Computers & Geosciences 158 (2022) 104945

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Distilling data to drive carbon storage insights

Paige Morkner [a,b,*], Jennifer Bauer [a], C. Gabriel Creason [a,b], Michael Sabbatino [a,b], Patrick Wingo [a,b], Randall Greenburg [c,1], Samuel Walker [c,2], David Yeates [c,3], Kelly Rose [a]

[a] National Energy Technology Laboratory, 1450 SW Queen Ave, Albany, OR, 97321, USA
[b] Leidos Research Support Team, 1450 SW Queen Ave, Albany, OR, 97321, USA
[c] Oak Ridge Institute for Science in Support of NETL, 1450 SW Queen Ave, Albany, OR, 97321, USA

ARTICLE INFO

Keywords:
Geologic carbon sequestration
Energy data exchange
Natural language processing
FAIR (Findable, Accessible, Interoperable, Reusable)
Data availability
Spatial data density analysis

ABSTRACT

Wide-spread implementation of carbon capture and storage has the potential to decrease carbon emissions and aid in meeting global climate change mitigation goals. Data availability is one of the biggest challenges faced by the carbon capture and storage (CCS) community for modeling risks associated with CCS, necessary for wide-spread implementation in coming years. Collecting, integrating, and intuitively managing data is a time-consuming process, but one which is fundamental to establishing necessary access to carbon storage data. The US Department of Energy (US DOE) has been a major supporter of energy research in the US, including significant investment into carbon capture and storage research and technology development over the last ten years. The US DOE investments into the Regional Carbon Sequestration Partnerships, the National Risk Assessment Partnership, and other CCS related research has resulted in a large volume of data, of which much has been made public through the National Energy Technology Laboratories data repository, the Energy Data eXchange (EDX). Researchers at the National Energy Technology Laboratory have developed workflows, tools, and other methods that leverage EDX, open-source software, machine learning, and natural language processing to discover, curate, label, organize and visualize available data. This paper describes the available data on EDX for carbon storage applications, describes the results of a spatial and temporal analysis of the data, describes where it is most geographically available, makes a general assessment of the quality of the available data, and discusses visualization tools and natural language processing tools developed for understanding, discovering and reusing the data.

1. Introduction

To meet climate change mitigation goals set out by the international community (Masson-Delmotte et al., 2021; *Paris Climate Agreement, 2015*; USGCRP, 2018), there is a need for wide-spread implementation of capture and storage of carbon dioxide (CO_2) emissions. Globally, carbon capture and storage (CCS) has the potential to reduce a sixth of CO_2 emissions and help prevent global temperature increases above 2 °C (Haszeldine et al., 2018). Recent technological advances have improved adsorption and absorption technologies capable of separating CO_2 from other byproducts at point sources (Cuéllar-Franca and Azapagic, 2015; Rodosta et al., 2017). Once captured, the CO_2 can be injected into porous rock formations deep enough to maintain the CO_2 in supercritical state, allowing long-term stable storage in the subsurface

Abbreviations: CCS, Carbon capture and storage; FAIR, findable, accessible, interoperable, reusable; NLP, Natural language processing; NATCARB, National Carbon Sequestration Partnership; RCSP, Regional Carbon Sequestration Partnerships; CSOD, Carbon Storage Open Database; NRAP, National Risk Assessment Partnership; ML, Machine Learning; U.S. DOE, United States Department of Energy; NETL, National Energy Technology Laboratory; LDA, Latent Dirichlet allocation; CO_2, Carbon dioxide.

* Corresponding author. 1450 SW Queen Ave, Albany, OR, 97321, USA.
E-mail addresses: Paige.Morkner@netl.doe.gov (P. Morkner), Jennifer.Bauer@netl.doe.gov (J. Bauer), Christopher.Creason@netl.doe.gov (C.G. Creason), Michael.Sabbatino@netl.doe.gov (M. Sabbatino), Patrick.Wingo@netl.doe.gov (P. Wingo), ragreenburg@gmail.com (R. Greenburg), srwalker6@crimson.ua.edu (S. Walker), Bart_Yeates@baylor.edu (D. Yeates), Kelly.Rose@netl.doe.gov (K. Rose).
[1] Present Address: G&D Chillers, 760 Bailey Hill Road, Eugene, OR 97402, USA.
[2] Present Address: University of Alabama, Tuscaloosa, AL, 95487, USA.
[3] Present Address: Baylor University, 1311 S 5th Street, Waco, TX, 76706, USA.

https://doi.org/10.1016/j.cageo.2021.104945
Received 27 July 2020; Received in revised form 31 August 2021; Accepted 16 September 2021
Available online 5 October 2021

(Cuéllar-Franca and Azapagic, 2015; Haszeldine et al., 2018; Rodosta et al., 2017). At present, one half of the U.S. CO_2 emissions originate from point sources (i.e., power plants and industrial manufacturing) where CO_2 can be captured for use or long-term storage, offering significant benefits for the U.S. economy and environment (Alcalde et al., 2018; Rodosta et al., 2017; U.S. Energy-Related Carbon Dioxide Emissions, 2019, 2020, p.; US EPA, 2015). With sufficient data and economic incentives, CCS has been effectively implemented at a number of sites worldwide. At one such site, Sleipner, an offshore sandstone reservoir in Norway, 0.9 Mt/year of CO_2 has been effectively injected since 1996 (Furre et al., 2017). Monitoring of the subsurface using four-dimensional (4D) seismic data and modeling has shown that CCS can be viable in the long term and an effective approach to decrease overall carbon emissions (Furre et al., 2017).

The United States Department of Energy (U.S. DOE) Carbon Storage program has been advancing CCS research and technology since 1997 (Rodosta et al., 2017). To meet outlined goals, the U.S. DOE has developed projects related to CCS data collection, curation, and risk modeling efforts to support future CCS project site screening including the Regional Carbon Sequestration Partnership (RCSP) Initiative consisting of seven regional partnerships spanning the U.S. and southern Canada, the National Carbon Sequestration atlas and database (NATCARB), and the National Risk Assessment Partnership (NRAP) (Fig. 1). The RCSP Initiative supports large and small-scale pilot CCS injection test projects and supports the development of the NATCARB atlas and database. NRAP aims to develop new methods and tools for science-based prediction of the environmental risks associated with long-term carbon storage.

The demand for access to high-quality geologic data to support CCS efforts is increasing. The data pyramid (Fig. 2) describes the process and levels of effort required to collect, store, curate, and utilize scientific data. Published literature consisting of heavily distilled and refined modeling, machine learning (ML) applications and data analysis makes up the top levels of the data pyramid. Significant time in research is spent on the lower levels of the data pyramid ("2016 Data Science Report - CrowdFlower," 2016) consisting of data discovery, acquisition, quality assessment, organization, labeling and curation. The lower levels are essential to support the analysis and conclusions made at the top. The necessary work that occurs in the lower levels of the data pyramid tasks can be completed in a way that develops and contributes tools and techniques for future researchers to leverage and build upon.

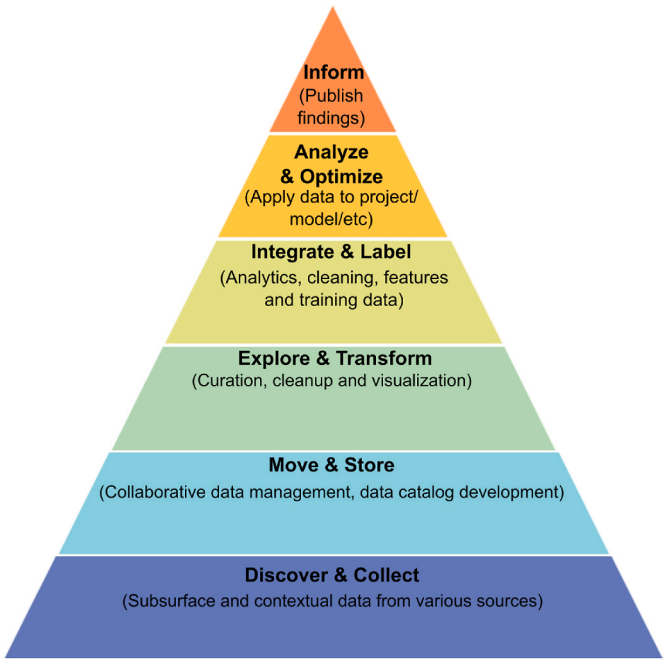

Fig. 2. The data pyramid, adapted from (Justman et al., 2020; Wenzlick et al., 2020). Each layer represents a group of tasks which must be completed prior to all layers above it. The width of the layer represents a relative effort commitment; the lower the layer, the more time is generally spent on tasks within ("2016 Data Science Report - CrowdFlower," 2016). Collectively the layers of the pyramid describe the steps from collecting data through using the data for publishing, for machine learning datasets, and for model inputs.

Efficient carbon storage modeling presently faces the challenge of low availability of data for site selection and hazard mitigation - a data collection and curation challenge represented by the lower levels of the

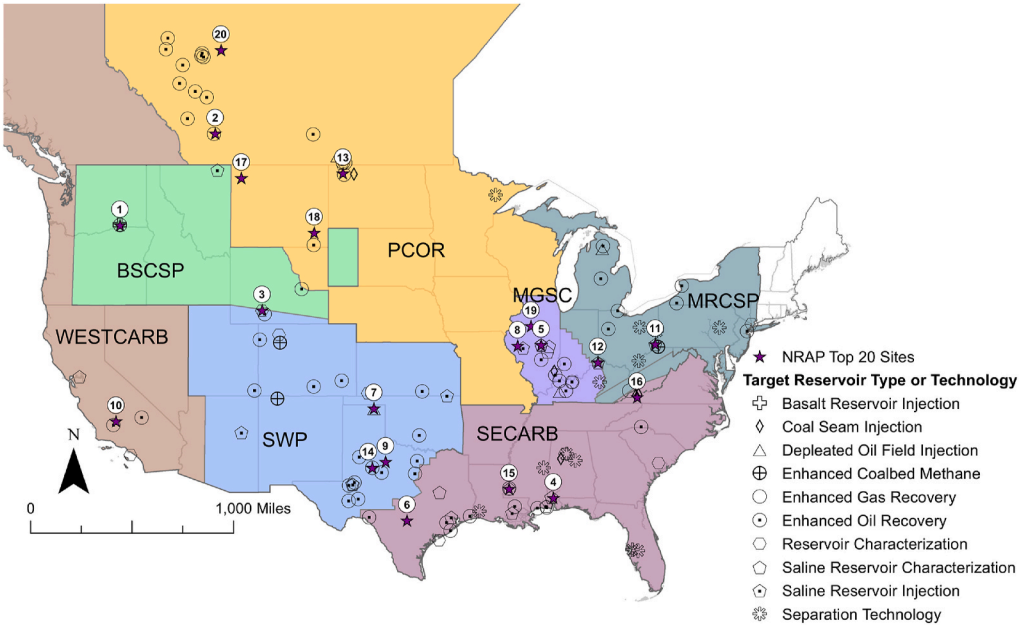

Fig. 1. Outline of RCSP boundaries labeled with partnership names, the NRAP top 20 sites for modeling CCS risk and N. American CCS sites. NRAP top 20 sites: 1. Big Sky Validation Phase - Wallula Basalt Pilot Project. 2. CaMI - Field Research Station. 3. CarbonSAFE - Wyoming. 4. Citronelle (SECARB). 5. Decatur. 6. Edwards Aquifer. 7. Farnsworth - Anadarko Basin. 8. FutureGen. 9. High Plains Aquifer. 10. Kimberlina (WESTCARB). 11. Appalachian Basin Test (MRCSP). 12. Cincinnati Arch Test (MRCSP). 13. Williston Basin Oil Field Test (PCOR). 14. Scurry Area Canyon Reef Operations. 15. Cranfield Site (SECARB). 16. Central Appalachian Basin Test (SECARB). 17. Kevin Dome. 18. Bell Creek. 19. CarbonSAFE Illinois Macon County. 20. Quest Canada.

