Curating Carbon Storage Data for Reuse: Enabling Research and Modeling from Earth's Surface to Subsurface

NATIONAL ENERGY TECHNOLOGY LABORATORY

Paige Morkner

Geologist + Geo-Data Scientist

New York Scientific Data Summit

September, 2024

Authors and Contact Information

Paige Morkner[1,2]; Abigail Choisser[1,2]; Michael Sabbatino[1,2]; Stephen Leveckis[1,2];
Kelly Rose[1]; Jennifer Bauer[1]

[1] National Energy Technology Laboratory, 1450 Queen Avenue SW, Albany, OR 97321, USA

[2] NETL Support Contractor, 1450 Queen Avenue SW, Albany, OR 97321, USA

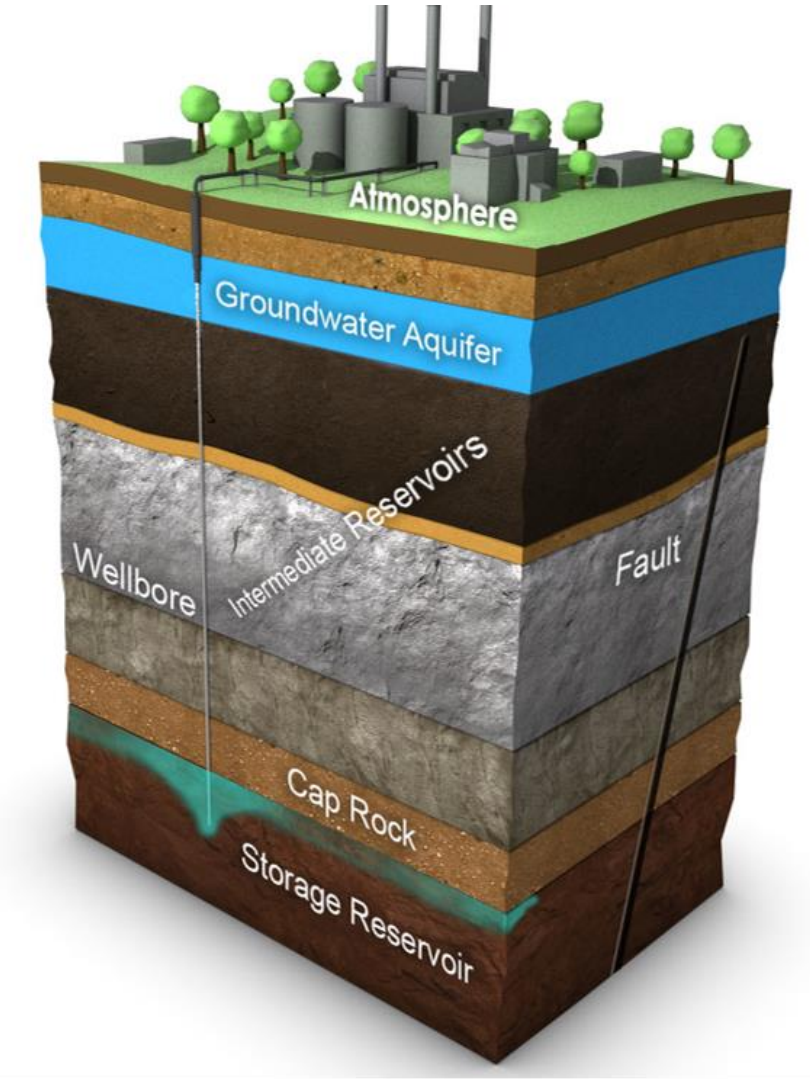
Disclaimer and Acknowledgement

This project was funded by the United States Department of Energy, National Energy Technology Laboratory.

This work was performed in support of the U.S. Department of Energy's Fossil Energy and Carbon Management's Advanced Storage R&D Carbon Transport and Storage Multiyear Research Plan and executed through the National Energy Technology Laboratory (NETL) Research & Innovation Center's FWP1025032.

Geologic Carbon Storage

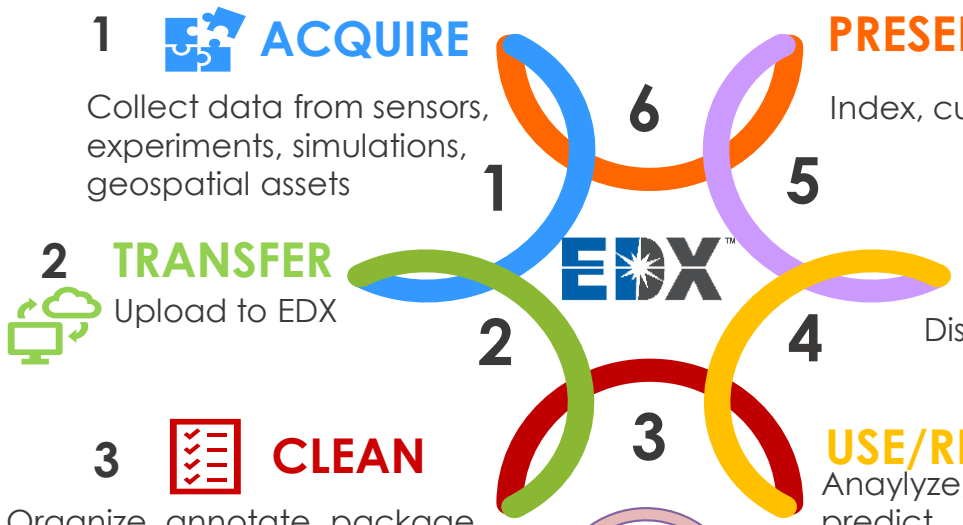
- Geologic carbon storage is the engineered process of injecting captured CO_2 into the deep subsurface for permanent storage
- To safely and confidently store CO_2, extensive subsurface and surface data is necessary quantify
  - Feasibility
  - Storage capacity
  - Risk quantification
  - Cost analysis
  - Impact to communities/environmental justice/social justice

