San Diego Supercomputer Center Director Offers Tips on Data Preservation In the Information Age

SDSC's Berman: Will your data be there when you need it?

December 10, 2008

Jan Zverina

The world has gone digital in just about everything we do. Almost every iota of information we access these days is stored in some kind of digital form and accessed electronically -- text, charts, images, video, music, you name it. The key questions are: Will your data be there when you need it? And who's going to preserve it?

In the December 2008 edition of *Communications of the ACM*, the monthly magazine of the Association for Computing Machinery, Dr. Fran Berman, director of the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, provides a guide for surviving what has become known as the "data deluge."

Managing this deluge and preserving what's important is what Berman refers to as one of the "grand challenges" of the Information Age. The amount of digital data is immense: A 2008 report by the International Data Corporation (IDC), a global provider of information technology intelligence based in Framingham, Mass., predicts that by 2011, our "digital universe" will be 10 times the size it was in 2006 - and almost half of this universe will not have a permanent home as the amount of digital information outstrips storage space.

"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," writes Berman, a longtime pioneer in cyberinfrastructure - an open but organized aggregate of information technologies including computers, data archives, networks, software, digital instruments, and other scientific endeavors that support 21st century life and work.

Berman is a strong advocate of cyberinfrastructure that supports the management and preservation of digital data in the Information Age - data cyberinfrastructure: "Just like the physical infrastructures all around us -- roads, bridges, water and electricity - we need a data cyberinfrastructure that is stable, predictable, and cost-effective."

In her article, Berman explores key trends and issues associated with preserving digital data, and what's required to keep it manageable, accessible, available, and secure. However, she warns that 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. What is needed are cross-sector economic partnerships," says Berman. She adds that the solution is to "take a comprehensive and coordinated approach to data cyberinfrastructure and treat the problem holistically, creating strategies that make sense from a technical, policy, regulatory, economic, security, and community perspective." Berman's ACM article closes with a set of "Top 10" guidelines for data stewardship:

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 becomes 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, for example, 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.

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 approach to stewardship and publication is compliant.

Berman is a national leader in this area and also co-chairs of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access with OCLC economist Brian Lavoie. The task force was formed late last year to explore and ultimately present a range of economic models, components, and actionable recommendations for sustainable preservation and access of digital data in the public interest. Commissioned for two years, the task force will publish an interim report outlining economic issues and systemic challenges associated with digital preservation later this month on its website.

For Berman's full *Communications of the ACM* article, please see: http://www.sdsc.edu/about/director/pubs/communications200812-DataDeluge.pdf

Media Contacts:

Jan Zverina, SDSC Communications, 858 534-5111 or jzverina@sdsc.edu