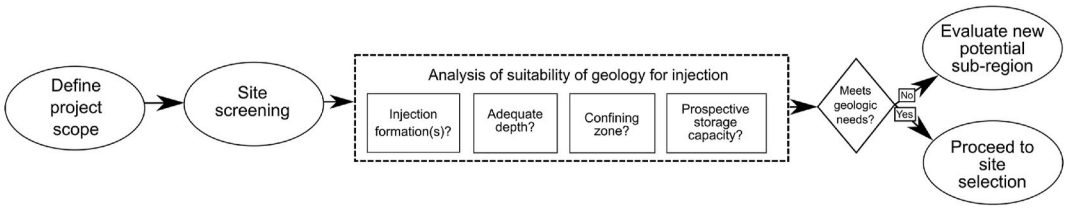

Fig. 3. Flow chart for CCS project site selection. Regional geologic data is essential to the site screening process. Adapted from Rodosta et al. (2011).

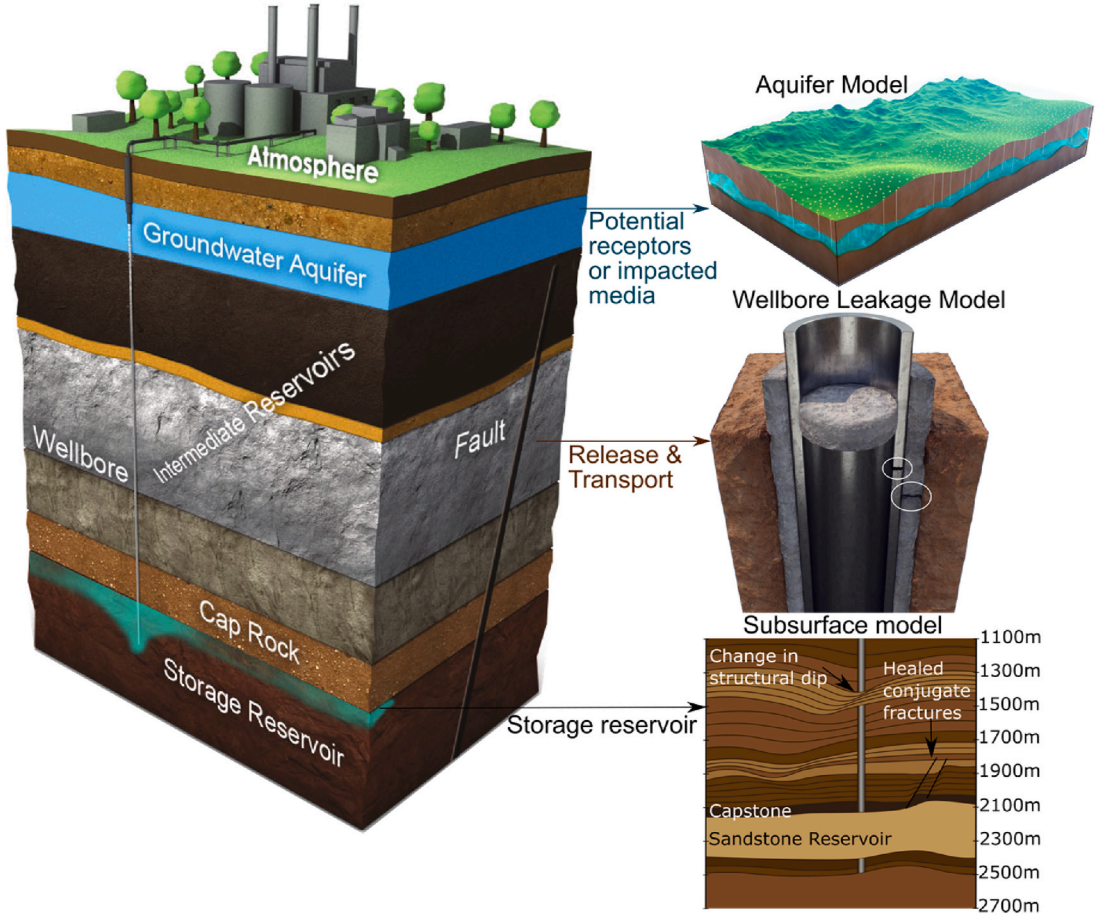

Fig. 4. NRAP tool modeling efforts target three distinctive parts of the subsurface to assess risk associated with CO_2 injection from the target reservoir to the surface. Modeling breaks down the subsurface at the storage reservoir level, release and transport (leakage) level, and the impacted media/receptor level (groundwater aquifer and surface atmosphere). On the right are examples of the types of models which model the different levels of the system.

data pyramid. Once the scope of a CCS project is defined, site screening is the first step in selecting an injection location (Fig. 3). Using available regional geologic data to analyze the suitability and capacity of storage is an essential step of site screening (Fig. 3) (Rodosta et al., 2011).

Once the geologic context of the potential CCS reservoir is defined, subsurface data is applied to hazard analysis and modeling (Fig. 4) (Rodosta et al., 2011). Necessary geologic data for hazard modeling varies, but generally consists of data related to the subsurface geology, groundwater aquifer, potential leakage pathways (e.g., faults), and seismicity potential (Fig. 4). If data are accessible for the subsurface in a region of interest, the site selection processes and hazard analysis can be simplified making CCS projects economically and physically achievable (Rodosta et al., 2011). By combining data and modeling, a more robust framework for future CCS becomes possible (Harp et al., 2019).

Overcoming challenges in data availability and heterogeneity is difficult. Lack of data and incentives in the U.S. has led to a delay in the number of CCS injection tests and projects, which are important to meet global climate change mitigation goals (Conway et al., 2013; DOE Data Crosscutting Requirements Review, 2013; Mackenzie, 2021). For industry to consider CCS a key technology for carbon management, the CCS community has stressed the need for better curated data to effectively simplify site screening and feasibility analyses (Accelerating

Breakthrough Innovation in Carbon Capture, Utilization, and Storage, 2017; Mackenzie, 2021; Page et al., 2020). Recently, the U.S. DOE carbon storage program has transitioned to preserving and publishing data primarily through the Energy Data eXchange (EDX), the National Energy Technology Laboratory's (NETL) data warehouse and collaboration platform, but data previously published on public websites and databases remained disparate and difficult to find. Consolidating resources into an easily accessible database allows users, such as researchers, industry, and government agencies, to access data necessary for geologic CCS site screening and selection, and to develop carbon storage hazard models. Developing data collection and review processes for public data ingestion into EDX ensures the interoperability, reuse and trustworthiness of additional resources.

Open-source and custom ML processes can be used to resolve the challenges faced with finding, transforming and managing data (Hey et al., 2009; Rose et al., 2015). Custom ML processes were applied when developing a workflow for collecting and curating available open source, disparate carbon storage data resources. This paper aims to describe the data collection and curation process, establish a spatial and temporal understanding of U.S. Department of Energy publicly available data for carbon storage applications, and discuss tools available for understanding, discovering, and reusing the data. While this paper specifically addresses the acquisition and curation of CCS data, the process and tools outlined here can be used as a template for data collection and curation for geothermal, waste-water injection, resource exploration, basin modeling, and many other computing science-geodata driven applications.

This paper will describe the methods developed to collect, catalog, and curate disparate, structured and unstructured, open-source carbon storage data, present the results of applying natural language processing for unstructured data curation, and present an analysis of U.S. DOE carbon storage data resource availability, spatial data distribution, and data quality.

2. Methods

A workflow (Fig. 5) was designed and implemented to help rapidly assimilate data from multiple sources using semi-automated collection, processing, formatting and review to produce more integrated and documented data. Each step and implemented method of the process is described. The data, once collected and organized, were used to develop a custom natural language processing carbon storage specific corpus and language library, produce a carbon-storage specific topic model for text-based resource classification and keyword identification, and derive advanced data analytics.

2.1. Data collection and assimilation

A systematic approach was developed for the collection, curation, metadata development, and publishing of carbon storage resources included in this analysis – the NATCARB database, the NRAP catalog database and the carbon storage open database. Publishing and curating the three resources together is intended to improve data availability and offer a more robust, comprehensive collection of data to support carbon storage analytics and modeling for site selection and hazard mitigation.

The disparate nature of open-source carbon storage data resources proved a significant challenge during organization and development of the databases. Data resources include, but are not limited to, spatial resources such as vector shape files and raster images, synthesized information from text-based resources, and reservoir, aquifer, leakage and seismic hazard simulations. Text-based resources include technical reports, conference proceedings, abstracts, informational pamphlets, site-specific reports, fact sheets, government reports, presentations, internal U.S. DOE reports, and public websites.

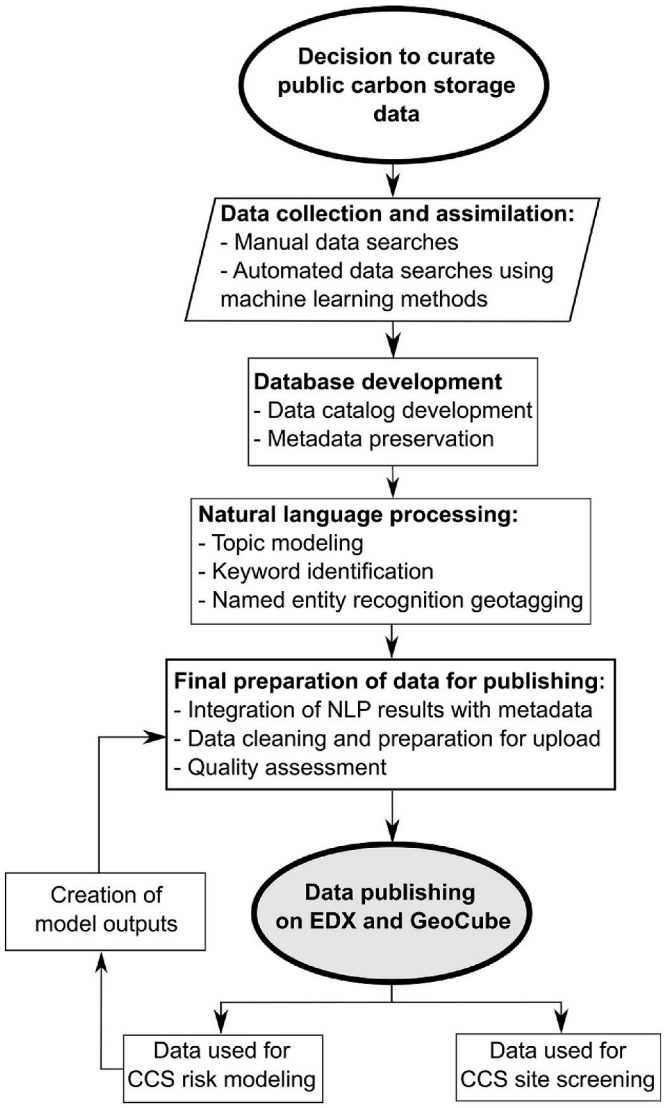

Fig. 5. Workflow diagram for data assimilation, processing, and assessment to support advanced analytics and modeling. The top circle indicates the decision to start the CCS data collection and curation effort. Parallelograms indicate data collection and assimilation processes, and rectangles indicate steps where data is moved, cataloged, curated, and applied. The rectangle titled "final preparation of data publishing" describes the process of final preparation of data through integration of new metadata produced from natural language processing, and the process of quality control and quality assessment prior to release. The bottom circle indicates the process of publishing the data publicly on EDX. EDX – Energy Data eXchange, CCS – Carbon capture and storage, NLP – Natural Language Processing.

