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.

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

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:

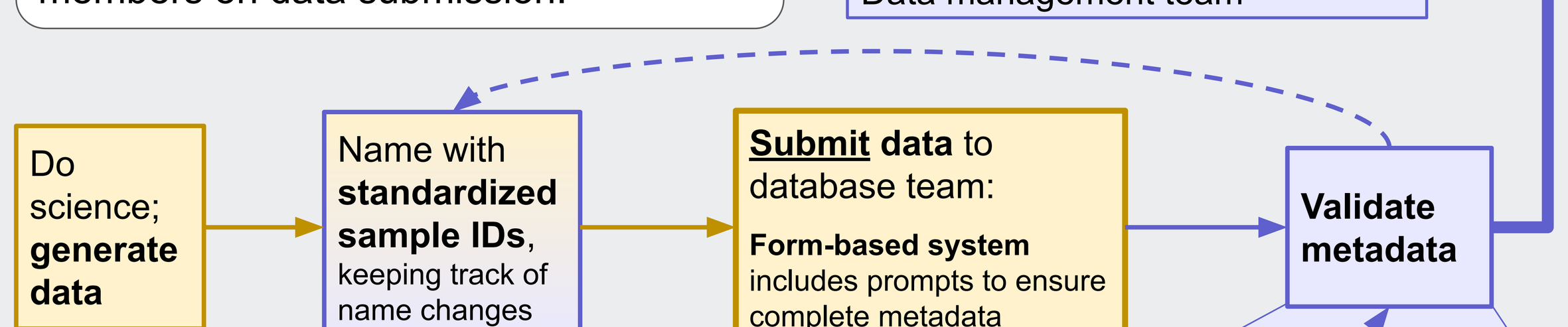
Data Integration within one graph framework

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.

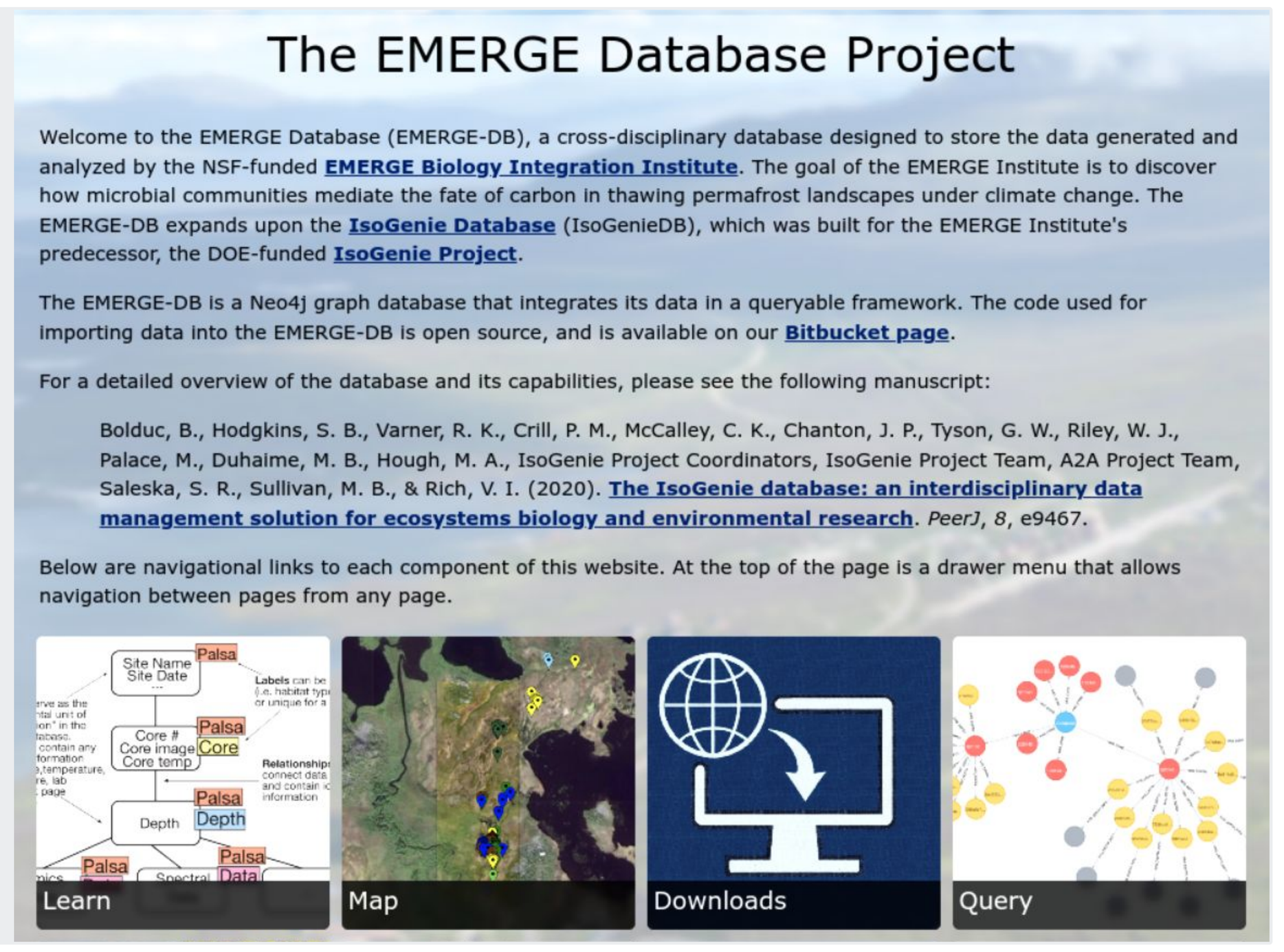
SampleID__	Site__	Core__	Date__	DepthMin__	DepthMax__	Habitat__	…
MainAutochamber.202107_P_3_30to34	Palsa Autochamber Site	3	2021-07-18	30	34	Palsa	
MainAutochamber.202107_S_1_1to5	Sphagnum Autochamber Site	1	2021-07-26	1	5	Bog	
IncubationMaterial.202107_IncE_2_10to14	Inc-Eriophorum	2	2021-07-23	10	14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

