Banking on a new community isotope database

January 16 2020, by Aaron Dubrow

Collecting tissue samples from a kill site in Argentina for isotopes. Credit: Pauli group, University of Wisconsin-Madison

Stable isotopes act like fingerprints or fibers in forensics, capturing details of where someone or something lived, what it ate or breathed, and how its environment changed over time.

Isotopes are variants of elements whose nuclei contain the same number of protons but a different number of neutrons. Some isotopes are unstable and short-lived, but there are approximately 300 known naturally occurring stable isotopes, including common elements like hydrogen, carbon, nitrogen, and oxygen, that possess different numbers of neutrons based on how they formed.

Stable isotope analysis is used in a large number of fields, from archeology, where they are used to date objects, to conservation biology, where they help understand why a species is struggling.

"It's a useful interdisciplinary tool that researchers can apply to just about anything," said Seth Newsome, an animal ecologist at the University of New Mexico who uses isotopes to study the resources that sustain mammals, birds, and fish. "Tell me a problem and I'll find a way to use isotopes to solve it."

Tens of thousands of researchers like Newsome use isotope analysis in their work, and frequently the same dataset can be useful to many researchers, even those working in disparate fields. However, until recently, no central data repository existed.

"Before, the process was largely independent," said Jonathan Pauli, associate professor of forestry and wildlife ecology at the University of Wisconsin-Madison. "Researchers would generate data to use in papers and then store it on laptops or servers, siloed away. There's two problems with that: First, it's dangerous—decades of data are possibly at risk. And second, the data is unavailable. Data should be available to others to be able to build on. Progress is at the heart of science."

From Discussion to Action

Starting in 2015, a handful of researchers set about trying to change this. They recognized that research was limited by a lack of access to, or knowledge of, existing data. They also saw how the creation of GenBank in the 1980s had accelerated the rate of discovery in genomics and become one of the most valuable resources in science—deemed so critical that it is now funded directly by Congress.

"When people talk about the impact that technology has had on bioinformatics, they talk about genetic sequencing getting cheaper, but without the data becoming available through GenBank, the cost wouldn't make much difference," said Chris Jordan, manager for Data Management and Collections at the Texas Advanced Computing Center (TACC). "Unless you can compare your genome to all the other genomes that have been sequenced, you can't make any sense of it. There is the potential here to do something similar with stable isotope data."

Meetings at conferences eventually led to a position paper in *BioScience*, where they asserted "[n]ow is the time to invest in a parallel special-purpose database for another burgeoning field of research with enormous promise: the use of stable isotopes."

In 2017, the National Science Foundation (NSF) funded a workshop that brought together a team that included both researchers who use isotopes and the staff who run the field stations where isotopes are analyzed. The meeting led to an opinion piece in Proceedings of the National Academy of Sciences, titled "Why we need a centralized repository for isotopic data," that outlined a vision of a shared resource for researchers using isotopes.

"Data is being generated at a tremendous rate," said Pauli. "It's time for

us to find a place to house all these data in a format that people can draw upon and ask bigger questions."

In 2018, NSF awarded a $1.5 million, three-year grant to the team of researchers and developers based at the University of Wisconsin-Madison, the University of New Mexico, the University of Utah, and TACC to develop a digital archive of data, and ultimately analytical tools, that could be used by researchers worldwide—an IsoBank.

The team held their first Principal Investigator's meeting at TACC in December 2018, spent 2019 consulting with the community through multidisciplinary workshops and developing metadata and data structures from the resulting feedback, and are now working with a group of initial data depositors to bring data into the system. The rich metadata resulting from this community process will provide crucial contextual information to enable rigorous re-use of stable isotope data retrieved from IsoBank.

Building an IsoBank for the Ages

The challenge in creating any large-scale repository is designing it so it can be maximally useful—organized, searchable, sustainable, and easy to access.

The long-term storage of the data associated with millions of samples is hard. However, much more challenging is identifying the metadata—the data that describes and gives information about each sample—that must accompany the raw data to make comparisons across time, space, and subject possible.

Corral, a storage and data management resource at TACC designed and optimized to support large-scale collections and collaborative research environments. Credit: TACC

Metadata helps classify where a sample came from, what it relates to, and how it was analyzed. Understanding what metadata is needed is the first step towards developing high-quality organizational schemas and accessible databases.

"We're isotope experts, but we're not experts in designing, building, and curating such a large database like this," said Pauli. "TACC is really central in providing the expertise for us to be able to build IsoBank the way it should be."

In addition to providing the compute power to support thousands of researchers each year, TACC hosts critical datasets for a range of communities and develops gateways and portals to make the data accessible.

It was TACC's experience developing and supporting the Arctos database of more than three million records from natural and cultural history collections that led Pauli and his collaborators to select TACC as a partner. Other comparable community databases based at TACC are those of the Billie L. Turner Plant Resources Center, DesignSafe (a resource for the natural hazards research community), and SD2E (a web-based analysis platform for the DARPA Synergistic Discovery and Design project).

"TACC's role is to figure out how to express and store all of this metadata; how to enable search in a meaningful way; and how to start adding in useful analysis tools," Jordan said. "Being able to eventually enable new types of research so researchers can ask questions they've never asked before and get useful answers—that's the kind of thing that TACC is here to do."

The long-term nature of TACC's storage and computing infrastructure also assures the sustainability of the project.

"We don't want an ephemeral product. We want to build something that can be used for many years. Being part of the TACC facility will allow us to do so," Pauli explained.

Building a Gateway to Discovery

While Pauli and his colleagues spread the word of the IsoBank and recruit researchers from a variety of subfields to determine the metadata requirements, Jordan and the team at TACC are translating the

community needs into a virtual framework meant to scale to many petabytes of data and last for decades.

"One of the things I always tell people about TACC is that it's not just big computers, it's about complex computing challenges," said Jordan. "Things that are difficult in a number of dimensions, that's what's going to need our expertise, as well as the ability to integrate interesting analysis and visualization tools that are developed in other contexts."

The goal is to develop a framework for computational analysis where researchers can bring together multiple datasets, including those created to answer very different questions, make comparisons and study changes across long distances or time-scales, and run statistical analyses.

For Newsome, whose personal research is in the area of biological conservation, this may mean combining isotope datasets to look at how isotopes of various types of organisms vary over space and time.

"The ability to see what's been done, or use pilot data to help make predictions or design projects—that will help push ecology and biology forward," he said.

Another benefit of central repositories is their ability to generate collaborations both within and between scientific fields.

"There isn't a field out there in natural science and medicine that's not interested in isotopes," Newsome said. "It'll be cool to see this lead to collaborations of truly distant fields, like pharmacology and ecology, and biochemistry and archeology."

With a team of domain specialists determining the requirements and gathering data, and TACC data management experts turning those requirements into useable technologies, the project has the potential to

transform science.

"Isotopes are a powerful tool across a diversity of disciplines and will continue to give us important insights into how the world works," Pauli said.

More information: Jonathan N. Pauli et al, Opinion: Why we need a centralized repository for isotopic data, *Proceedings of the National Academy of Sciences* (2017). DOI: 10.1073/pnas.1701742114

Provided by University of Texas at Austin