A combination of methods was used to collect data resources and produce metadata, including manual, expert-driven document search and automated collection using a custom, ML tool, Smart Search (Baker et al., 2016), and other Python scripts to scrape websites and download

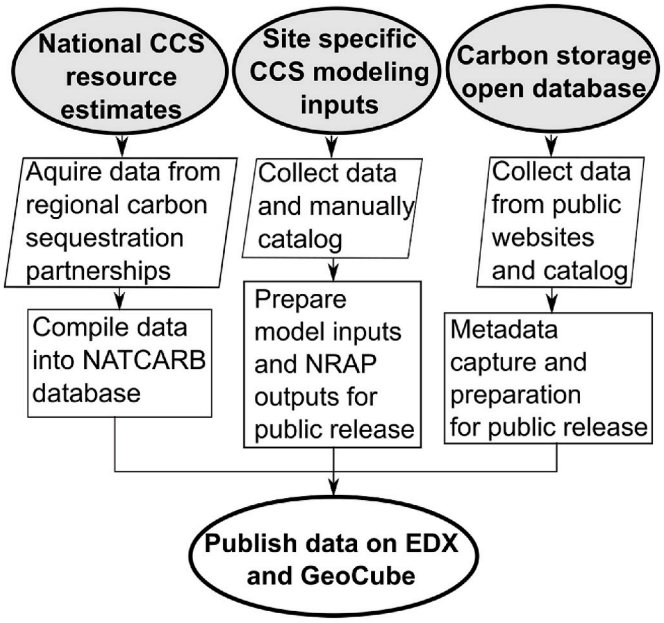

Fig. 6. Data collection workflow diagrams for NATCARB, NRAP, and open-source CCS-related projects. Symbology: circles indicate the beginning and end of processes or sub-processes, squares indicate a process of changing the value, form, or location of data, and parallelograms indicate when data is created or added. Acronyms used: NATCARB – national carbon sequestration partnership, NRAP – national risk assessment partnership, EDX – Energy Data Exchange.

metadata and data resources (Fig. 5). To rescue resources no longer available, the Internet Archive's Way back Machine website (https://archive.org/web/) was used to access known resources. Due to variability in the type and source of each database, collection and curation processes vary slightly for each; however, all data resources were collected according to the developed workflow, spatial resources collected were prepared and assimilated into collections, and text-based resources collected were used to develop the natural language processing model.

2.1.1. NATCARB data collection

Collaboration between NETL and the regional carbon sequestration partnerships (RCSPs) facilitated the generation of data for NATCARB (Fig. 6), from both spatial and text-based data collected and compiled from the RCSPs. Data are requested from the RCSPs based on a predefined list of data needed for subsurface CCS analysis and storage estimates (Goodman et al., 2011), as well as spatial data to support storage calculations and identify CO_2 emitters. The spatial data are created from a template to generate storage estimates for petroleum reservoirs, saline basins, and/or coal deposits as well as a 10 km × 10 km gridded shapefile for each saline and coal basin. The resources are uploaded into EDX (Fig. 6) and GeoCube, EDX's geospatial data display platform (Barkhurst et al., 2018) (Fig. 7). The last data call and update for the NATCARB database was in 2015 with NATCARB Atlas version 5.

2.1.2. NRAP data collection

An expert-driven manual approach was used to compile and catalog open-source CCS-related data resources for 20 priority CCS synthetic and demonstration sites in North America (Fig. 1). The catalog and resource collection were created to support internal development and testing of the NRAP risk modeling tools. Manual searching involved performing simple keyword queries and identifying relevant results (Fig. 5) and using general and research-specific web search engines and public data repositories (e.g., Google Scholar (scholar.google.com), Google Dataset Search (datasetsearch.research.google.com), U.S. DOE Office of Science and Technology (osti.gov), and EDX). Cataloged resources include both spatial and text-based datasets spanning wellbore, aquifer and reservoir characteristics, including geochemical properties, in situ stress conditions, and other CCS modeling parameters.

The NRAP data catalog also contains proprietary resources. These datasets, typically synthetic model simulations, were created and cataloged on a case-by-case basis to address unmet data needs for specific tools. Non-public datasets are safeguarded within a private yet community-accessible workspace on EDX until approved for public release. This approach ensures a comprehensive catalog that captures the data and information available for tool developers and users while honoring various intellectual property rights and restrictions.

2.1.3. Carbon storage open database

Additional CCS data was identified and compiled into the Carbon Storage Open Database (CSOD) (Morkner et al., 2020) from public sources, including the regional carbon sequestration partnership public websites and servers. A python script was developed using Environmental Science Research Institute's (ESRI's) ArcGIS (https://www.arcgis.com/) python extensions package (arcpy) to search for files and data with spatial information and collect the data into a single geodatabase. Read me files were prepared for each download package, capturing source metadata including data name, data format, date created, date published, date of last recorded update, spatial reference system, spatial extent, spatial resolution, brief description, layer legend, data source information, data source URL, data source restrictions, acquired by, acquisition date, and any processing steps. Spatial data was analyzed and classified into 12 distinct feature datasets: geology, bedrock geology, karst geology, mine, oil and gas, structure, geomorphology, geologic contours, surface hydrology, ground water, isotherms, and sedimentary basins.

2.2. Data cataloging, metadata capture, and database development

Data catalogs were developed to document metadata and preserve data provenance information. Catalogs are a methodical curation approach to preserve information about data during collection and subsequent changes to said data. Data catalogs were developed for each individual database: NATCARB, NRAP, and the carbon storage open database. Cataloged attributes where chosen based on the DataCite metadata schema 4.3 (DataCite, 2019), and customized to meet the needs of each project. The catalogs were partially autogenerated using python scripts developed using arcpy to extract information about each file type, file size, and record numbers. Information was manually extracted regarding data API, digital object identifier (DOI), original URL, date published, date last updated, data source, date acquired, local drive file path, spatial extent, summary of resource, licensing type, temporal data quality, and source data quality. These catalogs were used internally to prepare data for publishing, and in the case of the NRAP data catalog (DiGiulio et al., 2020), published publicly as a community resource.

Geospatial data resources from NATCARB and the carbon storage open database were converted from an ArcGIS Geodatabase (.GDB) into

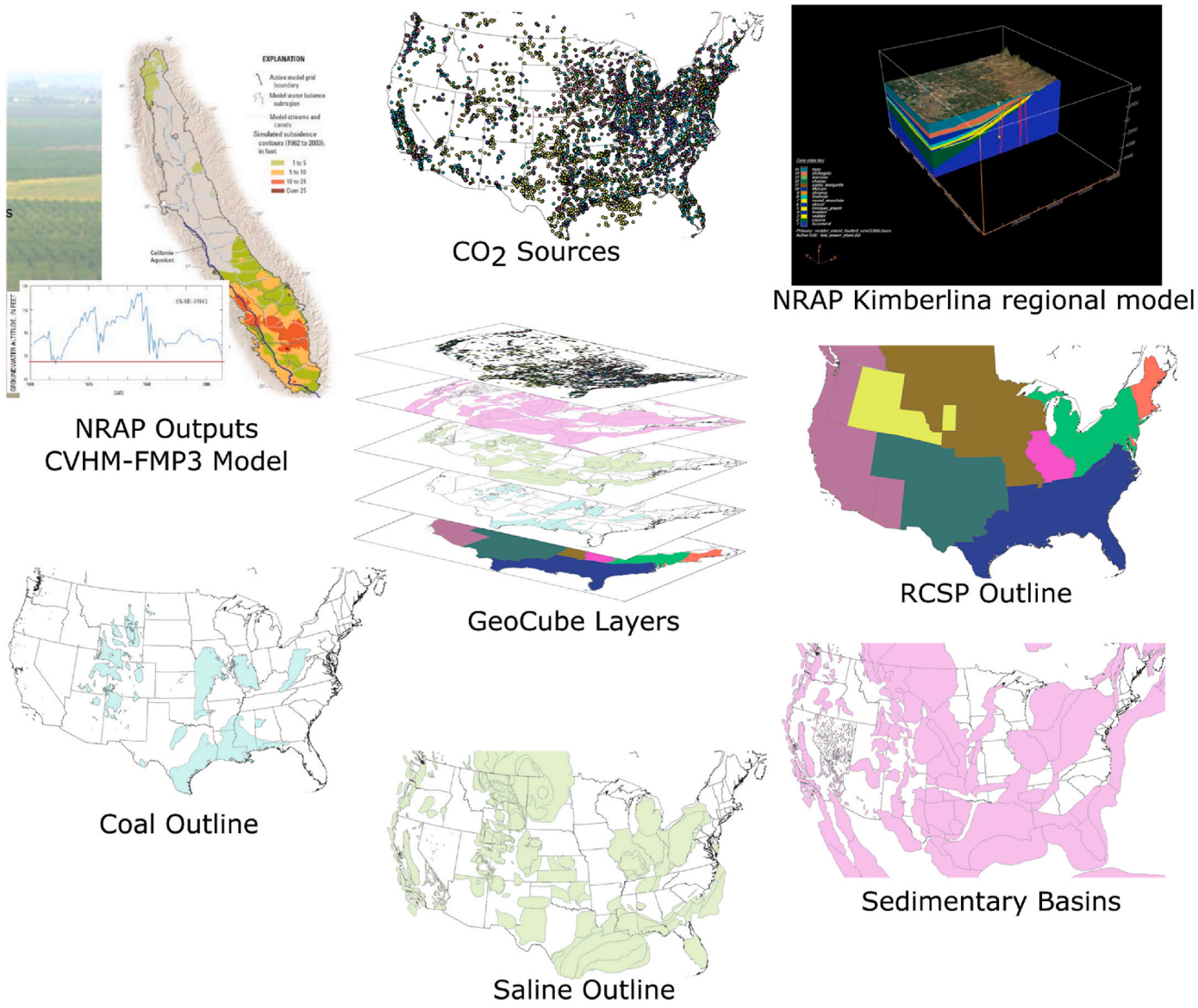

Fig. 7. Different layers of carbon storage data available, and how these are combined into GeoCube – EDX's geospatial data display platform. Layers shown are from the NATCARB database, and NRAP outputs CVHM-FMP3 Kimberlina model adapted from Faunt et al. (2016) and NRAP Kimberlina regional static model from Quinn et al. (2013).

a PostGIS database, a format compatible with data hosting on GeoCube. In addition, a metadata JavaScript Object Notation (JSON) file and a vector tile layer style was developed so data could be customized on the GeoCube web mapping application (Fig. 7).

