

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

12/08 VOL.51 NO.12

Surviving the Data Deluge

Open Information
Extraction
from the Web

CTOs on
Virtualization

Living Machines

High-Performance
Web Sites

DOI:10.1145/1409360.1409376

Tools for surviving a data deluge to ensure your data will be there when you need it.

BY FRANCINE BERMAN

Got Data? A Guide to Data Preservation in the Information Age

Imagine the modern world without digital data—anything that can be stored in digital form and accessed electronically, including numbers, text, images, video, audio, software, and sensor signals. We listen to digital music on our iPods, watch streaming video on YouTube, record events with digital cameras, and text our colleagues, family, and friends on BlackBerrys and cell phones. Many of our medical records, financial data, and other personal and professional information are in digital form. Moreover, the Internet and its digital products have become our library, shopping mall, classroom,

and boardroom. It is difficult to imagine the information age without unlimited access to and availability of the digital data that is its foundation.

Digital data is also fragile. For most of us, an underlying assumption is that our data will be accessible whenever we want it. We also regularly confront the fallacy of this assumption; most of us (or our friends) have had hard drives crash with the loss of valuable information or seen storage media become obsolete, rendering information unavailable (think floppy disks). Loss, damage, and unavailability of important digital business, historical, and official documents regularly make the news, further highlighting our dependence on electronic information.

As a supporting foundation for our efforts in the information age, digital data in the cyberworld is analogous to infrastructure in the physical world, including roads, bridges, water, and electricity. And like physical infrastructure, we want our data infrastructure to be stable, predictable, cost-effective, and sustainable. Creating systems with these and other critical characteristics in the cyberworld of information technology involves tackling a spectrum of technical, policy, economic, research, education, and social issues. The management, organization, access, and preservation of digital data is arguably a “grand challenge” of the information age.

As a society, we have only begun to address this challenge at a scale concomitant with the deluge of data available to us and its importance in the modern world. This article explores the key trends and issues associated with preserving the digital data that is the natural resource of the information age and what's needed to keep it manageable, accessible, available, and secure. (For common terms associated with digital data management and preservation, see the sidebar “Digital Data Terms and Definitions.”)

Data Cyberinfrastructure

The supporting framework for digital

PHOTOGRAPH BY THOMAS HERERICH

data is an example of cyberinfrastructure—the coordinated aggregate of information technologies and systems (including experts and organizations) enabling work, recreation, research, education, and life in the information age. The relationship between cyberinfrastructure in the cyberworld and infrastructure in the physical world was described in the U.S. National Science Foundation’s 2003 *Final Report of the Blue Ribbon Advisory Panel on Cyberinfrastructure*, commonly known as the “Atkins Report”² after its Chair, Dan Atkins: “The term *infrastructure* has been used since the 1920s to refer collectively to the roads, power grids, telephone systems, bridges, rail lines, and similar public works that are required for an industrial economy to function. Although good infrastructure is often taken for granted and noticed only when it stops functioning, it is among the most complex and expensive things that society creates. The newer term *cyberinfrastructure* refers to infrastructure based upon distributed computer, information, and communication technology. If infrastructure is required for an *industrial* economy, then we could say that cyberinfrastructure is required for a *knowledge* economy.”

The implication of the report is that like infrastructure in the physical world, data cyberinfrastructure, or data CI, should exhibit critical characteristics that render it useful, usable, cost-effective, and unremarkable. The innovation, development, prototyping, and deployment of CI with such

characteristics constitute a massive endeavor for all sectors, including the academic sector.^{1,3}

What are the components of data CI? In the research and education community, users want a coordinated environment that manages digital data from creation to preservation, accommodates data ingested from instruments, sensors, computers, laboratories, people, and other sources, and includes data management tools and resources, data storage, and data use facilities (such as computers for analysis, simulation, modeling, and visualization). Users want to store and use their data for periods spanning the short-term (days) to the long-term (decades and beyond), and they want it to be available to their collaborators and communities through portals and other environments. Figure 1 outlines the portfolio of coordinated components that constitute the data CI environment at the San Diego Supercomputer Center (www.sdsc.edu/). Such environments must be designed to meet the needs of the target user community while being continually maintained and evolved to support digital data over the long term.

