GCN Current issue | ▼ Subscribe | ▼ Blogs | Events | eSeminars | White papers | RSS / XML

June 19, 2006, Volume 25, No. 16 issue

Modern Relics

NIST and others work on how to preserve data for later use

By William Jackson, GCN Staff

“One of the big challenges of long-term archiving is deciding what to keep for how long. We’re not doing a very good job on that now. There’s a long way to go.” JOSH LUBELL, NIST

Image: Rick Steele

It’s a threat to any agency data-sharing initiative. It can render a knowledge management system obsolete. And in the future, it could mean the difference between a successful space launch and a trip back to the drawing board. It is, simply, data loss. Or data degradation. Or even data alteration. As more of what government creates in support of its mission is rendered digitally, experts must struggle with questions of how to ensure that digital data is available—and reliable—for future users.

The National Institute of Standards and Technology hosted a workshop in March on long-term knowledge retention, looking for answers to the questions of what digital data government, industry and academe should be saving, and how it should be saved.

They came up with no immediate answers, but they confirmed that a problem certainly exists and it’s growing, said Josh Lubell, a computer scientist in NIST’s Manufacturing Engineering Laboratory.

According to estimates offered at the workshop, the world churns out enough digital data to fill the Library of Congress every 15 minutes. Much of that is of no interest to anybody and can be discarded quickly. But in areas such as engineering, the production of information is outpacing our ability to ensure it will be available to those who may need and want to share it later on. As computer-aided evolves into computergenerated, terabytes of data are disappearing, or becoming inaccessible or corrupted every day.

“So much information is digital, and people are feeling the pain of losing access to their information,” Lubell said.

The attendees at the workshop agreed on the need to establish a business case for long-term data archiving and to develop standards for ensuring interoperability of data across time as well as across hardware and software platforms.

Specifically, engineering and design drawings are routinely saved today, said Doug Cheney, an interoperability consultant and product director for ITI TranscenData of Milford, Ohio. But unlike blueprints of the past, computer-aided design drawings, which are becoming increasingly important elements of geospatial initiatives, represent only the tip of the design iceberg.

“Digital data is not hardened, especially 3-D geometry,” Cheney said. “It is very interpretive, so it’s easy for unexpected things to creep in.”

The problem is compounded, because changes that creep in when data is moved from one platform to another often are not conspicuous.

“It’s what we don’t know that will kill us,” Cheney said.

This dilemma came to light in the aftermath of an industrial accident in Green Bay, Wis., when a pipe feeding a boiler exploded, scalding two men to death. The immediate problem was corrected, but in the next few years the company involved changed hands several times, said Crispin Hales, an engineering forensics investigator in Winnetka, Ill. In the process, a lot of institutional knowledge was lost.

“They almost had a repeat of the failure, and they had no records of what had happened only five years before,” Hales said.

Tacit knowledge goes missing

The problem is the disappearance of what Hales called tacit knowledge. Forty years ago, organizations were stable, people remained on the job for years and there was a community of knowledge that could be tapped.

“That is gone,” Hales said. It has been replaced by a more fluid environment, where access and exchange of complex data has replaced knowledge.

At the same time, data has become more complex. Too often that data, if it was retained, is not reliable because the systems used to generate it are no longer available.

“The government has this problem in spades,” said William Regli, associate professor of computer science at Drexel University.

Large engineering projects by the Defense and Energy departments generate huge amounts of data on systems that quickly become obsolete.

NASA’s space shuttle program, one of the most complex engineering projects in history, kicked off in 1981 and the last shuttle is not scheduled to be retired until 2009. The B-52 Stratofortress, for years the backbone of the Air Force’s nuclear armada, first flew in 1954, and its operational life span is expected to extend to 2040.

The preservation and maintenance of electronic data is a problem many agencies are wrestling with, including the National Archives and Records Administration and the Library of Congress. The library has for several years been digitally preserving information from other media, from audio recordings to parchment manuscripts.

Now, under the National Digital Information Infrastructure and Preservation Program authorized by Congress in December 2000, it’s working with other federal agencies and private-sector organizations to develop a national strategy for collecting, archiving and preserving the growing volume of materials created only in digital formats [GCN.com, Quickfind 584].

But NIST and others are particularly concerned about preserving the engineering data that goes into missions throughout government. The Library of Congress and NARA funded some of Regli’s work at Drexel University on digital preservation. “I got into the engineering side of it about five years ago,” he said. “The digital CAD process has a wealth of information that isn’t saved in any meaningful way.”

Engineering data had become too complex for humans to process on their own, he said. In the recent past, drawings produced by a draftsman with pen and paper were based on a human level of knowledge. Computer-aided designs are based on levels of complexity and assumptions that are not reproduced when the final drawings are printed or displayed. These often ephemeral assumptions and calculations can be as important as the drawing.

Beyond our comprehension

CAD systems today deal with higher orders of geometry beyond our comprehension, Cheney said.

“You can’t write an equation for the geometry,” he said. “The geometry is closely tied to the software that interprets it.” That software changes through a design’s lifecycle as it moves from one platform to another.

“Each time it is produced, it’s slightly different,” he said.

One solution offered now by ITI TranscenData is a CAD add-on tool that analyzes potential shape and fit problems in CAD models as they are translated between systems.

Industry is aware of the problem of design drift and corruption, but the solutions so far are mostly industry- or even company-specific.

“There is a lot of redundant effort in American industry,” Regli said. This translates into wasted effort and interoperability problems. And it does not address the question of long-term retention as systems become obsolete. “How do you ensure interoperability across time?”

These issues often get little attention in the private sector because they don’t produce revenue.

“The ultimate beneficiary is not the current business unit, but some future entity,” Regli said.

Current shareholders often are reluctant to pay for things future shareholders will bene- fit from, Lubell said. That is why government is getting involved in developing technical standards for longterm preservation. But government will not be able to solve the problems by itself.

“The business case has to be made first,” to convince people to use standards, Lubell said. “It’s a challenging area because it’s not dealing with an immediate issue,” Regli said.

Standards today

There are some standards in place, most notably the International Standards Organization’s Standard for the Exchange of Product Model Data. But none answer all the needs for interoperability with long-term integrity and availability.

Within government, NIST’s Manufacturing Engineering Lab has an exploratory program for developing data standards, and the IT Laboratory has a digital preservation program that is investigating how digital data degrades over time.

Efforts are under way by industry groups to extend STEP for archiving quality by adding geometry exchange specifications, and to include definitions for validating accuracy when data is exported. The European aerospace industry has begun a standards process for archiving 3-D images that at least some of the U.S. aerospace industry have joined.

Once adequate standards exist, the question will remain, what needs to be saved?

As storage becomes cheaper and access quicker, Regli said one philosophy is: “Why not just store everything?” But years down the road, that approach could produce the equivalent of digital archeology, as engineers and researchers sift through terabytes of data searching for the nugget they need.

“You have to strike a balance,” Lubell said. “One of the big challenges of longterm archiving is deciding what to keep for how long. We’re not doing a very good job on that now. There’s a long way to go.”

More news on related topics: Storage Management