



catalog.energy.gov

Extracting Information from Historic Oil & Gas Regulatory Documents

2024 NETL Resource Sustainability Review Meeting April 4th, 2024





Project Overview



catalog.energy.gov

- FECM Funded: *FWP-1025025*
- Undocumented Orphaned Well Program
- Collaborative effort under CATALOG
 - WP Leads: Greg Lackey (NETL), Dan O'Malley (LANL), David Buttler (LLNL), Dan Gunter (LBNL), Hamilton Link (Sandia)
- Performance period:
 - Five-year program started at the beginning of fiscal year 2022







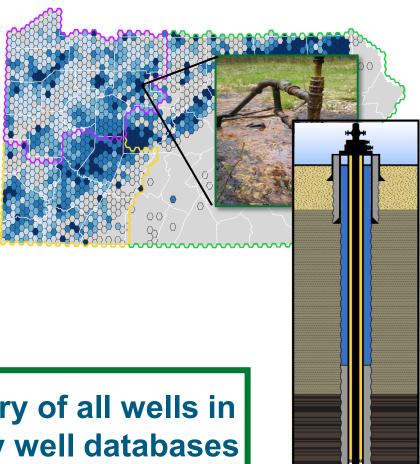






Most Well Information is Not Digital

- Obvious lack of information for undocumented orphaned wells (UOWs)
- Limited digital records for documented wells
- Hinders efforts to:
 - Prioritize and plan well plugging operations
 - Identify historic trends and practices
 - Focus characterization efforts

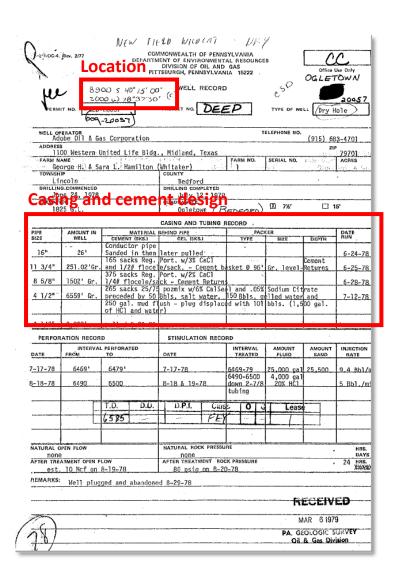


Details describing the construction and use history of all wells in the U.S. missing from most public and proprietary well databases

Well Regulatory Records

- Regulators manage millions of records
- Records document well permitting and construction process
- Many records have been scanned but are not in a machine-readable format

Oil & gas regulatory records contain valuable information about wells that remains "trapped" on scanned forms



Digitizing Well Records

- Quality and format of well regulatory records vary substantially over time and between jurisdictions
- Information contained in each record has generally increased
- Most agencies rely on manual review but would like to digitize all records

There is a critical need to digitize well records for both research and regulatory purposes

Pennsylvania - 1957

-56							
COLEONINALTH C DEPARTMENT	PENNSYLVAN OF MINES	IA					
OUADRANGE: RANSON 751 151	s Division ISEURG	PERMIT NO.	LAC-4 None req	uired SHALLOW			
MP REFERENCE: 8000' W. of 75° 47' 30' 3600' 5 of 41' 25' WELL	RECORD	KIND OF WE	IL: Gas	l, Gas, Other)			
COMPANY: Transcontinental Production Company	Size of Casing and Tubing	Used in Drilling	Left in Well	rackers: Type, Size and Depth			
ADDRESS: 744 Broad Street, Newark, N.J.	13 3/8"	40	, 40	• 42			
FARM Ellsworth Lacoe ACRES 130	8 5/8"	1165	1165				
WELL(FARM) NO. 1 CO. SERIAL NO. TPC-9	7"	1946	1946	c-300 Lark in bot			
ELEVATION: 991.11 LEASE: 54	2 7/8"	2760	2760	0-151 Lark in 7"			
TOWNSHIP: Ransom COUNTY: Lackawanna	1		· ·				
DRILLING ORBILLING COMPLETED: 5-12-57							
PROLUCTION: Gas - 200 MCF			1	PERFORATIONS AT:			
ROCK PRESSURE: 550 psig 48 hrs		<u> </u>	-				
WELL TREATMENT: (Shooting, Acidizing, Fracturing	1	1		- 4			
Etc.)							
Shot with 200 qts at 2853-2903	CEMENTING DA	TÀl (Size	Pipe. Dept	h. No. Bags, Date)			
	7" CSG @ 19	,					

Pennsylvania - 2002

											eta (EB	
		COMN	MONWE	ALTH O	F PENNSY	LVANI	A ((6474		Auth#	. C. DER USE	APS#
	DEP	ARTME			MENTAL P		CTION			Site#		Facility#
	w	ELL RE			ment Program		ORT			FIX Client	#	Sub-fac#
	W	ell Operato	r		DEP ID#	Well	API# (Pern	nt / Reg)	T	Project No	ımber	Acre
Atlas Resourses, Inc					6422	37-073-20333			n/a	, roject ivi	3111001	Acres
Adress						Well Farm Name			1 // U		Well#	Municipalit
101 McQuiston Drive						Moore					2	I
City						County				Municipality		
Jackson Ce	Jackson Center PA					Lawren	ce			Wilmington		
	Phone					ax USGS 7 5 min quadrangle map						
724-662-0300 724-66					2-3039 New Castle North							
	2223 93	WELL	RECOR	D	GUAS PON	Alad com	plete Lag o	Edmalina	e nhi A	ack (mana		Marie Vigo
Well Type	G	as	Oil	X Con	nbination Oil &			ction		Storage		Disposal
Drill Meth	N X	Rotary-	Air 🗀	Rotary-N	Aud 🗌				Cabl	e Toot		
Date Dn	ling Started		ate Drilling (Completed		e Elevatio	n			-Driller	Total	Depth-Logg
12/	12/21/01 12/29/01		01	1255				6409			6408	
	Casino	and Tul	bina		Cement return Cement ret				asına	X	No Yes	No X
		1	T			-						
Hole	Pipe		Thread/	Amount	Material Behind Pipe				Packer / Hardware / Centralizers			
Size	Size	Wt	Weld	ın Well	Type and Amount		Ty	pe	Size	Depth	Date R	
14 3/4	11 3/4	36	Threaded	250	400 sxs Standard							12/22/0
11	8 5/8	24	Threaded	940	160 sxs Light	110 sxs	Standard	Cer		8 5/8	940	12/23/0
7 7/8	4 1/2	10 5	Threaded	6195	230 50/50 Poz with 6% gel		Lato		4 1/2	6185	12/30/0	
	L				80 50/50 Poz	with 2% ge	el	5 Ce	ntral	4 1/2		

Technological Solutions Exist

Optical Character Recognition (OCR)

Algorithmic conversion of text into machine readable format

Large Language Models (LLMs)

 Neural networks trained on large volumes of text that can summarize and answer questions about documents

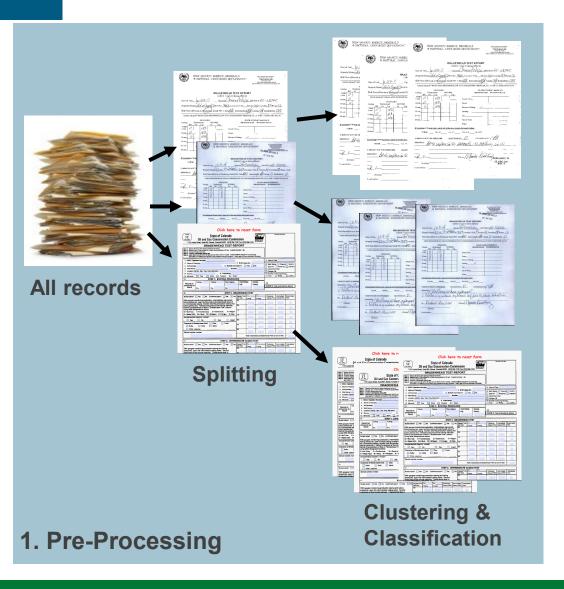
- Numerous public and proprietary tools available
- Performance relies on specific training