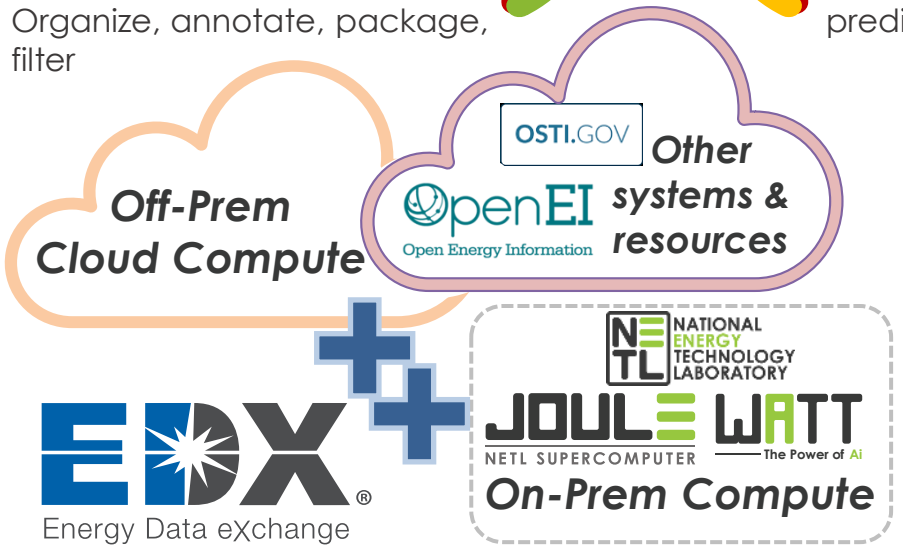


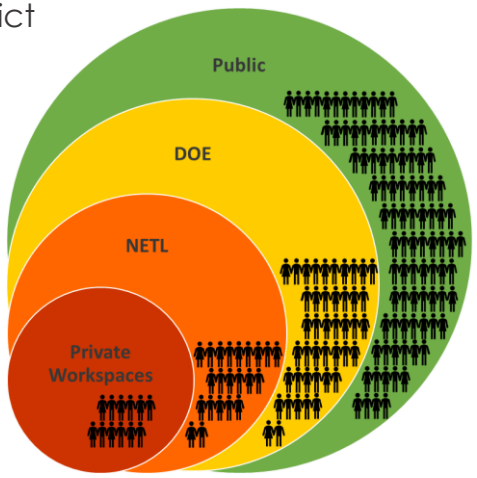
What is the Energy Data Exchange (EDX)?

- An online **virtual library and laboratory**
- A comprehensive data clearinghouse and system that **enables access to data, tools, presentations, publications, and promotes reuse**
- **Supports the entire life-cycle of data curation** through private collaboration and public dissemination
- Uses a **tiered access approach** and roll-based security
- **Provides tools** to enable data visualization and query

https://edx.netl.doe.gov/





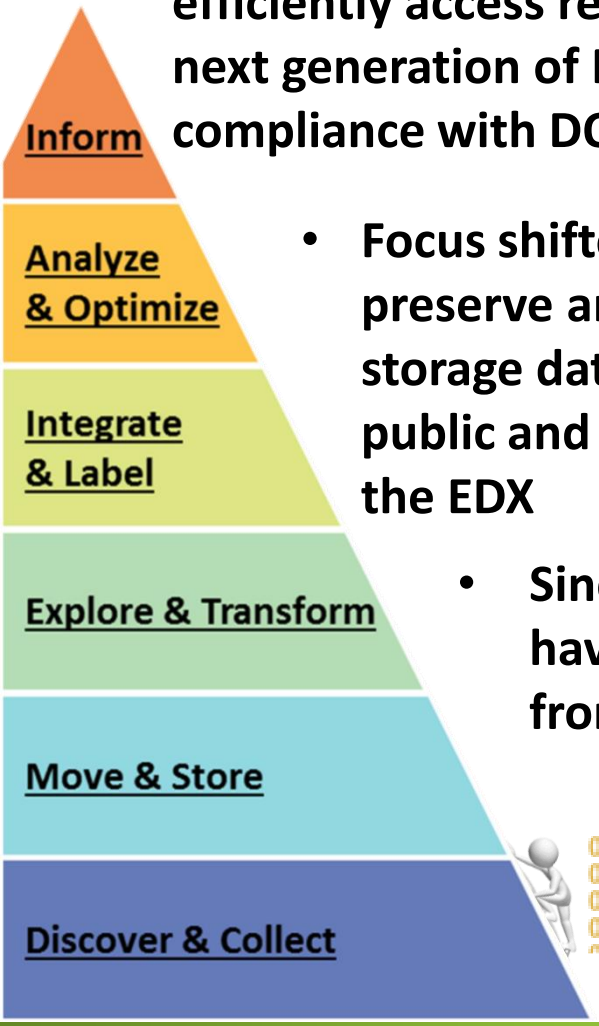


Tiered, roll-based security for access to data resources

Leveraging EDX for Carbon Storage Data

NATIONAL ENERGY TECHNOLOGY LABORATORY

- There is a need to preserve and efficiently access resources to drive the next generation of R&D while ensuring compliance with DOE regulations
  - Focus shifted in 2015 to preserve and curate carbon storage data products, both public and private through the EDX
    - Since then, millions of dollars in R&D products have been preserved and made publicly available from carbon storage research and projects



