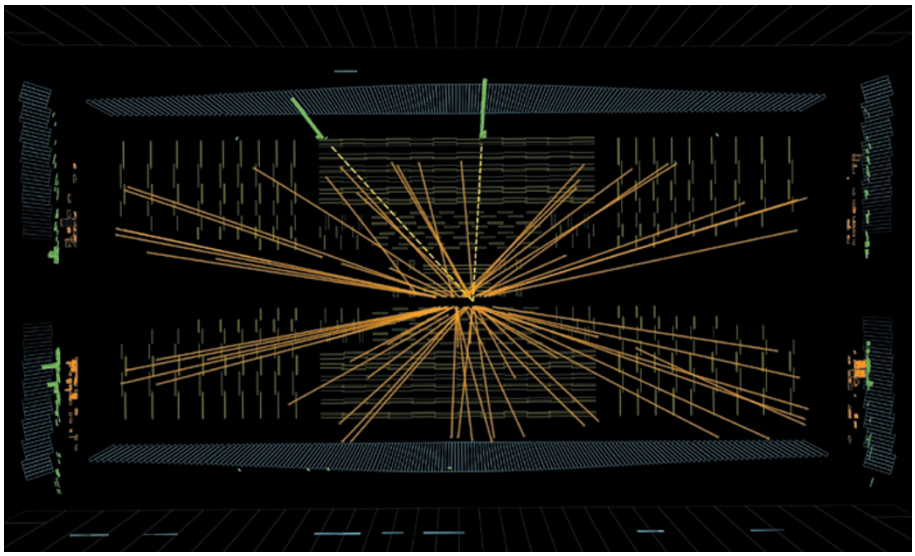

THOMAS MCCAULEY/LUCAS TAYLOR/CMS COLLECTION/CERN

Data from the Large Hadron Collider, such as this decay of a Higgs boson, could be made publicly available.

PHYSICS

LHC plans for open data future

Researchers share results to keep them accessible.

BY ELIZABETH GIBNEY

When the Large Hadron Collider (LHC) is humming along, the data come in a deluge. The four experimental detectors at the facility, based at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, collect some 25 petabytes of information each year.

Storing the data is not a problem: hard drives are cheap and getting cheaper. The challenge is preserving knowledge that is less commonly stored — the software, algorithms and reference plots specific to each experiment. These often degrade or disappear with time, says Cristinel Diaconu of the Marseilles Centre for Particle Physics in France, who is chair of the international Data Preservation in Long Term Analysis in High Energy Physics (DPHEP) study group. He worries that if the data continue to be stored in their current state, physicists trying to decipher them in 10 years' time will be unable to reconstruct the discovery of the Higgs boson. "When the LHC programme comes to an end, it will probably be the last data at this frontier for many years," he says. "We can't afford to lose it."

The DPHEP is therefore trying to push data-preservation efforts from mere storage to a system of open sharing. The thinking goes that data and the knowledge needed to interpret them are more likely to survive in the long term if many people outside an experiment are constantly trying to make sense of them.

Kati Lassila-Perini, a physicist at the Compact Muon Solenoid (CMS), one of the four experiments at the LHC, has a radical idea for this sort of sharing: giving data away to school pupils. Next year, a pilot scheme she leads will release 2010 CMS data, which the IT Center for Science in Espoo, Finland, will reformat and store. The centre will then share the data with pupils, who will recreate plots of particle decays using analysis tools adapted for the public. The CMS plans to make more data publicly available a few years after collection, and Lassila-Perini hopes that other data centres will adopt such schemes. "We are guaranteeing that the data we are not looking at any more remain accessible," she says.

The intent is not just to keep data for posterity. Old data can be mined to test new theories and provide crucial references for new experiments, says Diaconu. Before the Higgs boson was discovered in 2012, for example, the Large Electron–Positron collider — the LHC's predecessor at CERN — came back into the spotlight as physicists scoured its 1990s-era data, looking for an exotic type of Higgs that had not been theorized at the time the data had been gathered. In this way, the goals of keeping data alive and open are "enlightened self-interest", says Michael Hildreth, a physicist at the University of Notre Dame in Indiana and leader of the US-funded Data and Software Preservation for Open Science (DASPOS) effort, which has similar goals to the DPHEP.

DASPOS is building a template for preserving data — a checklist of items that should be stored, and how to do it. Next year, in a 'curation challenge', DASPOS will task physicists with recreating results from other experiments using only the information collected with this template. One test will almost certainly use LHC data — challenging, for example, CMS physicists to recreate results from the rival ATLAS experiment. Another test could come from a different field, such as astrophysics. If successful, the model could form a generic and simplified architecture for preserving data, says Hildreth.

Part of the challenge is coping with ever-changing algorithms, operating systems and data-analysis hardware. At the German Electron Synchrotron (DESY) in Hamburg, computing coordinator David South is leading a project that is already attempting to protect data in this way. His team has devised a system that will automatically comb through data and software from experiments on DESY's Hadron–Electron Ring Accelerator and test them for compatibility when hardware or operating systems change.

This plan to migrate data repeatedly onto new platforms stands in contrast to an approach at the BaBar experiment at the SLAC National Accelerator Laboratory in Menlo Park, California. There, versions of data and the operating systems needed to analyse them have been frozen in storage centres, where they are supposed to be accessible until at least 2018. South says that DESY's approach is more reliable. Although DESY's system needs monitoring — any incompatibilities must be fixed through human intervention — the goal is to deal with problems as they arise, rather than tackle them years later, when they may have compounded.

"When the LHC programme comes to an end, it will probably be the last data at this frontier for many years. We can't afford to lose it."

DESY scientists would know about that. In the 1990s, physicists wanted to take another look at data from a DESY collider that ran from 1979 to 1986, to further investigate the strong interaction that binds quarks together. They managed to measure it with increased precision, but Diaconu says that it took two years to reconstruct the data, which had not been maintained.

The data preservationists are quick to point out the expense associated with reconstruction efforts. Of course, preservation also costs money, but it is well worth it, says DPHEP project manager Jamie Shiers. He puts the bill for implementing good data-preservation at the LHC at around 1% of operating costs — just a few million dollars per year. "I think it's justified," he says. ■