Trends

A 2008 International Data Corporation (IDC) white paper sponsored by EMC

Corporation⁵ described the world we live in as awash in digital data—an estimated 281 exabytes (2.25×10^{21} bits) in 2007. This is equivalent to 281 trillion digitized novels but less than 1% of Avogadro’s number, or the number of atoms in 12 grams of carbon (6.022×10^{23}). By IDC estimates, the amount of digital data in our cyberworld will surpass Avogadro’s number by 2023.⁵ Even if these estimates are off significantly, storing, accessing, managing, preserving, and dealing with digital data is clearly a fundamental need and an immense challenge.

The development of data CI is greatly affected by both current and projected use-case scenarios, and our need to search, analyze, model, mine, and visualize digital data informs how we organize, present, store, and preserve it. More broadly, data CI is influenced by trends in technology, economics, policy, and law. Four significant trends reflect the larger environment in which data CI is evolving:

Trend 1. More digital data is being created than there is storage to host it. Estimates from the IDC white paper



indicate that 2007 marked the “cross-over” year in which more digital data was created than there is data storage to host it. At that point, the amount of digital data (information created, captured, or replicated in digital form) exceeded the amount of storage (all empty or usable space on hard drives, tapes, CDs, DVDs, and volatile and nonvolatile memory). At the crossover point, this amount was estimated to be around 264 exabytes (264×10^{18} bytes).⁵ This is almost a million times the amount of digital data hosted in 2008 by the U.S. Library of Congress (www.loc.gov/library/libarch-digital.html) and more than 20,000 times the aggregate of permanent electronic records projected to be stored in 2010 by the U.S. National Archives and Records Administration (www.archives.gov/era/). The IDC report further projected that by 2011 the amount of digital information created will be nearly 1.8 zettabytes (1.8×10^{21}), or more than twice the amount of available storage, estimated at 800+ exabytes.

The methodology under which these estimates were derived (what is counted and how it is calculated⁶) is fascinating and has generated considerable community discussion. However, even under alternative variations of the IDC methodology, the trend is unmistakable: We do not produce storage capacity at the same rate we produce digital information. Even if we wanted to, we cannot keep all of our digital data.

The thoughtful and methodical selection of which data sets to keep (called “appraisal” in the archival world) will be critical to communities used to keeping it all. In the research and education community, methods for community appraisal (coupled with the need for budgets to ensure adequate data stewardship and preservation for selected data sets) will likewise be more important over the next decade.

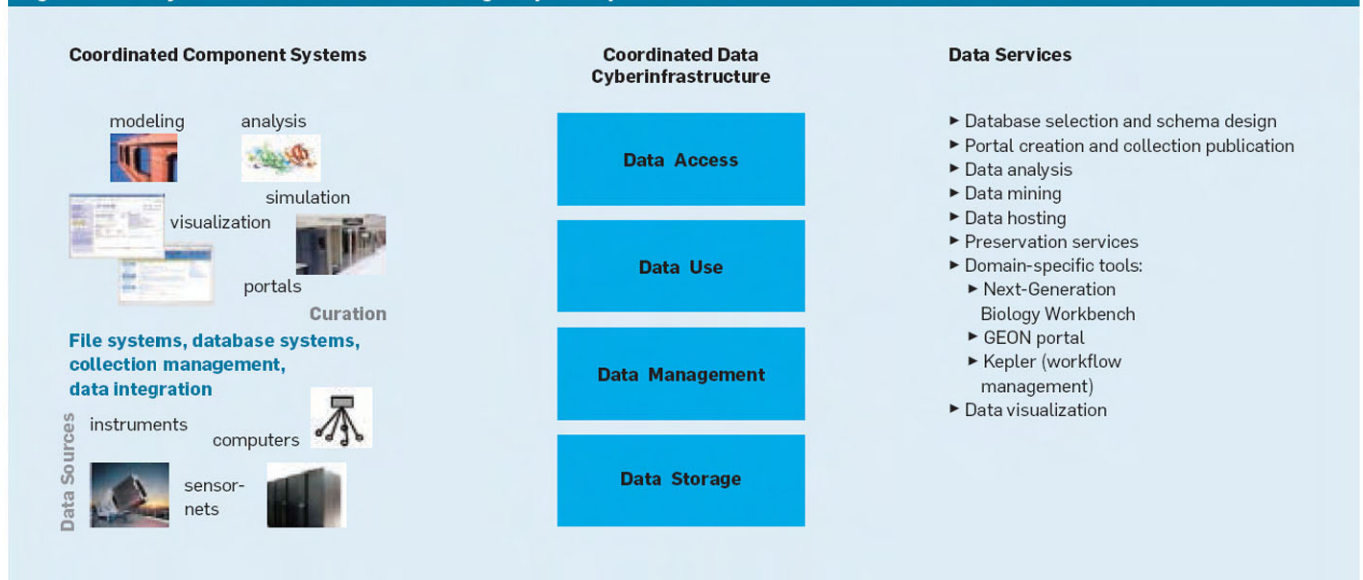
The need for community appraisal will push academic disciplines beyond individual stewardship, where project leaders decide which data is valuable, which should be preserved, and how long it should be preserved (except where regulation, policy, and/or publication protocols mandate specific stewardship and preservation timeframes). Some communities are beginning to develop explicit appraisal criteria and community stewardship models for valuable reference data collections (such as the Protein Data Bank, www.rcsb.org/pdb/home/home.do, in the life sciences and the Panel Study of Income Dynamics, psidonline.isr.umich.edu/Guide/, in the social sciences). Over the next decade, as more data is generated and the costs of data CI are incorporated into the “IT bill” at our institutions and enterprises, we can expect to devote more time and attention to the criteria and process through which we appraise data for stewardship and preservation.

