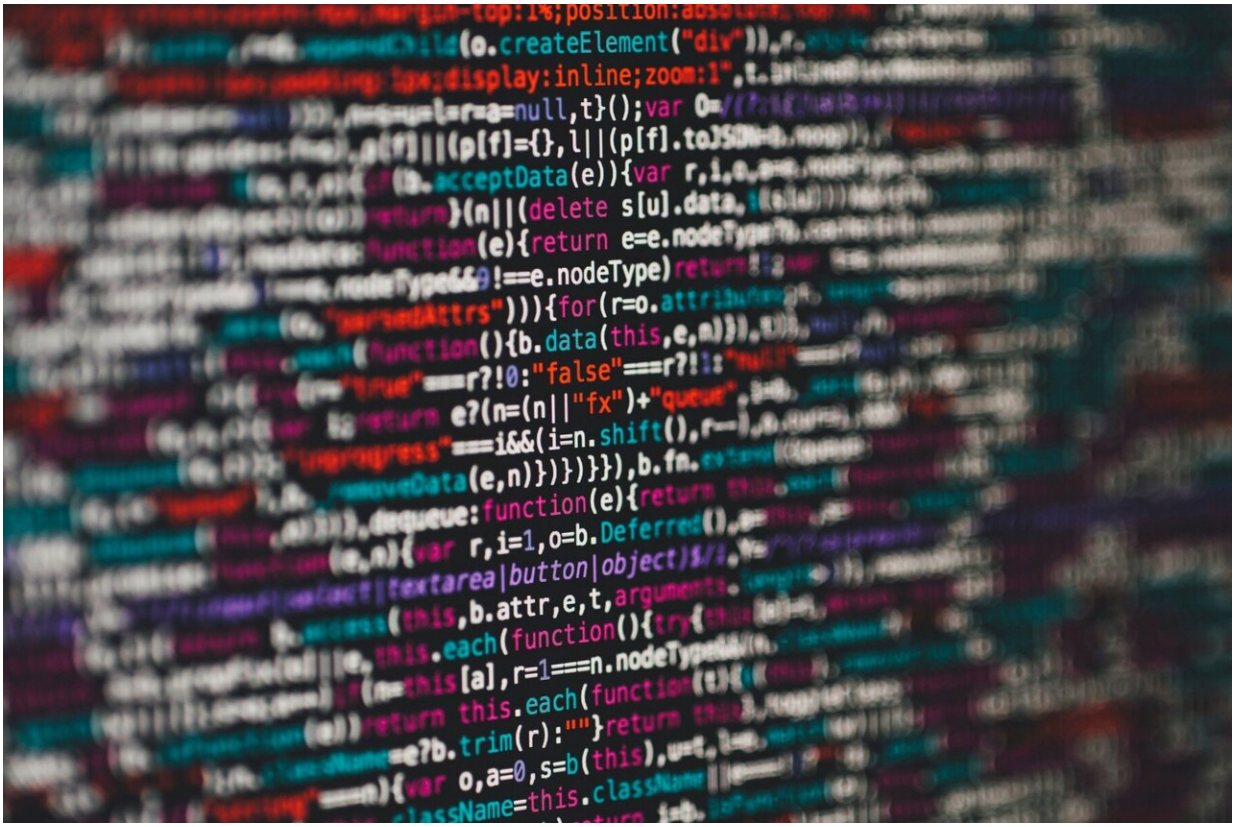
Researchers develop new open-source system to manage and share complex datasets

January 3 2020, by Laura Arenschield



Researchers have created a new open-source data-management system for scientists, with the hope that the system might make collaboration easier. Credit: Markus Spiske on Unsplash

Data is often at the heart of science—researchers track velocities,

measure light coming from stars, analyze heart rates and cholesterol levels and scan the human brain for electrical impulses.

But often, sharing that data with other scientists—or with peer-reviewed journal editors, or funders—is difficult. The software might be proprietary, and prohibitively expensive to purchase. It might take years of training for a person to be able to manage and understand the software. Or the company that created the software might have gone out of business.

A research team has developed an open-source data-management system that the scientists hope will solve all of those problems. The researchers outlined their system today in the journal *PLOS ONE*.

"We wanted to create a file format and a dataset model that would encapsulate the majority of datasets we work on, on all the instruments in a lab," said Philip Grandinetti, professor of chemistry at The Ohio State University and senior author of the paper. "There's this long-standing problem, pervasive among scientists, that you buy a multimillion-dollar instrument and the companies that make that instrument have their own proprietary format, and it's a nightmare to share with anyone else."

Large datasets are tricky to share, in part because software is often proprietary, but also in part because the files are often so large that they are hard to share in an email or through a cloud-based server. And even if the files can be exported as a file type that can be shared, important metadata—the things that explain what the dataset actually is—are often lost.

Their system, which Grandinetti and colleagues named the "Core Scientific Data Model," is designed to share complex datasets easily, without massive files that take up a lot of bandwidth and hard drive

space, and without losing metadata. Consider a dataset that includes air temperature, air pressure, wind velocity and solar flux—this system can handle it. Or consider the measurements and color of a light coming from a star in a distant galaxy—this system can handle it.

"You need a dataset that is incredibly flexible in its ability to hold all those things in one file format without losing information," Grandinetti said. "So the idea is we created a model that we thought was flexible enough to do that."

The Ohio State University team, in collaboration with Professor Thomas Vosegaard at the University of Aarhus in Denmark, and Dr. Dominique Massiot at the University of Orléans in France, built software that can run on a Mac or PC. They uploaded it to the web and made the code open-source (meaning anyone can look at it, use it, and download it for free.) The publication in *PLOS ONE* is intentional: The journal is also available to anyone, free of charge.

And, the researchers hope, the system could be a simple, free way to combine multiple types of data into one place.

"We study multiple datasets as scientists—and as a scientist myself, I'd like to be able to get the data from all those files and put them together in a way that I can work with," said Deepansh Srivastava, a postdoctoral researcher in Grandinetti's group.

"Instead of looking for data and plucking it from datasets, if we could simply export it as this one file type—as a core scientific data file type—we'd be able to work in a common system."

More information: Deepansh J. Srivastava et al. Core Scientific Dataset Model: A lightweight and portable model and file format for multi-dimensional scientific data, *PLOS ONE* (2020). DOI:

10.1371/journal.pone.0225953

Provided by The Ohio State University