STRATEGY FOR AMERICAN INNOVATION
Fostering Innovation for Sustainable Growth
and the Creation of High-Quality Jobs

This Grand Challenge was submitted to OSTP by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (brtf.sdsc.edu) on 4/15/10.

Grand Challenge: Ensure that the knowledge of today is available for use tomorrow.

Discoveries and innovation are built on the foundation of knowledge that has been passed on from generation to generation. The challenge for this generation is to retool and rebuild our knowledge infrastructure for the 21st century, where information is increasingly accessed and manipulated in digital form.

To accomplish this, we must ensure that today's most valuable digital information is available to future generations. Our knowledge infrastructure must be democratic in access, fully digital, and able to support the stewardship of the nation's discoveries and creative output for the benefit of current and future generations of citizens in American and beyond.

A 21st century knowledge infrastructure should be optimized to:

- Support research and education
- Spur creativity and problem-solving
- Encompass emerging communication forms unique to the 21st century
- Foster a literate, engaged, technology-enabled student body and work force
- Ensure access to the knowledge and discoveries of today for our children and future generations.

Should the United States make it a priority to achieve this grand challenge? Why or why not?

Current and future success in the United States will be driven by modern tools and technologies. In our data-intensive age, digital information drives progress for every strategic priority, including energy, health, the environment, economic stability, research and education, safety and security, and innovation.

The *access and preservation* of the digital information that drives our progress in these areas is a vital priority for the United States, one that will

- accelerate and build on ongoing cyberinfrastructure projects that foster achievement in science, engineering, technology, mathematics, and the arts, and
- ensure that the United States can maintain its historic role as a leader of knowledge discovery and innovation

From the breath-taking and visionary 3-D artistry of digital motion pictures such as *Avatar,* to the digital data streaming back to Earth from the Hubble Telescope, access today and in the future to digital information will drive innovation in America, and with it, the global knowledge economy.

What existing activities in the public and private sector could the United States build on to achieve this challenge?

We have already begun to build many component pieces that must be combined to develop a viable and sustainable national strategy for digital access and preservation. These include:

- Ongoing efforts by the National Coordination Office / NITRD with federal funding agencies in data cyberinfrastructure (http://www.nitrd.gov/PUBS/2011supplement/FY11NITRDSupp-FINAL-Web.pdf)
- Investigation of the issues by the Interagency Working Group on Digital Data (http://www.nitrd.gov/about/harnessing_power_web.pdf)

- Efforts within the federal funding agencies to foster STEM education initiatives
- Investigation of the economic issues for sustaining access to digital information by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (brtf.sdsc.edu)
- Efforts to develop technologies and standards by the Massive Storage and Systems Technical Committee (MSSTC/IEEE)
- Development of a national network for preservation partnerships by the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP)
- Investment in preservation cyberinfrastructure by the National Archives and Records Administration Electronic Records Archives (NARA ERA)
- Programs such as the National Science Foundation's DataNet Initiative for creating digital cyberinfrastructure for sustainable access and use
- Data needs associated with HPC and Exascale computing initiatives, and many more.

What are the most important scientific and technical challenges that would need to be addressed to realize this challenge?

The challenges involved in access and preservation of digital information require innovation in the technical, scientific, social and economic areas, along with regulation and policy that supports third-party archiving, data sharing, data management, and data stewardship.

Scientific and technical issues include research and development in:

- Robust long-term secure digital storage
- Robust next-generation rapid error-correcting code technology
- Automated intelligent systems for developing metadata and other parameters necessary for re-use of digital assets in the future
- Reliable data analysis techniques for predicting storage media aging
- Mining and analyzing of large-scale data collections to provide useful information (e.g. data.gov)

- Visualization of very large scale data sets
- Sustainable economic models to support digital access and preservation, etc.

What are the most important gaps in the nation's R&D portfolio that should be addressed?

New research in long-term storage technologies is critical:

- The design and development of long-term storage of preserved bits at low levels of curation that could be retrieved and curated at some point in the future is a key to economically efficient preservation of materials created today whose long-term value is not clear today.
- Progress in materials science, particularly in nanotechnology, that lead to lower failure rates and incorporate archival design considerations is crucial for long-term access and preservation.

New research in mining, analyzing, and visualizing very large scale data-sets is also critical. Such data sets already exist from the current generation of supercomputers, and will be even numerous as petascale supercomputers become prevalent and exascale supercomputers become possible.

New research in the economics of digital preservation is needed. The opportunities, constraints, and realities of sustainability are not fully understood from an economic perspective and require serious exploration and investigation.

What kinds of R&D investments (e.g. support for individual investigators, small teams, centers, research infrastructure, etc.) should the United States Government emphasize?

R&D investments are needed across the board, as digital preservation and access is an issue with technical, scientific, economic, social, policy, and regulatory components.

Research in storage systems and technologies will be critical for a continued trajectory of innovation. For example, research into the development of sustainable exascale archival storage systems by interdisciplinary teams with expertise in materials science, computer science, information science, and other disciplines will be critical for the sciences, but also for humanities, arts, and cultural materials creators focusing on digital media.

Investments in the development of preservation cyberinfrastructure by trusted third-party archival institutions are urgent. University libraries can play a major role in the preservation of research data. Programs and pilots that enable libraries and third-party archival institutions to develop the needed cyberinfrastructure capacity, workforce, and sustainable economic systems would enable them to take increasing responsibility for providing stewardship for digital data for the research and education community.