SAMI SCIENCE-BASED AI/ML INSTITUTE

IBM aws docker Microsoft Azure

WATT The Power of Ai

JOULE NETL SUPERCOMPUTER

EDX Energy Data eXchange

PUBLIC PRODUCTS*

PUBS | SPATIAL LAYERS	DATA & WEB TOOLS	PRODUCTS | GROUPS
1000+ | 700+	80+	>3500k | >20

Big Data

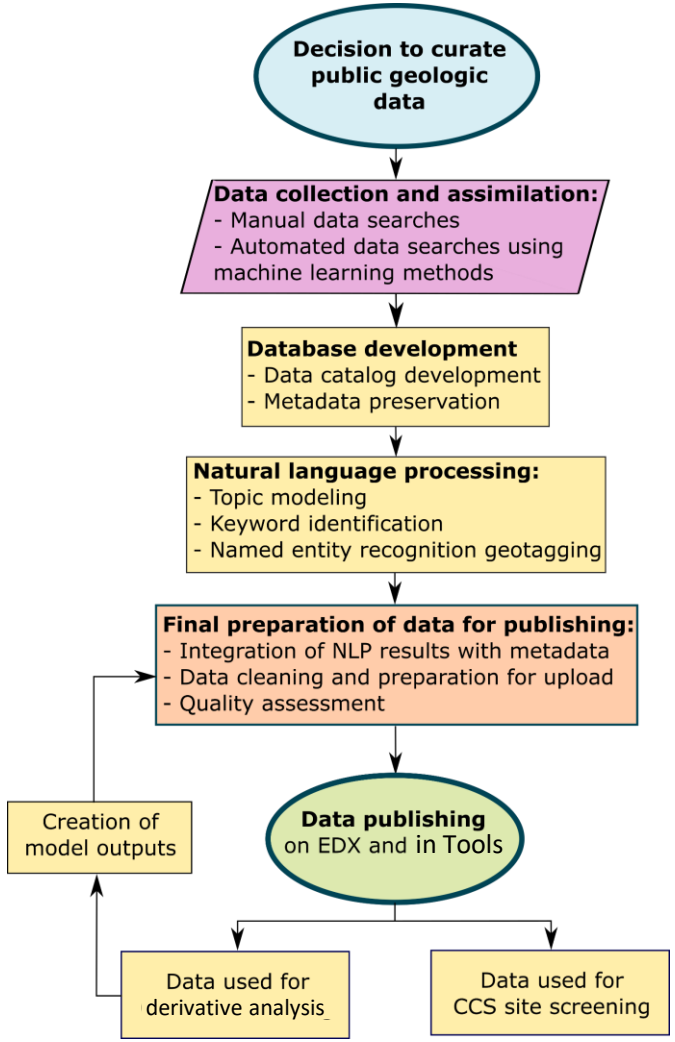
Addressing and Optimizing the Carbon Storage Data Lifecycle

Challenge: Finding open-source data to drive spatial and model-based derivative analysis is not easy nor intuitive

Solution: We have a standardized workflow for collecting, aggregating, curating, and producing data products from open-data resources

Why: At the intersection of geology and data science is the ability to help our end users and ourselves with data visualization, tool development, and derivative data analysis and modeling

How: Through deploying a data lifecycle workflow that integrates artificial intelligence (AI), machine learning (ML), and Natural Language Processing (NLP), and producing tools that enable data discovery and interaction



Carbon Storage Data Collections

NATIONAL ENERGY TECHNOLOGY LABORATORY

Data on EDX

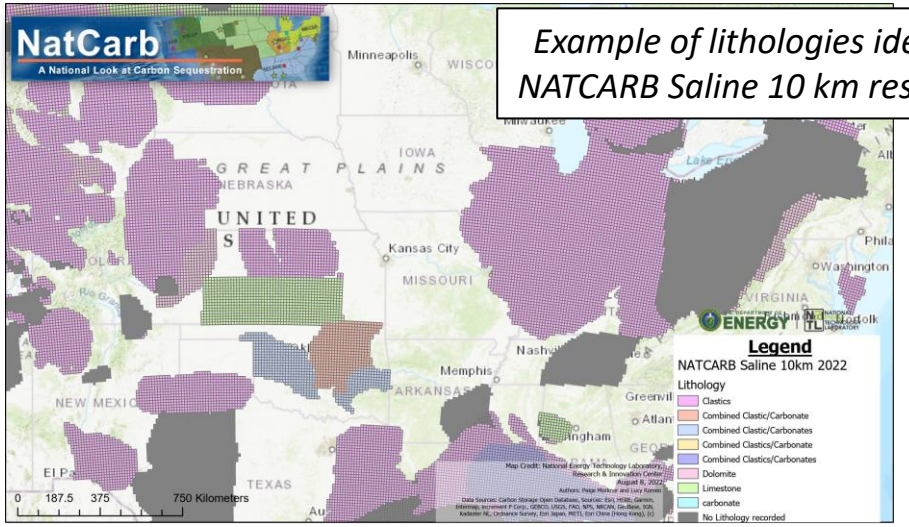
- 3185+ public data resources on EDX related to carbon storage
- 100+ TB of seismic data on private servers (going cloud)
- 8+ field projects represented, with two benchmark datasets:
  - Illinois Basin Decatur Project
  - FutureGen 2.0 Dataset

The National Carbon Sequestration Database (NATCARB)

