Putting Data to Work: Transforming Disparate, Open-Source Data for Engineered-Natural Systems and Models

Creason, C.G.[1,2], Romeo, L.[1,2], Bauer, J.[1], Rose, K.[1], Rowan, C.[1,3], and Sabbatino, M.[1,2]

[1] National Energy Technology Laboratory, 1450 Queen Ave. SW, Albany OR 97321, USA
[2] Leidos Research Support Team, 1450 Queen Ave. SW, Albany OR 97321, USA
[3] Attain, LLC, 3610 Collins Ferry Rd., Morgantown WV 26505, USA

Abstract #: IN51F-0689

NATIONAL ENERGY TECHNOLOGY LABORATORY
Research & Innovation Center

Abstract: As hydrocarbon exploration and production expands in the U.S. offshore, researchers at DOE's NETL have worked to build, manage, and maintain databases to help inform future research, regulatory, and commercial data-driven needs. Initial regional databases were designed as baselines to guide the prevention and preparation of hydrocarbon spills. These data baselines have been expanded and are currently being applied to support development of novel tools for assessing risk related to long-term CO_2 storage, estimating offshore subsurface CO_2 storage potential and associated risks, as well as regional-scale evaluation of gas hydrate deposits in onshore and offshore settings. These custom databases, which contain spatio-temporal information spanning the subsurface to surface in both onshore and offshore environments, have been built, rebuilt, redesigned, and transformed to ensure proper representation and retention of key knowledge and information for each system. This poster highlights obstacles encountered in building a baseline of open-source and big data, leveraging databases for multiple uses via scripting and manual processing, and lessons learned about attaining and maintaining accurate information from geospatial data. v

Building a Data Baseline:

The events of 2010's Deepwater Horizon and the resulting impacts introduced knowledge, technology, and data gaps pertinenant to better predictions, prevention, and preparedness of hydrocarbon release events in offshore regions. Within the scope of these needs, researchers collected over 5 terabytes of disparate data representing ambient conditions, socio-economic and environmental impacts, and response availability. Through a rigorous, multi-step process the data was cleaned, edited, and queried, into regional geodatabases for the Gulf of Mexico, offshore Southern California, and Gulf of Alaska. Since development, these databases required updates, management, and curration for a range of additional projects - costing more time and effort than originally expected.

Initial Application: Offshore Risk Modeling Suite

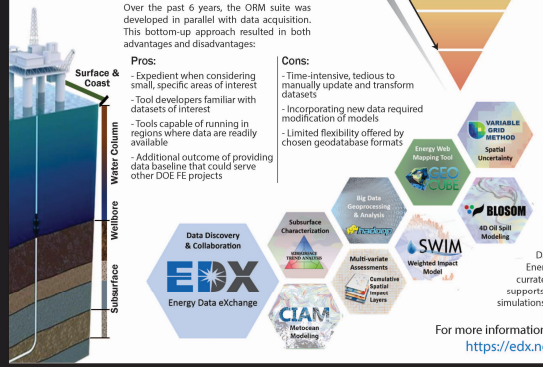
Over the past six years, researchers at NETL developed the Offshore Risk Modeling (ORM) suite, which is comprised of eight innovative data-driven computational tools and models designed to predict, prevent, and prepared for future hydrocarbon release events. The R&D 100 award-winning suite applies a novel approach spanning the full engineered and natural offshore system from the subsurface, through the water column, and to the coast.

R&D 100 AWARDS

Over the past 6 years, the ORM suite was developed in parallel with data acquisition. This bottom-up approach resulted in both advantages and disadvantages:

Pros:
- Expedient when considering small, specific areas of interest
- Tool developers familiar with datasets of interest
- Tools capable of running in regions where data are readily available
- Additional outcome of providing data baseline that could serve other DOE FE projects

Cons:
- Time-intensive, tedious to manually update and transform datasets
- Incorporating new data required modification of models
- Limited flexibility offered by chosen geodatabase formats



Flexible Protocol for Data Use & Reuse

Define Objectives
- Identify project outcomes and end-user expectations
- Clearly define data needs based on:
  - * Expert knowledge
  - * Comprehensive literature review
  - * Data accessibility