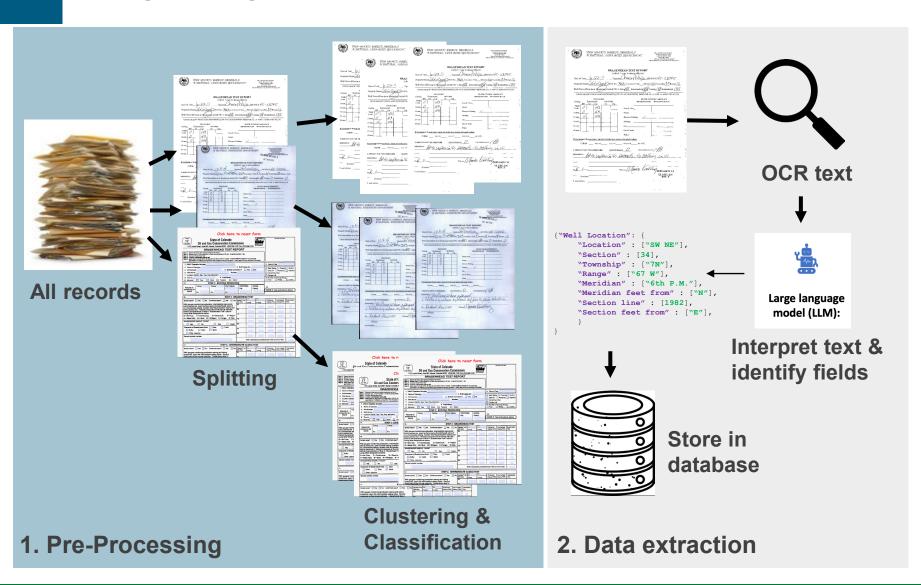
OCR and LLM approaches can reliably digitize information from well records but need to be trained for specific purpose

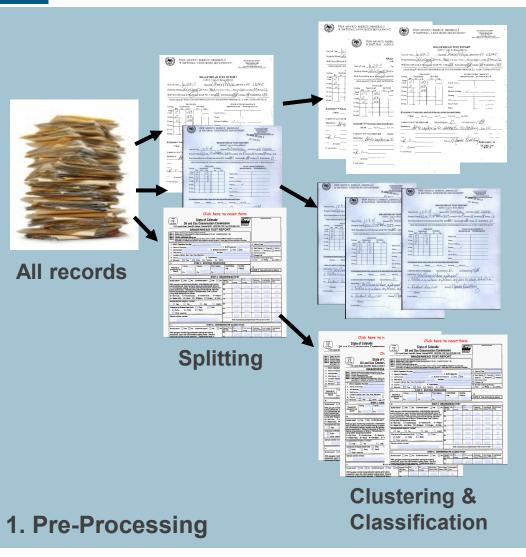
Project Goals

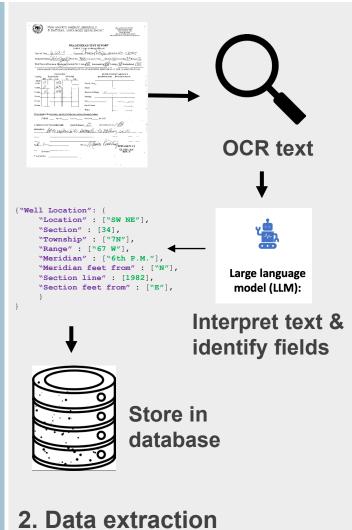
- Design custom OCR and LLM approaches for extracting data from oil and gas regulatory records
- 2. Build a platform that allows users to leverage OCR and LLM models for rapid data extraction from records

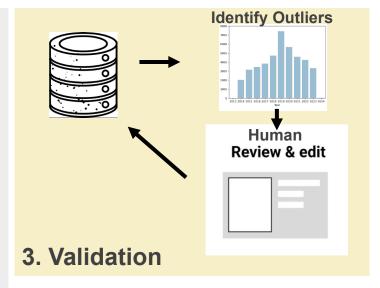
Team will deliver a working tool prototype and summary report in the first year

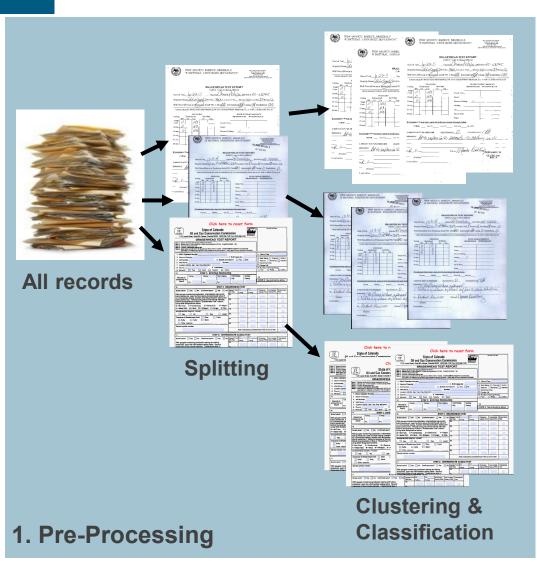


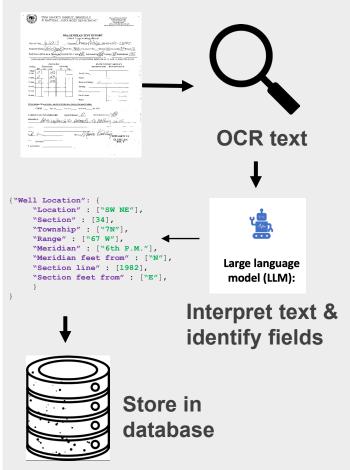




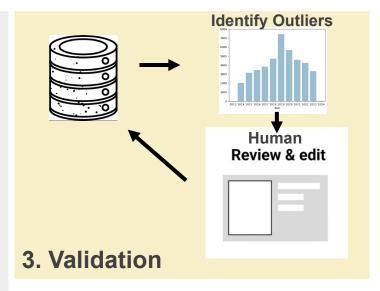


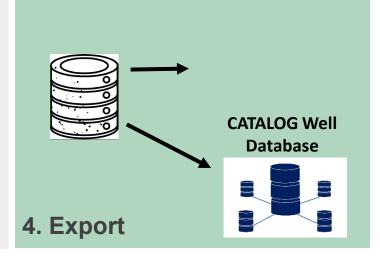












Development Approach

Tool Development

LBNL: Prototype design and development

NETL: record collection; stakeholder engagement; backend design

OCR & LLM Development

LLNL: splitting, clustering, field ID

SNL: OCR LANL: LLM

development development



Final Product

Parallel efforts to develop tool alongside OCR and LLM approaches

Record Collection & Stakeholder Engagement

- Gathered ~440K well records from four states (CA, CO, PA, NM, OK)
- Interviewed prospective users from:
 - Five oil & gas regulatory agencies
 - Three research institutions
 - One oil & gas major
- Documented feedback for tool development

Questions for stakeholders

Management questions

- What does your team do? What are the main goals & tasks?
- What teams handle well records (historic and modern)? What do these teams do?
- What are the main goals, responsibilities of the teams handling the well records?

Questions pertaining to well records' usage

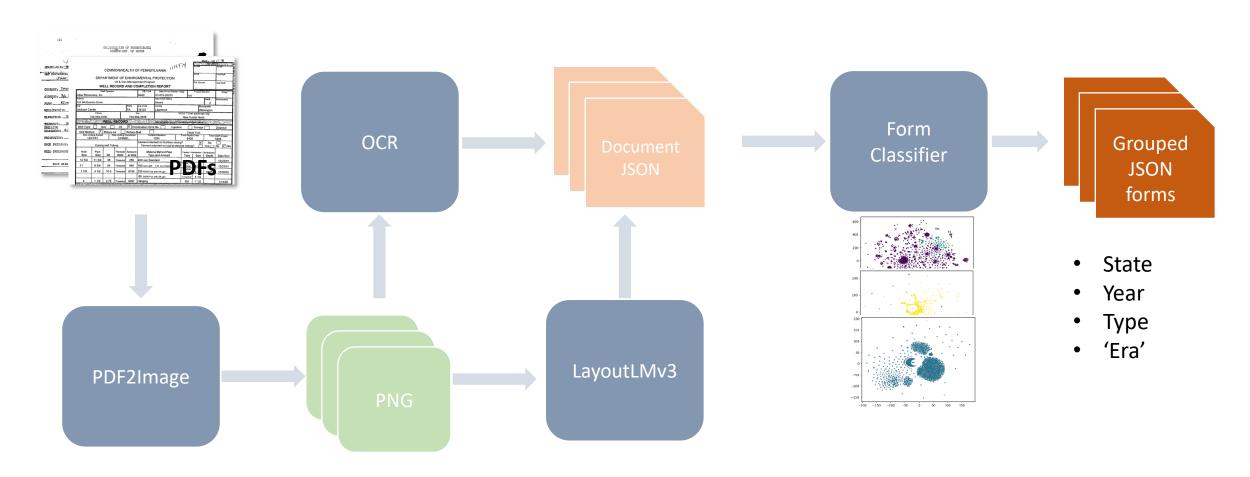
- What are the different tasks that well records are generally used for?
- Are there any processes you follow to digitize the manual records?
- Could you show us an example of how you do your work?
- What are the current gaps, pain points, and needs?
- Are there any upcoming changes we should know about?