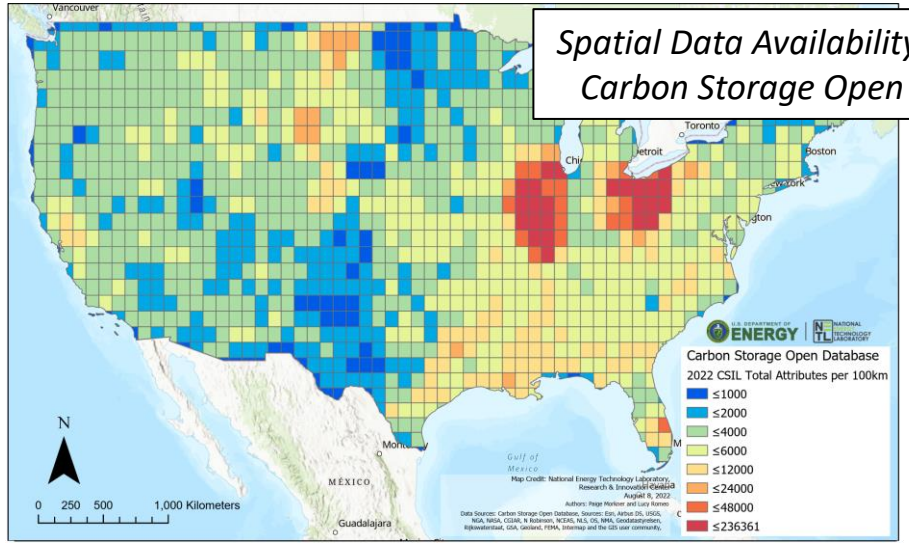
- Reservoir data for deep saline, unminable coal, and depleted oil and gas reservoirs
- 10km grid layer containing properties for each formation
- Key reservoir property data at different spatial scales

Carbon Storage Open Database:

- 800+ Shapefiles (Lines, Polygons, Points) and Rasters available on EDX's GeoCube
- 1846 text-based documents on EDX
- 1000s of additional resources published to EDX Energy Data eXchange



Example of lithologies identified for NATCARB Saline 10 km reservoir units



Spatial Data Availability Heat Map Carbon Storage Open Database

The Carbon Storage Open Database

- How do we leverage and curate all the different types of data into a single resource?
- ML and NLP can be used to improve collection, movement, storing, exploring, transforming, aggregating, and labeling of data

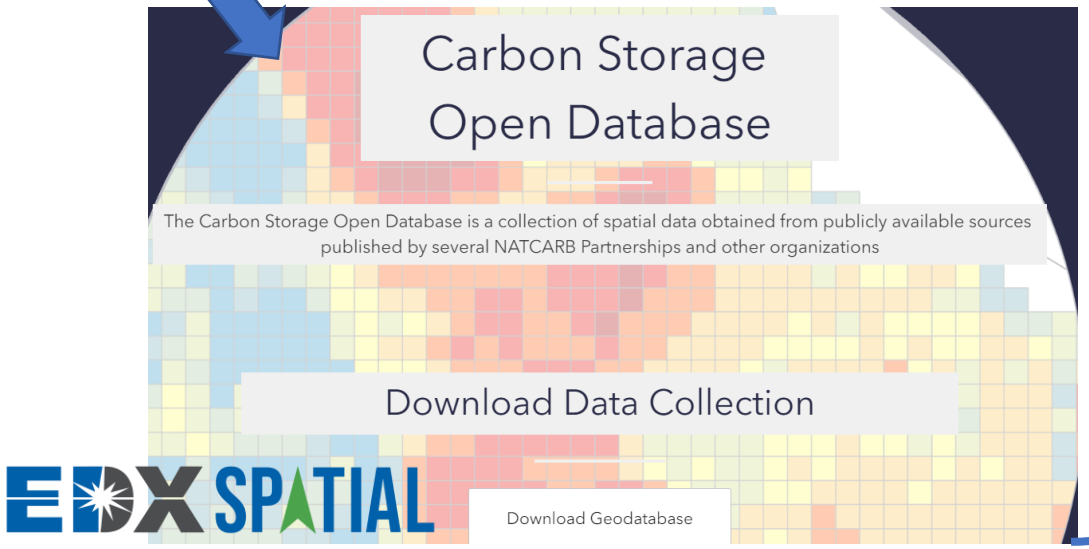
Public data collection from expert-driven research and leveraging "smarter" ML/NLP search tools

ML/NLP tools to increase data discovery

Text-based data

Spatial data

Text sources parsed with natural language processing for keywords, topic modeling



Information and data is extracted for further use in analytical tools and platforms

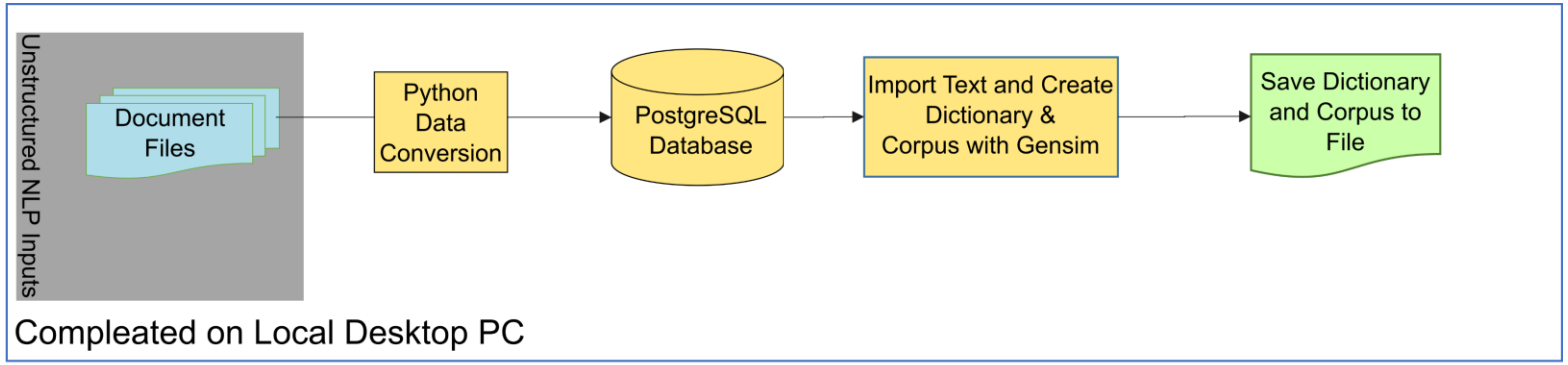
NLP Unsupervised ML for Document Classification

