Automated Extraction of Well Coordinates from Oklahoma 1002A Records

Priyank Jaiswal, PhD
Professor, Boone Pickens School of Geology
Editor-in-Chief, Sustainable Geosciences
Oklahoma State University, USA
https://experts.okstate.edu/priyank.jaiswal
priyank.jaiswal@okstate.edu

Akshay Kumar Mundrathi, MS
Oklahoma State University, USA
https://www.linkedin.com/in/akshay-kumar-mundrathi/
amundra@okstate.edu

AAPG 4th Annual Orphan, Abandoned, Idle, and Marginal Wells: Opportunities with Legacy Assets

Problem Statement

Oklahoma 1002A well records exist as scanned PDFs where critical location information (County, Section-Township-Range (STR), and grid position) is embedded visually and often degraded by scan artifacts.

Manual STR-based latitude/longitude computation is slow, error-prone, and difficult to scale.

An automated, validated system is required to extract and convert these records into accurate geospatial coordinates.

Input Example (1002A Form)

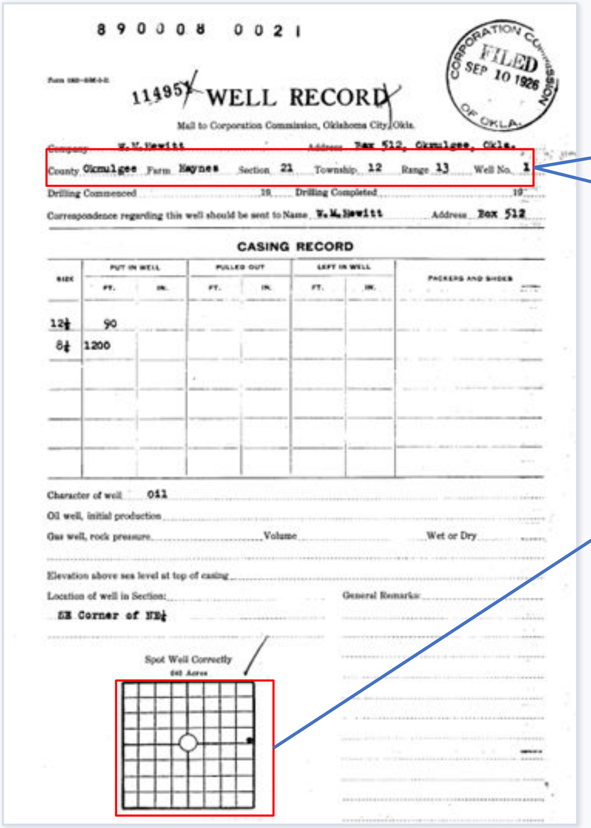

Figure 1: Scanned well form with mixed text and "spot well correctly" grid.

Core Extraction Targets

County: Identifies the regional administrative area.

STR: Section, Township, and Range identifiers.

Well Grid: The "Spot well correctly" coordinate grid.

The pipeline transforms unstructured scans into machine-readable spatial data.

Grid Interpretation

Scanned PDFs are converted into high-resolution images to enable precise grid analysis. Image processing techniques correct for skew, noise, and stamps to reliably detect and standardize the "spot well correctly" grid across varying document conditions.

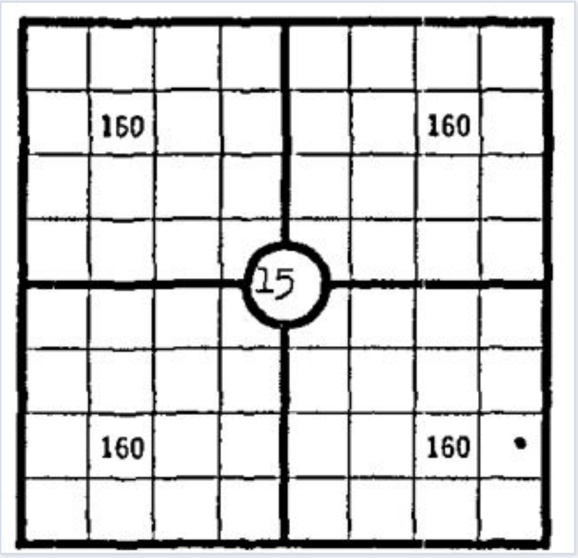

Figure 2: Extracted "spot well correctly" grid from the record.

Ongoing Work: U-Net (segmentation model) Based Grid Dot Detection

An ongoing effort uses a U-Net segmentation model to improve dot localization by learning from automatically generated image-mask pairs. The predicted dot centroid will be integrated into the geospatial pipeline for accurate latitude and longitude computation.

Pipeline Overview

PDF rendering → 2× high-resolution page images

Text branch

- Unified crop of county and STR block
- OCR (Optical Character Recognition) (Google Vision) → raw text
- LLM (Large Language Model) based vision model (Gemini) → normalized county/STR and confidence

Grid branch

- Detect & crop well grid region PDF → image processing ensemble
- Dot centroid → mapped to (row, column) on 8×8 grid
- (row, column) → PLSS subdivision label

Merge branches by image_id → combined CSV

SQL PLSS lookup (AWS RDS) → retrieve section corner coordinates

Bilinear interpolation → latitude/longitude

Outputs: combined CSV with computed coordinates and interactive map

Methodology

Well record PDFs are processed to extract both structured text (County and STR) and grid-based well position. OCR with validation ensures standardized STR values, which are then matched to a PLSS database to compute latitude and longitude. Final well locations are exported to a CSV and visualized on an interactive map for quality control.