Trend 2. More and more policies and

regulations require the access, stewardship, and/or preservation of digital data. Even before the information age, the Copyright Clause (Article 1, Section 8) of the U.S. Constitution and subsequent regulation set the stage for policy with respect to the rights and dissemination of information in the U.S. Today, many forms of digital rights management and a broad range of public policies govern the access, stewardship, and preservation of digital data around the world. In the U.S., the Sarbanes-Oxley Act of 2002 promotes appropriate responsible management and preservation of digital financial and other records for publicly owned companies, and the Health Insurance Portability and Accountability Act of 1996 ensures the privacy of digital medical records. On the research front, investigators at the U.S. National Institutes of Health are required to submit digital copies of their publications to PubMed Central (pubmedcentral.nih.gov/), and the U.S. National Science Foundation’s data-sharing policy “expects its awardees to share results of NSF-assisted research and education projects with others both within and outside the scientific and engineering research and education community.”¹⁰

Increased emphasis on the access, preservation, and use of digital materials is not limited to the U.S. For example in the U.K., the Joint Information Systems Committee (www.jisc.ac.uk/) and the British Library (www.bl.uk/npo)

Figure 1: Data cyberinfrastructure at the San Diego Supercomputer Center.



have been leaders in data curation, access, and preservation issues. DigitalPreservationEurope (www.digitalpreservationeurope.eu) in the E.U., the National Library of Australia (www.nla.gov.au/policy/digpres.html), Koninklijke Bibliotheek (the National Library of The Netherlands, www.kb.nl/index-en.html), and others around the world are contributing to an increasing body of knowledge and infrastructure to support data preservation and access for efforts enabled by technology.

The digital data generated by research, industry, and governments over the next decade will be subject to increased regulation and evolving community formats, standards, and policies. This means the CI developed to host and preserve it will need to incorporate mechanisms to enforce community policies and procedures like auditing, authentication, monitoring, and association of affiliated metadata. (As an unconventional example, think tagging of Facebook and Flickr photos.) Emerging data CI and management environments and systems, including IRODS (www.irods.org/), LOCKSS (www.lockss.org/lockss/Home), the Fedora Commons (www.fedora-commons.org/), and D-Space (www.dspace.org/), are beginning to develop and incorporate mechanisms that implement relevant policies and procedures. Over the next decade, the ability to automatically address the requirements of policy and regulation will be needed to ensure that our data

CI empowers rather than limits us.