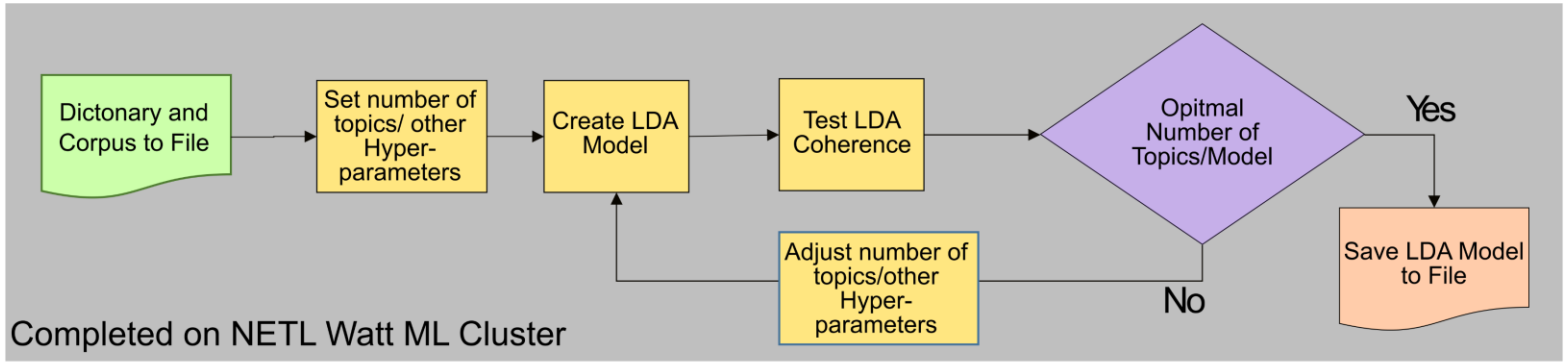
Problem: 1800+ documents were collected from public resources but needed to be organized prior to publishing on EDX

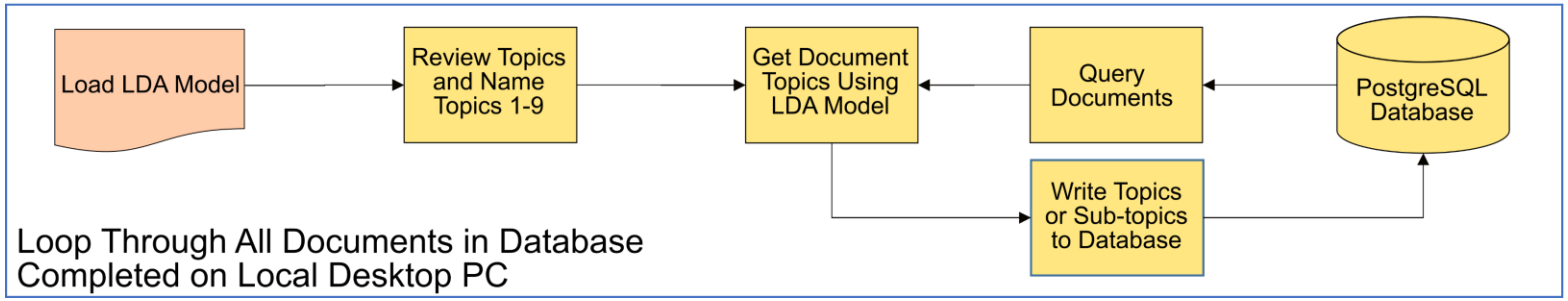
Solution: NLP topic modeling

- Latent Dirichlet allocation **(LDA) model based on corpus of 2071 text-based documents**
- Topic names assigned by subject-matter experts
- **Each document is classified** by % of each topic it's associated with
- **Each document has 50+ keywords identified** and can be **associated with metadata on EDX**
- **Parse geographic location to associate with each document** – when possible

