Science | DOI:10.1145/2133806.2133811 Gregory Goth

Preserving Digital Data

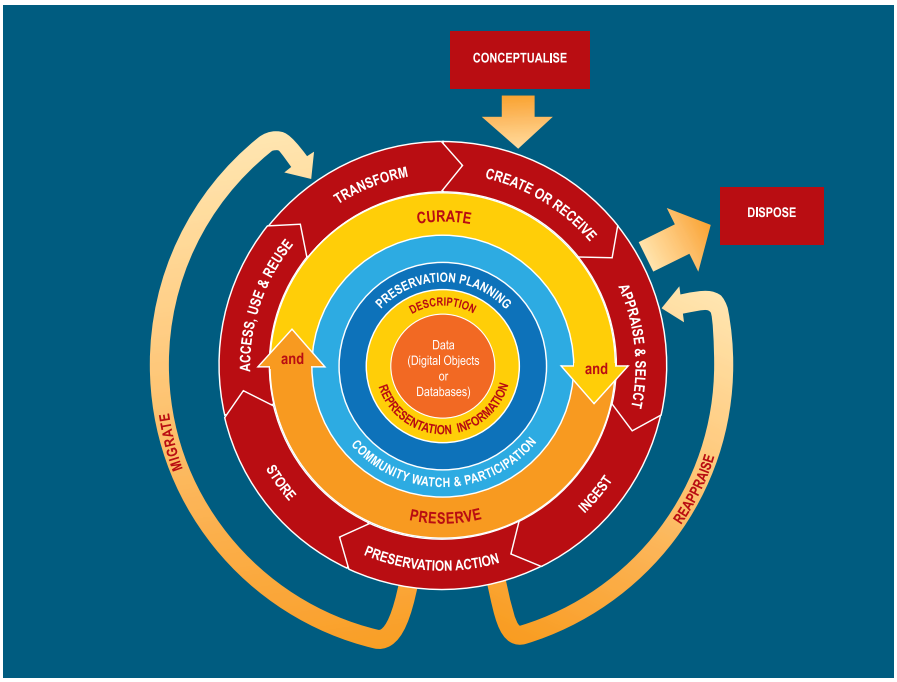
Scientific data is expanding at an unprecedented rate. While new tools are helping preserve this data, funding must be increased and policy coordination needs improvement.

DATA PRESERVATION EXPERT Jeff Rothenberg says digital data has a behavior problem. Rothenberg does not mean digital artifacts are purposely causing mayhem across platforms and applications, but rather the questions of how to make these artifacts useful across space and time are running into perception issues.

For centuries, Rothenberg says, archivists could properly do their jobs by preserving static artifacts such as printed pages and photographs for even the most complex undertakings of a society, but that has irrevocably changed.

On the far end of the spectrum from such "archaic" forms as paper, stone, and wood "is what I call inherently digital artifacts," says Rothenberg, "something that really requires a computer to even render it, and increasingly we are producing things like that. Ph.D. theses, particularly in the sciences, are becoming inherently digital, including not just text but models, which show behavior and are part of the result of the thesis. The real crux of preserving this data is how to preserve the behavioral aspects of artifacts, which generates the image in the first place."

Agencies, researchers, and technologists are beginning to seriously address this crux, but the sheer number of agencies and efforts tackling it also seem to highlight the enormity of the task.



In the U.K., the Digital Curation Center's Curation Lifecycle Model provides a graphical, high-level overview of the necessary stages for the curation and preservation of data.

FIGURE COURTESY OF THE DIGITAL CURATION CENTRE/UNIVERSITY OF EDINBURGH

Data Discovery and Preservation

Perhaps the foremost irony facing those dealing with scientific data preservation is the entrenched cultural emphasis on the next discovery and a less appreciative attitude toward old data.

"Scientists are living in a world that is spinning around very fast, writing papers, publishing results, and going on to the next project, and data is not preserved," says Stefan Winkler-

Nees, program officer in the Department of Scientific Library Services and Information Systems at the German Research Foundation (DFG), the country's largest funder of research. "The big 'A-ha!' effect in Germany was that somebody realized we are losing up to 90% to 95% of the data produced in public science over time because it is not accessible anymore."

Winkler-Nees's observations are echoed by others, including Rothenberg, Gary King, director of the Institute for Quantitative Social Science at Harvard University, and Sylvia Spengler, program director in the U.S. National Science Foundation's Computer Science and Engineering directorate.

