# Integrating interdisciplinary data: The EMERGE Database and its broader lessons for data management best practices

Suzanne Hodgkins[1], Benjamin Bolduc[1], Dustin Miller[2], Virginia Rich[1], and EMERGE Biology Integration Institute[3]

[1]Department of Microbiology, The Ohio State University, Columbus, OH, United States
[2]College of Arts and Sciences Technology Services (ASCTech), The Ohio State University, Columbus, OH, United States
[3]https://emerge-bii.github.io/

April 19, 2024

**Abstract**

In environmental research, cross-disciplinary analyses enable the discovery of novel insights that may not otherwise be evident. Doing these analyses efficiently requires integration of heterogeneous data into a common data structure; however, this type of data integration represents a major challenge, especially for large, multi-institutional projects. Not only should the sharing of individual datasets follow FAIR principles (Findable, Accessible, Interoperable, Reusable), but the ideal data management system should also include a central multidisciplinary data organization framework.

The EMERGE Database (EMERGE-DB; https://emerge-db.asc.ohio-state.edu/) is the central data hub of the EMERGE Biology Integration Institute (NSF award # 2022070), which investigates the changing dynamics of a thawing permafrost ecosystem in Stordalen Mire, northern Sweden. The EMERGE-DB accomplishes the essential tasks of data management (i.e., data storage and sharing), while also offering more advanced functionality to facilitate interdisciplinary collaboration. Data and standardized metadata—including both sample and file metadata—are integrated within a Neo4j graph database, which allows combined datasets from different source files to be obtained via efficient custom queries. A front-end web portal provides access to this data for both the public and for EMERGE project members (who can access non-public data via login), with different pages providing different "views" of the database for different common use cases. Although data are still deposited to external community repositories (e.g. Zenodo, NCBI databases) to ensure cost-effective long-term accessibility, these depositions are tracked within the EMERGE-DB's standardized metadata system, with all internally- and externally-stored datasets displayed within a centralized page on the web portal. Although this data integration and sharing framework is customized for the EMERGE project's needs, many of its guiding principles—such as the centralized web access point for all datasets, and general file formatting standards to streamline the detailed integration of sample metadata—are broadly applicable as "best practices" that other projects can apply in their own data management systems.

# Integrating Interdisciplinary Data: The EMERGE Database and its Broader Lessons for Data Management Best Practices

Suzanne Hodgkins[1] (hodgkins.3@osu.edu), Benjamin Bolduc[1], Dustin Miller[2], Virginia Rich[1], and EMERGE Biology Integration Institute*

(1) Department of Microbiology, The Ohio State University, Columbus, OH, United States; (2) College of Arts and Sciences Technology Services (ASCTech), The Ohio State University, Columbus, OH, United States. *https://emerge-bii.github.io/

## Introduction

- Interdisciplinary research enables the exploration of emergent phenomena, broadening the horizons of scientific discovery.
- To enable different disciplines to effectively "speak" to one another, interdisciplinary research data must be organized, integrated, and shared based on **FAIR principles** (Findable, Accessible, Interoperable, Reusable).
- Interdisciplinary data integration faces several major challenges:
  - Broader scale, more interdisciplinary projects = larger, more numerous, and more heterogeneous datasets.
  - Different disciplines and labs use different terminologies.
  - Multiple levels of data processing, each representing different information "quality"; and these vary across disciplines.

### EMERGE (EMergent Ecosystem Response to ChanGE)

- An in-depth interdisciplinary study of ecosystem-climate feedbacks in the thawing permafrost peatland Stordalen Mire (northern Sweden).
- Builds on a over a decade of work:

*Data heavily leveraged by EMERGE.* IsoGenie

10-year DOE-funded study of permafrost-carbon feedbacks at Stordalen

Archaea 2 Atmosphere

3-year NASA-funded study scaling IsoGenie findings to regional and pan-Arctic levels

… plus numerous other projects, including climate data spanning over a century

The **EMERGE Database (EMERGE-DB)**, the project's central data archive, accomplishes the **essential tasks** of data management:

**Data Storage** on fully-RAIDed OSU server

**Data Sharing** via web portal

while also offering more **advanced functionality** to facilitate interdisciplinary collaboration:

**Data Integration** within one graph framework

**Data Exploration** with complex custom queries

## Data import workflow

EMERGE has a **Data Policy** that further explains this process and guides project members on data submission.

**KEY - Roles & responsibilities:**

Project members / data generators

Data management team

Do science; **generate data**

Name with **standardized sample IDs**, keeping track of name changes

**Submit data** to database team:

**Form-based system** includes prompts to ensure complete metadata

**Validate metadata**

Import data (python)

Once submitted, all source datasets are:

- Assigned basic file metadata in consultation with data generators:
  - title, authors & contact info, version #, quality level, access rights
- Recorded in the EMERGE-DB's Metadata nodes for sharing via the website's Downloads page.
- Shared via external repositories (for publication-ready data).