NETL SmartParse



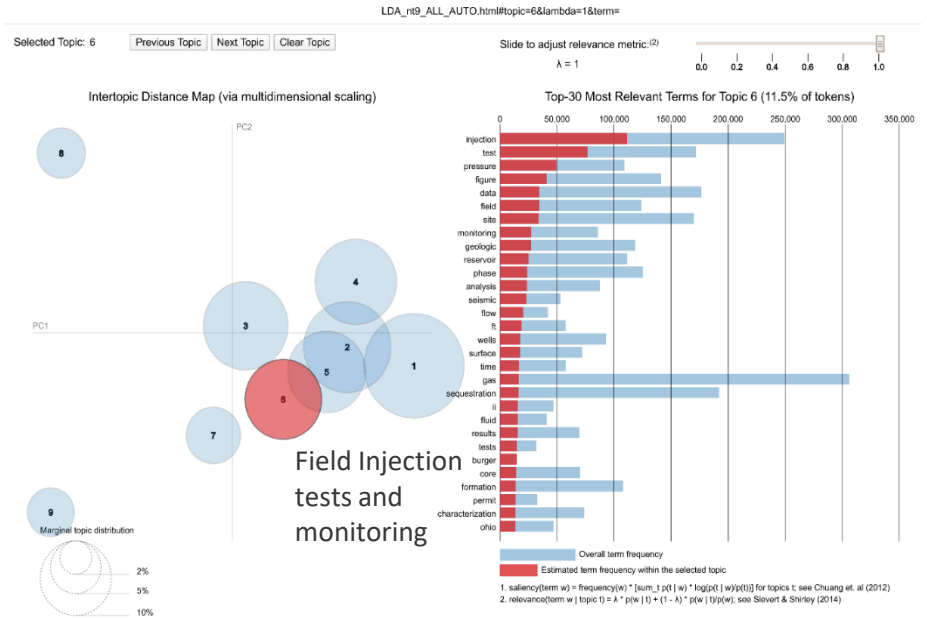




Morkner et al. 2022. Distilling data to drive carbon storage insights. Computers & Geosciences, https://doi.org/10.1016/j.cageo.2021.104945

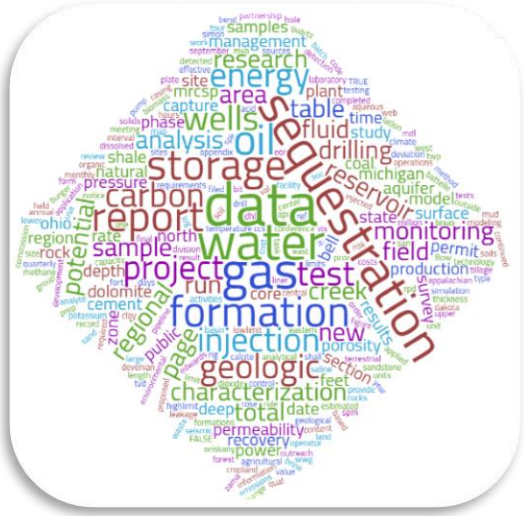
Results: NLP

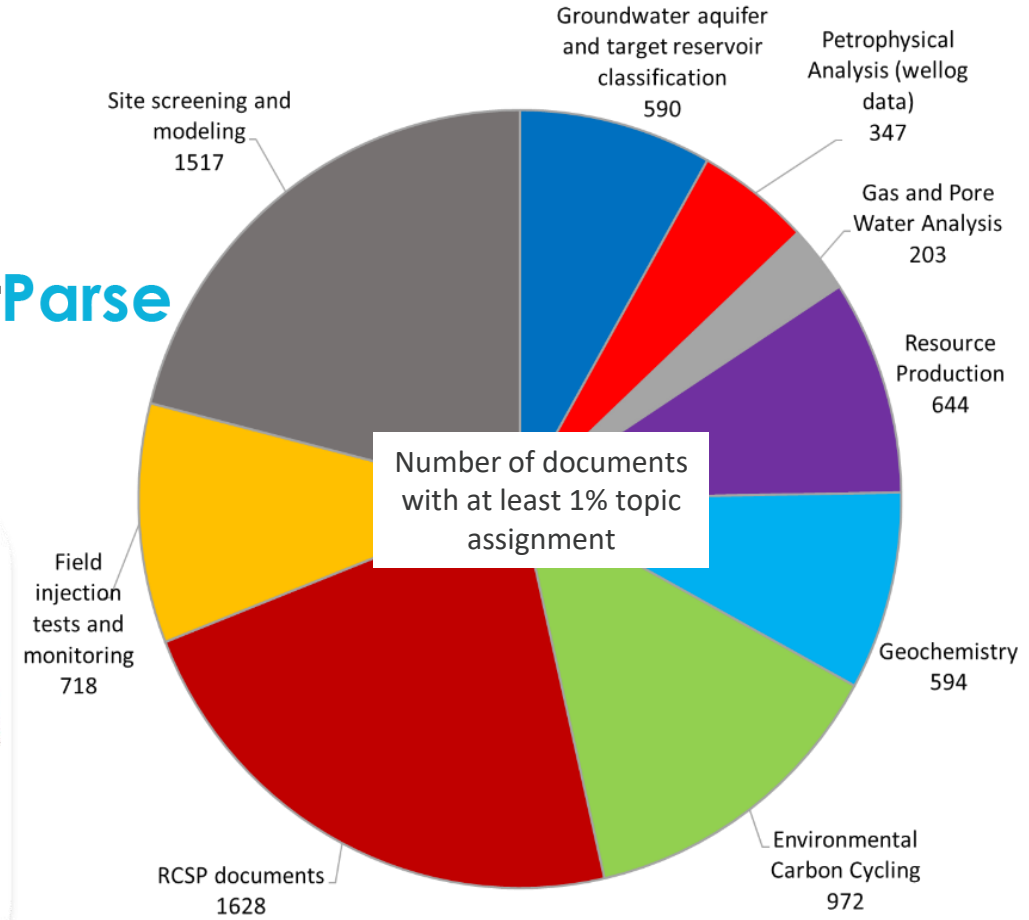
- Produced a **nine topic LDA model** – grouping similar papers
- Produced **keywords** associated with resources
- Geographic location recognition
- Papers published in groupings to EDX