Import data (python)

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.

sample metadata

measurement data

file metadata

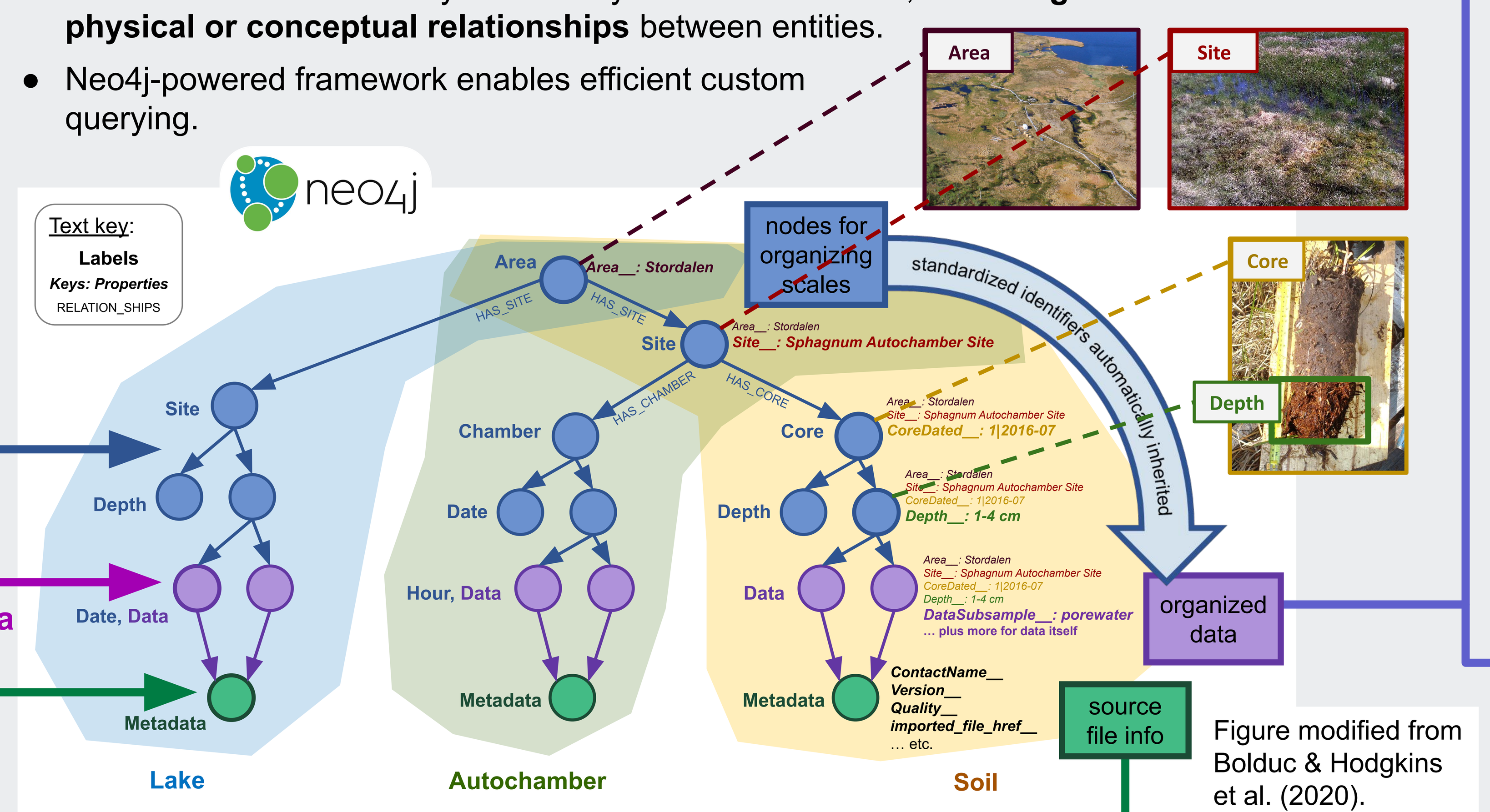


