PRESERVING THE DATA HARVEST

Canning, pickling, drying, freezing–physicists wish there were an easy way to preserve their hard-won data so future generations of scientists, armed with more powerful tools, can take advantage of it. They've launched an international search for solutions.

By Nicholas Bock

Photos: Reidar Hahn, Fermilab

When the BaBar experiment at SLAC National Accelerator Laboratory shut down in April 2008, it brought an end to almost nine years of taking data on the decays of subatomic particles called *B* mesons. But that was hardly the end of the story for the 500 scientists working on the experiment. In November they celebrated the publication of their 400th paper, and they expect the next few years will yield at least 100 more.

These BaBar results and discoveries stem from more than two million megabytes of data. As impressive as this number is, it's only a fraction of the data that will come out of the next generation of high-energy physics experiments. For instance, the ATLAS detector at CERN's Large Hadron Collider will produce a whopping 320 megabytes of data every second, surpassing BaBar's total output within three months.

BaBar's treasure trove of data, which may contain answers to questions we don't even know how to ask yet, raises an increasingly important question in high-energy physics: When the party's over, what do you do with the data?

In the past, this was not so much of a concern. New experiments came along in a regular drumbeat, regularly superseding one another in terms of what could be done with the data they produced. Today, as experiments get bigger, more complex, and much more expensive, the drumbeat has slowed considerably, and physicists are starting to realize the value of wringing as much insight out of every experiment as they possibly can.

But without a conscious effort to preserve them, data slowly become the hieroglyphs of the future. Data preservation takes a lot of work, and with that, a lot of resources. Researchers have to think not only about where to store the data, but also how to preserve it in a way that it can still be used as technology and software change and experts familiar with the data move on or retire.

"Preserving the bits for all time is probably not difficult, but the data themselves become very, very rapidly an arcane, dead language," says Richard Mount, SLAC's head of scientific computing. "Preserving the ability to fully understand the nuances of a dead language is not without its cost."

It's an investment, though, that a growing number of physicists and collaborations are seriously considering. A study group known as DPHEP, for Data Preservation and Long Term Analysis in High Energy Physics, has been holding workshops to look at the issue. The BaBar collaboration has also emerged as an important force in the effort to solve the puzzle, with members striving to provide a working model of how data preservation can be done.

DIGGING FOR TREASURE

The value of old data was recently demonstrated by Siegfried Bethke of the Max Planck Institute for Physics in Munich. His group resurrected 25-year-old data from the JADE experiment at DESY, the Deutsches Elektronen-Synchrotron laboratory in Hamburg, Germany, and combined it with more recent data taken by the OPAL experiment at CERN. They wanted to study the strong coupling constant—a value that reflects the strength of the strong nuclear force that binds quarks and gluons together into protons and neutrons. Equipped with improved analysis tools and refined theory, Bethke and his colleagues were able to get a better feel for how the strong coupling constant changes at different energy levels.

Old data have also proven to be an invaluable teaching tool. Physics professors can use simplified sets of real experimental data to guide students through the basic steps of conducting an analysis. BaBar physicist Matt Bellis has taken this idea one step further, working with colleagues to develop outreach programs for high school and college students that use old data to introduce high-energy physics.

But some physicists report that older data sets have been getting harder and harder to access. Files have become damaged or lost, old software is incompatible with newer operating systems, and the knowledge needed to put everything together for analysis has dissipated. As time goes on, trying to do an analysis with the data becomes more of an archaeological dig than a physics experiment. In Bethke's case, the analysis required the help of experts who had originally worked on the JADE experiment. It also required a lot of hard work and luck. Whole chunks of the data thought to have been lost were recovered in the form of computer printouts. While this was a great find, members of Bethke's team had to go through the printouts page by page and re-enter the data by hand. It was worth the effort, though, since the JADE data had been taken at an energy level where no other data existed.

