

# Team Meeting Notes – Oil Wells Data Wrangling

Team Name: [Your Team Name]

Members: [Member 1], [Member 2], [Member 3]

Assignment: Lab 6 – Oil Wells Data Wrangling

---

## Thursday – Initial Setup and Planning

Mode: Virtual

Duration: ~45 minutes

What we did: - Carefully reviewed the full assignment instructions. - Identified major components: PDF extraction, database storage, web scraping, and mapping. - Agreed on using Python (OCR + parsing), MySQL, and OpenLayers for visualization. - Set up shared development environment and project structure.

Main takeaway: We clarified that each PDF represents one well-specific dataset and structured our scripts accordingly.

---

## Sunday – Core Development (Heavy Work Day)

Mode: Virtual

Duration: ~3–4 hours

What we did: - Finalized database schema for wells and stimulations tables. - Implemented full PDF ingestion script with OCR processing. - Removed early-stop OCR logic and ensured all pages are processed. - Built robust parsing logic for: - API number - Well name - Latitude / Longitude (DMS + decimal) - Operator, county, address - Implemented complete stimulation data extraction including: - Date stimulated - Formation - Top/Bottom depth - Stages - Volume and units - Treatment type - Lbs proppant - Acid % - Max treatment pressure - Max treatment rate - Ensured safe UPSERT logic to avoid duplicate API crashes.

Issues encountered: - OCR inconsistencies across pages. - Overly strict regex patterns missing fields. - Address field exceeding DB size.

Fixes: - Made parsing more flexible. - Added truncation safeguards. - Reworked stimulation parsing to match assignment snapshot exactly.

---

## **Monday – Integration and Web Scraping (Major Work Day)**

Mode: Virtual

Duration: ~4+ hours

What we did: - Integrated PDF parsing with MySQL database fully. - Verified that all required stimulation fields are saved. - Implemented web scraping using API # and well name. - Extracted: - Well status - Well type - Closest city - Oil production - Gas production - Added preprocessing to clean text and handle missing values. - Ensured database entries are consistent and complete. - Tested multiple PDFs end-to-end.

We also verified: - All PDFs are processed completely. - No early termination during OCR. - All relevant fields from assignment figures are extracted and stored.

---

## **Overall Summary**

Most of the development and debugging work was completed on Sunday and Monday.

Thursday was used for planning and setup.

We ensured: - Every PDF is fully processed. - All required well-specific and stimulation data fields are extracted. - All extracted data is stored correctly in the database. - Web scraping integrates properly with stored records.

The system is complete and ready for visualization in Part 2.