Figure modified from Bolduc & Hodgkins et al. (2020).

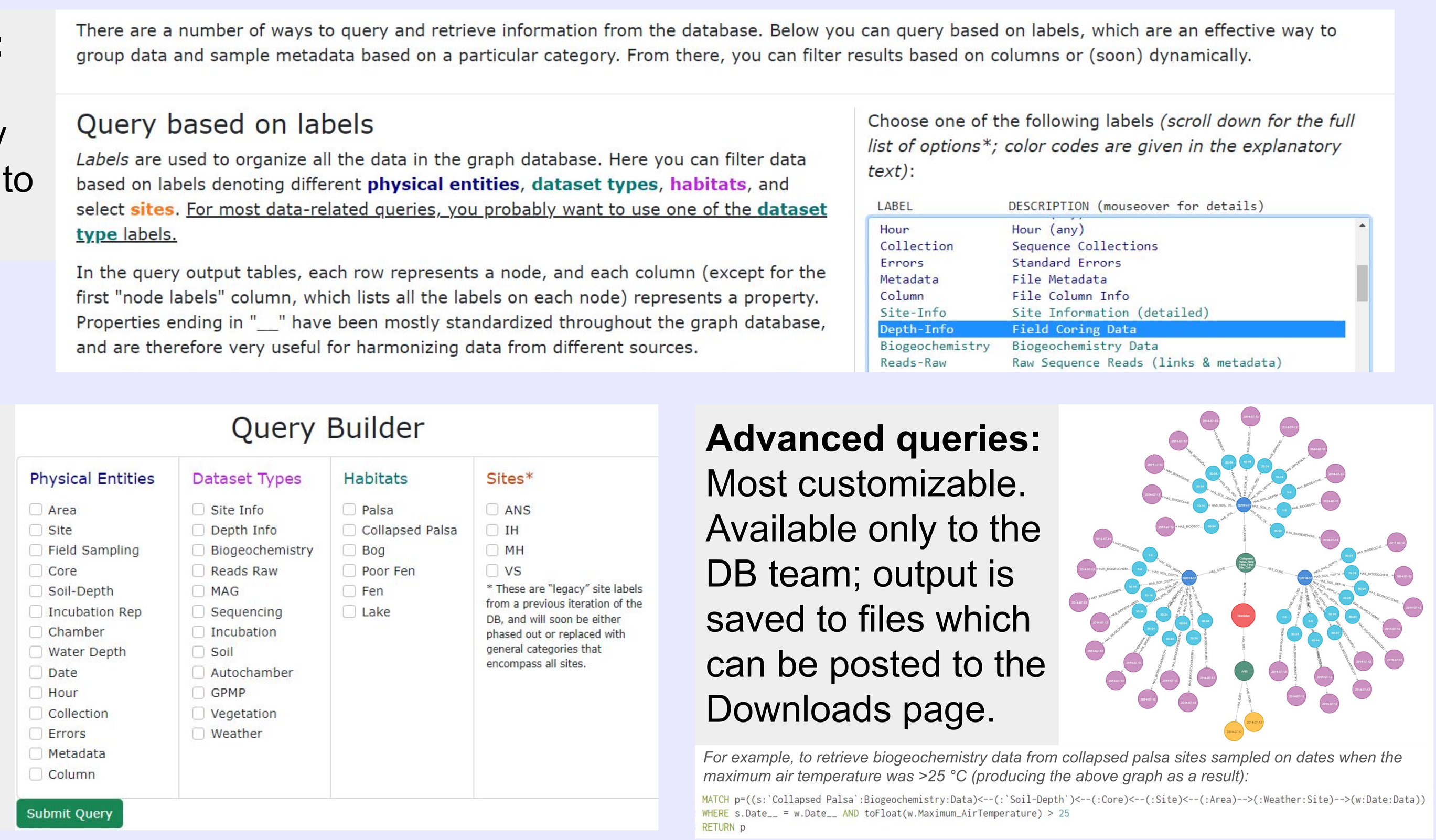
auto-population of Downloads page from **Metadata** nodes