Submitted sample metadata is standardized for cross-dataset consistency.

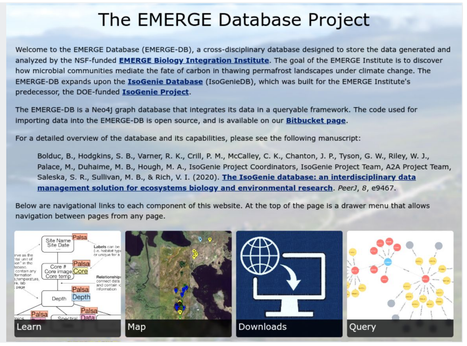
These **standardized properties** (indicated with "___") then guide detailed data import.

| SampleID__ | Site__ | Core__ | Date__ | DepthMin__ | DepthMax__ | Habitat__ | … |
|---|---|---|---|---|---|---|---|
| MainAutochamber.202107_P_3_30to34 | Palsa Autochamber Site | 3 | 2021-07-18 | 30 | 34 | Palsa | |
| MainAutochamber.202107_S_1_1to5 | Sphagnum Autochamber Site | 1 | 2021-07-26 | 1 | 5 | Bog | |
| IncubationMaterial.202107_IncE_2_10to14 | Inc-Eriophorum | 2 | 2021-07-23 | 10 | 14 | Fen | |

## Web Portal

**emerge-db.asc.ohio-state.edu**

- Provides both public and within-project data access.
- Different pages provide different "views" of the data (see screenshots connected from Graph Database below).




## Graph Database

- Connects data thematically in a flexibly-structured network, **mirroring physical or conceptual relationships** between entities.
- Neo4j-powered framework enables efficient custom querying.

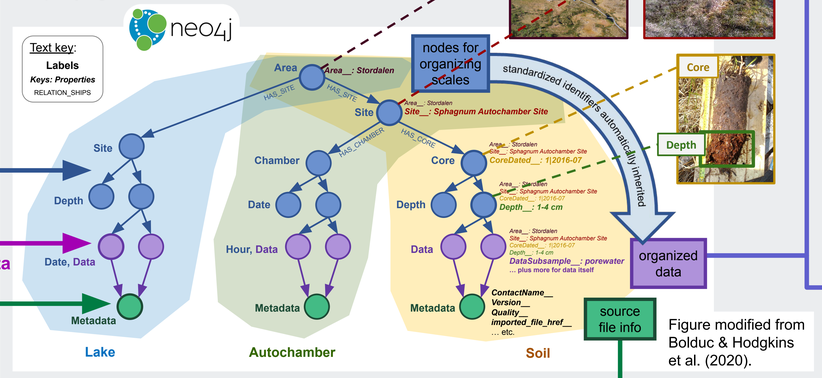
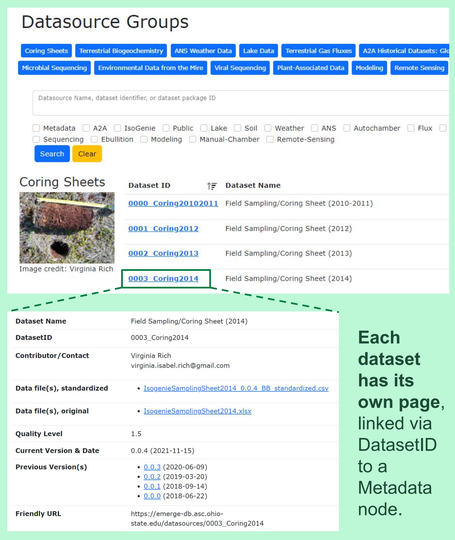



Figure modified from Bolduc & Hodgkins et al. (2020).

auto-population of Downloads page from **Metadata** nodes

## Downloads: Source data files




Each dataset has its own page, linked via DatasetID to a Metadata node.

**Where applicable, Downloads pages include links to community data repositories.**

From EMERGE Data Management Plan:

**"The EMERGE-DB is the hub of the DMP, while public data repositories and partner institutions' internal databases are the spokes."**

GitHub

NCBI National Center for Biotechnology Information

zenodo

ProteomeXchange

Stockholm University | Bolin Centre for Climate Research

## Queries: Retrieve subsets of integrated data

**Cached queries:** Snapshots of single-label query results; available to the public.

There are a number of ways to query and retrieve information from the database. Below you can query based on labels, which are an effective way to group data and sample metadata based on a particular category. From there, you can filter results based on columns or (soon) dynamically.