Trend 3. Storage costs for digital data are decreasing (but that's not the whole story). One of the most important trends affecting digital data is the decrease in price over time for a terabyte (10^{12} B) of data storage. According to IDC¹¹, a terabyte of “enterprise” storage was priced at roughly \$440,000 in 1997. A decade later, the price for a terabyte of enterprise storage averaged around \$5,400. In 2008, terabyte drives cost approximately \$200 (OEM cost). In addition, holographic memory and other new technologies promise even better performance per price unit.

With storage so affordable, one would expect the “data bill” of institutions and enterprises to be equivalently affordable. However, as storage costs decrease, critical components of the data bill (such as power, curation/annotation, and professional expertise) are not decreasing. Today’s companies and institutions are investing in enterprise data centers in locations selected to minimize the power bill. Google, Microsoft, and other technology companies spend billions on such data centers—the heart of their businesses—and the cost savings rendered through strategic placement can be immense. Storage costs may be going down, but the number of data centers and the cost of powering them are taking a bigger and bigger bite out of current and projected data budgets. (See Moore et al.⁸ for a 2007 assessment of the San Diego Supercomputer Cen-

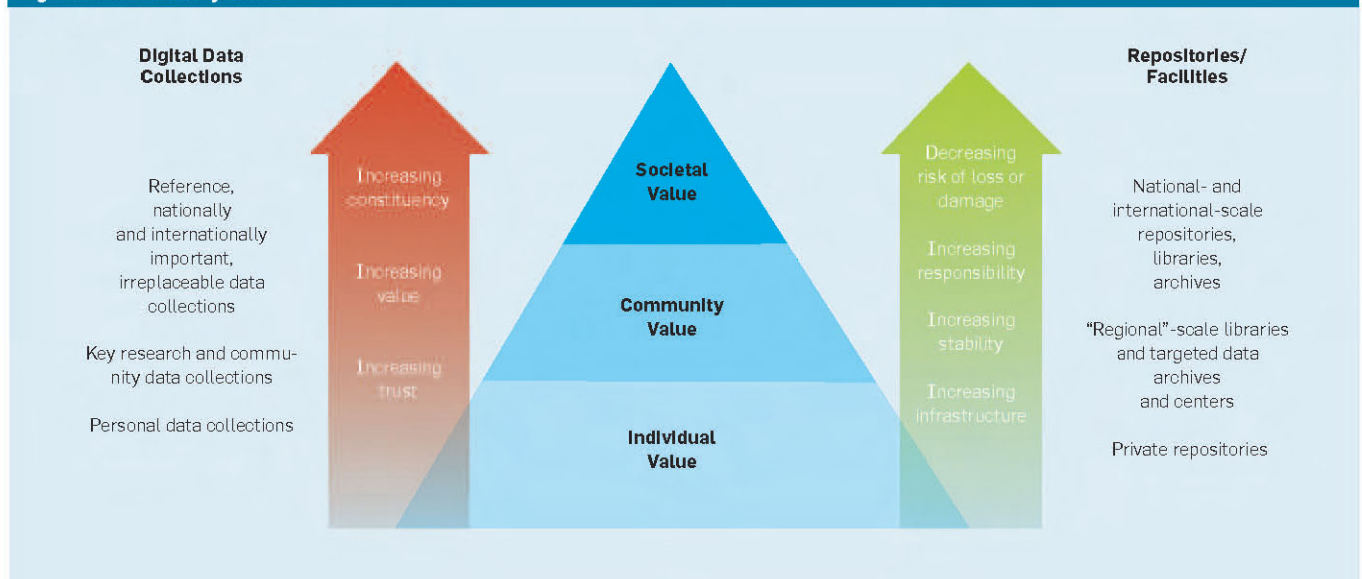
ter’s total cost of providing storage infrastructure.)

In addition, most data centers employ a knowledgeable, professional work force to ensure appropriate curation and annotation of digital data for the smooth running of the data center infrastructure and to plan ahead for future institutional and enterprise data needs. A capable data work force is important for all sectors and will likely increase as a percentage of the overall IT work force, along with the increasing need for a well-managed, sustainable digital data CI.

Finally, data centers must also factor in the cost of compliance with current and future regulations (possibly requiring additional physical and/or cyberinfrastructure for power backup and monitoring) and the need for replication of valuable data sets. (Data with long-term or substantive value is commonly stored with at least three copies, some hosted off-site.) We should expect the overall costs of data centers to continue to be substantial for the foreseeable future.