Queries: Retrieve subsets of integrated data

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

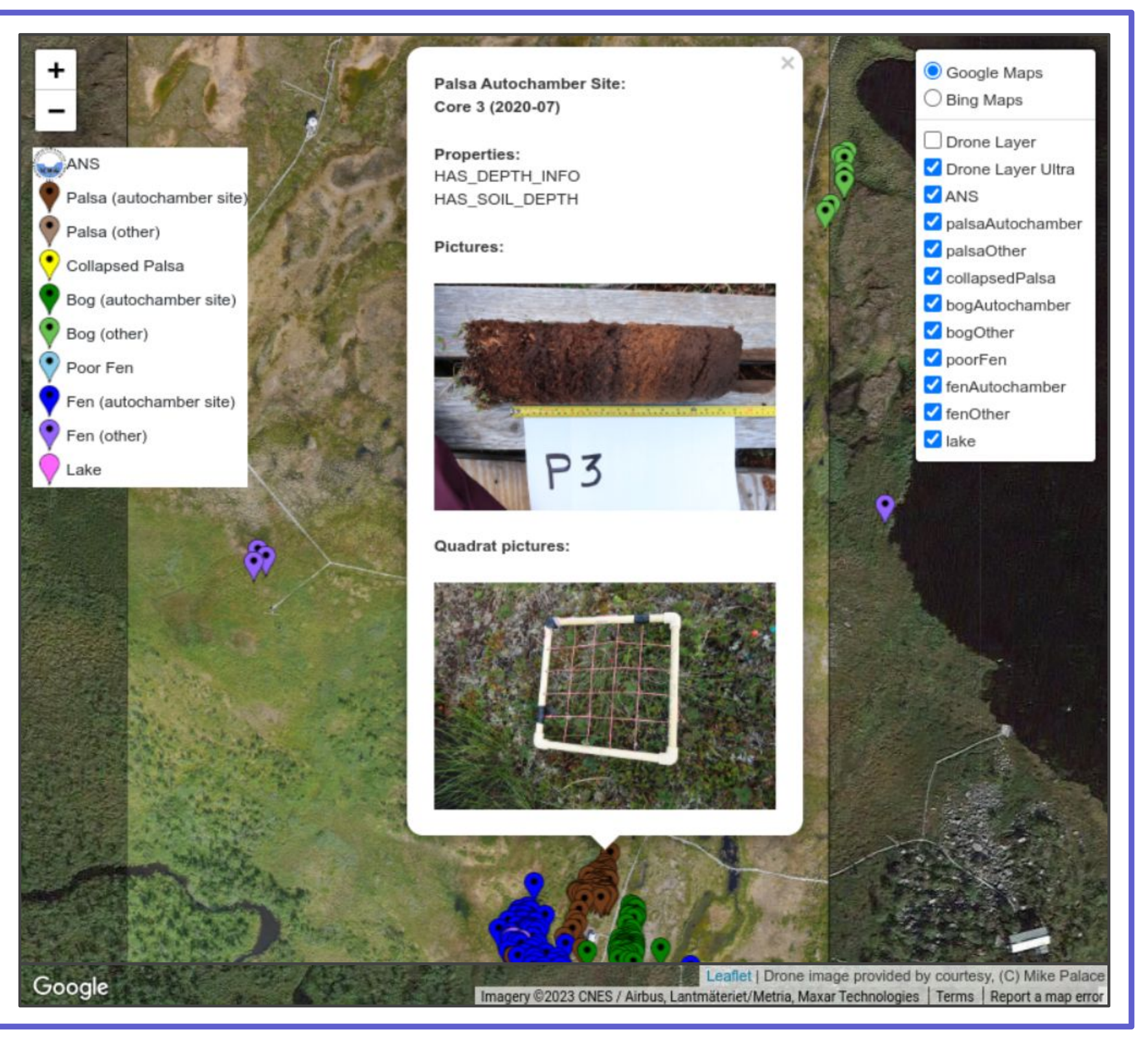
Live queries: Real-time queries on one or more labels. Available to EMERGE members (public access is forthcoming).