Query based on labels

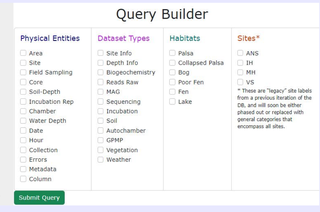
*Labels* are used to organize all the data in the graph database. Here you can filter data based on labels denoting different **physical entities**, **dataset types**, **habitats**, and select **sites**. For most data-related queries, you probably want to use one of the **dataset type** labels.

In the query output tables, each row represents a node, and each column (except for the first "node labels" column, which lists all the labels on each node) represents a property. Properties ending in "___" have been mostly standardized throughout the graph database, and are therefore very useful for harmonizing data from different sources.

Choose one of the following labels *(scroll down for the full list of options*; color codes are given in the explanatory text)*:

| LABEL | DESCRIPTION (mouseover for details) |
|---|---|
| Hour | Hour (any) |
| Collection | Sequence Collections |
| Errors | Standard Errors |
| Metadata | File Metadata |
| Column | File Column Info |
| Site-Info | Site Information (detailed) |
| Depth-Info | Field Coring Data |
| Biogeochemistry | Biogeochemistry Data |
| Reads-Raw | Raw Sequence Reads (links & metadata) |

**Live queries:** Real-time queries on one or more labels. Available to EMERGE members (public access is forthcoming).




**Advanced queries:** Most customizable. Available only to the DB team; output is saved to files which can be posted to the Downloads page.

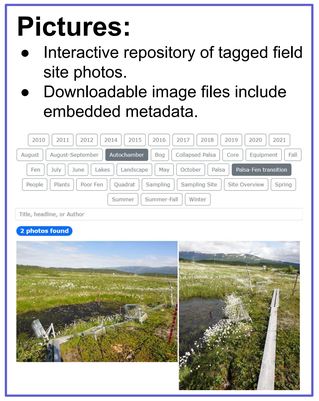
*For example, to retrieve biogeochemistry data from collapsed palsa sites sampled on dates when the maximum air temperature was >25 °C (producing the above graph as a result):*

```
WITH p{(n:`Collapsed Palsa`:BiogeochemistryData)<-[:Soil*0..]-(:Core)<-[:CORE]-(:Area)<-[:HAS_AREA]-(:Weather:Site)}->(x:Date:Data)
WHERE n.Date__ = x.Date__ AND toFloat(x.Maximum_AirTemperature) > 25
RETURN p
```

**Map:** Graphical information on cores and other geo-referenced entities.

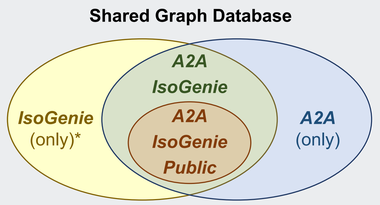

**Pictures:**

- Interactive repository of tagged field site photos.
- Downloadable image files include embedded metadata.




## Managing Data Access

- The EMERGE-DB framework is shared by multiple related projects with both public and private data, necessitating a system for managing data access.
- **Access labels** are assigned to all nodes and used by the website to filter shared data by access rights.




* For historical reasons, **EMERGE** data uses the **IsoGenie** access label.

Figure modified from Bolduc & Hodgkins et al. (2020).

## Conclusions

- Flexible data integration can be balanced with long-term data sharing via:
  - **Metadata standardization workflows** to facilitate data interoperability & reusability for integration.
  - **A central integrated data structure** that can be explored with custom queries.
  - **Links to records in community repositories** for long-term accessibility of the original datasets.
  - **A web portal providing access** to both the integrated data and the versioned original datasets, improving their findability.
- These broadly-applicable lessons for data management best practices from the EMERGE-DB team can provide a roadmap for other interdisciplinary teams building data management systems.

## Reference

- Bolduc, B., Hodgkins, S. B., … & Rich, V. I. (2020). The IsoGenie database: an interdisciplinary data management solution for ecosystems biology and environmental research. *PeerJ*, *8*, e9467. https://doi.org/10.7717/peerj.9467

## Acknowledgments

This research is a contribution of the EMERGE Biology Integration Institute, funded by the National Science Foundation, Biology Integration Institutes Program, Award # 2022070. We thank the Swedish Polar Research Secretariat and SITES for the support of the work done at the Abisko Scientific Research Station. SITES is supported by the Swedish Research Council's grant 4.3-2021-00164. The IsoGenie Project was funded by the Genomic Science Program of the United States Department of Energy Office of Biological and Environmental Research, grants DE-SC0004632, DE-SC0010580, and DE-SC0016440. The A2A Project was funded by the NASA Interdisciplinary Research in Earth Science (IDS) program, grant # NNX17AK10G.