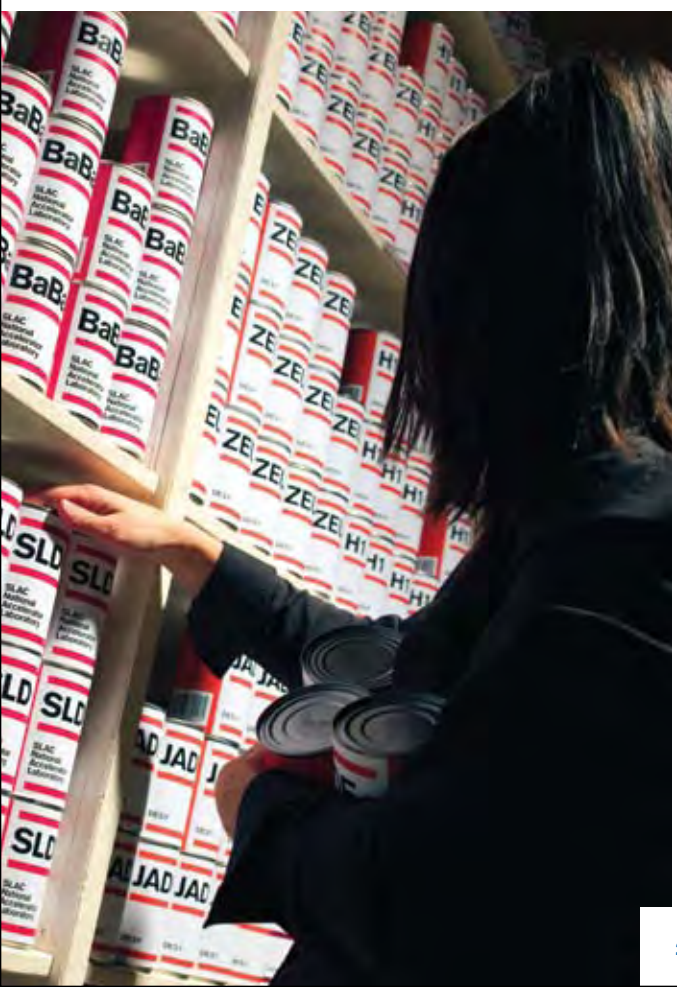
In an October 2008 survey of more than 1000 physicists, an overwhelming majority indicated that data preservation is important. More than 43 percent reported that access to old data sets could have improved their more recent results, and more than 46 percent expressed concern that important data had been lost in the past.

The survey was conducted by a group of CERN scientists involved in PARSE.Insight, a project funded by the European Commission to provide insight into data preservation across fields of science. CERN physicist Salvatore Mele, who runs the CERN group and is also a member of the DPHEP study group, offers a blunt analysis of the issue, describing high-energy physics as a worst-case scenario in terms of data access and preservation.

"We have funding streams to build accelerators, we have funding streams to build experiments, we have funding streams to operate experiments, we have funding streams to write software, we have funding streams to analyze data," he says, "but we don't have funding streams to preserve data."

HOLDING DATA CLOSE

For particle physicists, it's a surprising place to be; the field has been at the forefront of information technology and open access for decades. Since the 1960s, particle physicists eager to share results with colleagues have distributed preprint manuscripts of their results, initially on paper and then electronically on the arXiv repository, which has accrued more than half a million documents on its servers. In the late 1980s, CERN software consultant Tim Berners-Lee invented the World Wide Web to facilitate communication within the high-energy physics community. And in 1991, the particle-physics article database SPIRES became the first Web site outside Europe, helping set into motion events that have revolutionized information technology.



Other fields have run with this model, using online databases to share not only results and publications, but also raw data. Bioinformatics is buoyed by the efforts of thousands of researchers who make genetic sequence data publicly available, and data from NASA experiments goes public within a year after it's taken. But this exuberance for open-source data seems to have passed over high-energy physics, with collaborations tightly guarding the data from their experiments.

"Typically, in high-energy physics, if you have any interest in somebody's data, you have no access to it unless you know somebody on the team and they're willing to work with you," SLAC astrophysicist Richard Dubois says.

A MATTER OF CULTURE

Dubois found a much different attitude toward data-sharing and preservation when he left particle-physics experiments at SLAC in 1999 to work on the Fermi Gamma-ray Space Telescope, a joint mission of the Department of Energy and NASA (see "Fermi's excellent adventure" in this issue). The space telescope's detectors are similar to those used throughout high-energy physics to record the trajectories of subatomic particles. But in return for NASA's collaboration, Fermi scientists had to make some concessions. They agreed to organize their data in a format called FITS, which is used by all NASA science missions. And they pledged to make all their data public one year into the mission, along with any tools needed to analyze it.

NASA "basically says, 'If you want to build this thing, these are the conditions under which we'll let you do it,'" Dubois says. "Their goal is that anybody can make scientific discoveries with this public data."

For astrophysicists, the model has its pros and cons. The simplicity of the FITS format can limit the breadth of analyses collaborations are able to undertake. But at the same time, because all NASA experiments use FITS, it becomes easy for researchers to use data across different collaborations. NASA also promises to store and maintain FITS data indefinitely.

There is no analogous organization for high-energy physics. The Department of Energy underwrites many high-energy physics experiments conducted in the United States, but does little in the way of mandating what is to be done with the data produced. The responsibility of preserving data often falls on the collaborations themselves and, as a result, can become something of a bugbear, channeling money away from research projects and offering little in the way of immediate returns.

