Developing a Virtual Data Subsurface Framework for the U.S.-Combining Products of Fossil Energy R&D with Advanced Data Computing

Abstract: The data revolution has resulted in a proliferation of resources that span beyond commercial and social networking domains. Research, scientific, and engineering data resources, including subsurface characterization, modeling, and analytical datasets, are increasingly available through online portals, warehouses, and systems. Data for subsurface systems is still challenging to access, discontinuous, and varies in resolution. However, with the proliferation of online data there are significant opportunities to advance access and knowledge of subsurface systems. DOE’s Energy Data Exchange (EDX) is an online data computing platform designed to improve access to fossil energy R&D products, support multi-organizational R&D through an increasing suite of virtual sharing and analytical capabilities, through public and private online workspaces.

The EDX team has been developing a virtual subsurface digital data framework to support FE R&D for subsurface energy research. The system curates both structured and unstructured data about subsurface systems. These resources span petrophysical, geologic, engineering, geophysical, interpretations, models, and analyses associated with carbon storage, water, oil, gas, geothermal, induced seismicity and other subsurface systems to support the development of a virtual subsurface data framework. The EDX team is also currently developing custom machine learning algorithms and capabilities to enhance user experience, make access and connection to relevant, open-source, subsurface data resources more efficient for research teams to use, analyze and manage. EDX content and capabilities are continuously evolving. The development of a virtual subsurface data framework to support DOE FE data science and computing is the next step in the public and private resources in EDX seek to make subsurface energy research more efficient, reduce redundancy, and drive innovation.

Data Science Hierarchy of Needs

Scientific data poses unique challenges for organizations that need to store, share, and curate data. The Data Science Hierarchy of Needs outlines the fundamental concepts of data management that NETL is applying to its ever increasing data stores. This includes developing the tools, data, and procedures needed to move data from disparate data sets and schemas into a unified Living Database. This Living Database includes a continuous cycle of data collection and analysis to satisfy the needs of data science.

Next Steps

• Next 18 months focused curation of DOE FE data to seed virtual subsurface data framework further, including resources from RCEPS, Carbon Safe, oil/gas field projects, etc.
• Implementing NETL “Smart Search” tool via EDX to facilitate adaptive, real-time integration of open, online data from externally hosted sources
• Implementing advanced computing solutions to enhance data storage and scalability
• Potential to coordinate or actively integrate other authoritative federal subsurface systems (e.g. USGS, NSF, DOE, etc.) on private side of system

Need - Employ Data for New Insights

• Volume of data is growing: Scientific data is projected to exceed more than 40,000 exabytes by 2020.
• Finding older R&D data is hard: As published research ages, access to the underlying datasets decreases.
• 20% of world’s data are stored online while 80% are being privately held