Trend 4. Increasing commercialization of digital data storage and services. The 2006 introduction of Amazon Simple Storage Solutions (www.amazon.com/gp/browse.html?node=16427261) was a high-profile example of the trend toward commercialization of data storage and data services. Today, there is considerable activity in the private sector around data storage and services for the consumer; for example, we share

Figure 2: The Data Pyramid.




and store digital photos through Flickr, employ Apple's Time Capsule for regular personal computer backup, and use LexisNexis for online legal services.


The commercialization of data storage and services contributes an important component of the data CI environment needed to harness the potential of our information-rich world. However, private-sector storage and services are not the solution to all digital data needs. For some digital data considered to be "in the public interest" (such as census data, official records, critical scientific data collections, and a variety of irreplaceable data), a greater level of trust, monitoring, replication, and accountability is required to minimize the likelihood of loss or damage and ensure the data will be there for a very long time. For such community data sets, stewardship by a trusted entity (such as libraries, archives, museums, universities, and institutional repositories), whose mission is the public good rather than profit, is generally required.

There is no one-size-fits-all solution for data stewardship and preservation. The "free rider" solution of "Let someone else do it"—whether that someone else is the government, a library, a museum, an archive, Google, Microsoft, the data creator, or the data user—is unrealistic and pushes responsibility to a single company, institution, or sector when what is needed are cross-sector economic partnerships. Sustainable economic models for digital data in the public interest are the focus of an international Blue Ribbon Task Force for Sustainable Digital Preservation and Access (brtf.sdsc.edu), whose goal is to examine digital preservation as an economic activity and explore cost frameworks for various institutional scenarios. The Task Force's final report, due at the end of 2009, will focus on economic models, components, and actionable recommendations for sustainable digital preservation and access, though it is already clear is that a diverse set of economic approaches are necessary.

In aggregate, these four trends point to the need to take a comprehensive and coordinated approach to data CI and treat the problem of sustainability holistically, creating strategies that make sense from a technical,



We do not produce storage capacity at the same rate we produce digital information. Even if we wanted to, we cannot keep all of our digital data.



policy, regulatory, economic, security, and community perspective.

Value and Sustainability

In developing effective models for data CI, perhaps the greatest challenge is economic sustainability. A key question is: Who is responsible for supporting the preservation of valued digital data? Critical to answering is the recognition that "value" means different things to different people. There is general agreement that official digital government records (such as presidential email and videos of congressional hearings in the U.S.) are preservation-worthy and of great political and historical value to society, but the video of your niece's voice recital is likely to be of value to a much smaller family group (unless, of course, your niece is, say, Tina Turner).

Sustainability solutions for digital data are inextricably related to who values it and who is willing to support its preservation. Governments worldwide are willing to support the preservation of digital content of national value, a substantial undertaking that involves hosting multiple copies of the same data, migration of the data from one generation of storage media to the next to ensure it lives in perpetuity, and protection of its integrity and authenticity. Your niece's voice recital may live on the hard drives of one or more family members, but there is rarely an explicit plan for how such a treasured family artifact will be preserved for the next decade and beyond.

How might we distinguish among all the data use, stewardship, and preservation scenarios to create and identify the data CI solutions needed to support them? One way is to borrow from the world of computation and adapt the Branscomb Pyramid model to today's data-use and data-stewardship scenarios. In 1993, the NSF asked Lewis Branscomb to lead a distinguished committee to consider the future of high-performance computing in the U.S. The final report included a useful framework, now known as the Branscomb Pyramid, where the base of the Pyramid associated the least-powerful computational platforms with users needing computation for "everyday" applications, the middle associated more powerful computational plat-

forms with users whose applications require greater performance, and the tip associated the most powerful computational platforms with the users requiring the greatest performance for “hero” applications. The same approach can be used to create a Data Pyramid (see Figure 2) to frame today’s digital information and stewardship options.

The Data Pyramid outlines the spectrum of data-collection and data-stewardship alternatives. The bottom includes data of individual (“local”) value whose stewards focus primarily on individual needs (such as personal tax records and digital family photographs and videos). We back this up on our hard drives, with an additional copy off-site if we are methodical, but little of this data will ever be considered of great societal value.