"You don't get tenure because you invested six years of your life preserving data," Mele says. "You get tenure because you do physics."

WHAT YOU SEEK IS WHAT YOU GET

In high-energy physics, data collection starts with a detector—an oftentimes enormous piece of scientific equipment

that measures the outcomes of collisions between subatomic particles.

To get an idea of what a collision looked like, researchers reconstruct it from tracks the debris left in the detectors. Then they compare the actual data with the results of simulations based on what they had expected to find. Only then can they draw conclusions about what it all means. Further complicating things, their reconstructions depend on the kind of event they're looking for; using the same data set, a researcher studying tau physics might come up with a completely different reconstruction than a researcher studying beta decay.

"There's no way to analyze a high-energy physics data set and get all the information out of it. That doesn't happen," Mount says. "If you make completely different guesses about what you might be looking for, you might get different things."

Reconstructing the data and running the simulations require a thorough familiarity with how the detector works—something that's developed over years of working within a collaboration. Because tools tend to be so unique, it is difficult for physicists from different collaborations to work with each other's data.

A COMPUTING CHALLENGE

When it comes to figuring out how to make this kind of exchange work, though, the data itself plays a relatively small role. The real challenge comes from preserving the software used to access and analyze the data, the software needed to run simulations, the operating system needed to run the software, and, perhaps most importantly, the knowledge of how to use the software with the data to produce results.

It's something that BaBar Computing Coordinator Homer Neal spends a lot of time thinking about. Neal oversees all of BaBar's computing tasks, from how the collaboration goes about doing computer-intensive analysis jobs to which operating systems it uses. He is also heavily involved with the data preservation effort, participating in DPHEP workshops and helping draft a report on the group's findings.

The challenges involved in data preservation are very real for Neal, who last summer helped oversee the ultimate reprocessing of BaBar's data. This reprocessed data and the corresponding raw data are being transferred to newer, high-density tapes, a process that will take about a year. By about 2012, he says, BaBar's archival system will store several copies of the data on disk and/or solid-state drives.

"This is our legacy data, because it is very unlikely that we will ever have the resources again to reprocess all the data," Neal says. The archival system will ensure that it remains in a form that can be analyzed.

Dealing with the operating system could prove to be a lot trickier. Once a software distributor stops providing security upgrades for an operating system, running equipment on it quickly becomes a liability. To keep servers safe, the operating system has to be upgraded regularly. But all the analysis and simulation software must be upgraded with it, lest researchers be overwhelmed with a deluge of error messages when they try to access the data.

EXPLORING SOLUTIONS

During the past year, members of DPHEP have been meeting to try to hash out some solutions. The group's first workshop was held in January 2009 at DESY. The group met at SLAC in May, and is scheduled to meet again this month at CERN. In August 2009 its efforts were endorsed by the International Committee for Future Accelerators, which provides a forum for discussions of particle physics worldwide.

One topic at these workshops has been virtualization, a technique that takes advantage of multi-core processors to run likenesses of old machines on new operating systems. The main computer would run on the latest operating system available, while the old, or virtual, machines securely run old software without having to worry about the periodic cycle of updates.

Neal and other BaBar researchers hope that data storage and virtualization methods will soon progress to the point where all the experiment's data and software can be housed in a self-contained archival system. According to Neal, new members of the collaboration could be working exclusively from the system as early as 2013. Another model, proposed by BaBar members at the University of Victoria, operates on the same idea, but with data and software housed at computer centers around the globe rather than at a central location. The team recently got $577,000 from Canada's Advanced Innovation and Research Network to design a prototype of the system.

BaBar researchers are also working to find easier ways to share data. While the complexity of most data sets eliminates any hope of creating a universal data format like FITS, researchers have been trying to develop data formats that at least let similar experiments exchange data. This can decrease statistical uncertainty in analyses and provide a good way to double-check results. Researchers at DESY have already started combining data sets from that lab's H1 and Zeus experiments, and recently submitted publications signed by both collaborations.

Physicists at BaBar are trying to build on that model. They are exploring ways their data can be combined with data from the Belle experiment at Japan's KEK lab, which also investigates decays of *B* mesons.

Amidst all the uncertainties of the data preservation debate, two things seem clear: the amount of data produced by high-energy physics experiments will continue to increase. And finding ways to preserve that data and make it reusable will be a challenge.

"There are voices saying that this will never be possible for high-energy physics," says H1 Spokesperson and DPHEP Chair Cristinel Diaconu. "It is very complicated, but it is not impossible. We have to just aim toward some form of open access and see how far we can get."


H1
DESY
CLE
ornell