BIRNing to Create a New Standard in Scientific Collaboration

by Stephanie Sides

John Wooley, associate vice chancellor at UC San Diego, was joking about the term *data mining*, which describes the process of analyzing massive amounts of data so as to discover patterns, relationships, or fundamental principles. "It really refers to data *mine*, not yours," says Wooley with a laugh. He is referring to the unfortunate scientific standard that, if you don't create the data, you can't hope to access it for your research. That common attitude has held scientific progress in check.

Fortunately, that tradition is beginning to reverse course, literally, thanks to visionaries like Mark Ellisman who has a different view of the future. He's a neuroscientist and bioengineer with joint appointments in UC San Diego's School of Medicine and the Jacobs School of Engineering. He's also a member of the California Institute for Telecommunications and Information Technology (Calit2).

Ellisman, a year ago, embarked on a multi-campus project known as the Biomedical Informatics Research Network (BIRN) funded by the National Institutes of Health's National Center for Research Resources (NCRR). BIRN's scientific goal is to integrate a wide variety of types of data, acquired by the most advanced biomedical imaging and General Clinical Research Centers around the U.S., into an extensible knowledge system. To make this possible, BIRN is designing and implementing a hardware and software infrastructure, supported by the necessary expertise, to manage, provide access to, and support analysis across geographically distributed data sets.

Why does this project matter? Imagine you wanted to do a study of all female patients of a certain age that have certain symptoms. And suppose you wanted to investigate the correlation between this set of characteristics and the onset or progression of a particular disease. To do that, you would want to access *all* relevant data sets no matter where they were, who had collected them, or by what technology they had been acquired. You'd want all this data to be calibrated in a scientifically credible way so you could access it seamlessly and analyze it quickly, then move on to the next step of your research.

BIRN is using this type of scenario as a carrot to motivate collaboration as a scientific virtue: BIRN hopes to convert convinced data *mine*-ers into data sharers.

Data sharing will enable better understanding of the morphology of neurological disease, including multiple sclerosis, schizophrenia, Alzheimer's, and Parkinson's, some of the diseases slated for study during the early years of this project. If you were to add in other diseases that pose long-term health problems and seriously impact medical providers, insurance companies, and costs, you could see that, yes, this project *matters*. And, as it grows, it will matter *more*.

The goal sounds straightforward, but the technical challenges in making this work are daunting. Data are being collected on different types of brain activity, over a range of scales (molecular to the whole brain), over a range of time periods depending on relevance to the image acquisition

technology and the topic of study, and using different laboratory methodologies (such as positron emission tomography, high-voltage electron microscopy, and magnetic resonance imaging).

Surprisingly, even like models of laboratory equipment made by a given manufacturer need to be calibrated so that a data set collected by one machine can be correlated accurately with that collected by another. Additional complexity comes from the fact that the image files use different filetypes and live on different storage technologies, based on different computers, using different operating systems – and in different places!

Some think Ellisman was *crazy* to take this problem on.

But he insists just the opposite: It made absolute sense. His motivation arose from a lot of hard work over many years, building relationships with respected colleagues and exploring what technologies could advance their mutual scientific interests. In the tradition of breakthrough science that gathers together the best minds to tackle what seems an overwhelming problem, this project built carefully on long-nurtured, successful collaborations supported by NIH and the National Science Foundation (NSF).

This national partnership is anchored at UC San Diego by the BIRN Coordinating Center. This center leverages the National Center for Microscopy and Imaging Research, Ellisman's home lab; the National Biomedical Computation Resource at UC San Diego, which coordinates this grant; the San Diego Supercomputer Center (SDSC), which staffs the BIRN Network Operations Center and Help Desk; and the National Partnership for Advanced Computational Infrastructure, an NSF grant centered on SDSC that has integrated many important, long-standing research collaborations across a variety of institutions and the breadth of the country.

That's a lot of talent to throw at such an ambitious, complicated scientific problem. But that's exactly the kind of partnership needed to make headway.

Scientific research continues to build on innovation arising from individual principal investigator-driven research. However, to attack problems requiring development of *national* infrastructure, federal agencies are supporting larger, longer grants for ultra-complex research projects that require a broad range of expertise spanning computer science, disciplinary science, and social science. It's this combination of talents that makes it possible to address larger-scale issues.

"Scientific progress is being driven at the interfaces of disciplines working together to address more complex problems," says Larry Smarr, Calit2 director. "We consider BIRN a showcase for this new kind of science."

Francine Berman, director of SDSC, underscores this point: "BIRN is already paying dividends by serving as a model for yet another disciplinary domain, geoscience, through our GEON project. In fact," she emphasizes, "this model really applies to large-scale projects and cyber infrastructure *in general*."

Day-to-day operations for BIRN are overseen by project manager Mark James. He merges a background in computer science (simulation and modeling) with private sector experience

running data centers and software projects. “We’re trying to define processes and procedures, and establish best practices to build a standardized system that’s reliable and scalable to support the work of thousands of researchers,” he says. By “thousands,” though, he’s talking only about neuroscientists.

