Information, writ widely

August 24 2015, by Clea Simon

Initially unveiled in 2007 as a resource primarily for the social sciences, the Dataverse has been continuously improved since, according to Mercè Crosas, director of data science at the Institute for Quantitative Social Science. “After almost a decade we have built a community — an international community — of universities across the world,” she said. Credit: Kris Snibbe/Harvard Staff Photographer

Imagine an online catalog for data from researchers worldwide, a resource allowing scholars to not only read journal articles, but view the complete data sets behind the studies, a way for them to draw from, reproduce, cite, or build on this trove of findings without translating across formats or risking security or accuracy. Now imagine that catalog made better, more accessible—and increasingly applicable across the spectrum of disciplines.

That's what happened this spring when the Harvard Dataverse 4 was unveiled. The latest iteration of the Dataverse, an open-source Web application developed at the Institute for Quantitative Social Science (IQSS), in collaboration with the Harvard Libraries and Harvard University Information Technology (HUIT), not only tackles some of the initial problems of the application but also broadens its scope, making it useful to researchers of everything from elections to Ebola.

Initially unveiled in 2007 as a resource primarily for the social sciences, the Dataverse has been continuously improved since, according to Mercè Crosas, director of data science at IQSS. In this iteration, the system's user interface has been improved with better search functions and full data citation support, and the technology has been updated. But the biggest change is in the Dataverse's scope. By adding structured and searchable metadata, says Crosas, the system now supports data sharing in a much wider range of fields.

"Making this available across different disciplines has made a difference," says Crosas, who credits the IQSS data science team as well as the larger Dataverse community for the ongoing improvements. "After almost a decade we have built a community—an international community—of universities across the world."

That community has been expanding exponentially since the project was conceived by Gary King, director of the Institute for Computational

Social Science, in 2006. "The number of new dataverses [virtual archives or containers of data sets] and data sets in the Harvard Dataverse per month has increased by a factor of 10," on average, says Crosas. That is, in 2007 and 2008, there were around 50 new data sets published per month. In the last few months there were closer to 500, she says.

In practical terms, these new developments mean that researchers from the social to the physical sciences can now share information both easily and securely. (Humanities are also invited to take part, and future upgrades will address their specific concerns.) Already, government Professor Stephen Ansolabehere has collaborated with more than 50 universities and colleges across the country for the Cooperative Congressional Election Study. This study, on which Ansolabehere is the principle investigator, organized teams from each institution, which in turn created questionnaires to produce a 50,000-person common survey that has been released and archived through the Dataverse.

Working with the Dataverse, says Ansolabehere, his team (with the support of the dean of Faculty of Arts and Sciences and the Sloan Foundation) also has created the Harvard Election Data Archive, using an online workspace, which in turn allowed a team of six researchers at Harvard and researchers at other universities to collaborate. The resulting database, he says, has already been downloaded approximately 75,000 times since 2011 and is updated annually.

"The Dataverse is invaluable," says Ansolabehere. "There is no other workspace like this. Dataverse isn't just an electronic world. The support team is incredible, and they constantly work to improve and expand the Dataverse."

With the latest version, which has had smaller improvements every few weeks, the natural sciences can now enjoy the same access, says Crosas. For example, Pardis Sabeti, associate professor at the Center for

Systems Biology, is using the Dataverse to share research on 213 cases of Ebola in Sierra Leone. Sabeti's study was published in the *New England Journal of Medicine* in November 2014. Thanks to the Dataverse, her data sets already have been downloaded more than 40 times.

The Dataverse is not done growing. The next complete overhaul, due in 2016, will address the challenges of confidentiality. Now being developed in conjunction with the Center for Research on Computation and Society at the Harvard School of Engineering and Applied Sciences, IQSS, the Data Privacy Lab, and the Berkman Center, Dataverse 5 will negate the need to strip confidential material — names, personal medical histories, etc. — from research before it is uploaded. This next version will assign different levels of sensitivity to the material, based on a five-level color spectrum from blue (lowest) to crimson (highest). Accessing each level would require a specific security clearance and involve escalating levels of encryption. To simplify: Researchers will be able to upload their complete data sets, without editing. Then, depending on a users' clearance (and requirements), that information would be made available at an appropriately secure level — from the basic numbers to those with the lowest clearance to the full, raw data to those with the highest.

Farther down the line, Dataverse 6 will tackle the problems of large-scale data and streaming, with additional improvements to come as more and more users sign on and share the specific requirements of their fields.

"Making data easily accessible and reusable helps accelerate and validate research output," says Crosas. "We are unchaining data."

This story is published courtesy of the Harvard Gazette, Harvard University's official newspaper. For additional university news, visit