At the top is data of widespread and/or societal value whose stewards are primarily public-interest institutions (such as government agencies, libraries, museums, archives, and universities). Included are official records, data infeasible or too expensive to replace (such as the Shoah Collection of holocaust survivor testimony, college.usc.edu/vhi/, and digital photographs from the most recent NASA space voyage). Much of it must be preserved over the long term by trusted institutions. It is typically replicated many times, the focus of explicit plans for preservation, and hosted by only the most reliable cyberinfrastructure.

In the middle of the Pyramid is data of value to a specific community whose stewards range from individuals to community groups to companies to public-interest institutions. It includes digital records from your local hospital, scientific research data preserved in community repositories, and digital copies of motion pictures preserved for decades, commercially valuable in the future in the form of, say, the “director’s cut.” In every sector, groups are beginning to grapple with the responsibility of creating plans for data stewardship that are cost-effective, support reliable digital preservation, and are not subject to the whims of markets and/or community social dynamics.

The Data Pyramid makes it easy to see that multiple solutions for sustainable digital preservation must be

Digital Data Terms and Conditions

The following definitions are derived from a number of sources, including the American Library Association (www.lita.org/ala/), National Information Assurance Glossary (www.cnss.gov/), and Joint Information Systems Committee Digital Information Briefing Paper (www.jisc.ac.uk/):

APPRAISAL

Evaluation and selection of digital material for long-term curation and preservation, documented policies, guidance, and legal requirements may require that it be done securely;

AUTHENTICATION

Security measure designed to establish the validity of a transmission, message, or originator or a means of verifying an individual’s authority to receive specific categories of information;

CURATION

Digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for current and future use. It builds on the underlying concepts of digital preservation while emphasizing opportunities for added value and knowledge through annotation and continuing resource management;

DIGITAL RIGHTS MANAGEMENT

The use of technologies to control how digital content is used and reused;

INGEST

Controlled or secure transfer of material to an archive, repository, data center, or other custodial environment in adherence to documented guidance, policies, or legal requirements;

INTEGRITY

The condition when data is unchanged from its source and has not been accidentally or maliciously modified, altered, or destroyed;

METADATA

Documentation relating to data content, structure, provenance (history), and context (such as experimental parameters and environmental conditions). Standards for metadata provide a basis for widespread community data sharing; and

PRESERVATION ACTION

Actions undertaken to ensure the long-term viability and availability of the authoritative nature of digital material. Preservation actions should ensure the material remains authentic, reliable, and usable while its integrity is maintained; such actions include validation, assigning preservation metadata, assigning representation information, and ensuring acceptable data structures and file formats.

devised. At the bottom, commercial services fill the need for primary, additional, or backup sites for collections of individual or private value. At the top, stewardship is primarily in the hands of libraries, museums, archives, government funding agencies, and other trusted institutions. In the middle, institutions, communities, enterprises, and others are the primary stewards of data, wrestling with institutional and community solutions for stable and sustainable digital stewardship and preservation. The next decade will likely see more creative partnerships among all the players in the Pyramid, as well as more attention to who actually pays the data bill and how its costs are managed.

Creating an economically viable Data Pyramid must also be complemented with continued research into and development of solutions that address the technical challenges of data management and preservation, resulting in the ability to utilize and create new knowledge from the data being stored. For example, the process of searching and mining data depends on how it is organized, what additional information (metadata) is associated with it, and what information might be included about the relationship (ontological structure) of data items to one another in a collection. All these functions are active and important areas for research. Privacy and policy controls for data collections and security of the supporting infrastructure are also critical research areas. Addressing the technical, economic, and social aspects of digital preservation will be critical to ensuring that the information age has the foundation required to achieve its potential.