The biggest challenge facing data preservationists, King says, is "the infrastructure to make it all happen. It's straightforward to get grants to do research. The agencies that could provide the funding tend to provide short-term funding—primarily for research, sometimes to build infrastructure, but virtually never for preservation. Preservation is the promise of keeping things in perpetuity. That's a long time. Figuring out that is really hard."

"The kinds of domains which do think about long-lived collections have an inevitable tension between the acquisition of the data and the preservation of the data," Spengler says. "And the funds for that, at least currently, are all within one budget. So it has to be an active decision within a community how to deal with the preservation issues for that community."

However, these existing cultural models are being challenged somewhat by more comprehensive national initiatives. In Germany, the DFG recently funded 28 projects with €9.9 million under the aegis of data accessibility, including preservation; in the U.S., the National Science Foundation published requirements effective January 2011 that every funded project must include a two-page data management plan, including retention periods, on a directorate-specific basis; and in the U.K., JISC (originally the Joint Information Services Committee) has long coordinated data management and preservation policies among the nation's funders and research institutions.

Recent research into the efficacy of archived data also indicates a substantial return on investment. Researchers receiving support from the Dryad data repository project estimated the repository could ingest and curate 10,000 publications annually for about U.S. $400,000—and that "a $400,000 investment would contribute to more than 1,000 papers within four years, far greater than the accepted value of a research dollar."

Yet there is little to no national level coordination of data preservation standards, even in the U.K.'s well-documented research council guidelines. The U.K.'s Digital Curation Center, for example, links to each of the country's nine primary research funders' policies; some include very amorphous wording on data preservation while others stipulate a 10-year retention period.

Simon Hodson, program manager for digital infrastructure at JISC's Managing Research Data program, says funders are well aware of the potential for retaining and repurposing data in future experiments, but that there must be some discipline-specific autonomy in setting policy.

"That's going to have to go forward on a case-by-case basis," Hodson says. "You can't keep everything, and I think researchers and institutions have to be in a position where they can make these decisions on what they retain and what they throw away."

Rothenberg says a deep cultural divide between scientists and archivists must be addressed to forge a consensus on how this data should be preserved and curated.

"Scientific data archives are not archives in the same sense as a national archive," he says. "People may have been trained in archival studies, but they don't have the same mind set [as scientists]. Each of those communities has, to some extent, its own history and philosophy and vision. Those may be realigning under the mandate of digital convergence, but it's a slow process."

Gaining Momentum

Fitful as progress may be, the global effort at addressing data preservation is gathering momentum. The International Organization for Standardization (ISO) is actively pursuing a Digital Preservation Interoperability Framework specification; the JISC-sponsored *Keeping Research Data Safe* report, published in two parts in 2008 and 2010 respectively, features numerous best practice recommendations; Rothenberg and RAND Europe researcher Stijn Hoorens in 2010 co-authored a comprehensive report, which was sponsored by the British Library, that explored the library's possible role in preserving scientific, technical, and medical data; and the White House Of-

Technology

Africa Increases Its Internet Usage

Africa's Internet usage has grown dramatically during the last decade due to new information and communication technologies, including improved fiber-optic networks and increased availability of computers and mobile phones.

Africa had nearly 140 million Internet users by the end of 2011, compared to only 4.5 million users at the end of 2000, according to Internet World Stats. The continent's Internet penetration is 13.5%, compared to 26.2% in Asia, 35.6% in the Middle East, 39.5% in Latin America and the Caribbean, 61.3% in Europe, and 78.6% in North America.

The African nations with most Internet users are Nigeria (45 million), Egypt (21.7 million), Morocco (15.7 million), Kenya (10.5 million), and South Africa (6.8 million). Nigeria's 29% Internet penetration is slightly below the global average of 32.7%.

After Google, Facebook is the most popular Web site in Africa. Egypt boasts the largest number of Facebook users (9.4 million), followed by South Africa (4.8 million), Nigeria (4.4 million), Morocco (4.1 million), and Kenya (1.3 million). Africa's Facebook penetration is 3.6%, compared to 4.7% in Asia, 8.4% in the Middle East, 25.5% in Latin America, 27.4% in Europe, and 50.3% in North America. Egypt's 11.4% Facebook penetration is almost equal to the global average of 11.5%.

—*Jack Rosenberger*

fice of Science and Technology Policy released a Request for Information, soliciting comments regarding public access to digital data resulting from federally funded research in November 2011.