What he *doesn’t* say is that this type of system could potentially be used, with perhaps only minor modification, by nearly *any* discipline. That’s the grander glory of this project: scalability across not just numbers of people, but numbers of *disciplines*. So multiply the thousands of neuroscience researchers he mentions by some additional number of disciplines that could benefit, and you gain a sense of the potential long-term impact of this work.

The development of this cyberinfrastructure, invariably, is also encouraging research on the infrastructure itself: new techniques in databasing, information retrieval, visualization, and computational processing. The infrastructure initially is perceived as a “data grid,” but, in the future, as the demands of computational models grow, the emphasis is likely to include more parity with computation.

The infrastructure is based on a “standard-issue” local node designed and provided to each site by the coordinating center. It includes a grid computing rack with three Linux-based computers, 1-10 terabytes of storage, an Oracle database, the SDSC Storage Resource Broker, and gigabit network access via Internet2’s Abilene network. The cost of each rack is on the order of $90,000, which is paid for by the grant.

Local campus, medical school, or hospital connectivity to the Internet2 gateway is assumed to be gigabit-per-second, which each institution commits to as part of its participation. Each site is also required to provide a network administrator and half-time equivalents each of a database programmer and an applications programmer trained by the coordinating center to support the local node.

The coordinating center is responsible for deploying the network infrastructure, developing software tools (such as for structural and functional analysis) and a common web portal interface, sharing tools/code across institutions, converting data among tools, and providing support services, such as network monitoring, performance measurement, statistical analysis of how the infrastructure is being used, a help desk, problem tracking and resolution, and documentation.

Commonality of the infrastructure for each site is the key to BIRN’s rapid national-scale buildout. This standard, though, is not only supporting NIH biomedical researchers. Perhaps more importantly, it is helping drive the requirements of the NSF Middleware Initiative, which will support even more far-reaching impact across a wider range of disciplines.

Just as important as the definition of the hardware, software, and networking required for the local nodes is the coordinating center’s role in training and installation. Staff from the center conduct a survey and site visit prior to each installation to match their expectations to the site’s needs. Then the local node is pre-configured at SDSC, shipped, and installed. The devil, as they

say, can be found in the details at the site in question: What is the available power supply? Is the door large enough to accommodate the equipment? Can the flooring support 1,500 pounds?

With the local node installed, users can access data through a standardized web portal that accommodates use of individual sites' applications. And an important part of the bargain to gain *access* to the data is that each site *contribute* data.

One main advance of this project is the decision *not* to ask sites to modify the format of their data before contributing it. That, Ellisman realized, would have been horribly time-consuming and probably a show stopper. Rather, the choice was made to deploy software to enable the various data sets to "talk" to each other seamlessly as is, regardless of data format, storage medium, hardware platform, or location. That's where the SDSC Storage Resource Broker makes such a formidable contribution to this collaboration.

Inevitably, this project bumps up against issues of data security and confidentiality. Clinical data is governed by the Health Insurance Portability and Accountability Act (HIPAA) of 1996: If you're a patient, your doctor can't release information about you to another person without your permission. The act further stipulates that your identity be "anonymized." So the needs of this project are pressing the federal granting agencies and institutional review boards (that oversee research projects with human subjects) to adapt their guidelines on the use of human clinical data to the modern world of data grids while preserving patient rights.

BIRN is also exploring the changing sociology of research management with respect to the various *levels* of management (individual project management, coordinating center management, and funding agency oversight) and the *mechanisms* (read: technologies) by which management is supported.

National BIRN participants meet frequently in various groupings: all members of a given project, all project managers, coordinating center staff with participants at a site experiencing a problem, and so forth. These groups meet by phone, videoteleconferencing, and in-person visits.

Weekly status reports are also provided to NCRR. While this frequency is unheard of in federal granting circles, NCRR considers this degree of communication important to monitor the pulse of the project, and identify and help BIRN management address problem areas to maintain the project's momentum and ensure smooth growth of the infrastructure.

Other questions remain to be answered of the kind that plague many efforts to bridge communities or special interest groups, including how to minimize the barriers of potential mistrust or perceived "competitive advantage," how to determine who gets credit for accomplishment in a shared environment, and how to integrate new participants, both technically and scientifically, as the project grows.

Last year, five institutions joined BIRN. Now the overall infrastructure counts 10. And BIRN expects strong growth going forward: NIH expects to add some 35 General Clinical Research Centers to the project over the next three years. And demand will increase to incorporate data about other organs, diseases, and species.

Looking further down the road, as visionaries do, Ellisman says, “The next big push is building on BIRN to support clinical informatics.” BIRN, in its fullest incarnation, he says, will help provide an information infrastructure supporting *individualized* health care by aggregating and enabling access to *relevant* information obtained from all biomedical science, clinical practice, demographic characteristics, and treatment histories. Now that’s a vision that might provide a conceptual framework for the current debate in Washington about healthcare reform.