2.3. Natural language processing

Once data were collected and assimilated into a database, additional processing was performed to generate information and metadata to further integrate and connect unstructured data resources (Fig. 5). Open-source natural language processing applications were customized to parse, categorize, and assign key terms for each resource. Additional machine learning techniques were used to recognize locations named within each resource when possible.

2.3.1. Topic modeling and keyword identification

Natural language processing (NLP) was used to parse common terms associated with each resource, to be integrated into the metadata for better key term search capabilities (Fig. 8), and to develop a topic model for paper classification. For NLP topic modeling, PDF resources collected for the carbon storage open database and the NRAP catalog database were assembled into a literature corpus. To build topic models, the corpus was cleaned and parsed using Python scripts produced at NETL into a PostgreSQL database, analyzed for common terms, and grouped (Fig. 8). The Python library Gensim (https://radimrehurek.com/gensim/) was used to develop a Latent Dirichlet allocation (LDA) model that takes an input of a corpus and an integer representing a target number of topics (Fig. 8). This library then converts the collected text into an unsupervised model that classifies the most common phrases and groups them into logical topics. To select the number of topics for the final model, multiple models with different topic numbers were produced and the coherency scores were compared. The topic coherence score is a measure of the semantic similarity between high scoring words in each topic. Coherency scores increase as topic numbers increase, but above a certain number of topics no longer improves. When the increase in topic number no longer improves coherency, the model is considered optimized. A qualitative check is then performed by subject matter experts to ensure that keywords determined for each topic make sense. To

optimize the performance of the output the models were tested using the NETL machine learning cluster (https://ml.netl.doe.gov/). The high-performance cluster was able to create multiple models with various topic sizes within hours, as opposed to days of processing on a traditional desktop computer.

Once the model was optimized for the number of topics, each topic was assigned a category name by subject experts at NETL (Fig. 8). Papers from the corpus were then run through the optimized model and assigned a percentage for each relevant topic. The common terms from each topic were used as additional metadata associated with each paper. The python codes developed for the NLP and topic modeling process (Fig. 8) are available on Github (https://github.com/NETL-RIC/Carbon-Storage-Distillation-NLP–Release).

2.3.2. Named entity recognition

EDX and GeoCube enable the user to find data with both keyword and spatial search functions. To enable spatial searches for text-based resources, each resource was geotagged – the process of adding geographical information to metadata – using SpaCy (https://spacy.io/), a natural language processing model and library in Python. Named entity recognition (NER), preformed using the SpaCy library (https://spacy.io/api/entityrecognizer), was used to classify text strings as geopolitical entities (city, state, country) and other spatial location names, including the custom entities of geologic basin names and CCS field site names. An additional Python script was used to identify the most common locations mentioned; the locations are then geotagged with latitude and longitude coordinates using the Google Maps API. The coordinates are associated with the metadata of each resource, enabling spatial search to filter PDF resources for an area of interest.

2.4. Data cleaning and metadata capture

Metadata is fundamental to ensuring the integrity, timeliness and reusability of data. It is commonly used to describe the purpose, extent, source, processing, and quality of a data resource. When integrating data from multiple sources and formats, additional metadata fields may need to be added to ensure consistency for each data record within the collection. When publishing data on EDX, metadata is prepared for each database, spatial data resource, and text-based resource submission based on the DataCite.org metadata citation standards (DataCite, 2019). For database collections (i.e., NATCARB database and Carbon Storage Open Database) this includes DOI number, resource type, title, authors/contributors, description, citation, keywords, publisher, publication year, point of contact, spatial extent, and maintainer contact information. Python scripts were used to automate the process of metadata extraction and preparation for each individual spatial data resource, including data category, description, licensing, source URL, keywords, shapefile type, layer name, spatial extent and database collection name. The metadata was prepared in a JSON file, uploaded with the PostGIS database, and used to customize the data for display in GeoCube.

In addition to source-provided metadata, topics and keywords generated from natural language processing were added to each

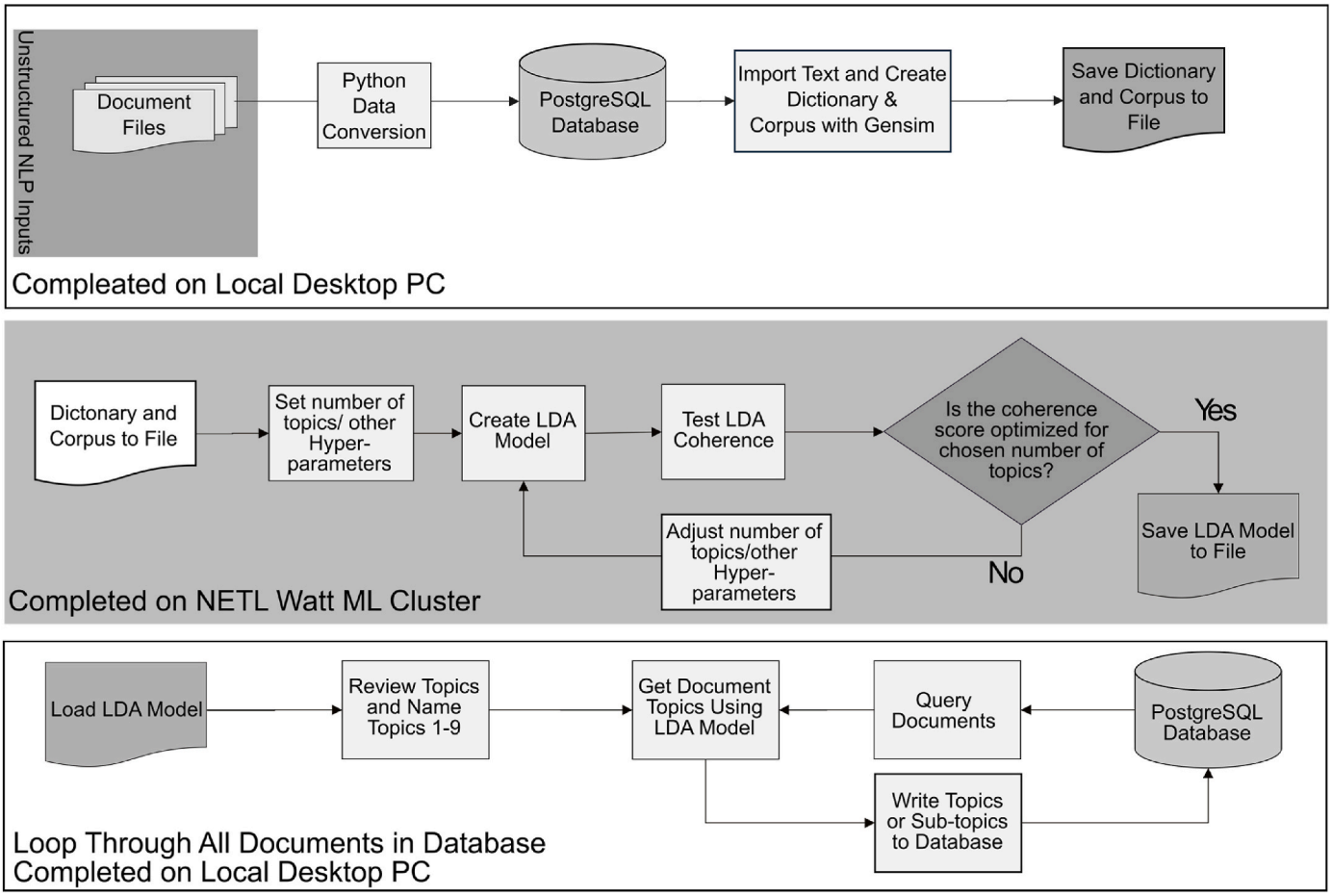

Fig. 8. Workflow diagram describing multistep process for natural language processing. Starting with NLP inputs (top left), a literature corpus is input as the document files. This undergoes a conversion and parsing process, resulting in a dictionary and corpus file. The dictionary and corpus file is then used to develop the LDA model on the NETL Machine learning Cluster with a variable number of topics. Once an optimal number of topics is found, the model is then analyzed for coherence. The LDA model topics are assigned names, and each document desired to be parsed is then run through the model and assigned percentages within each topic, and key terms are defined.

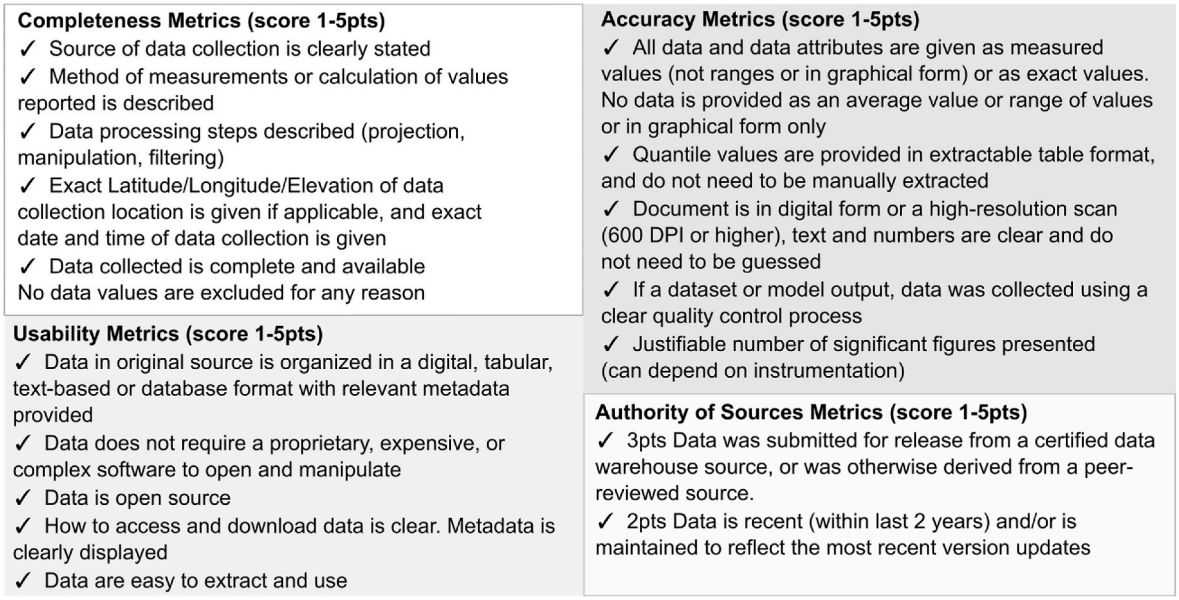

Fig. 9. Data quality assessment metrics and criteria. Each facet of the data quality assessment method aids the producer and user of the data in understanding the value of a dataset.

resource based on the frequency of the terms contained within the text. Location specific metadata tags were added from the named entity recognition geotagging.

To support interoperability with other non-EDX data products, data resources within EDX and GeoCube have an API reference code produced when published, available upon request from EDX support. This enables data users to sync GeoCube data resource updates with outside data products. In addition, users can upload their own data resources to GeoCube from a local drive to display with available data collections.

2.5. Data quality assessment

Understanding the quality of data is essential to the reliability of scientific predictions. Data quality assessment during database implementation is essential to future database usability. Due to the disparate nature of the datasets included in this analysis, the quality control process had to be multifaceted, integrating past quality control metrics into a data quality assessment method to analyze geologic and contextual data in both spatial and text-based formats.