Also, some of the research funded under NSF's Sustainable Digital Data Preservation and Access Network Partners (DataNet) program is beginning to bear significant fruit. Although Spengler says the DataNet projects were and are intended to be exemplars and fairly restrictive prototypes due to limited funding, her NSF colleague Rob Pennington says DataNet awardees are working with other researchers eager to find ways to share data across domains and disciplines.

One standout example of this is iRODS (integrated Rule-Oriented Data System), developed by the Data Intensive Cyber Environments (DICE) research group at the University of North Carolina (UNC) and the University of California, San Diego. Institutions spanning disciplines from climatology to social sciences are adopting the innovative data grid tool. The NSF awarded iRODS developers $8 million in September 2011 to build a policy-driven national data management infrastructure, motivated by the discrete data management requirements of the NSF's Ocean Observatories Initiative, NSF's Consortium of Universities for Advancement of Hydrologic Science, engineering projects in education, CAD/CAM/CAE archives, the genomic databases of the iPlant collaborative, the H.W. Odum Institute for Research in Social Science at UNC, and NSF's Science of Learning Centers.

iRODS has also been adopted by scientific data centers worldwide, including astronomical observatories in Canada and France, climate centers in the U.S., and at the Sanger Institute genomics databases in the U.K. It is also in use in the U.S. National Archives and Records Administration's Transcontinental Persistent Archives Prototype.

iRODS is the successor to the pioneering Storage Resource Broker (SRB) architecture. DICE Director Reagan Moore says iRODS's rule engine-based architecture makes the distributed management of a data grid much simpler than the hard-coded SRB architecture, can serve as the sort of reporting tool that demonstrates researchers are meeting their mandated data management plan outlines—and is also the sort of policy-based system that melds the concepts of data management and data preservation for whatever duration is required.

There is little to no national-level coordination of data preservation standards, even in the U.K.'s well-documented research council guidelines.

"We make the assertion that any data-management application really consists of the policies you're applying in order to validate assertions about what you've done," Moore says. "So if I build a preservation environment, my policies are related to authenticity, integrity, chain of custody, and original arrangement."

Moore says the rule-based iRODS architecture makes it possible for users to tailor which policies apply to a given action without having to rewrite any code, whereas server-side commands in SRB were hard coded. These rules are applied in a platform-agnostic manner through any number of 254 microservices selected as pertinent by any user community.

At least one project has already successfully replicated a recognized sample archive, the Harvard IQSS-developed Dataverse, using iRODS. Researchers at UNC's Odum Institute performed a Dataverse-to-iRODS transfer using the Open Archives Initiative Protocol for Metadata Harvesting and the compatible Data Documentation Initiative standard, plus XML.

"The result is an accurate copy of a Dataverse archive inside iRODS," according to the UNC authors, "which data grid administrators can preserve over the long term by, for example, replicating the information to many geographically distributed storage resources."

Phil Butcher, head of information technology at the Sanger Institute, says the organization's iRODS installation has run smoothly. He believes funding agencies should make themselves aware of the details of such groundbreaking technologies, even if comprehensive national and international management and preservation policies are not possible.

"Unless you start to deploy tools like this, you'll be in trouble," Butcher says. "The funding bodies in particular have a real opportunity to understand a bit more about the technology. If there is a list of two or three tools people could use, the best would come to the surface. We're getting to the point where some of these decisions have to be made. Otherwise a lot of groups will be in a lot of trouble." C

Further Reading

Beagrie, N., Lavoie, B., and Woollard, M.
***Keeping Research Data Safe 2*, JISC, Bristol, U.K., April 2010..**

Chiang, G., Clapham, P., Qi, G., Sale, K., and Coates, G.
Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute, *BMC Bioinformatics 12*, 361, Sept. 9, 2011.

Neuroth, H., Strathmann, S., and Vlaeminck, S.
Digital preservation needs of scientific communities: The example of Göttingen University, *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, Aarhus, Denmark, Sept. 14–19, 2008.

Rothenberg, J. and Hoorens, S.
***Enabling Long-term Access to Scientific, Technical and Medical Data Collections*, RAND Europe, Cambridge, U.K., 2010.**

Ward, J.H., de Torcy, A., Chua, M., and Crabtree, J.
Extracting and ingesting DDI metadata and digital objects from a data archive into the iRODS extension of the NARA TPAP using the OAI-PMH, 5th IEEE International Conference on e-Science, Oxford, UK, Dec. 9–11, 2009.

Gregory Goth is an Oakville, CT-based writer who specializes in science and technology.