The IsoGenieDB: An Integrated Solution to Cross-Disciplinary Data Management

Benjamin Bolduc[1]* (bolduc.10@osu.edu);-Mike Palace[2], Ruth Varner[2], Gene Tyson[3], Jeff Chanton[4], Patrick Crill[5], **Scott R. Saleska**[6], Malak M. Tfaily[7], Carmody K. McCalley[8],-Matthew B. Sullivan[1] and **Virginia Rich**[1]

[1]Ohio State University, Columbus; [2]University of New Hampshire, Durham; [3]University of Queensland, Australia; [4]Florida State University, Tallahassee; [5]University of Stockholm, Sweden; [6]University of Arizona, Tucson, [7]Pacific Northwest National Laboratory, Richland and [8]Rochester Institute of Technology, Rochester.

https://isogenie-web-dev.asc.ohio-state.edu

http:// isogenie.osu.edu

Project Goals: The objective of the IsoGenie2 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) molecular microbial and viral ecology, (b) molecular organic chemistry and stable and radiocarbon isotopes, and (c) state-of-the-art modeling, along an interconnected chronosequence of permafrost thaw and post-glacial lakes in subarctic Sweden. Data management across disciplines and scales is key to success.

Abstract Text: Understanding systems-scale data – from geochemistry measurements of isotopic data, microbial ecological and biochemistry to vegetation surveys and climate data – is essential to identifying statistical relationships, and to model and predict biogeochemical cycling in any system. Integration of and access to this systems-scale data presents an exceptional challenge in data management, exacerbated by technological advancements in all areas of environmental data generation. This includes non-traditional investigative methods such as high-resolution drone imaging. The IsoGenieDB – a graph-based database with a web portal – provides an efficient means of not only organizing and storing the highly diverse data types that exist within the IsoGenie Project, but also offers exceptional querying capabilities that can leverage cutting-edge network-based analytics. The results of these analytical methods can reveal underlying or overarching patterns of interaction typically invisible within ecosystems data due to incomplete integration of the relationships within and between data types. The IsoGenieDB therefore seeks to directly solve the often paradoxical challenge of integrating increasingly complex and large datasets while simultaneously allowing end-users the ability to easily access and query against stored data generated by the IsoGenie Team. This data includes extensive, multi-year datasets including meta-omics, high resolution chemical and geochemical data,

vegetation “ground cover,” satellite and unmanned aerial imagery and many other data types, spanning temporal (seconds to decades) and spatial (depth profiles, and site locations) data series. This data is augmented by the 100-year historical meteorological measurements of the Abisko Scientific Research Station. A web interface allows non-technical users the ability to query against the database for specific datatypes, temporal or depth-associated information, geospatial grouping or a combination of the above. More sophisticated queries – including network-based methods and pattern recognition algorithms – can be performed by accessing the underlying database server directly. The web interface provides both private and public access, with private data shared among all members of the IsoGenie group, with public data release of bundled cross-disciplinary datasets upon publication.

Taken together, the IsoGenieDB is a novel data management and data interrogation system developed from the ground up, built to more accurately represent the data transformations within all levels of an ecosystem, regardless of data type or origin, e.g. temporal, spatial, geochemistry, imagery, abundances, minutes, hours, etc. We believe the IsoGenieDB will provide (a) a platform for IsoGenie members to explore their data through data and relationships provided by themselves and other collaborators to more fully describe their system and address their particular focuses and goals, and (b) a model for solving such data management challenges in other systems-scale projects, where ease of integration, in addition to data integration itself, can fundamentally change how scientists view interdisciplinary work and approach problems.

Funded by the DOE Genomic Science Program of the United States Department of Energy Office of Biological and Environmental Research, grant DE-SC0010580.