Tool

- What kind of tool/service would be most helpful for projects that use well records?
- What would you want the tool/service to do?
- Is there anything you wish it could do, but seems too hard?

Records sufficient for tool development gathered and key partnerships established

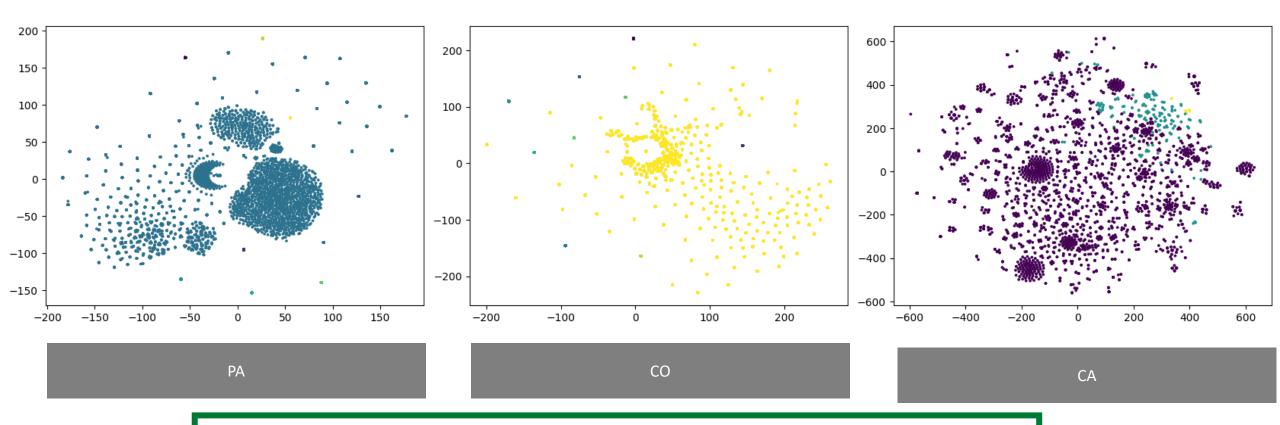
Form Splitting and Clustering Approach



Form Classifier Combines Multiple Embeddings



Clustering Embeddings Shows Clear Separation Across States



Clustering embeddings show clear separation across states

Large Language Modeling Approach



Large language model (LLM):

- DocQuery^[1], developed/fine-tuned by Impira
- Based on Microsoft's LayoutLM model
- Used two dataset, i.e., SQuAD2.0 and DocVQA
- Document Query Engine Powered LLMs
- Able to analyze semistructured and unstructured documents (PDFs, scanned images, etc.)
- Zero-shot learning (notraining is used for this task at the current stage)



Historical documents

- 150 Drilling Completion Reports from Colorado
- Text-based PDFs
- Information of well location, depth, etc.



Model performance:

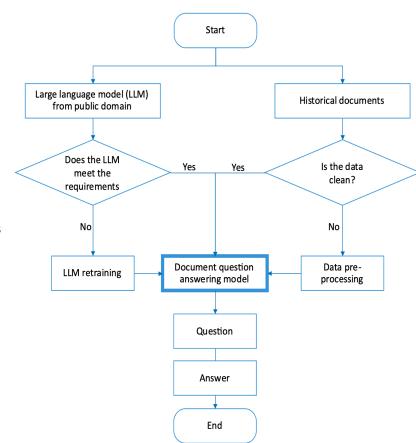
- Extraction time: within two seconds per document
- Accuracy: 100% on a simple dataset
- Struggles with a more complex dataset



Future direction:

- Generate our own dataset for with questions and answers based on historical documents
- Fine-tuning LLM models
- Use more powerful LLMs, which show more promising preliminary results on the complex dataset

-104.567670



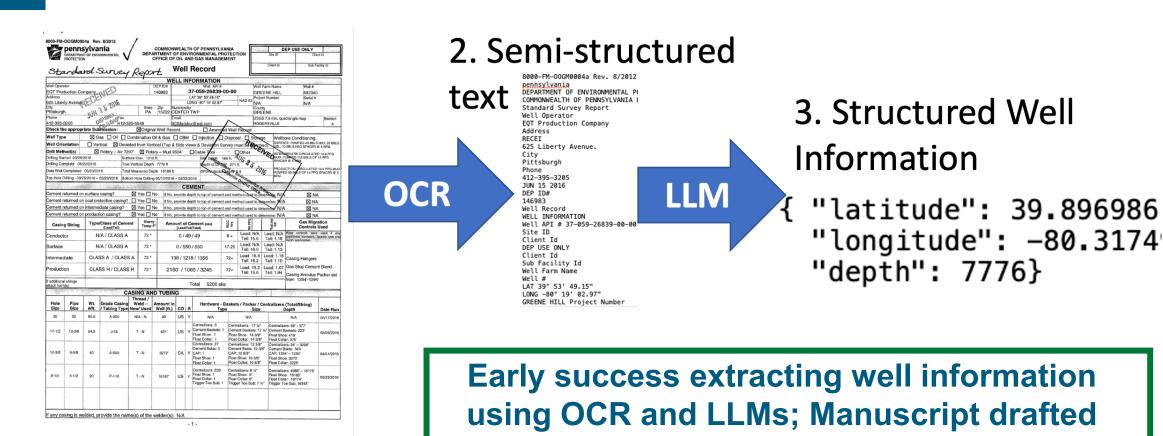


DRILLING COMPLETION REPORT

By using Large Language Models, we can obtain correct well location information from historical documents

[1] https://github.com/impira/docquery

Large Language Modeling Results



1. Document image

Validation and verification will be critical to this effort

Tool Prototype

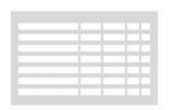
New project



Upload records



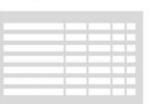
Digitized records



Review & edit



Export records

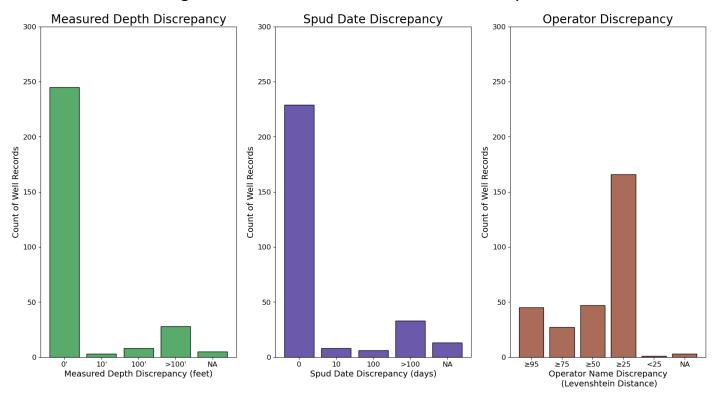


Google Document Al Case Study

- 1. Selected 165 CO completion reports (115 training, 50 test)
- Trained to extract Operator Name, Spud-Date and Measured Depth
- 3. Google Reported Metrics:
 - F1: 0.905 Precision: 92.1% Recall: 89%
- 4. Selected Another 289 reports separate from above 165
 - Performed custom analysis post Google Document AI processing
 - Compared Doc AI results vs state database records