Top 10 Guidelines for Data Stewardship

Whether your data portfolio is of personal, community, or societal value (or some combination), its viability and usefulness to you will result from how you plan for stewardship and preservation over its lifetime. The following guidelines help promote effective stewardship and preservation of digital data:

1. *Make a plan.* Create an explicit strategy for stewardship and preservation for your data, from its inception

to the end of its lifetime; explicitly consider what that lifetime may be;

2. *Be aware of data costs and include them in your overall IT budget.* Ensure that all costs are factored in, including hardware, software, expert support, and time. Determine whether it is more cost-effective to regenerate some of your information rather than preserve it over a long period;

3. *Associate metadata with your data.* Metadata is needed to be able to find and use your data immediately and for years to come. Identify relevant standards for data/metadata content and format, following them to ensure the data can be used by others;

4. *Make multiple copies of valuable data.* Store some of them off-site and in different systems;

5. *Plan for the transition of digital data to new storage media ahead of time.* Include budgetary planning for new storage and software technologies, file format migrations, and time. Migration must be an ongoing process. Migrate data to new technologies before your storage media goes obsolete;

6. *Plan for transitions in data stewardship.* If the data will eventually be turned over to a formal repository, institution, or other custodial environment, ensure it meets the requirements of the new environment and that the new steward indeed agrees to take it on;

7. *Determine the level of "trust" required when choosing how to archive data.* Are the resources of the U.S. National Archives and Records Administration necessary or will Google do?;

8. *Tailor plans for preservation and access to the expected use.* Gene-sequence data used daily by hundreds of thousands of researchers worldwide may need a different preservation and access infrastructure from, say, digital photos viewed occasionally by family members;

9. *Pay attention to security.* Be aware of what you must do to maintain the integrity of your data; and

10. *Know the regulations.* Know whether copyright, the Health Insurance Portability and Accountability Act of 1996, the Sarbanes-Oxley Act of 2002, the U.S. National Institutes of Health publishing expectations, or other policies and/or regulations are relevant to your data, ensuring your ap-

proach to stewardship and publication is compliant.

While adherence is not a magic bullet guaranteeing the long-term safety and accessibility of fragile digital data, these guidelines help focus appropriate attention, effort, and support on the maintenance and preservation of our valued digital information. Such attention is critical to our ability to harness the immense potential of the information age to illuminate and empower us in our changing world.

Acknowledgment

I am grateful to John Gantz, Chris Greer, Nancy McGovern, David Minor, David Reinsel, Brian Schottlaender, Jan Zverina, and the reviewers for their useful comments and generous help with this article. □

References

- Alvarez, R. *Developing and Extending a Cyberinfrastructure Model*. Research Bulletin 5. Educause Center for Applied Research, Boulder, CO, 2008.
- Atkins, D. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure*. NSF report. NSF, Arlington, VA, 2003; www.nsf.gov/jod/oci/reports/toc.jsp.
- Berman, F. Making cyberinfrastructure real. *Educause Review* 43, 4 (July/Aug. 2008), 18–32.
- Branscomb, L. et al. *From Desktop to TeraFlop: Exploiting the U.S. Lead in High-Performance Computing. Final Report of the National Science Foundation Blue Ribbon Panel on High-Performance Computing*. National Science Foundation, Arlington, VA, 1993; www.nsf.gov/pubs/ftis1993/nsb93205/nsb93205.txt.
- Gantz, J. *The Diverse and Exploding Digital Universe*. White paper. International Data Corporation, Framingham, MA, Mar. 2008; www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.
- Gantz, J. *The Expanding Digital Universe*. White paper. International Data Corporation, Framingham, MA, Mar. 2007; www.emc.com/collateral/analyst-reports/expanding-digital-ido-white-paper.pdf (methodology discussion begins on 17).
- Higgins, S. Draft DCC curation model. *International Journal of Digital Curation* 2, 2 (2007), 82–87.
- Moore, R., DAoust, J., McDonald, R., and Minor, D. Disk and tape storage cost models. In *Proceedings of the Society for Imaging Science and Technology's Archiving Conference* (Arlington, VA, 2007), 29–32; users.sdsc.edu/~mcdonald/content/papers/dt_cost.pdf.
- National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Arlington, VA, Sept. 2005; www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.
- National Science Foundation. *NSF Data Sharing Policy*. Arlington, VA, 2001; www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf.
- Reinsel, D. Personal communication. Group Vice President, Storage and Semiconductors, International Data Corporation, July, 2008.

Francine Berman (berman@sdsc.edu) is Director of the San Diego Supercomputer Center, Professor of Computer Science and Engineering, and HPC High Performance Computing Endowed Chair in the Jacobs School of Engineering at the University of California, San Diego, and is also co-chair of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access.