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



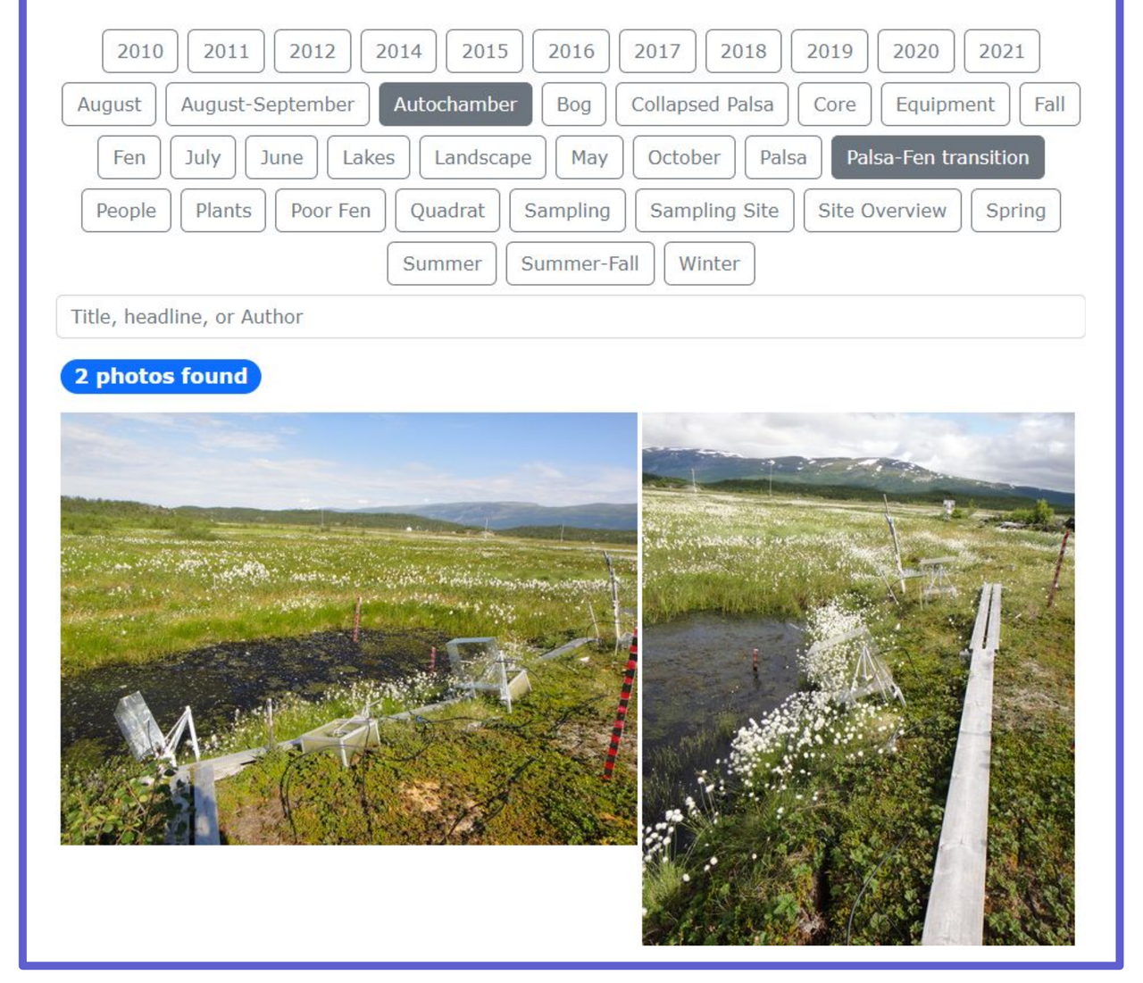
Map:

