintaining compliance with the county's security policies.

Processes and organizes key information from reports automatically, converting unstructured data into structured, usable outputs, all within a secure, Box-based environment.

Discussion & Future Work

Future work will expand data extraction to more Discovery types, fully transition to the Box interface by removing the API, and enable direct CMS export while maintaining human verification in the workflow

Scan the QR code for the demo video

