Solving Multi-disciplinary, Multi-Scale, Interactive Data Management for Heterogeneous Users: The IsoGenieDB

Suzanne Hodgkins[1†], Benjamin Bolduc[1†], Matthew Sullivan[1*] (sullivan.948@osu.edu), Ruth Varner[2], The IsoGenie Team[3], and **Scott Saleska**[4], **Virginia Rich**[1*] (rich.270@osu.edu)

[1]Ohio State University, Columbus; [2]University of New Hampshire, Durham; [3]10 co-investigator institutions; [4]University of Arizona, Tucson. [†]These authors contributed equally.

https://isogenie-db.asc.ohio-state.edu/

Project Goals: The objective of the IsoGenie3 Project is to discover how microbial communities mediate the fate of carbon in thawing permafrost landscapes under climate change. We are engaged in a systems approach integrating (a) microbial and viral ecology, (b) organic chemistry and stable and radiocarbon isotopes, and (c) state-of-the-art modeling, across an interconnected system of thawing permafrost and post-glacial lakes in Arctic Sweden. To facilitate interdisciplinary collaboration in the IsoGenie Project, and to ensure broader community access to its near-decade of system-scale datasets, the IsoGenie Database (IsoGenieDB) integrates these datasets into a cohesive graph database framework. IsoGenieDB is not only a data repository for the project, providing both public and private access, but it also allows custom querying to explore connections between different datasets. The code base for building the database is open access and can be adapted by other projects to suit their own unique needs.

Organization of disparate environmental data types isn't a new challenge, with storage systems designed for this purpose including ESS-Dive, NEON, LTER, Pangaea, DataDryad, DataONE, etc. These systems are powerful and have their respective strengths, but they do not generally integrate data into frameworks that can be queried in a unified, detailed way. Doing so would require assimilation of data from different sources, which may exist in different formats and involve differences in calibration methods, into a single framework. This framework's structure would ideally reflect that of the system itself, in that the relationships connecting research sites, sampling locations, and different assay methods are recapitulated in the data structure. Such a structure would move beyond data organization, to facilitate exploration of ecological, organismal, and physicochemical interactions occurring both horizontally (between different systems at the same scale) and vertically (across scales).

We sought to address this data integration challenge for the IsoGenie Project by developing the IsoGenie Database (IsoGenieDB), a novel data management and exploration platform. IsoGenieDB has been built as a Neo4j graph database (the same platform used by eBay, Cisco, NASA, and numerous others, helping ensure its reliable longevity) that leverages the inherent relationships within the data to build the basic framework of the database. The graph data structure is a natural mimic of biological and conceptual relationships. The fundamental design of the database follows a property graph model, where *nodes* serve as the primary unit of

organization. Nodes can have *labels*, which serve as a high-level means of categorizing nodes for fast access and classification, while data is stored within the nodes as a set of key:value *properties*. These properties can contain any numeric or text-based information (e.g., time, temperature, pH), as well as links to flat files that store non-text information. Nodes are connected to other nodes through *relationships* (also known as edges), which store information about how the two nodes are related.

Most of the nodes in IsoGenieDB are organized hierarchically by location and data type, starting with a "root" node representing the whole of Stordalen Mire. This root node is then connected by branching relationships to other nodes representing increasingly specific temporal and/or spatial resolution, and finally to data nodes for each data type existing at any given point in time or space. While this base structure is tree-like, the flexibility of the graph database format allows other conceptual patterns to be traced with other relationship structures. These include the storage of file-level metadata (e.g. file links, version, and quality information) in dedicated metadata nodes with relationships to each of their corresponding data nodes, as well as more complex networks of relationships, such as those describing data processing pipelines.

The code base for building the graph database, written in Python, is designed to be easily re-usable for new datasets with only minimal data cleaning of the contributed data files. This allows for automated building of node networks that illustrate common conceptual patterns in ecological research. For instance, within the base tree structure, each child node automatically inherits identifying metadata properties from its parent or "upstream" node (e.g., a node for a soil sample is automatically labeled with the core name and sampling date). This automation has the additional benefit that common node types are given consistent labels regardless of dataset origin; this allows for greater ease and flexibility of querying the database.

A front-end web portal, accessed at https://isogenie-db.asc.ohio-state.edu/, provides data access for non-coding project members and the general public. This web interface includes general information about the IsoGenie project, a Data Downloads page with metadata and downloads of source data files, a map interface for georeferenced data (e.g. core locations and remote sensing images), a tagged photo repository, and a dynamic querying page.

The flexibility of the IsoGenieDB codebase, and availability of its construction scripts and web portal, allows the database to be adapted and modified by other projects. One of these is the A2A-DB, a database for the NASA-funded Archaea to Atmosphere (A2A) project (PI: Ruth Varner, co-I on the IsoGenie Project), whose goal is to use modeling and remote sensing to upscale carbon cycle feedback findings from five Arctic peatlands (including that examined by IsoGenie) to the pan-Arctic. The A2A-DB is in development but will include a public-facing side like the IsoGenie-DB, and is designed to be further expandable to house data from other related studies.

This study was funded by the Genomic Science Program of the United States Department of Energy Office of Biological and Environmental Research, grants DE-SC0010580 and DE-SC001644. Additional funding for the expanded A2A-DB was provided by the NASA Interdisciplinary Studies in Earth Science program (Award # NNX17AK10G).