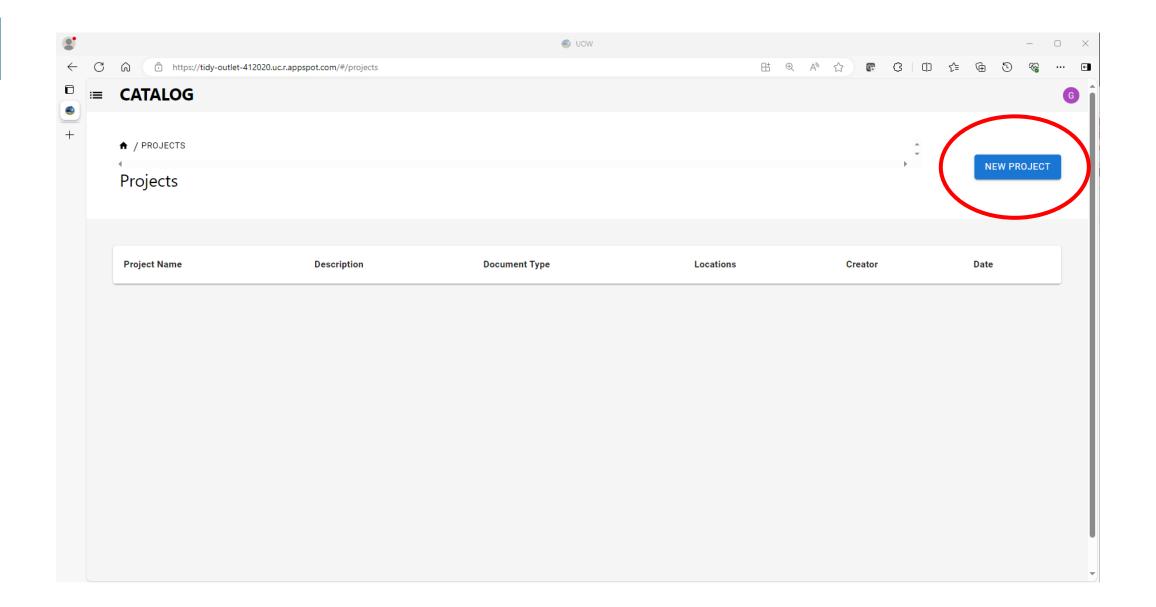
Case Study Results

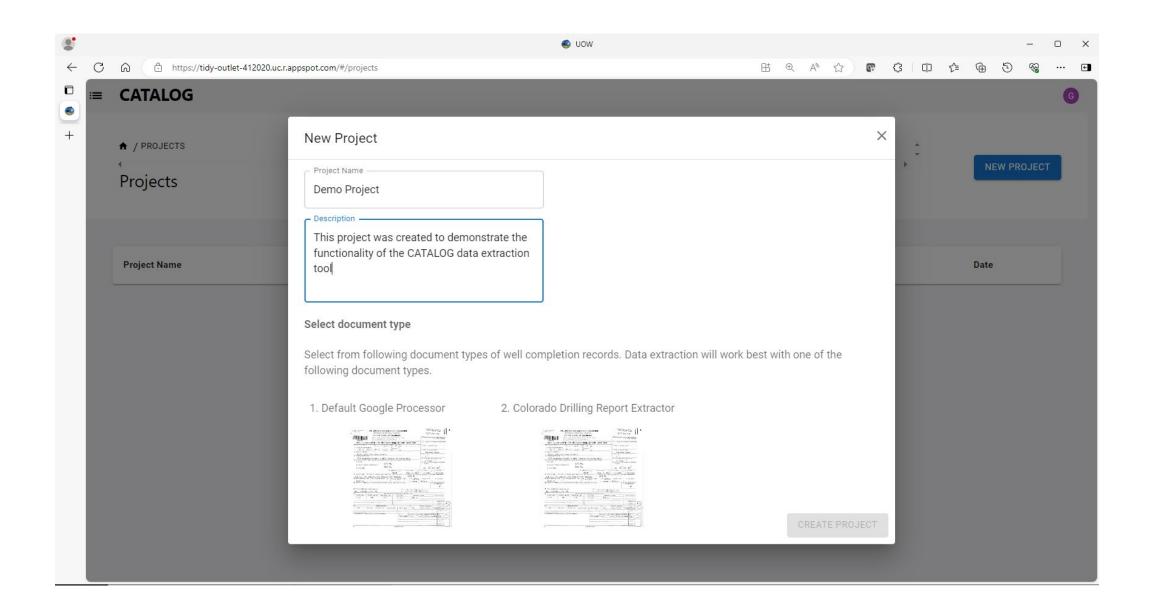
Google Doucment AI & State Record Comparison

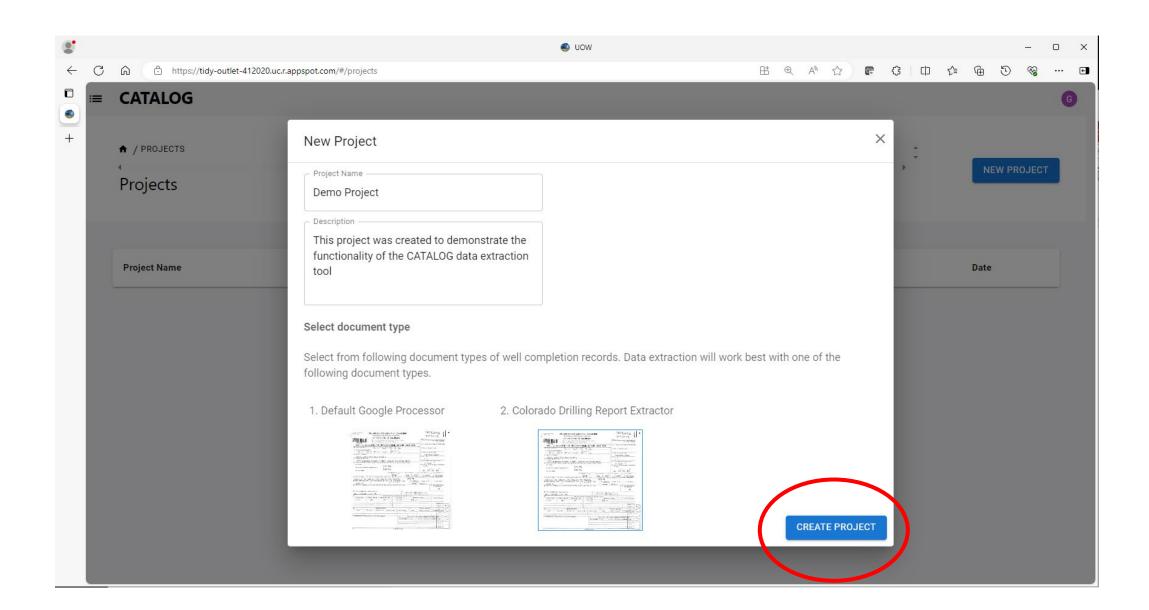


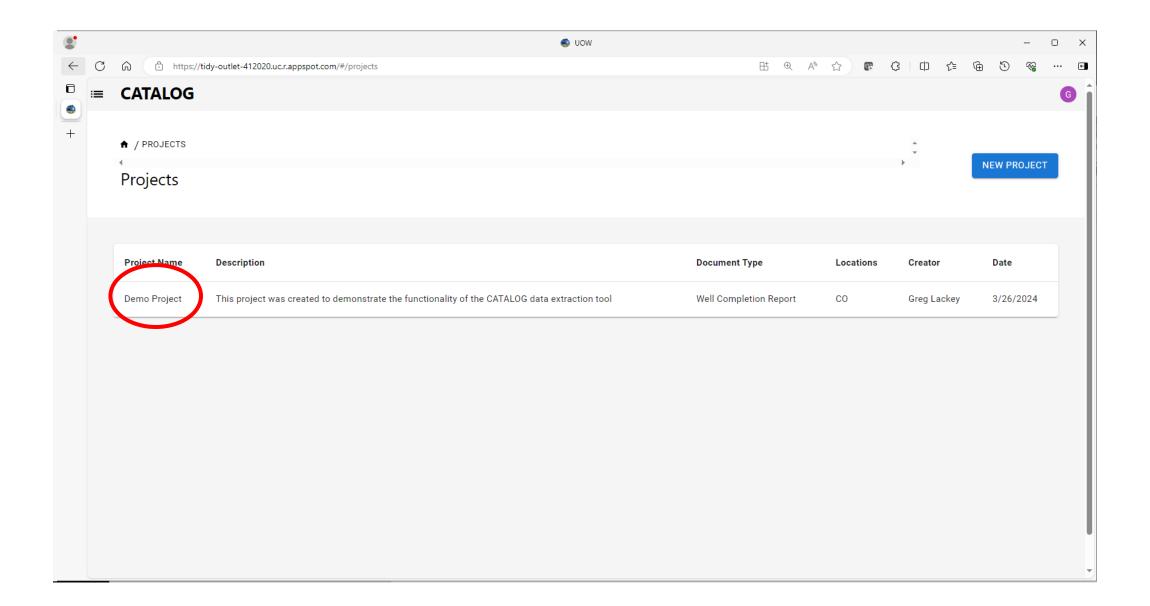
Google Doc AI results matched database ~90% of time for depth and spud date