Graphical information on cores and other geo-referenced entities.

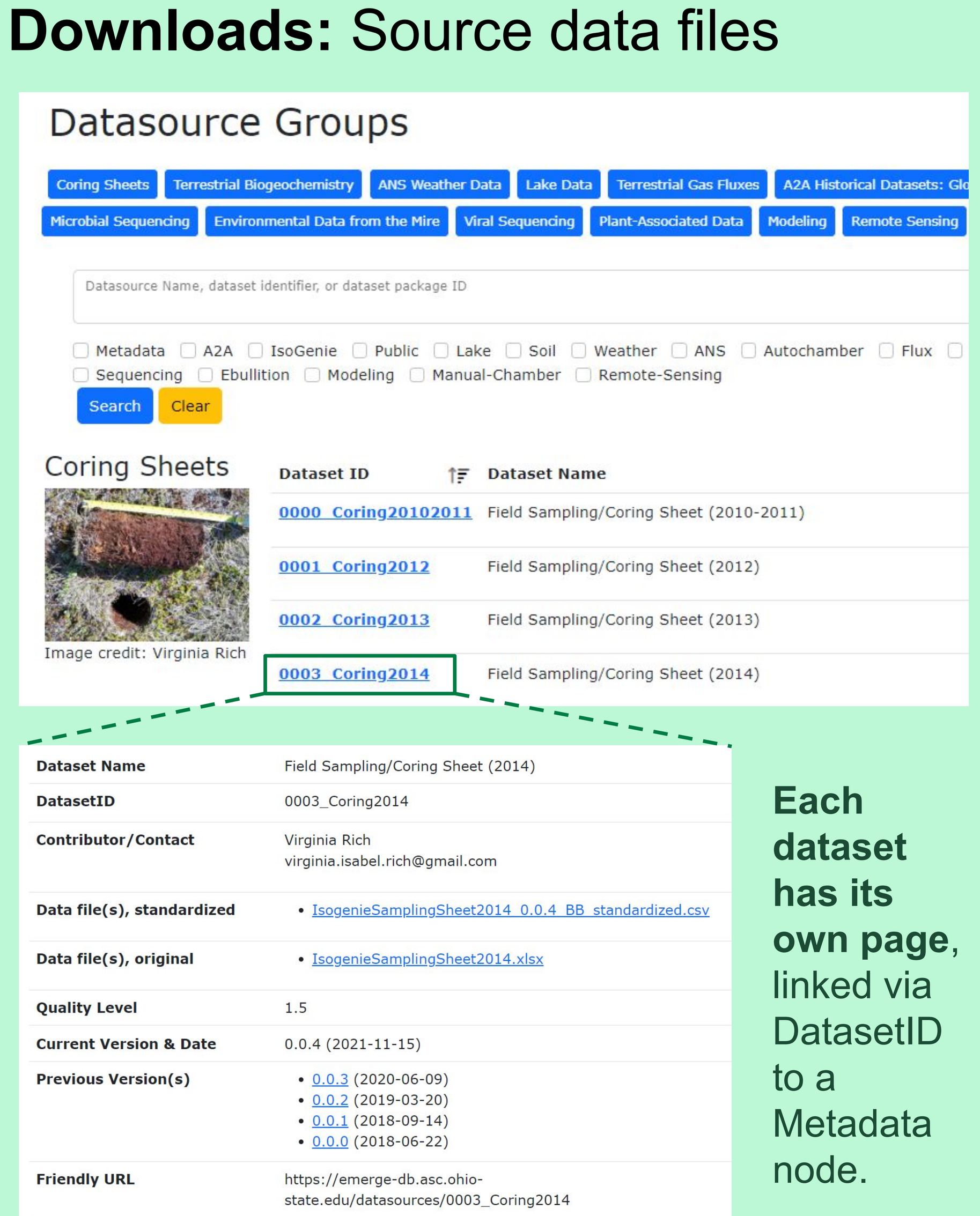


Pictures:

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



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.