NETL SmartParse

spaCy

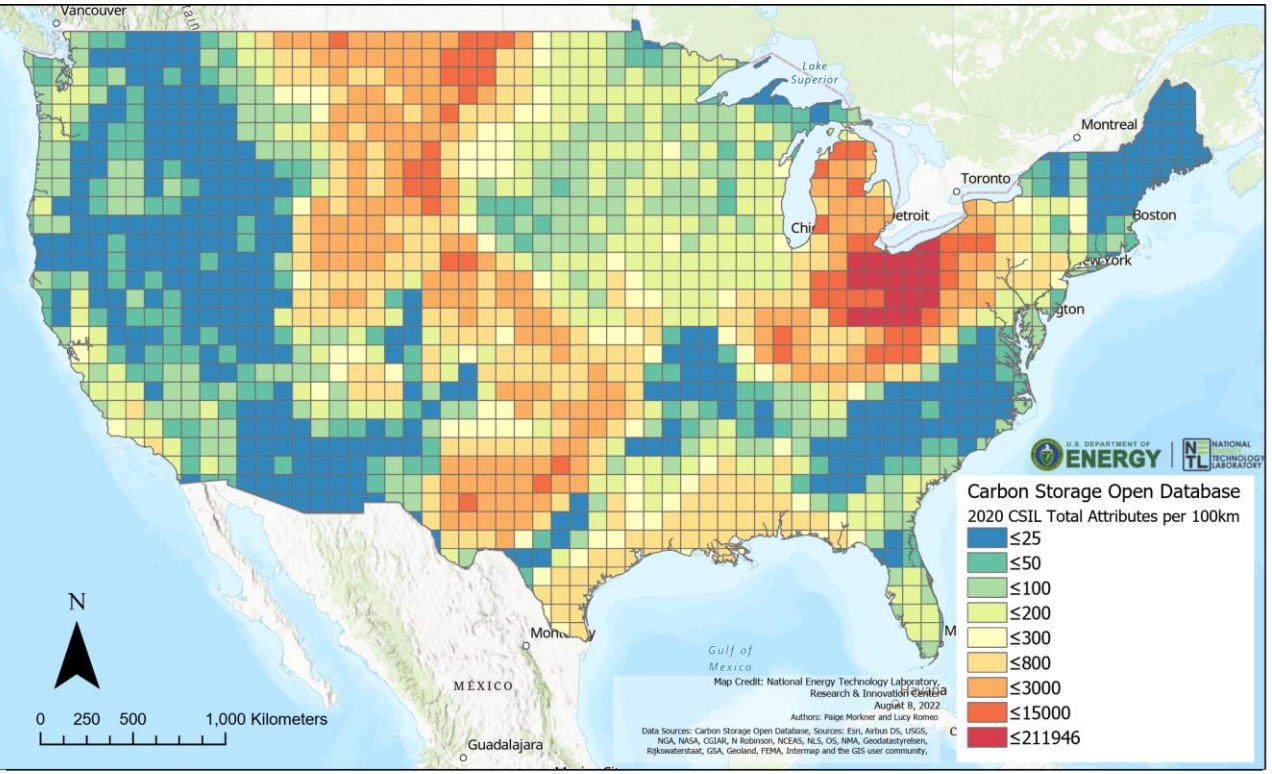




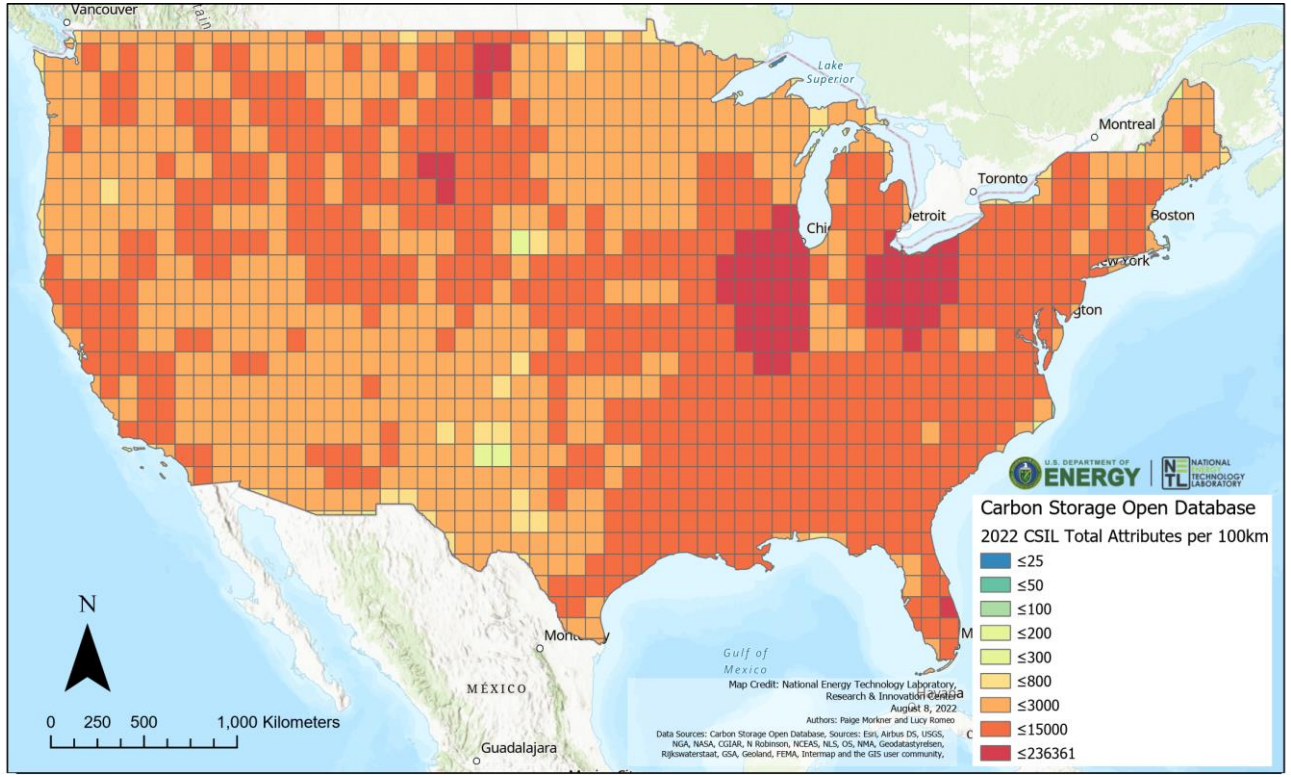
Morkner et al. 2022. *Distilling data to drive carbon storage insights*. Computers & Geosciences, https://doi.org/10.1016/j.cageo.2021.104945