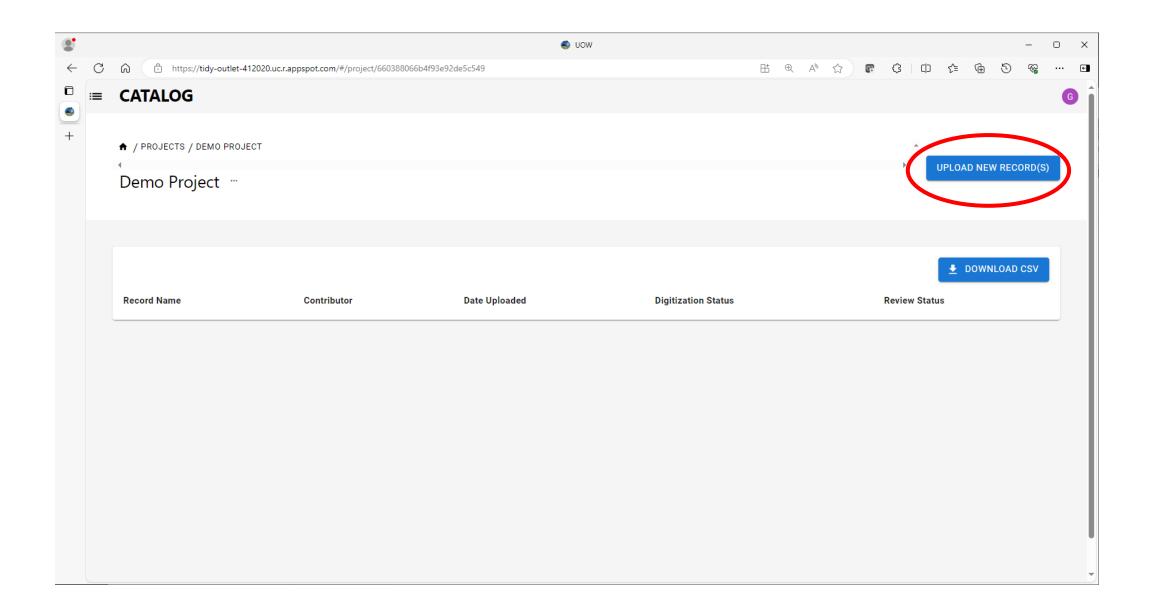
Tool Prototype Demo

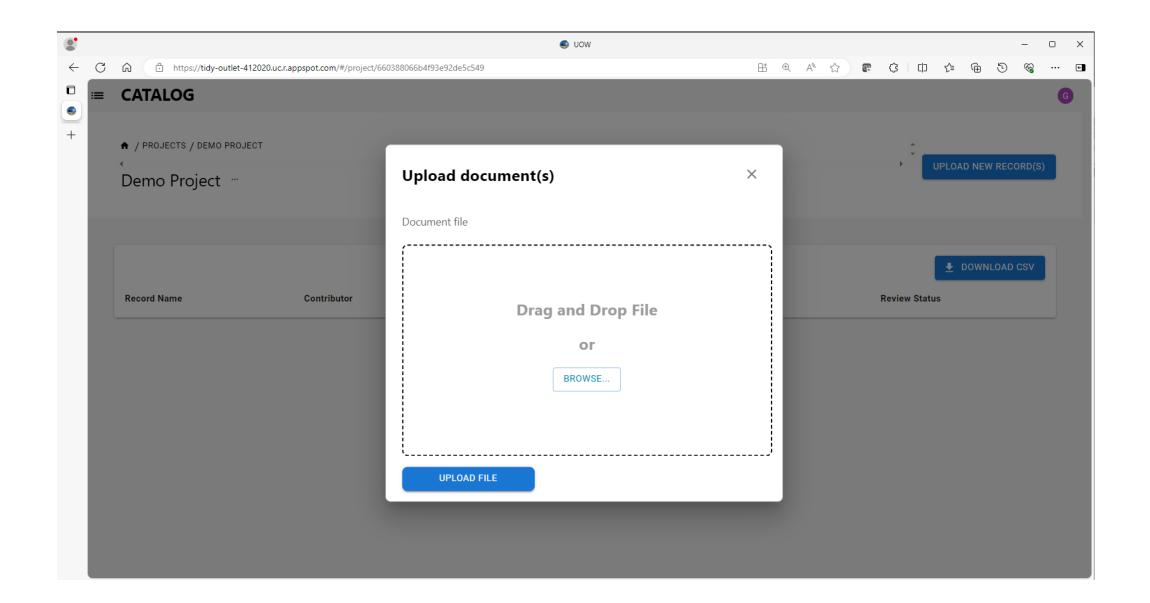


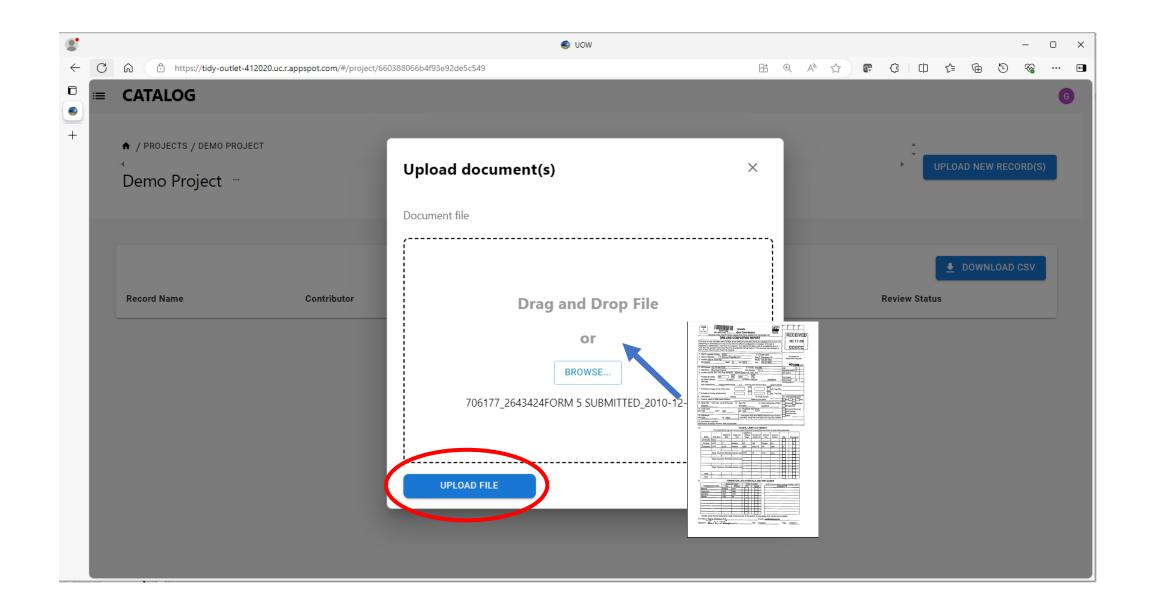


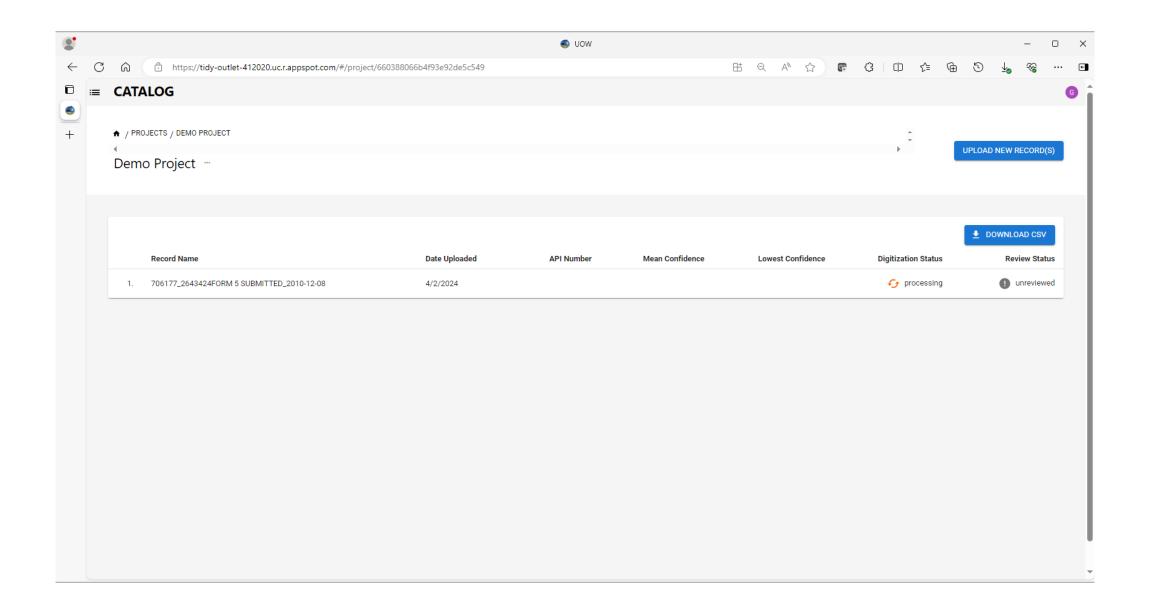


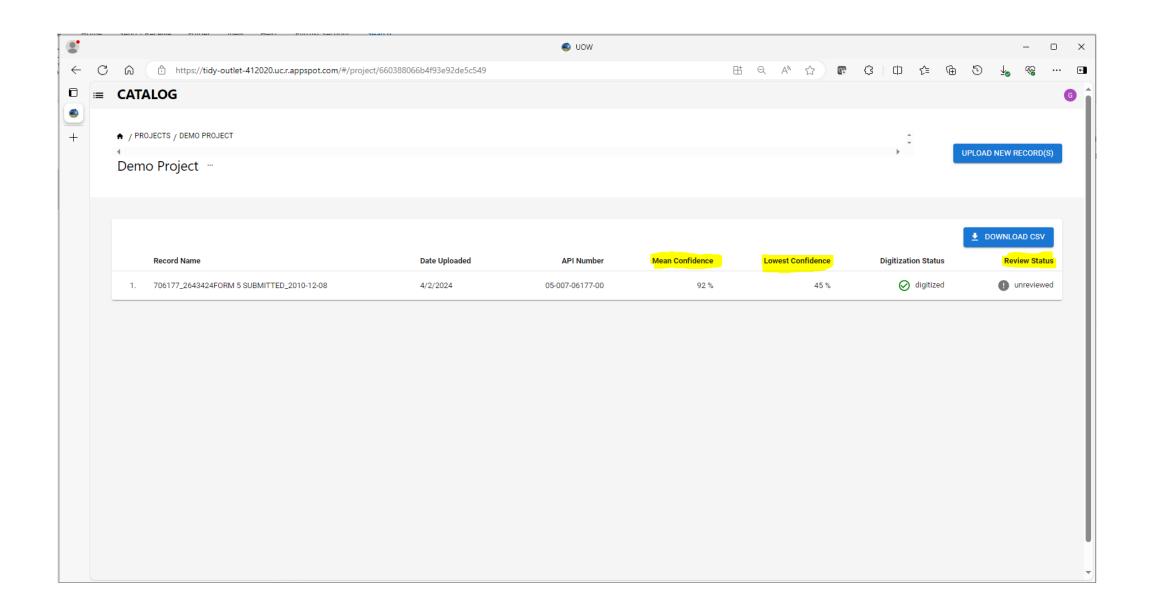


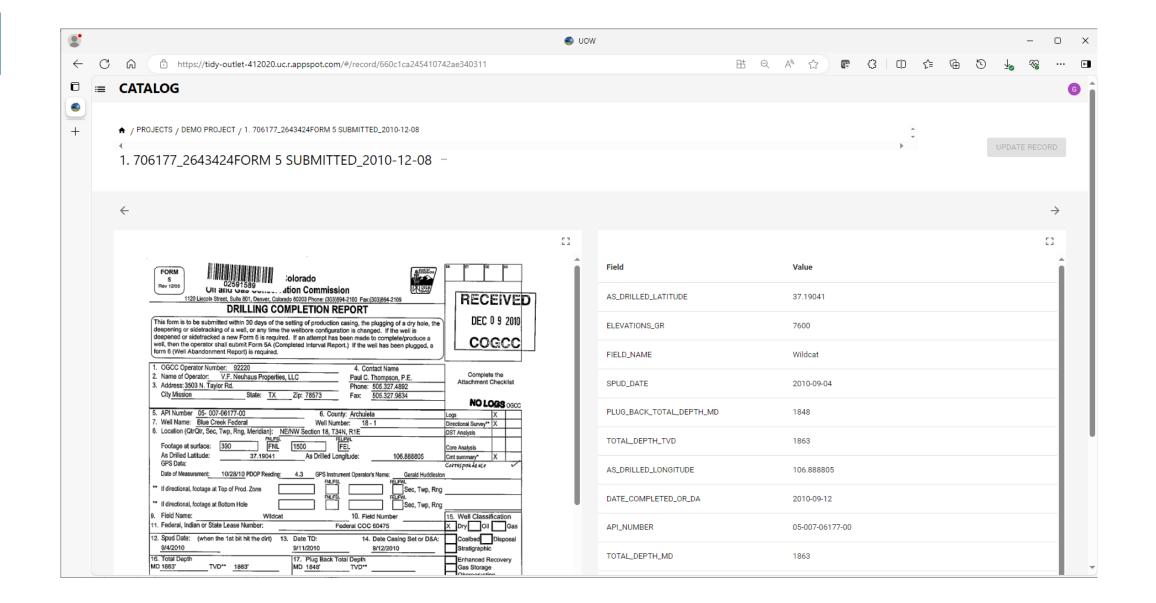


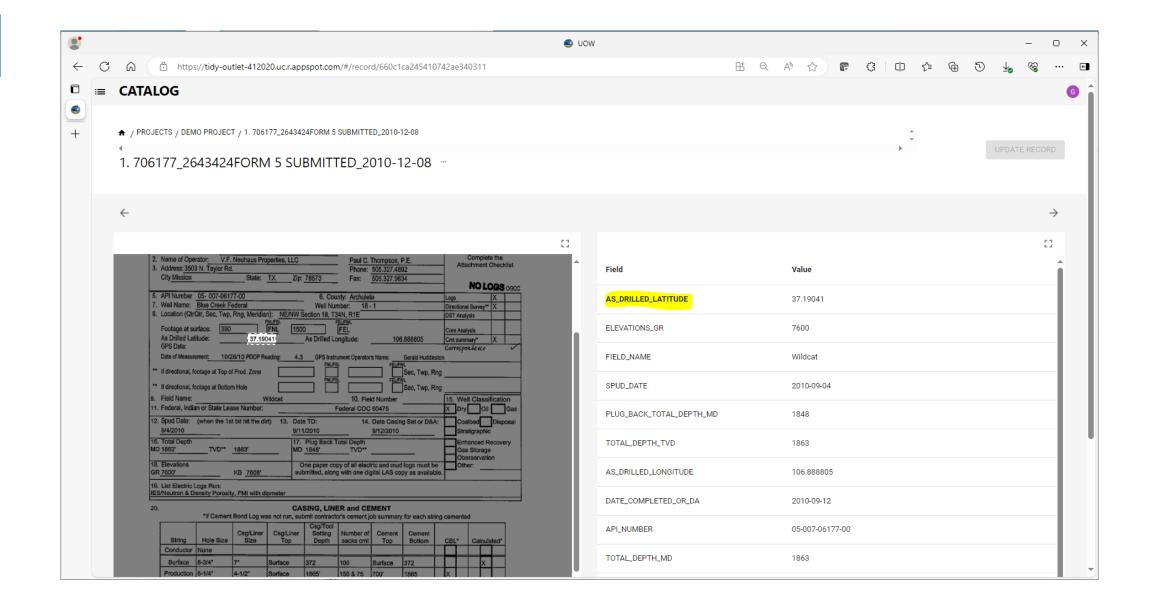


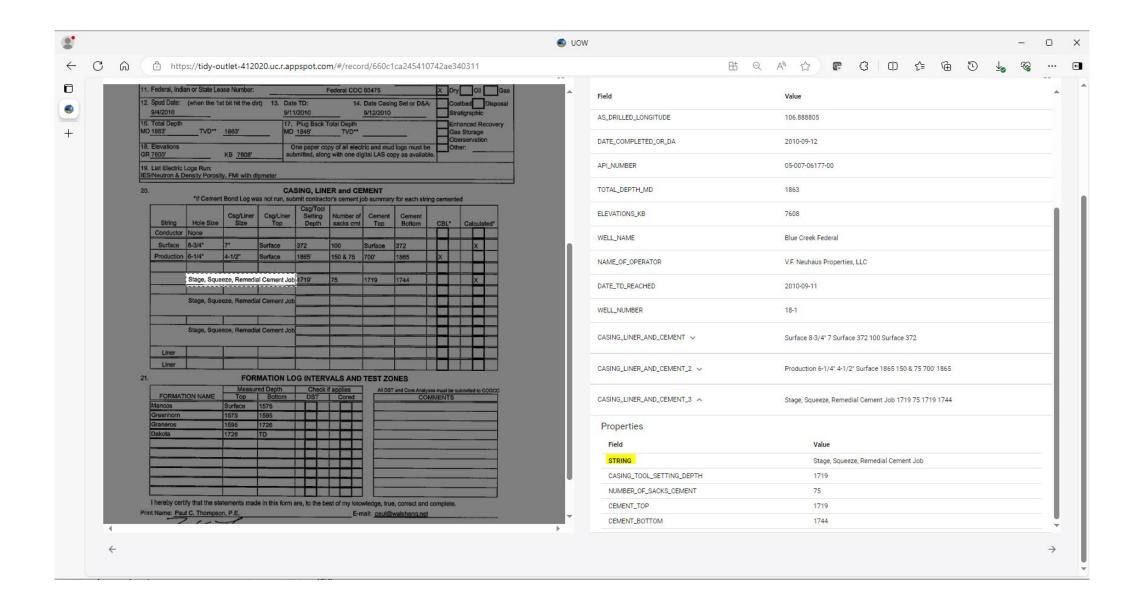