A five-point value was determined for each data resource based on the average score of four data quality metrics: completeness, accuracy, usability, authority of source (Fig. 9). Completeness scores were assigned based on the presence or absence of desired attributes, including complete reporting of data values and metadata. Accuracy scores were dependent on how values are reported, the data format and resolution, quality control processes, and reporting of significant figures. Usability scores were dependent on the format of data in relation to the ability to download, open, understand, and apply the data. Authority of source scores were dependent on the accountability of a dataset to have been reported to the standards of the community in which it is going to be applied. The metrics allow the data producer and user to understand the reliability of data for use and reuse purposes, and guide users to evaluate data useability.

For the metrics of completeness, accuracy, and usability, if specific defined attributes from each metric are missing (Fig. 9), a point is removed, or if the attribute is incomplete, then partial points are removed. To score for the authority of source metrics, one point is removed if the data was submitted to a non-certified data warehouse, two points removed if submitted to a website, half a point removed if produced and/or updated in last 2–5 years, and a point removed if produced more than five years ago without having been updated.

2.6. Descriptive summary analytics

After the completion of the workflow, a spatio-temporal analysis was completed on the combined resources from NATCARB, NRAP and the carbon storage open databases to summarize and communicate the density of publicly available data resources, their spatial coverage, and data quality. These descriptive summary analytics were generated using the Cumulative Spatial Impact Layers (CSIL) tool developed at NETL for ArcGIS (Romeo et al., 2019a,b). The CSIL tool uses geographic data (such as shapefiles) and summarizes the number of data features present within defined size grids of an overall area. Spatial data layers were analyzed for data density and average data quality using CSIL.

3. Results and discussion

3.1. U.S. DOE carbon storage data collections and enhancements in spatial search capabilities

The NATCARB database, Carbon Storage Open Database, and NRAP Community Datasets CCS Site Catalog are publicly available on EDX, and spatial data layers were assimilated into GeoCube. As of early 2021, 13 tools, 148 datasets, and 1996 text-based resources have been published publicly through the Carbon Storage Open Database and by the regional carbon sequestration partnerships, ("Carbon Storage Open Database Group,") and 380 spatial data layers have been published to GeoCube from the NATCARB database and the Carbon Storage Open

Database. An additional 70 spatial data layers have been collected from open sources to support the NRAP catalog database. This corpus continues to grow and evolve as new resources are added to EDX and GeoCube. Data assimilated into the Carbon Storage Open Database increased the overall number of resources available, both text-based and spatial (Fig. 10).

The National Carbon Sequestration Database (NATCARB) offers a collection of geographic information system (GIS) data, curated from each regional carbon sequestration partnership (Fig. 1) (Rodosta et al., 2017) and other sources, to support the evaluation of CCS potential across the U.S. Resources in NATCARB are used to support the publication of the National Carbon Storage Atlas (Carbon Storage Atlas, 2015) and are hosted on EDX's GeoCube (Barkhurst et al., 2018; Bauer et al., 2018), an online spatial data web mapping application, for broad accessibility and usability. In GeoCube, NATCARB serves as an online tool to track CO_2 emitters and carbon storage reservoir potential throughout the partnership regions (Fig. 1) in saline basins, post-production oil and gas fields, and unmineable coal seams. NATCARB consists of geologic, contextual, and capacity calculation datasets required for CCS site screening. In addition to the NATCARB spatial dataset, the regional carbon sequestration partnerships also publish data on EDX including wellbore logs, geophysical data, geologic data, and oil and gas field production information. From the data, lithological subsurface information can be derived, such as depths, porosity, permeability, ground water aquifer information, fault and fracture data, and subsurface reservoir depths. NATCARB data consists of 519 MB of carbon storage spatial resources information: saline basins, unminable coal seams, oil and gas reservoirs, CO_2 point sources information, and 10 km square grids describing storage information for coal and saline basins including porosity, permeability, thickness, depth, and storage estimations.

The National Risk Assessment Partnership (NRAP) is an initiative from the U.S. DOE Office of Fossil Energy that aims to develop new methods and tools for science-based prediction of the environmental risks associated with long-term carbon storage (Rodosta et al., 2017). This multi-institutional effort, led by the National Energy Technology Laboratory, leverages U.S. DOE's capabilities of quantifying risks and uncertainties in engineered-natural systems to quantitatively evaluate geologic carbon storage risks, such as the release of CO_2 or brine from a storage reservoir and potential ground motion impacts due to high volume fluid injection (Fig. 4). A key outcome of NRAP has been development of a suite of tools for modeling different components of the carbon storage subsurface containment system (e.g., reservoir, seals, wells, aquifers) and associated risks, at both individual component and full storage-systems scales. Underpinning the NRAP tools, models, and scientific foundations are numerous data resources, spanning geologic, geochemical, geophysical, and engineering disciplines at various spatial scales. These resources consist of publicly available data, including those from NATCARB, the regional carbon sequestration partnerships, U.S. DOE-FE oil/gas and coal research efforts, and other federal and state government datasets, as well as proprietary and synthetic datasets used for in-house tool and method development and validation. Resources for the NRAP data catalog consist of 20 field demonstration and synthetic CCS injection sites cataloging 277 total records, 70 of which are spatial records, and 19 of which are model simulation results.

The Carbon Storage Open Database is a data collection from the regional carbon sequestration partnership public websites and ArcREST servers and consists of 315 spatial data layers consisting of 2.8 GB and 1848 text-based resources consisting of 8.9 GB of data. Spatial data layers are available in GeoCube (Morkner et al., 2020) and text-based resources have been published on EDX.

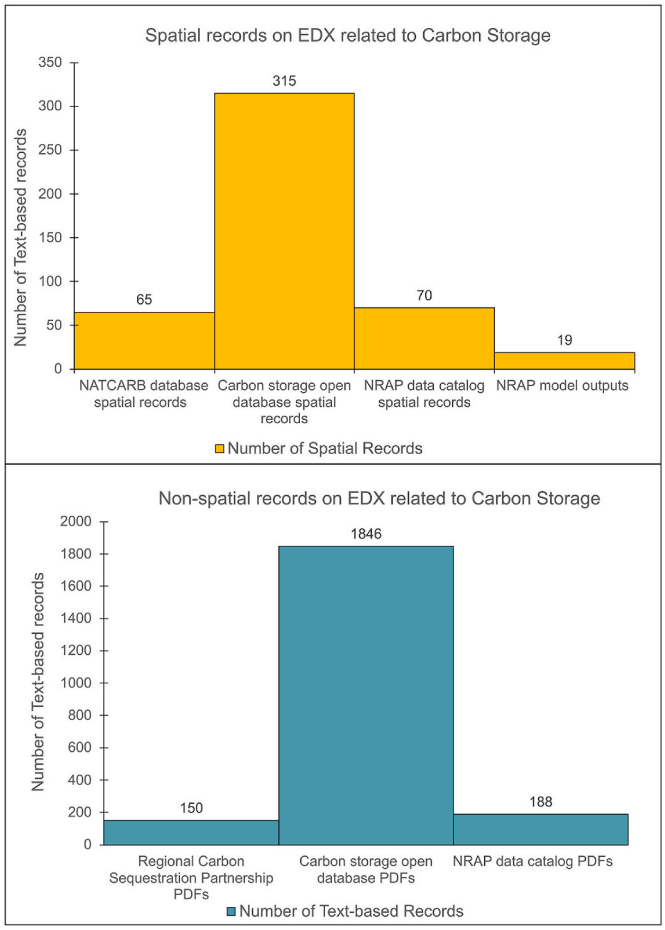

Fig. 10. Bar graph representing the total counts of public resources based on database, and type (spatial or text-based). Spatial (GeoCube and spatial data developed from PDFs and model outputs) and non-spatial PDF record numbers displayed by database, either from NATCARB, the Carbon Storage Open Database, or NRAP.

In addition to the NATCARB database, the NRAP catalog database and the Carbon Storage Open Database, as of early 2021 the regional carbon sequestration partnerships have publicly published 879 resources totaling 633 GB to EDX. All public carbon storage resources on EDX have been curated into the Carbon Storage Open Database group ("Carbon Storage Open Database Group,"), to aid researchers in efficiently finding CCS data.

A spatial search capability was developed for GeoCube (Fig. 11) to improve visibility, discoverability and access to spatial datasets in this analysis. By defining an area, a spatial search yields map layers tied to the defined location (Fig. 11). Future updates will allow the user to use the spatial search to also find relevant geotagged documents, model simulations, and datasets within EDX. Future updates will also expand functionality of GeoCube to include searches for resources outside of EDX, such as hyperlinks and databases using spatial and keyword criteria.

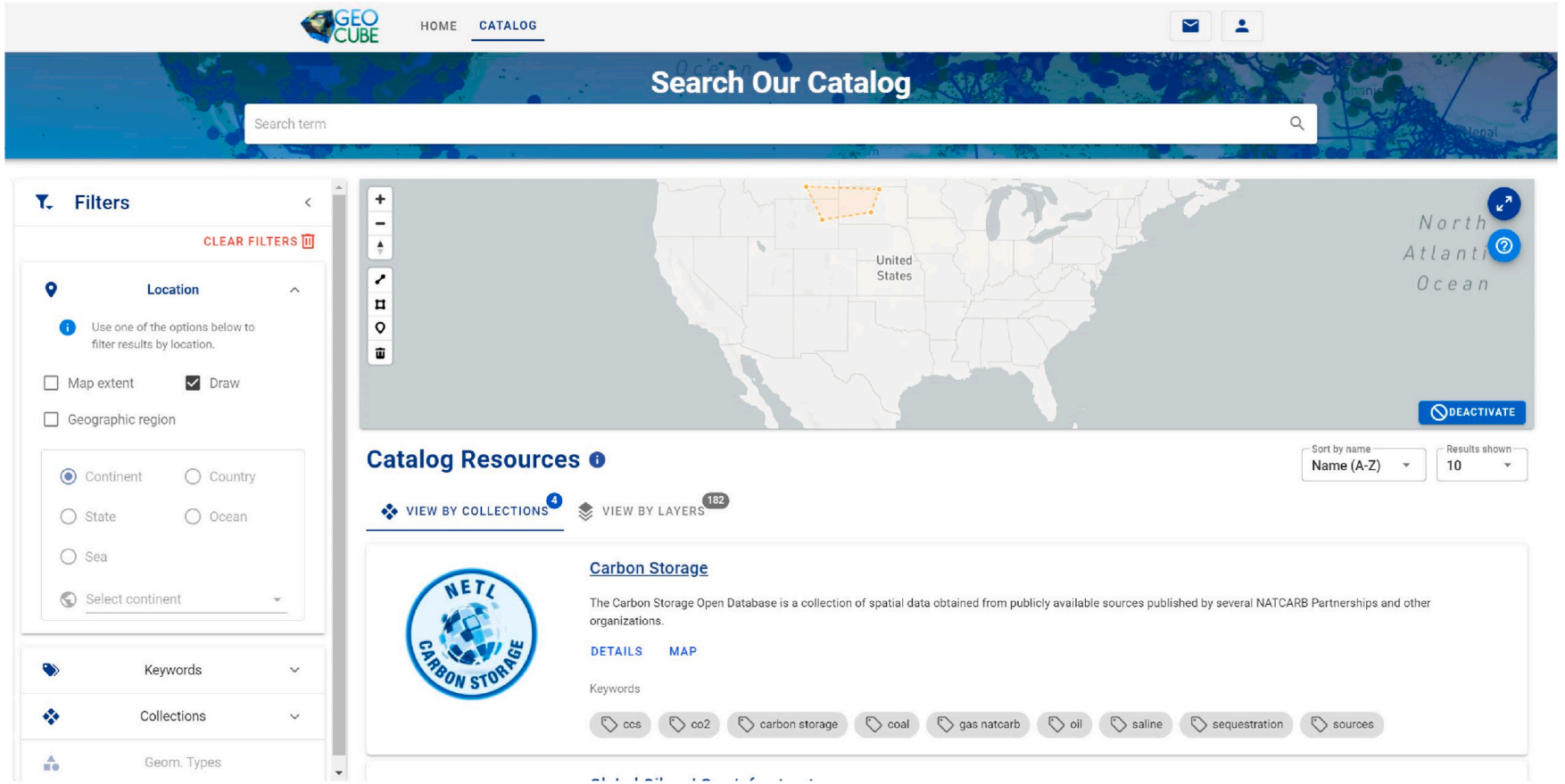

Fig. 11. GeoCube spatial search, Barkhurst et al., (2018), https://edx.netl.doe.gov/geocube/. On map is box drawn by user, and under map is results of search, showing datasets that are associated with the location.

