Preserving the Nation's Endangered Digital Information

As the digital world evolves, the quantity of information churned out in service of science, medicine, scholarship, private enterprise, personal interests, and the entire social infrastructure is almost beyond comprehension. According to UCSB's internationally recognized geographic information systems expert Michael Goodchild, geospatial data alone is acquired at an annual global rate of one petabyte—the equivalent of 2^{50} (1 quadrillion) bytes, or about a billion megabytes.

Storage is one problem—not because technical solutions can't be designed, but because digital data is fragile, especially "born-digital" information that has no paper counterpart. The Library of Congress has noted that masses of historically significant digital content—from aerial photography and public-television programs to opinion polls and voting records—are at risk of becoming lost or altered. Millions of items have already disappeared.
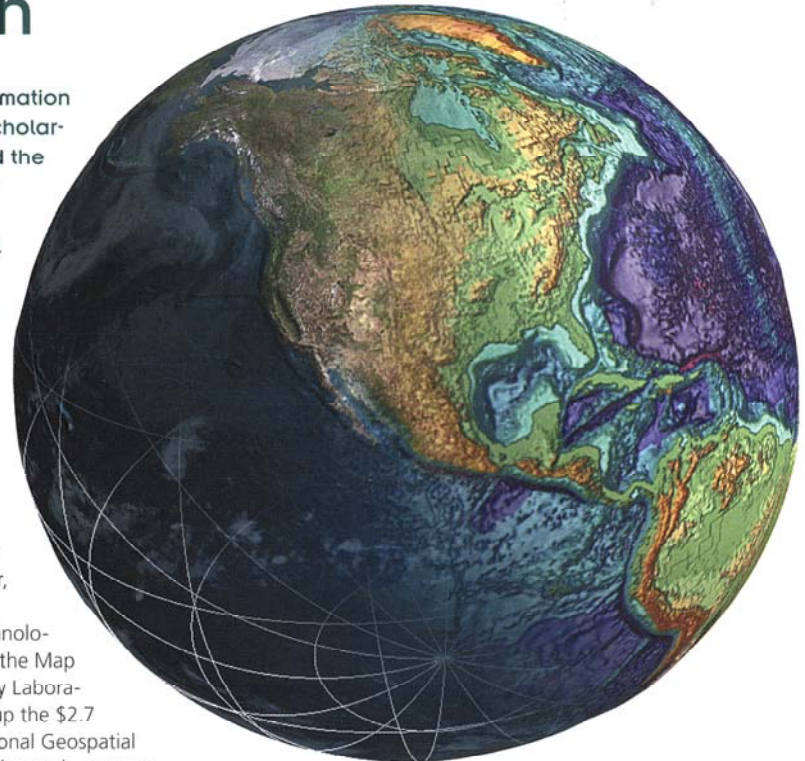
"Hard drives, mag tape, digital tape, old floppies, aging CDs—all have problems," says University Librarian Sarah Pritchard. "This stuff is even worse than paper for durability. Recopying is one familiar strategy, but when information is transferred from one format to another, errors are likely to be introduced."

To identify, collect, and preserve the nation's digital heritage for the long term, the Library of Congress (LC) has awarded nearly $14 million in matching funds to eight lead institutions and partners through its National Digital Information Infrastructure

Global view showing land elevation and ocean-bottom depth.

and Preservation Program. Pritchard and co-principal investigator Larry Carver, director of Library Technologies and of the Map and Imagery Laboratory, head up the $2.7 million National Geospatial Digital Archive node, partnering with Stanford University. Carver was operational architect of the now-thriving Alexandria Digital Library project, a foundational repository of geospatial data initially supported by a five-year grant from the National Science Foundation. Pritchard was instrumental in making the ADL permanent and positioning it to support these more advanced digital library initiatives.

UCSB's focus in the LC project is on designing an innovative infrastructure and collecting born-digital geospatial materials ranging from satellite imagery to websites to other mapping content from university, government, and corporate resources. "Since technology 50 or 100 years from now will be so different, we have to either preserve documents with their software or preserve enough information about the document's creation to recreate it with different software in the future," Pritchard says. Preserving such metadata for every digital object is a tremendous challenge: Imagine binary data in volumes amounting to yottabites (2^{80}—1 septillion—bytes, or a million trillion megabytes), all of which must be verified after successive conversions.

In addition to preserving over 15 terabytes of geographic content, another task is building software that automatically takes materials into the archive, so that users can contribute to and access deeply layered information without a bottleneck. Still another need is building an automated registry of the hundreds of data formats in order to move data from one technology to another—together with fully documented descriptions of the data types and content so a programmer could accurately reassemble the information and put the object back together.

Adds Carver: "We will decouple what the archive contains from the interfaces that sit on top and the storage that fits behind it. Storage can come and go, and people can write new interfaces to better access and analyze from the top—and yet the content, the format registry, and the documentation will be migrated along the way without having to reinvent it all the time."
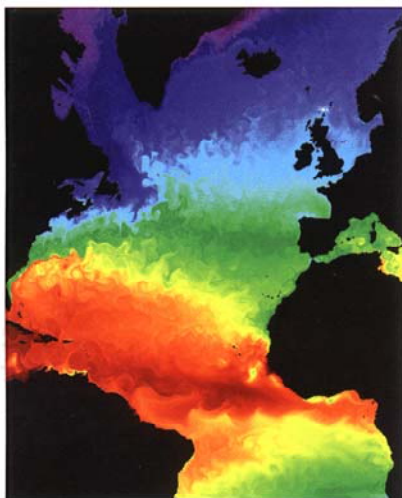
Color-coded water-surface temperatures in the Atlantic Ocean.