Acquisition
- Identify key search terms/phrases
- Search public/private data sources online or in scientific literature
- Digitize or download structure and unstructured data
- Utilize AI/Natural Language Processing
- Deploy automated or semi-automated web/text crawler
- Search public/private data sources online or in literature

Pre-processing
- Systematically scrub data
- Capture and record data quality and analytical uncertainty
- Record metadata with pre-set protocol; adhere to a standard metadata format (e.g., FGDC)
- Convert unstructured into structured format

Processing

Manual
- Query, transform data for specific project needs
- Integrate data into structured database
- Document processing steps to ensure consistency and reproducibility

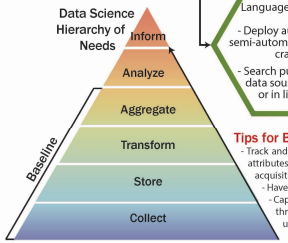
Automated
- Query, transform data as needed
- Integrate data into structured database

Post-processing
- Append data onto databases
- Break out for individual project needs
- Re-incorporate project results back into database

Update data given results

Automatic data updates (ex. real-time data)

Data Science Hierarchy of Needs



Tips for Building a Better Data Baseline:

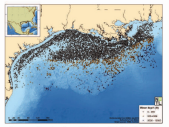
- Track and **ensure data accuracy**, both within the data attributes and spatial certainty throughout the acquisition, management, and exploration steps.
- Have and **enforce a QA/QC procedure** at each step.
- Capture and **record various types of uncertainty** throughout each step. This enables a better understanding of the data, analysis, and results for future use.

Reusing Baseline Data:

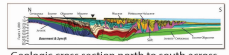
Leveraging data for multiple projects often requires multi-step reprocessing and additional time. Data can be made reusable more efficiently with the implementation of a flexible data management protocol (workflow above) at the onset of project.

Potential Challenges/Pitfalls:

- Limited data accessibility/availability
- Handling of proprietary data, ensuring conformance to use-restrictions from different sources
- Limited computational capabilities
- Neglecting to develop a post-project data management plan to ensure appropriate curation of data and metadata
- Personnel turnover, inadequate documentation to ensure project continuity



Petrophysical well logs (DiGiulio et al. 2019) and borehole data (BOEM, 2015) from the Gulf of Mexico.



Geologic cross section north to south across Louisiana coastal plain, shelf, and slope. (Modified from Morris et al., 2015).

Predicting Regional Gas Hydrate Deposits

Applying Offshore Data & Tools

This new project focuses on developing and applying an approach for a regional-scale prediction of natural gas hydrate occurences that can be tested in areas such as the Alaska North Slope and U.S. Gulf of Mexico. This effort aims to bring together existing, disparate datasets into a cohesive framework that supports data-driven predictions of offshore and onshore gas hydrate systems.

METHANE HYDRATES

For more information on NETL's Hydrates Program:
https://edx.netl.doe.gov/hydrates

Applications

Offshore CO2 Storage Methodology & Calculator

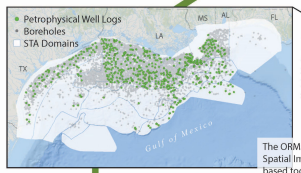
Applying Offshore Data & Tools

NETL's Offshore CO2 Storage project applies a DOE methodology to calculate potential CO2 storage capacity in offshore regions, starting in the Gulf of Mexico.

This project applies baseline data and information gained from the ORM suite's Subsurface Trend Analysis (STA) approach turned tool, including the resulting domain boundaries. These boundaries represent spatially-distinct geologic domains based on a regions lithologic, structural, and alteration histories (Mark-Moser et al. 2018; Rose 2016).

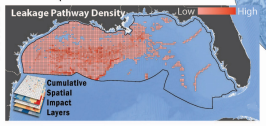
NETL CARBON STORAGE

SUBSURFACE TREND ANALYSIS



Resulting from this project, the Petrophysical Well Log dataset (DiGiulio et al. 2019) (left) and CO2 Storage Calculator (Romeo et al. 2019) were released via the ORM suite's EDX for easy access.

