In the mid-1990s, Siegfried Bethke decided to take another look at an experiment he participated in around 2 decades earlier as a young particle physicist at DESY, Germany’s high-energy physics lab near Hamburg. Called JADE, it was one of five experiments at DESY’s PETRA collider, which smashed positrons and electrons into each other. Looking at the strength of the force that binds quarks and gluons into protons and neutrons, JADE finished in 1986 when DESY closed down PETRA to build a more powerful collider. In the decades since, new theoretical insights had come along, and Bethke hoped the old data from JADE—taken at lower collision energies—would yield fresh information.

What the physicist found was a disaster. Since JADE shut down and the experiment’s funding ended, the data had been scattered across the globe, stored haphazardly on old tapes, or lost entirely. The fate of the JADE data is, however, typical for the field: “There’s funding to build, collect, analyze, and publish data, but not to preserve data,” says Salvatore Mele, a physicist and data preservation expert at the CERN particle physics lab near Geneva, Switzerland.

This tendency has prompted some in the field to call for better care to be taken of data after an experiment has finished. For a very small fraction of the experiment’s budget, they argue, data could be preserved in a form usable by later generations of physicists. To promote this strategy, researchers from a half-dozen major labs around the world, including CERN, formed a working group in 2009 called Data Preservation in High Energy Physics (DPHEP). One of the group’s aims is to create the new post of “data archivist,” someone within each experimental team who will ensure that information is properly managed.

Physics archaeology
For the founders of DPHEP, Bethke’s struggles with the JADE data are both an inspiration and a cautionary tale. It took Bethke, now the head of the Max Planck Institute for Physics in Munich, Germany, nearly 2 years—and a lot of luck—to reconstruct the data. Originally stored on magnetic tapes and cartridges from old-style mainframes, most of it had been saved by a sentimental colleague who copied the few gigabytes of data to new storage media every few years. Other data files turned up at the University of Tokyo. A stack of 9-track magnetic storage tapes was stashed in a Heidelberg physics lab. One critical set of calibration numbers survived only as ASCII text printed on reams of green printer paper found when a DESY building was being cleaned out. Bethke’s secretary spent 4 weeks reentering the numbers by hand.

Even then, much of the data couldn’t be read. Software routines written in arcane IBM assembler codes such as SHELTRAN and MORTRAN, tweaked for ’70s-era computers for which memory was at a premium, and stored on long-deactivated personal accounts, were lost forever. A graduate student spent a year recreating code used to run the numbers.

The recovery work was motivated by more than Bethke’s nostalgia. In the years since JADE ended, new theories about what physicists call the strong coupling strength had emerged. These predict phenomena that can best be seen at lower energies than today’s colliders are able to replicate. By reanalyzing the old data, Bethke’s team squeezed more than a dozen high-impact scientific publications out of the resurrected JADE data. Some of the data helped confirm quantum chromodynamics, the theory governing the interior of atomic nuclei, and was cited by the committee that awarded the 2004 Nobel Prize in physics to David Gross, David Politzer, and Frank Wilczek. “It was like physics archaeology,” Bethke says today. “It took a lot of work. It shouldn’t be like that. If this was properly planned before the end of the experiment, it could have all been saved.”

The usefulness of JADE’s old data may not be an isolated occurrence. “Big installations are more high-energy, but they don’t replace data taken at lower energy levels,” says Cristinel Diaconu, a particle physicist at DESY. “The reality is a lot of experiments done in the past are unique; they’re not going to be repeated at that energy.”

If anything, the need to better preserve...
particle physics data has grown more urgent in the past few years as CERN’s Large Hadron Collider (LHC) captured the world’s attention and a handful of other high-profile projects—BaBar at the SLAC National Accelerator Laboratory, Japan’s KEK collider, and the latest DESY experiments—wrap up work and prepare to disband. “In the past, experiments were smaller and more frequent. Now we build very big devices that cost a lot of money and person power over a number of years,” says Diaconu. “Each experiment is one application, built specifically for the task.” The LHC alone represents nearly a half-century of work, with 20 years invested in design and construction and 20 years of scheduled operation. There will never be another experiment like it.

The issue, experts say, isn’t data degradation. “The problem starts when the experiment is over, and the data used by one group of people is only understood by those people,” Diaconu says. “When they go off and do other things, the data is orphaned; it has no parents anymore.” The orphan metaphor only goes so far: After a certain point, orphaned data can’t be adopted by later researchers who weren’t part of the original team. Even given the raw data, only someone intimately involved in the original experiment can make sense of it. “The analysis is so complex that to understand the data you have to be there with it, working on the experiment,” says SLAC database manager Travis Brooks. “There’s a whole spectrum of things you need to keep around if you want petabytes [10¹⁵ bytes] of data to be useful.”

That spectrum includes everything from internal notes that explain the ins and outs of specific experiments. And then there’s the fuzzy-sounding “metadata,” the hacks and undocumented software tweaks made by a team in the midst of a project and then quickly forgotten.

Making it worse, particle physicists don’t usually share their data outside their collaborations the way most peer-reviewed scientists do. “We don’t publish the data, because it’s something like a petabyte—you can’t just attach the raw data in a ZIP file,” Brooks says. As a result, there’s been no incentive to find a standard format for the raw information that would be readable to outsiders.

A data librarian

To give shuttered experiments a future, the DPHEP working group is looking for ways to keep data in working order long after the original collaboration has disbanded. Typically, software that can make sense of the data is custom-made to run on servers that are optimized for the experiment and shut down when funding runs out. And the constant churn of technology can make software and storage media obsolete within a matter of years. “The data can’t be read if the software can’t be run,” Brooks says.

One option is to “virtualize” the software, creating a digital layer that simulates the computers the experiment was originally run on. With regular updates and maintenance, software designed to run on the UNIX machines of today could be rerun on the computers of the future the same way people nostalgically play old Atari games on new PCs, for example.

To capture and preserve the less tangible aspects of a particle physics experiment, the working group has suggested the job of data archivist. The archivist would be in charge of baby-sitting the data and standardizing the software used to read it, helping to justify huge investments in the big machines of physics by making data usable by future researchers or useful as a teaching tool. The idea has been endorsed by the International Committee for Future Accelerators, an advisory group that helps coordinate international physics experiments. DPHEP is also pushing data preservation among funding agencies, arguing that the physics experiments of the future should be designed with a data-preservation component to help justify their cost.

Diaconu admits that the idea has a way to go before it captures the minds of young physicists focused on publishing new data. “Some people say, ‘Can you imagine how boring, to sit and look at old data for 20 years?'” he says. “But look at a librarian. Part of their job is taking care of books and making sure you can access them.” A data archivist would be a mix of librarian, IT expert, and physicist, with the computing skills to keep porting data to new formats but savvy enough about the physics to be able to crosscheck old results on new computer systems.

The DPHEP group estimates that archivists—and the computing and storage resources they’d need to keep data current long after an experiment ended—would cost 1% of a collider’s total budget. That can be a hefty financial commitment: It would amount to $90 million for CERN. But keeping data in a usable form would provide a return on the investment in the form of later analyses, the group argues. Says Diaconu: “Data collection may stop, but it’s not true that’s the end of the experiment.”

—ANDREW CURRY
Andrew Curry is a freelance writer based in Berlin.