3.2. Natural language processing

NLP produced two separate functions that work with the new search capabilities in GeoCube. Enhanced keyword lists were produced from the top 100 words within each document parsed with NLP. The topic modeling produced groupings of documents based on similar subject matter (topic). The topics were integrated into the resource metadata so the user can search for documents within a specific topic.

The optimal number of topics identified for the LDA model was 9 topics (Fig. 12). The nine topics were assigned the following names based on the most common keywords: 1. Groundwater aquifer and target reservoir classification 2. Petrophysical analysis (well log data), 3. Gas and pore water analysis, 4. Resource production, 5. Geochemistry, 6. Environmental carbon cycling (e.g. biomass), 7. Regional carbon sequestration partnership documents, 8. Field injection tests and monitoring, 9. Site screening and modeling (Fig. 12). Of 2071 total PDF resources parsed by NLP, Fig. 12 describes the number of papers that registered as >1% assignment to each topic. Most documents were assigned >1% of one to four different topics, with few results falling into more than four topics. The final model output results for the corpus

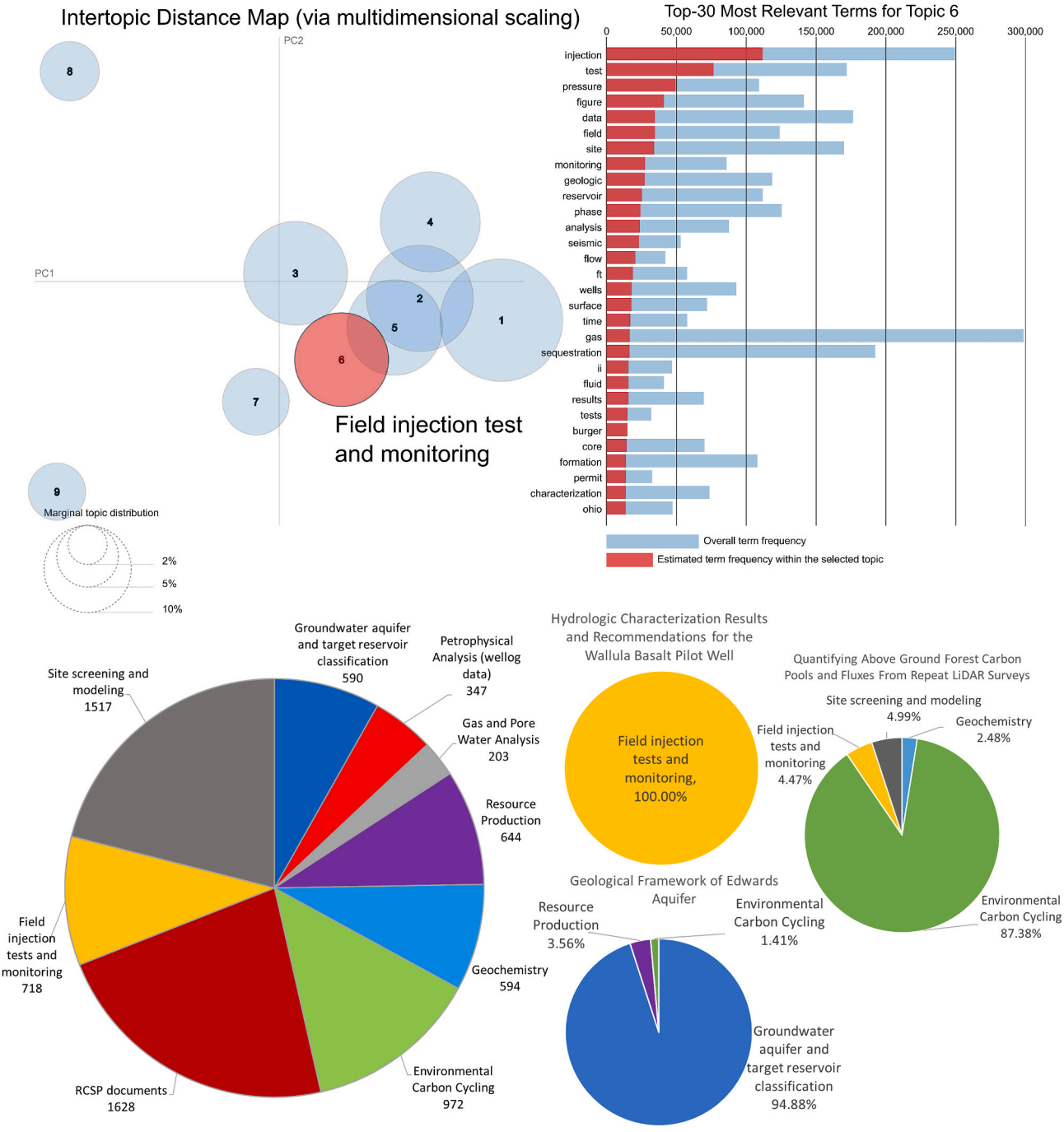

Fig. 12. Summary of the topic groupings produced from the NLP process. Top shows 9-topic LDA model distribution, with left being the intertopic distance model for all nine topics, with topic 6 "Field injection test and monitoring" highlighted in red and top terms frequency for topic 6 shown on the right. The bottom left pie chart shows all topics identified with NLP from 9-topic model with number of documents within corpus registering 1% or more within each topic. Bottom right is examples of document classifications of % each topic for three different documents. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

classified are available in the GitHub repository. The 9-topic model results were used to organize and produce metadata for the collection of 1846 (Fig. 10) resources from the carbon storage open database collection. Documents were grouped by topic, and by geographic region (RCSP), and then assigned keywords based on the terms extracted from the LDA modeling. The collection of documents was published on EDX. The topic modeling and keyword extraction demonstrates how NLP tools can be used for large corpus organization, keyword extraction, and geographic named entity recognition, and presents an opportunity to develop and expand NLP use for data in EDX.

Machine learning will play a fundamental roll in increasing the value of text-based documents in the future. One key difference between spatial and text-based data is the actionability of the data for further use. Currently, text-based documents can be classified, organized, and made searchable through the implementation of NLP. Traditionally, text-based resources require manual extraction or parsing to use data locked in text strings, tables, graphs and images. The growing body of work in machine learning will enable the use of NLP and other forms of machine learning to process text-based documents and convert them into actionable data through automated data (text, tabular, graphical and image) extraction. Such actions will benefit large, text-based, databases such as the open carbon storage database. These nascent machine learning/NLP capabilities continue to grow as the scientific community implements machine learning, maximizing the collection and organization of data in the future, especially in the case of maximizing sparse datasets.

3.3. Descriptive summary analytics

To analyze the spatial distribution and presence of publicly available carbon storage data from NATCARB, NRAP and the Carbon Storage Open Database, the scale of the different data sources must first be considered. NRAP data is, almost exclusively, on a field-based scale (typically within less than a 10 km area) focusing on a specific subregion of the subsurface, such as the target reservoir, groundwater reservoir, caprock/capstone, or some integration of these areas together. NATCARB and the Carbon Storage Open Database data are almost exclusively basin and regional scale. With this difference of scale in mind, spatial distribution and density of data is dependent on the scale. Another important factor to note here is that all numbers presented only capture a moment in time, as additional public resources are continually growing.

Over 380 spatial data layers were used to produce the CSIL (Fig. 13) consisting of 100 km-by-100km grids. Because of the large volume of attributes in some shapefiles, some CSIL grids have over 15000 overlapping attributes. Data density is high in the eastern U.S., central U.S. and in south-