The ORM suite's Cumulative Spatial Impact Layer (CSIL) is a GIS-based tool that spatially assess data density and presence. As shown below, CSIL can be used to summarize infrastructure, geologic, and environmental data optimize the selection of injection sites based on risk and available infrastructure.

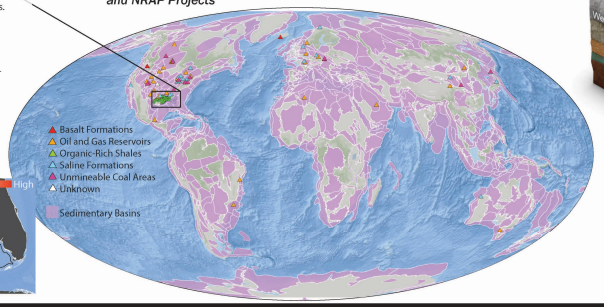


Data Library & Laboratory: EDX (Energy Data Exchange) is the DOE's Office of Fossil Energy's online data computing platform. EDX curate products, including tools and data, and supports the execution of gigabyte- and petabyte-scale simulations, including those performed by the ORM suite.

For more information on NETL's ORM Suite:
https://edx.netl.doe.gov/offshore

Potential Areas for CO2 Capture/Storage

Data from Offshore Carbon Storage and NRAP Projects



NRAP Data Model

National Risk Assessment Partnership

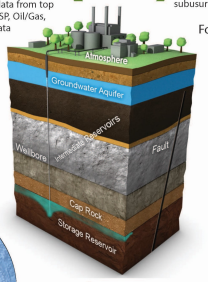
Supporting NRAP Data & Tools

The NETL-led NRAP project involves developing new tools, models, and methodologies to assess risk and quantify uncertainty related to large-scale CO2 capture and storage (CCS). This multi-institutional effort has produced related carbon storage resources, including geologic, engineering, geophysical, modeling, and analytical datasets spanning surface-subsurface systems.

Our role in NRAP is to: 1) provide a summary catalog of available data from top CCS sites in the US; 2) synchronize with other DOE FE resources (RCSP, Oil/Gas, Coal, REE, Geothermal, etc.) to coordinate additional leveragable data resources; and 3) synthesize a CCS "community dataset" for a single US site (or sites) to aid development and validation NRAP CCS modeling tools, and support application by external tool users.

NRAP National Risk Assessment Partnership

For more information on NRAP:
https://edx.netl.doe.gov/nrap



Developing a Virtual Subsurface Data Framework

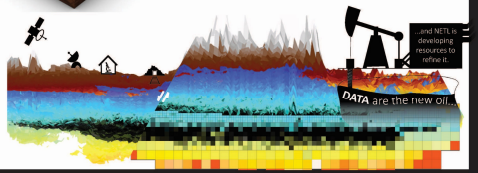
Addressing the Needs of Subsurface R&D

Building upon the baseline data presented herein and other ongoing federally funded efforts, NETL geo-data science researchers are refining and building tools to address needs specific to the subsurface community. In pursuit of this, we are leveraging lessons and best-practices learned from developing, using, and reusing the inherently large, disparate datasets for characterizing the subsurface.

For more information, see companion presentation by Rose et al., (Abstract: IN51F-0701)

Baseline data include:

- Depositional Environment
- Tectonic setting
- Faults/structure
- Seismic data
- Lithology
- Oil & gas infrastructure
- Well locations
- Well header info
- Production status
- Temperature
- Temperature gradient
- Pressure
- Pressure gradient
- Permeability
- Porosity
- Effective Porosity
- Net & gross sand thickness
- Areal extents
- Depth to top of target Formation
- Depth to protected groundwater
- Potential cap rock unit
- Salinity
- Atmospheric data (emissions)
- Geographic & surface systems data
- Lab & experimental data
- Numerical modeling data



Acknowledgement: This work was performed in support of the US Department of Energy's Fossil Energy Crosscutting Technology Research Program. The Research was executed through the NETL Research and Innovation Center's National Risk Assessment Partnership project. Research performed by Leidos Research Support Team staff was conducted under the RSS contract 89243318CFE000003.

U.S. DEPARTMENT OF ENERGY

LRST