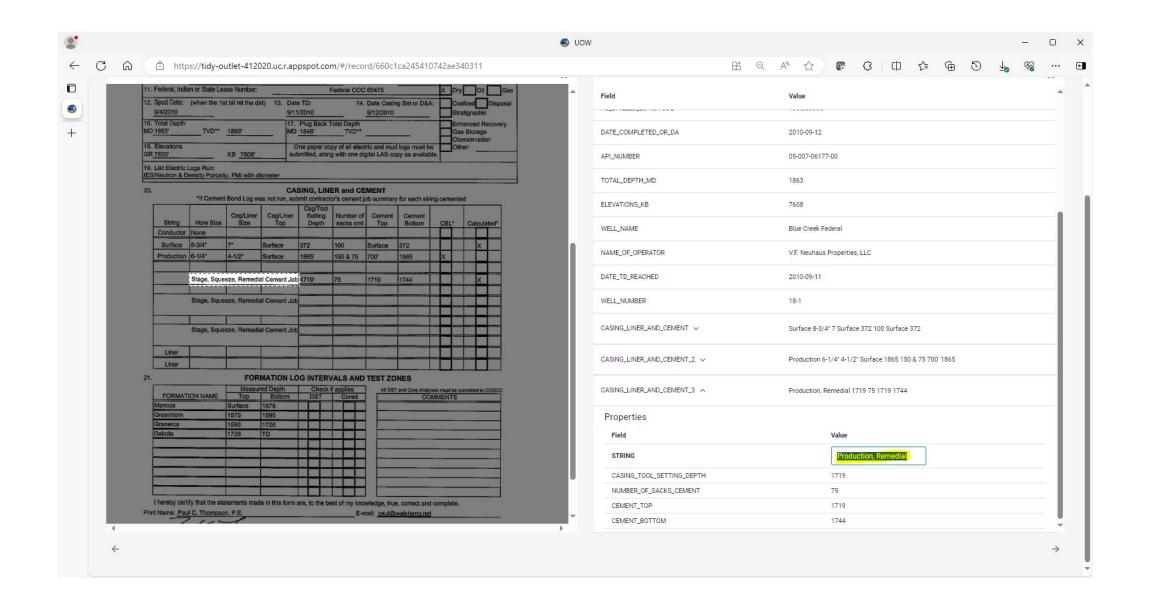


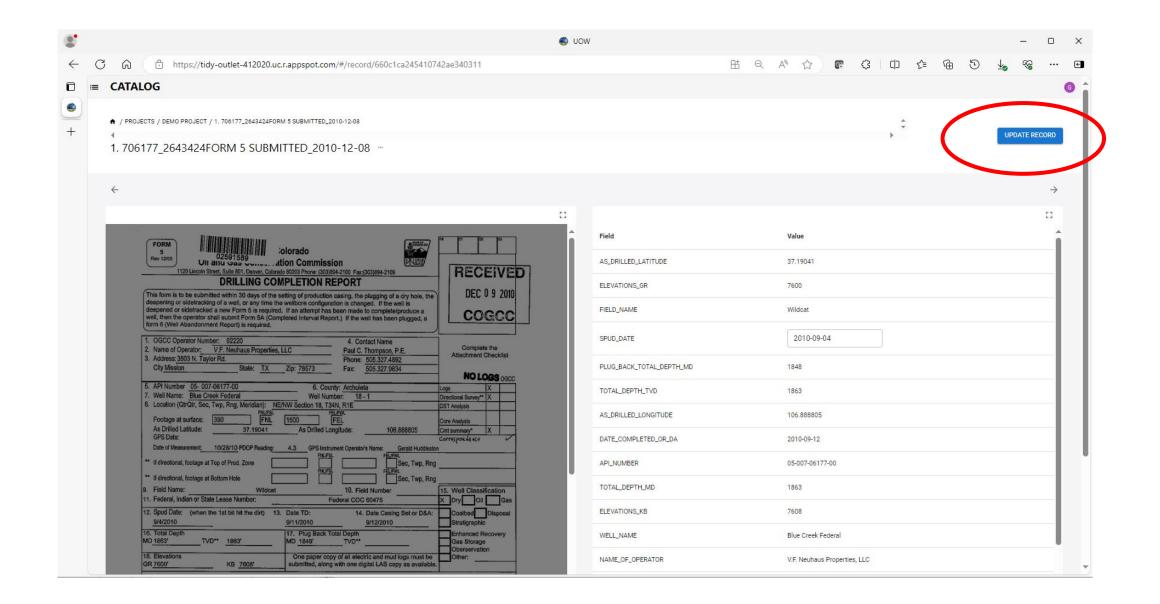


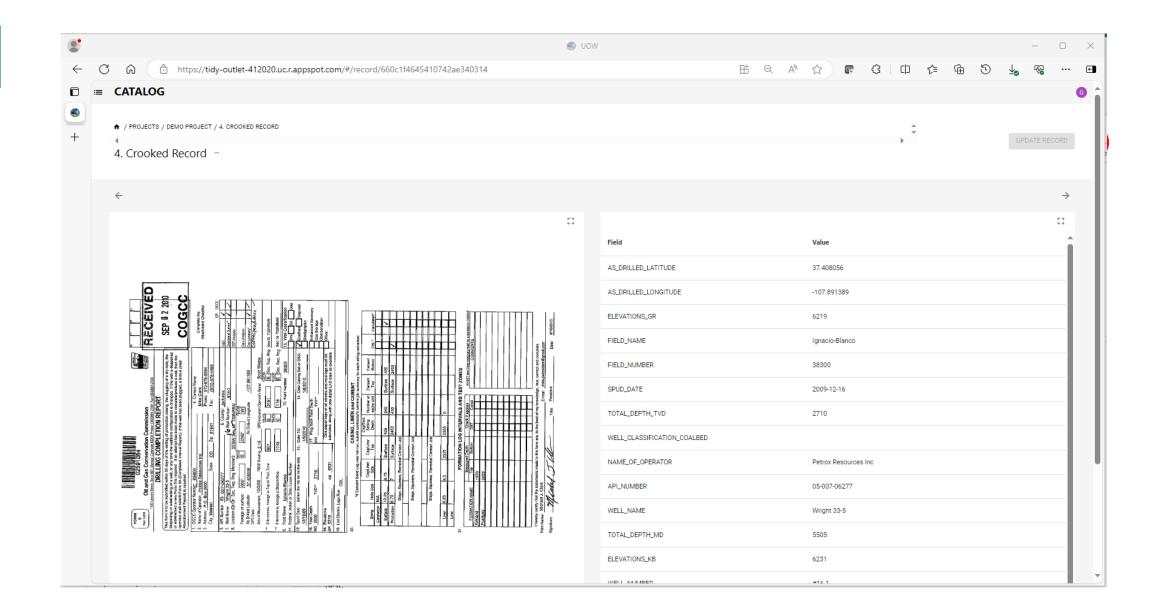


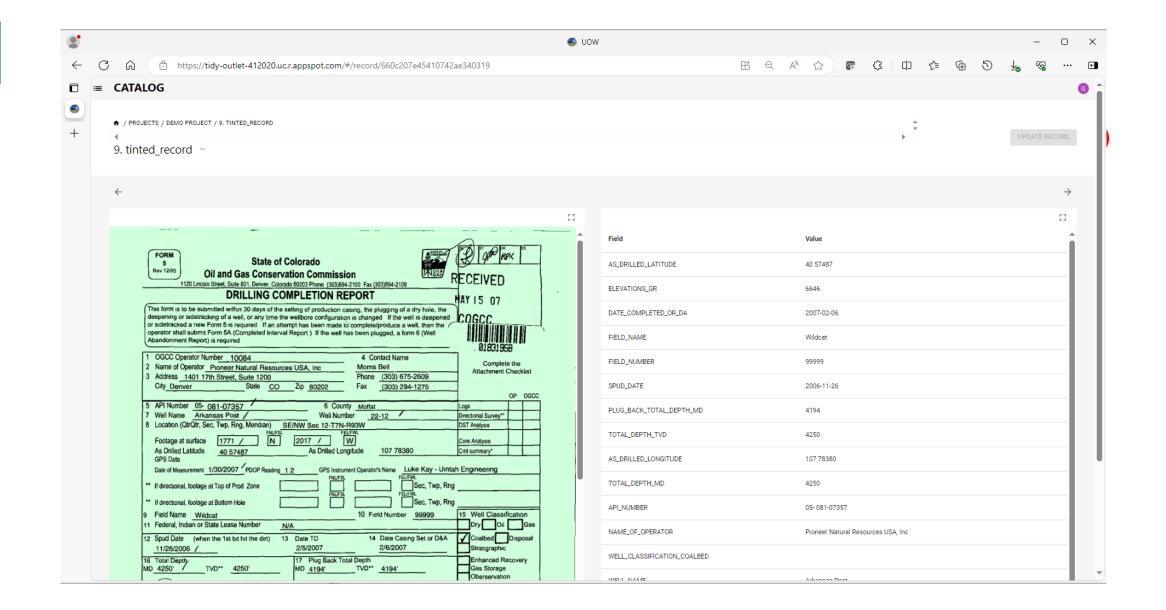


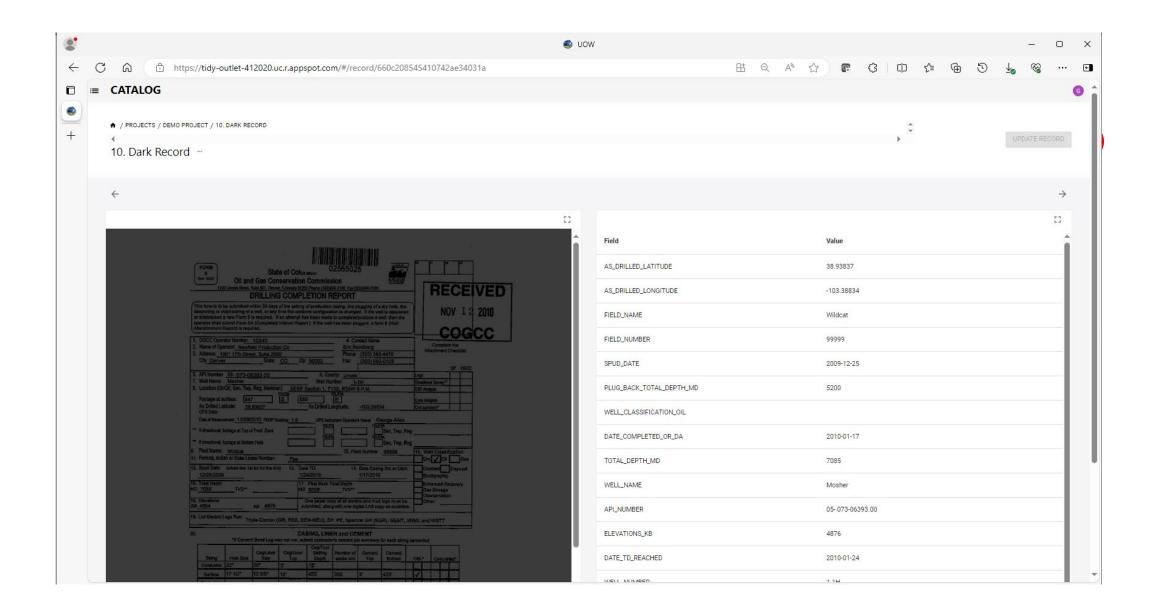


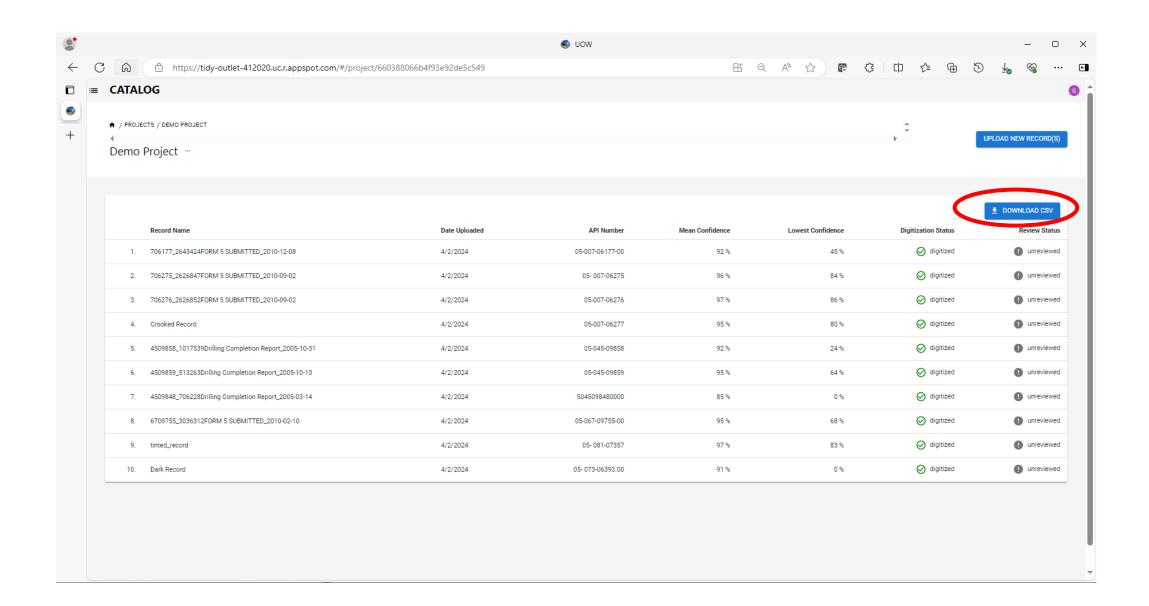


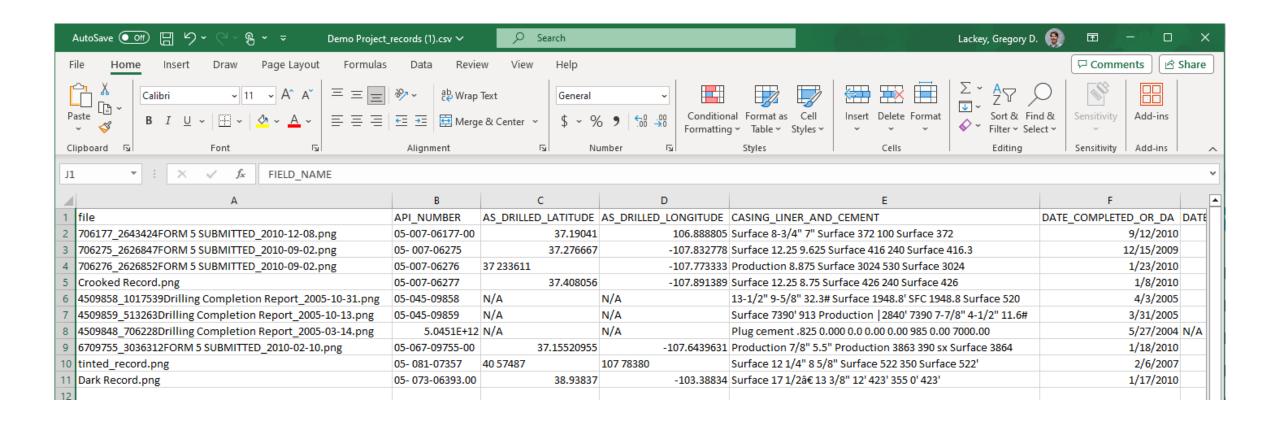












Next Steps

- 1. Iterate tool prototype design and functionality internally
- 2. Partner with a small number of stakeholders (researchers, state agencies) that will use tool to digitize well records
- 3. Collect feedback and continue to improve tool design
- 4. Incorporate document splitting and clustering workflow
- 5. Progress open-source OCR and LLM approaches and replace Google Document AI backend
- 6. Expand collaborations and release general version of tool

Contacts

Greg Lackey

Research Engineer
National Energy Technology Laboratory
<u>Gregory.Lackey@netl.doe.gov</u>
412-386-7590 (office)

Dan O'Malley

Subsurface Flow & Transport Team Lead Los Alamos National Laboratory omalled@lanl.gov