Disclaimer

CSV Output Fields

The CSV includes image ID (filename), normalized STR values, PLSS subdivision labels, and computed latitude/longitude coordinates of the well.

filename	section	township	north_south	range	east_west	county_name	latitude	longitude
00000000_No Data_548897	21	12	N	13	E	Okmulgee County	35.50734379	-95.92929207
00000000_No Data_1030447	11	14	N	12	E	Okmulgee County	35.7110032	-95.99923203
00000000_No Data_548922	25	27	N	13	E	Washington County	36.80033860298848	-95.86499563
00000000_No Data_1220867	34	19	N	12	E	Tulsa County	36.08842026968229	-96.01372407
00000000_No Data_689375	34	10	N	3	E	Pottawatomie County		

Table 1: Final CSV file with required columns.

Map Visualization

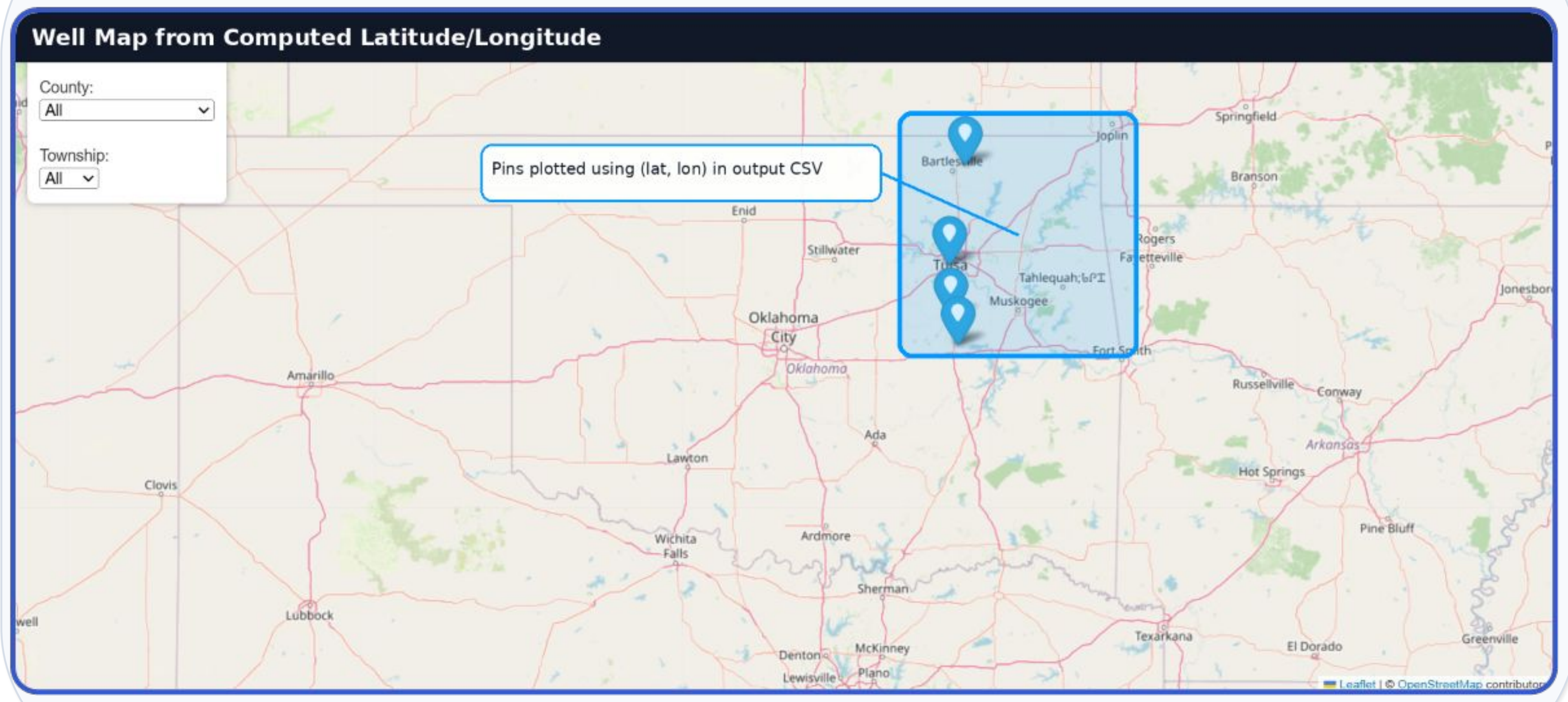

Figure 3: Interactive map of computed well locations.

Limitations

Dot localization remains the primary bottleneck due to scan noise and distortions, despite reliable PLSS-based coordinate computation.

Future Work and Conclusion

Future Work
Improve grid dot localization using U-Net segmentation and enhanced training data, replacing classical computer vision methods with a more robust learning-based approach.

Conclusion
This end-to-end system converts scanned 1002A well records into validated, GIS-ready coordinates, enabling scalable transformation of legacy land records into structured geospatial datasets.

Acknowledgements

This material is based upon work supported by the Department of Energy Office of Fossil Energy and Carbon Management under the following:
Award number: DE-FE0032362; Award Title: Anadarko Basin Carbon Management Hub;
Award Start and End dates: 07/01/2024 – 06/30/2026.
Data and Material were provided by Devon Energy