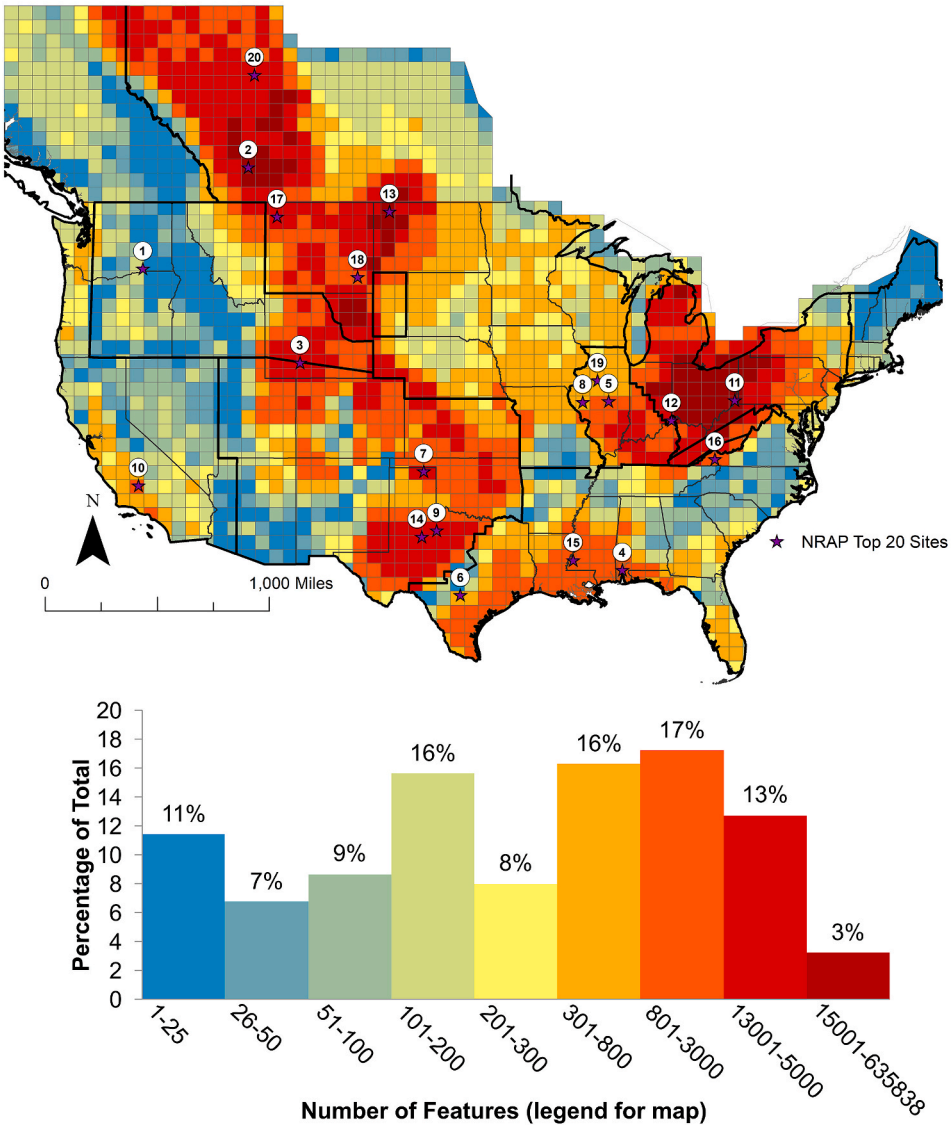

Fig. 13. CSIL heat map showing the number of spatial resources available from the NATCARB atlas 5 v3, the Carbon Storage Open Database, and NRAP Data Catalog and model output resources. Each spatial layer contains a variable number of attributes, which are summarized by the CSIL. The CSIL consists of 100 km-by-100km square grid, and the histogram describes the distribution of the number of attributes per grid cell. NRAP top 20 sites: 1. Big Sky Validation Phase - Wallula Basalt Pilot Project. 2. CaMI - Field Research Station. 3. CarbonSAFE - Wyoming. 4. Citronelle (SECARB). 5. Decatur. 6. Edwards Aquifer. 7. Farnsworth - Anadarko Basin. 8. FutureGen. 9. High Plains Aquifer. 10. Kimberlina (WESTCARB). 11. Appalachian Basin Test (MRCSP). 12. Cincinnati Arch Test (MRCSP). 13. Williston Basin Oil Field Test (PCOR). 14. Scurry Area Canyon Reef Operations. 15. Cranfield Site (SECARB). 16. Central Appalachian Basin Test (SECARB). 17. Kevin Dome. 18. Bell Creek. 19. CarbonSAFE Illinois Macon County. 20. Quest Canada.

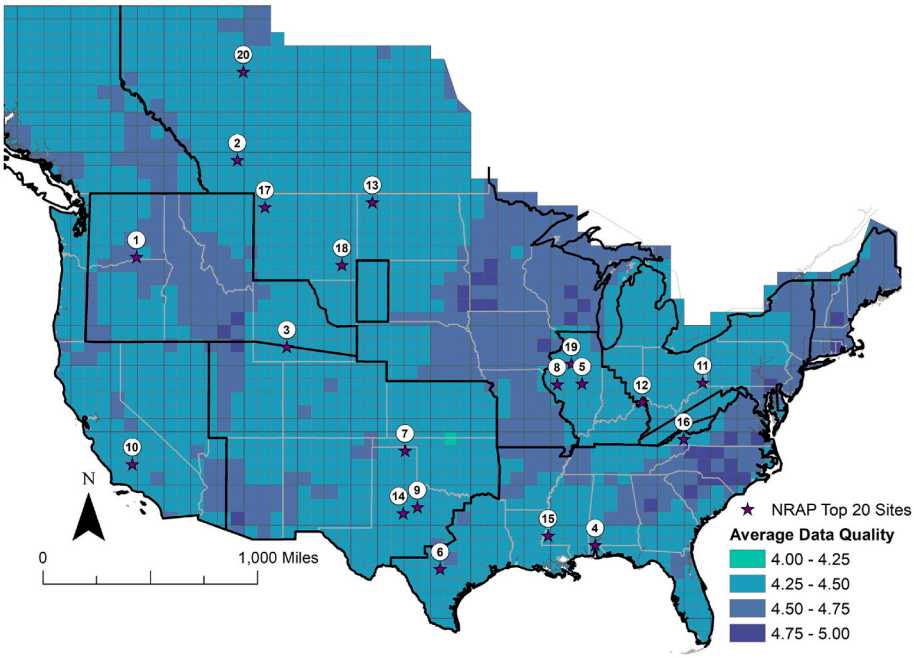

Fig. 14. Heat map of the average data quality for 100 km-by-100km grid of U.S. based on the spatial data from NATCARB and the Carbon Storage Open Database. Regional carbon sequestration partnership (RCSP) boundaries are outlined in black, refer to Fig. 1 for exact RCSPs. NRAP top 20 sites: 1. Big Sky Validation Phase - Wallula Basalt Pilot Project. 2. CaMI - Field Research Station. 3. CarbonSAFE - Wyoming. 4. Citronelle (SECARB). 5. Decatur. 6. Edwards Aquifer. 7. Farnsworth - Anadarko Basin. 8. FutureGen. 9. High Plains Aquifer. 10. Kimberlina (WESTCARB). 11. Appalachian Basin Test (MRCSP). 12. Cincinnati Arch Test (MRCSP). 13. Williston Basin Oil Field Test (PCOR). 14. Scurry Area Canyon Reef Operations. 15. Cranfield Site (SECARB). 16. Central Appalachian Basin Test (SECARB). 17. Kevin Dome. 18. Bell Creek. 19. CarbonSAFE Illinois Macon County. 20. Quest Canada.

central Canada. In the eastern US, data density is highest in the Appalachian Basin (West Virginia, Kentucky, Pennsylvania, Ohio, Tennessee), and in the states of Illinois and Michigan. Data is concentrated especially around three CCS site: The Appalachian Basin Test, Cincinnati Arch Test, and the Central Appalachian Basin Test (Figs. 13 and 1).

In the central U.S. and south-central Canada, data density is highest within Nebraska, South Dakota, North Dakota, and Canada. Data is especially concentrated around the Containment and Monitoring Institute Field research site (CaMI) in southern Canada, and around the Williston Basin Oil Field Test Site in North Dakota. In the southern U.S. data is concentrated within the Permian Basin and Wyoming CarbonSAFE field site. Concentration of data in the Permian Basin is likely high because of oil and gas drilling operations and the presence of multiple CCS projects including Farnsworth-Anadarko site, High Planes Aquifer site, and the Scurry Area Canyon Reef operations site. Data density is high around the coast of the Gulf of Mexico, correlating to Edwards Aquifer, Cranfield, and Citronelle field sites.

Assimilation of disparate datasets offers an improved understanding of carbon storage data across the U.S. and Canada; highlighting areas where a robust catalog of data exists to assist in carbon storage applications and others, such as geothermal, waste-water injection, resource exploration, and basin modeling. The majority of higher data density areas correlate with CCS test injection sites and their general proximity. Anomalous locations of data density are present in central Minnesota and Southern California. These data hot spots correspond with the Kimberlina oil field site in southern California that has been the focus of additional, advanced synthetic modeling for subsurface analysis and risk assessments. The grid cell demonstrating high data density in Minnesota is due to the high data reporting of regional data, even though no major CCS sites are located there (Figs. 1 and 13).

Though most of the United States and central Canada has moderate to high data density - greater than 300 geospatial layer attributes - the western US is generally data sparse. WESTCARB, the partnership region containing most of the western US territory, ended prior to the implementation of good data management and curation practices and only a portion of their data was still accessible when this curation and integration effort was initiated. The challenges experienced with WESTCARB data access and recovery, exemplifies the now recognized need for robust collection and curation of data, particularly for when an individual project is part of a large data program, such as NATCARB.

3.4. Data quality

Data density alone does not define the quality of a dataset. Characterizing spatial data quality is important as it provides the means to analyze the completeness, accuracy, usability and authority of the source of the data relative to the data density (Rose et al., 2018).

The average data quality is generally high for all resources (Fig. 14), with each 100 km grid cell averaging 4-5 pts out of a 5 pts in the data quality ranking system, based on the data quality metrics defined in Fig. 9. The high spatial data quality is due to the high authority of source (RCSP provided, hosted on a reputable data warehouse, and regularly being updated with new data), high usability (spatial data formats), high accuracy (metadata provided and methods of data collection understood) and relatively high completeness (few <null> or empty values).

Calculating data quality by hand can be a time-consuming and arduous process. Machine learning/natural language processing can be used to assist in data quality assessment processes in the future. For example, EDX has built data quality analysis included as part of the submission process and metadata building for extreme materials (Wenzlick et al., 2020), and for spatial datasets for the Global Oil and Gas Infrastructure database (Rose et al., 2018). The connection between computing and data quality is a novel occurrence, but one that is fundamental to understanding reliability and usability of datasets.

4. Conclusion

This paper outlines best practices for acquiring and curating carbon storage data, highlighting tools and protocols that ensure carbon storage data are accessible and useful for both modeling and regulatory decision support. Bringing disparate datasets together is a fundamental to creating a cohesive database from big data projects such as NATCARB and NRAP, and for use by researchers, government agencies and industry. Data warehouses conforming to high metadata standards, such as EDX, support the movement within the scientific community to publish high-quality data. In addition to enhanced spatial and key term searchability, EDX is a credible data warehouse that adheres to international data curation and custodianship standards, FAIR (findability, accessibility, interoperability and reuse) data principles (Wilkinson et al., 2016), publishing data with a unique, trackable citation which can be referenced within published literature. The relationship between the

National Energy Technology Laboratory, NATCARB and NRAP makes EDX the ideal platform to continue to collect and display data for both projects. A central location for data curation ensures that as data products aligned to the carbon capture and storage community continue to be released publicly, they can be distilled and integrated with the existing body of knowledge and data presently curated within the EDX platform.

Data discoverability for reuse is fundamental to driving scientific research and insight. CCS studies are more achievable with access to searchable, extractable, high quality data. Tools for data curation enhance access to data necessary for scientific predictions needed to implement CCS at a large scale and enhance access to data for a variety of subsurface and surface modeling applications.

The horizon of data management is forever evolving, moving towards wide-spread implementation of machine learning to aid in data collection and curation. Data drives analytics and modeling, and tools that leverage computing and artificial intelligence capabilities accelerate the usability and curation of big data projects. The tools described in this research – GeoCube, natural language processing, geotagging, automated data quality analysis – all leverage machine learning capabilities, applicable in a wide variety of subsurface applications. Supporting a shift in time spent on the base of the data pyramid collecting, labeling, and transforming data (Fig. 2) to the top of the data pyramid, where data is analyzed, optimized and used to inform, will drive greater scientific insights in the entirety of the subsurface geologic data community.

Disclaimer

This project was funded by the United States Department of Energy, National Energy Technology Laboratory, in part, through a site support contract. Neither the United States Government nor any agency thereof, nor any of their employees, nor the support contractor, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Computer code availability

Natural Language Processing code is available on GitHub. “Carbon-Storage-Distillation-NLP—Release”, (GPL-3.0 License) developed by Michael Sabbatino, David Yeates, Samuel Walker, and Randall Greenburg, at the National Energy Technology Laboratory, Albany, Oregon. 1450 SW Queen Ave, Albany, OR, 97321. Point of contact: Michael Sabbatino, Michael.Sabbatino@netl.doe.gov, 541-967-5960 (office). First available in 2020, requires a PC, requires a platform to run python (such as Anaconda), as code is written in python. Can be accessed through GitHub repository: https://github.com/NETL-RIC/Carbon-Storage-Distillation-NLP–Release.

Cumulative Spatial Impact Layers, available on the Energy Data eXchange, is an ArcGIS plug-in for download. Developed by Lucy Romeo, Patrick Wingo, Jake Nelson, Jennifer Bauer and Kelly Rose, at the National Energy Technology Laboratory, Albany, Oregon. 1450 SW Queen Ave, Albany, OR, 97321. POC: Lucy Romeo, Lucy.Romeo@netl.doe.gov, 304-285-4545 (office). First released in 2019. Requires ArcGIS software for use. Toolbox can be downloaded at: https://edx.netl.doe.gov/dataset/cumulative-spatial-impact-layers.