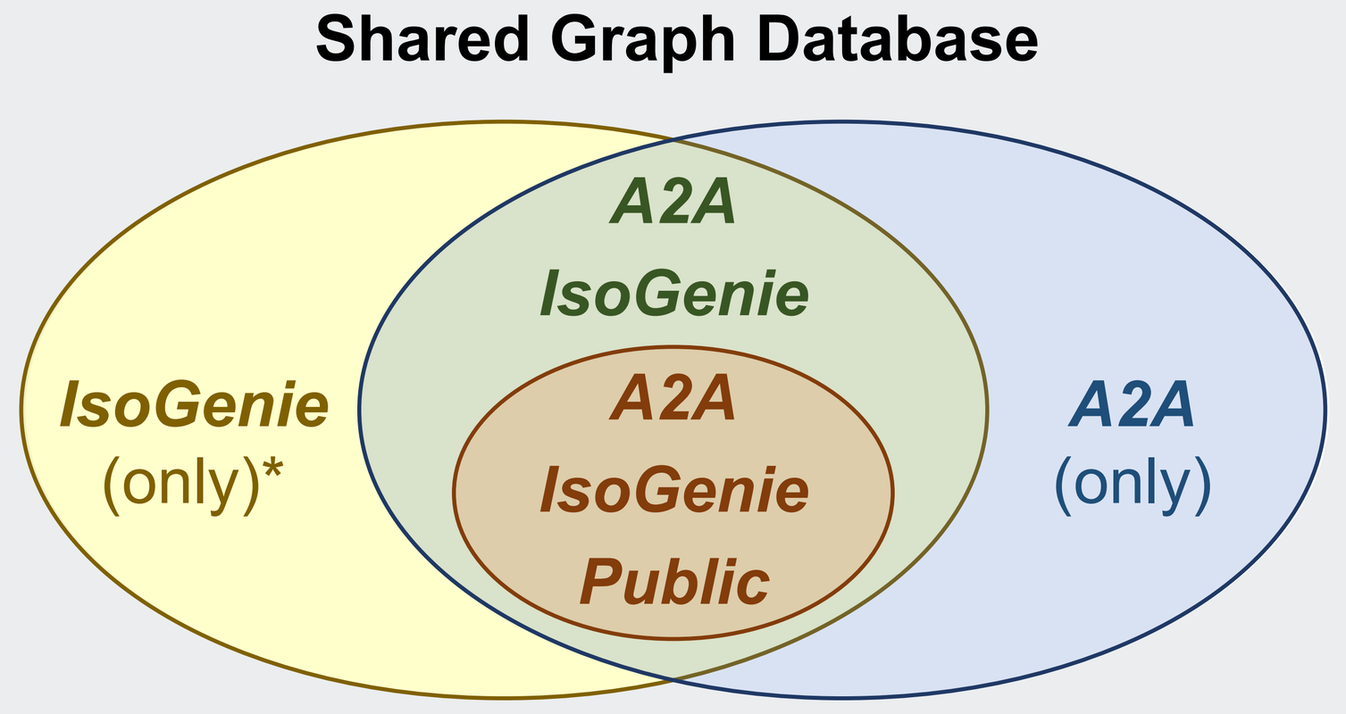
From EMERGE Data Management Plan:

"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

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.