



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



Suzanne Hodgkins¹ (hodgkins.3@osu.edu), Benjamin Bolduc¹, Dustin Miller², Virginia Rich¹, 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.

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

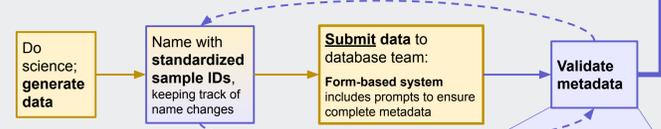
Data Exploration
with complex custom queries

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



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).

The EMERGE Database Project

Welcome to the EMERGE Database (EMERGE-DB), a cross-disciplinary database designed to store the data generated and analyzed by the NSF-funded **EMERGE Biology Integration Institute**. The goal of the EMERGE Institute is to discover how microbial communities mediate the fate of carbon in thawing permafrost landscapes under climate change. The EMERGE-DB expands upon the **IsoGenie Database** (IsoGenieDB), which was built for the EMERGE Institute’s predecessor, the DOE-funded **IsoGenie Project**.

The EMERGE-DB is a Neo4j graph database that integrates its data in a queryable framework. The code used for importing data into the EMERGE-DB is open source, and is available on our **Bitbucket page**.

For a detailed overview of the database and its capabilities, please see the following manuscript:

Bolduc, B., Hodgkins, S. B., Varner, R. K., Crill, P. M., McCalley, C. K., Chanton, J. P., Tyson, G. W., Riley, W. J., Palace, M., Duhalme, M. B., Hough, M. A., IsoGenie Project Coordinators, IsoGenie Project Team, A2A Project Team, Salsieka, S. R., Sullivan, M. B., & Rich, V. I. (2020). *The IsoGenie database: an interdisciplinary data management solution for ecosystems biology and environmental research*. *PeerJ*, 8, e9467.

Below are navigational links to each component of this website. At the top of the page is a drawer menu that allows navigation between pages from any page.

Graph Database

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

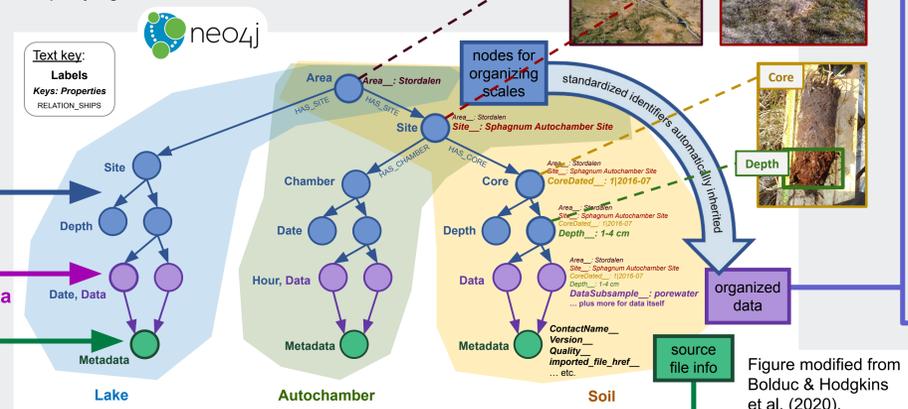


Figure modified from Bolduc & Hodgkins et al. (2020).

sample metadata
measurement data
file metadata
auto-population of Downloads page from Metadata nodes

Downloads: Source data files

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**.”

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

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
Columns	File Column Info
Site-Info	Site Information (details)
Site-Name	Site Name
Biogeochemistry	Biogeochemistry Data
Read-File	Raw Sequence Reads (Links & metadata)

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

Query Builder

Physical Entities: Area, Site, Depth Info, Field Sampling, Core, Soil-Depth, Incubation Rep, Chamber, Water Depth, Date, Hour, Collection, Errors, Metadata, Column

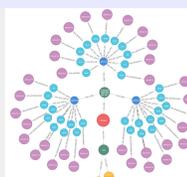
Dataset Types: Site Info, Biogeochemistry, Reads Raw, MAG, Sequencing, Incubation, SPMS, Autochamber, GPRF, Weather

Habitats: Palsa, Collapsed Palsa, Bog, Fen, Lake

Sites*: ANS, SH, MH, VS

*These are “legacy” site labels from a previous iteration of the DB, and will soon be either phased out or replaced with general categories that encompass all sites.

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):

Map:

Graphical information on cores and other geo-referenced entities.

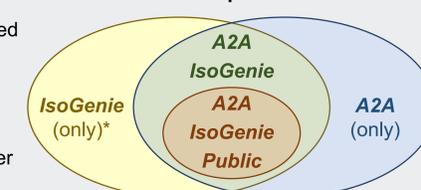
Pictures:

- 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.

Shared Graph Database



* 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.