Updating public data collections in EDX SPATIAL

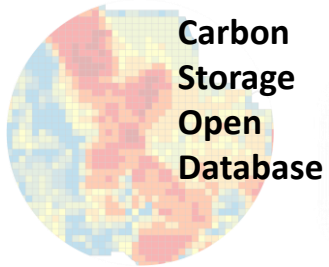
NATIONAL ENERGY TECHNOLOGY LABORATORY



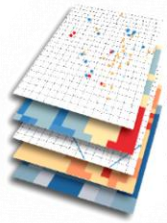
2019 Carbon Storage Open Database



2022 Carbon Storage Open Database with 580 additional shapefiles and rasters

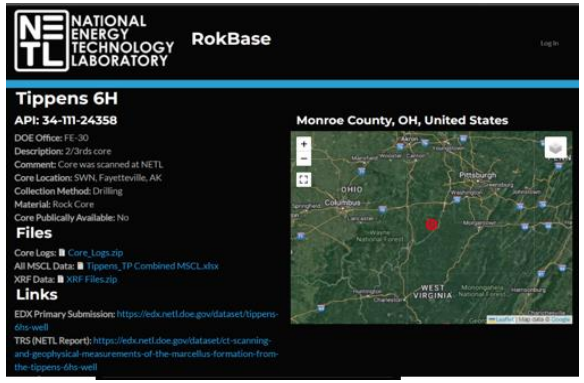


Carbon Storage Open Database



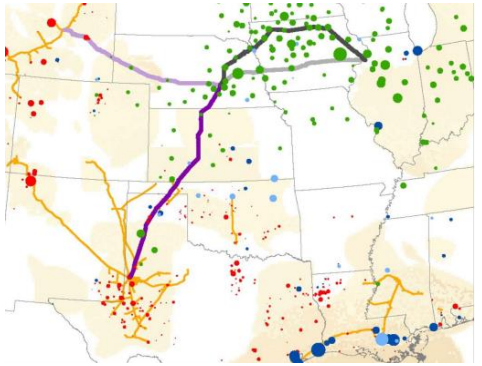
Cumulative Spatial Impact Layers

Workflows such as this have enabled action-orientated science under BIL work



RokBase

Virtual portal and tool for users to access, explore and query available core and rock property data collected from carbon storage spots across the U.S.

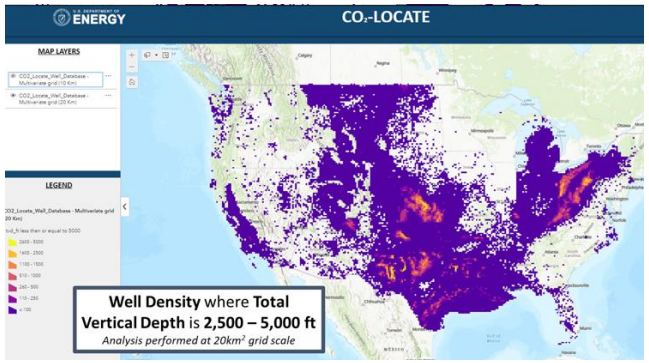


CCS Pipeline Route Planning Database

Containing >90 gigabytes of data representing critical considerations for the spatial routing of pipelines and transport of CO2

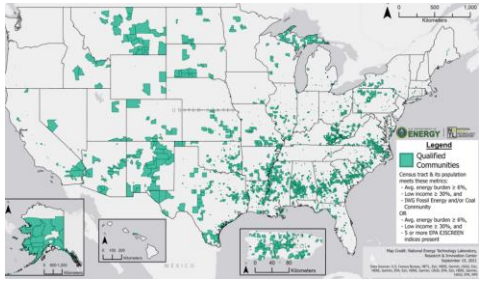
CO_2-Locate Database

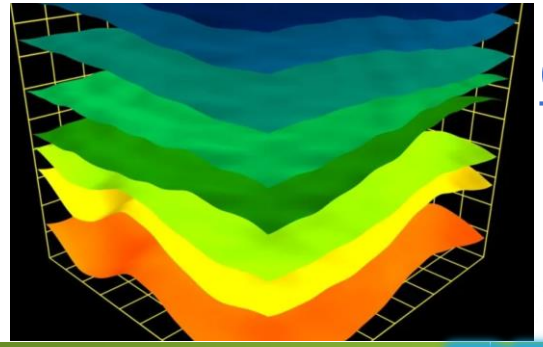
National Well Database & Class II Well Reuse and Regional Evaluation Tool built to support site selection, risk analysis and other stakeholder needs



CCS EJ-SJ Database

Social and environmental justice database aligned to carbon storage systems to aid in decision making around key future infrastructure placement & more





Class VI Data Support Tool

Geodatabase of spatial data layers to support the initial development of EPA Underground Injection Control (UIC) Class VI permits

CS Technical Viability Database V0.1

Geodatabase of geologic, geophysical, structural, hydrologic, energy extraction, transportation infrastructure, local political boundaries, community and environmental data

NETL RESOURCES

VISIT US AT: **www.NETL.DOE.gov**

@NETL_DOE

@NETL_DOE

@NationalEnergyTechnologyLaboratory