Data availability

Data for NATCARB is available https://edx.netl.doe.gov/dataset/natcarb. Data from the carbon storage open database is available https://edx.netl.doe.gov/dataset/carbon-storage-open-database. NRAP data is available https://edx.netl.doe.gov/group/nrap as well as from RCSPs https://edx.netl.doe.gov/group/rcsp-big-sky, https://edx.netl.doe.gov/group/rcsp-mgsc, https://edx.netl.doe.gov/group/rcsp-mrcsp, https://edx.netl.doe.gov/group/rcsp-pcor, https://edx.netl.doe.gov/group/rcsp-secarb, https://edx.netl.doe.gov/group/rcsp-swp, and https://edx.netl.doe.gov/group/rcsp-westcarb. Most carbon storage data mentioned in this can be found in the Carbon Storage Open Database Group on EDX: https://edx.netl.doe.gov/group/carbon-storage-open-database.

Authorship statement

Paige Morkner is responsible for conceptualizing and refining research ideas, data collection, interpretation and analysis, drafting the manuscript, and editing. Jennifer Bauer is the principle investigator for NATCARB database management, and is responsible for conceptualizing and refining research ideas, data collection and preparation, and editing. C. Gabriel Creason is responsible for conceptualizing and refining research ideas, data collection and preparation, drafting, and editing. Michael Sabbatino is responsible for developing and implementing python scripts for collection of data and metadata, interpretation of results, drafting, and editing. Patrick Wingo is responsible for developing and implementing python scripts for data and metadata collection, development of figures, reviewing, and editing. Randall Greenburg is responsible for developing and implementing python scripts for natural language processing, data collection, and reviewing. Samuel Walker and David Yeates are responsible for development of python scripts for natural language processing, data collection, and interpretation of analysis. Kelly Rose is the principle investigator and lead senior scientist on the project implementing natural language processing and machine learning tools for carbon storage research, and is responsible for the research concept development, conceptualizing and refining research ideas, reviewing and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to thank a number of collaborators and colleagues who provided insights and support for the work represented in this manuscript. These include, Andrew Bean, Aaron Barkhurst, Na Hyung Choi, Jennifer DiGiulio, Brad Gooch, Jason Guinan (NETL MultiMedia), Katherine Jones, Lucy Romeo, Chad Rowan, Alexander Tong, and Grant Zoch. This work was performed in support of the US Department of Energy’s Fossil Energy and Carbon Management’s Carbon Storage Program.

References

Accelerating Breakthrough Innovation in Carbon Capture, Utilization, and Storage, 2017. Report of the Mission Innovation Carbon Capture, Utilization, and Storage Experts’ Workshop. US Department of Energy. https://www.energy.gov/fecm/downloads/accelerating-breakthrough-innovation-carbon-capture-utilization-and-storage.

Alcalde, J., Flude, S., Wilkinson, M., Johnson, G., Edlmann, K., Bond, C.E., Scott, V., Gilfillan, S.M.V., Ogaya, X., Haszeldine, R.S., 2018. Estimating geological CO2 storage security to deliver on climate mitigation. Nat. Commun. 9, 2201. https://doi.org/10.1038/s41467-018-04423-1.

Baker, D.V., Kelly, R., Bauer, J., Rager, D., 2016. Computational advances and data analytics to reduce subsurface uncertainty. In: Presented at the 50th U.S. Rock Mechanics/Geomechanics Symposium. American Rock Mechanics Association.
Barkhurst, A., Bauer, J., Rose, K., Chittum, J., Rowan, C., Romeo, L., 2018. GeoCube. https://doi.org/10.18141/1471973.
Bauer, J., Rowan, C., Barkhurst, A., Digiulio, J., Jones, K., Sabbatino, M., Rose, K., Wingo, P., 2018. National Carbon Sequestration Database (NATCARB). https://doi.org/10.18141/1474110.
Carbon Storage Atlas, fifth ed., 2015. National Energy Technology Laboratory.
Carbon Storage Open Database Group, 2020. Energy Data eXchange. https://edx.netl.doe.gov/group/carbon-storage-open-database. accessed 12.8.20.
Conway, E., Pepler, S., Garland, W., Hooper, D., Marelli, F., Liberti, L., Piervitali, E., Molch, K., Glaves, H., Badiali, L., 2013. Ensuring the long term impact of earth science data through data curation and preservation. ISQ 25, 28. https://doi.org/10.3789/isqv25no3.2013.05.
Cuéllar-Franca, R.M., Azapagic, A., 2015. Carbon capture, storage and utilisation technologies: a critical analysis and comparison of their life cycle environmental impacts. Journal of CO2 Utilization 9, 82–102. https://doi.org/10.1016/j.jcou.2014.12.001.
2016 Data Science Report - CrowdFlower, 2016..
DataCite metadata schema documentation for the publication and citation of research data v4.3, 2019. 73 pages. https://doi.org/10.14454/7XQ3-ZF69.
DiGiulio, J., Creason, G., Morkner, P., Sabbatino, M., Wentworth, A., Jones, K., Cameron, E., Bean, A., Rose, K., 2020. NRAP Community Datasets CCS Site Catalog.
DOE Data Crosscutting Requirements Review, 2013. U.S. Department of Energy.
Faunt, C.C., Sneed, M., Traum, J., Brandt, J.T., 2016. Water availability and land subsidence in the Central Valley, California, USA. Hydrogeology Journal 24, 675–684. https://doi.org/10.1007/s10040-015-1339-x.
Furre, A.-K., Eiken, O., Alnes, H., Vevatne, J.N., Kiær, A.F., 2017. 20 Years of monitoring CO2-injection at Sleipner. Energy Procedia 114, 3916–3926. https://doi.org/10.1016/j.egypro.2017.03.1523.
Goodman, A., Hakala, A., Bromhal, G., Deel, D., Rodosta, T., Frailey, S., Small, M., Allen, D., Romanov, V., Fazio, J., Huerta, N., McIntyre, D., Kutchko, B., Guthrie, G., 2011. U.S. DOE methodology for the development of geologic storage potential for carbon dioxide at the national and regional scale. Int. J. Greenh. Gas Contr. 5, 952–965. https://doi.org/10.1016/j.ijggc.2011.03.010.
Harp, D.R., Oldenburg, C.M., Pawar, R., 2019. A metric for evaluating conformance robustness during geologic CO2 sequestration operations. Int. J. Greenh. Gas Contr. 85, 100–108. https://doi.org/10.1016/j.ijggc.2019.03.023.
Haszeldine, R.S., Flude, S., Johnson, G., Scott, V., 2018. Negative emissions technologies and carbon capture and storage to achieve the Paris Agreement commitments. https://doi.org/10.6084/M9.FIGSHARE.C.4009930.
Hey, T., Tansley, S., Tolle, K., 2009Hey (Ed.), The Fourth Paradigm: Data-Intensive Scientific Discovery, vol. 287. Microsoft. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.
Justman, D.M., Romeo, L.F., Barkhurst, A.A., Tucker, D., Bauer, J., Duran, R., Dyer, A.S., Nelson, J.R., Sabbatino, M., Wingo, P., Wenzlick, M.Z., Zaengle, D., Rose, K., 2020. Advanced Geospatial Analytics and Machine Learning for Offshore and Onshore Oil & Natural Gas Infrastructure (No. RSS579). National Energy Technology Laboratory (NETL), Pittsburgh, PA, Morgantown, WV, and Albany, OR (United States).
Mackenzie, W., 2021. How to Scale Up Carbon Capture and Storage. Forbes. https://www.forbes.com/sites/woodmackenzie/2021/07/08/how-to-scale-up-carbon-capture-and-storage/. accessed 8.12.21.
Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekci, O., Yu, R., Zhou, B., 2021. IPCC, 2021: Climate Change 2021: the Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Morkner, P., Creason, C.G., Sabbatino, M., Wingo, P., DiGiulio, J., Jones, K., Greenberg, R., Bauer, J., Rose, K., 2020. Carbon Storage Open Database. https://doi.org/10.18141/1671320.
Page, B., Turan, G., Zapantis, A., 2020. Global Status of CCS 2020. Global CCS Institute.
Paris Climate Agreement, 2015.
Quinn, N., Wainwright, H., Jordan, P., Zhou, Q., Berkholzer, J., 2013. Potential Impacts of Future Geological Storage of CO2 on the Groundwater Resources in California's Central Valley (No. CEC-500-2014-028). Lawrence Berkeley National Laboratory.
Rodosta, T.D., Litynski, J.T., Plasynski, S.I., Hickman, S., Frailey, S., Myer, L., 2011. U.S. Department of energy's site screening, site selection, and initial characterization for storage of CO2 in deep geological formations. Energy Procedia 4, 4664–4671. https://doi.org/10.1016/j.egypro.2011.02.427, 10th International Conference on Greenhouse Gas Control Technologies.
Rodosta, T., Bromhal, G., Damiani, D., 2017. U.S. DOE/NETL carbon storage program: advancing science and technology to support commercial deployment. In: Energy Procedia, 13th International Conference on Greenhouse Gas Control Technologies, GHGT-13, 14-18 November 2016, Lausanne, Switzerland 114, pp. 5933–5947. https://doi.org/10.1016/j.egypro.2017.03.1730.
Romeo, Lucy, Nelson, Jake, Wingo, Patrick, Bauer, Jennifer, Justman, Devin, Rose, Kelly, 2019. Cumulative spatial impact layers: A novel multivariate spatio-temporal analytical summarization tool. Transactions in GIS 23 (5). https://doi.org/10.1111/tgis.12558.
Romeo, Lucy, Wingo, Patrick, Nelson, Jake, Bauer, Jennifer, Rose, Kelly, 2019. Cumulative Spatial Impact LayersTM. Energy Data eXchange. https://doi.org/10.18141/1491843. https://edx.netl.doe.gov/dataset/cumulative-spatial-impact-layers.
Rose, K., Bauer, J.R., Baker, D.V., 2015. Big data geo-analytical tool development for spatial analysis uncertainty visualization and quantification needs. AGU Fall Meeting Abstracts 43, IN43B–1731.
Rose, K., Bauer, J., Baker, V., Bean, A., DiGiulio, J., Jones, K., Justman, D., Miller, R.M., Romeo, L., Sabbatino, M., Tong, A., 2018. Development of an open global oil and gas infrastructure inventory and Geodatabase. https://doi.org/10.18141/1427573.
U.S. Energy-Related Carbon Dioxide Emissions, 2019., U.S. Energy Information Administration.
US EPA, 2015. Sources of Greenhouse Gas Emissions. US EPA. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions. accessed 12.4.20.
USGCRP, 2018. Fourth National Climate Assessment. U.S. Global Change Research Program, Washington, DC.
Wenzlick, M., Bauer, J.R., Rose, K., Hawk, J., Devanathan, R., 2020. Data assessment method to support the development of creep-resistant alloys. Integrating Materials and Manufacturing Innovation 9, 89–102. https://doi.org/10.1007/s40192-020-00167-3.
Wilkinson, M.D., Dumontier, M., Aalbersberg, IjJ., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Santos, L.B. da S., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Schaik, R. van, Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Lei, J. van der, Mulligen, E. van, Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 1–9. https://doi.org/10.1038/sdata.2016.18.