An important issue for R&D is coordination of the many stove-piped applications and data silos that result from diverse R&D investments. Along with this, interoperability, the development of standards, and creative approaches to combining diverse efforts to engender useful information is needed.

Many issues addressed in the National Academies 2004 report, *Facilitating Interdisciplinary Research*, still remain unaddressed by academic institutions. In addition to the collaborative challenges in science and engineering explored in the report, there are the equally complex and potentially fruitful collaborations between the arts and sciences that should be prioritized for research and prototyping.

Additional issues are also discussed in the Interim Report and the Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (both at brtf.sdsc.edu).

What are the appropriate roles of the government, industry, academia, and other stakeholders in achieving this challenge, and what new forms

of collaboration should be explored? What are the appropriate roles for pre-competitive collaboration and market-based competition?

No one sector can address the entire digital access and preservation problem. Partnership between sectors is critical to a solution. Broad collaborations between academic research libraries, archival institutions, and private-sector companies willing to support digital preservation for the public good should be modeled and tested. In addition, policy and regulation that mandate preservation of valued digital information is needed to ensure future access (and hence preservation) of valuable digital content, and to create a market for curation and archiving services as well as support for archival institutions.

Each sector has important roles to play in order to attack the digital access and preservation problem. These include:

Government Roles

Stimulate R&D and the Development of Necessary Technical Infrastructure

1. Increased bandwidth for efficient, timely backup and delivery
2. Secure storage that survives long periods of "benign neglect"
3. Distributed networks of data storage and curation centers
4. WEB 3.0 and beyond tools to support data creation, curation, and preservation
5. Foster innovative uses of social networking tools for education and libraries

Stimulate the Development of Human Infrastructure

1. Invest in STEAM (Science, Technology, Engineering, the Arts, and Math) and STEM (Science, Technology, Engineering, and Math) Education to prepare tomorrow's workers
2. Create new fellowship programs for cross-disciplinary initiatives in library, archival, museum, and information sciences and related disciplines
3. Extend engineering fellowships to engineers working in digital archives and the arts

Develop Policy Infrastructure

1. Foster adoption of open, well-documented, "archives-ready" formats
2. Provide incentives or mandates as appropriate for researchers to publish and archive data
3. Develop policies and procedures for data management and stewardship mandates for publicly funded knowledge creation
4. Develop policies and procedures such as mandates for data curation and stewardship in all federally funded research-based activities, from scientific exploration to industrial development of technologies. Match these mandates to development of repositories that can scale to meet data deposit requirements
5. Develop financial incentives for private enterprise to archive valued information in the public good

Public-private Partnerships and Industry Roles

1. Stimulate cross-sector collaboration among public and private data creators and distributors to develop standards and best practices
2. Provide access to test beds and data sets for research and development of new technologies
3. Create and encourage consortia of industry/academia and government to develop hardware technologies and software tools
4. Stimulate the creation of cloud-computing repositories for digital storage and preservation as a resource for multiple agencies; leverage economies of scale and scope among different knowledge domains and sectors to avoid unnecessary duplication of effort
5. Provide start-up capital and operational funding to assist libraries and archival institutions in building digital preservation capacity

Academic Roles

1. Foster adoption of open, well-documented, "archives-ready" formats
2. Provide incentives or mandates as appropriate for researchers to publish and archive data

3. Develop policies and procedures for data management and stewardship mandates for research and publication
4. Provide incentives for researchers to publish data and preserve it using open standards and formats
5. Create test beds and data sets, and fund pilot programs to model and test sustainable stewardship strategies
6. Sponsor development of assessment and selection of scholarly materials for long-term stewardship.
7. Provide stewardship for digital data generated by the academic community, particularly as the result of sponsored research

What are the economic, ethical, legal, and societal issues raised by pursuit of this challenge? What roles are there for researchers and scholars in the humanities and the social and behavioral sciences?

Ethical issues arise around security of information, privacy, data integrity, etc. Legal and social issues include ownership of, and rights to disseminate information. As the rise of social networking, crowd sourcing, blogs, wikis, Facebook, Twitter, and other new forms of communication and information become increasingly prevalent, who owns digital information, and who has the right to preserve, access, and use digital information will become increasingly thorny.

Experts in the humanities, arts, and social and behavioral sciences are critical players in this. In the arts, music, movies, and other media are increasingly in digital form and require new approaches that enable their future access and preservation. The economics and social dynamics of information acquisition, selection, management and preservation are fundamental to a successful knowledge economy.

In addition to investment in R&D – what are other policies should the United States Government be considering to achieve this challenge and to

realize the broader economic and societal benefits associated with related scientific and technological advances?

Regulatory authorities should bring current requirements for mandatory copyright deposit into harmony with the demands of digital preservation and access.

In analog preservation and access, copyright law provides incentives to create valuable content by giving owners the exclusive right to provide access for specified periods of time. But it also grants limited rights to qualifying public institutions to preserve the same content in the public interest. This alignment between the public and private interest does not exist in the digital realm, for a number of technical reasons. The Copyright Office and the Library of Congress have made recommendations to change the law in light of digital technology realities

We also advocate mandated deposit of copyrighted electronic content into authorized public institutions to secure their long-term preservation. In countries with mandatory copyright deposit, that mechanism has been very effective for ensuring some public provision of preservation for privately held materials. In the United States, the moving image record has been better and more complete since the mid-twentieth century, when the Copyright Office in the Library of Congress began demanding deposit of film. Deposit in authorized libraries and archives have preserved materials that commercial firms did not keep safe. While there is debate about how effective mandatory deposit will be in a digital environment, it does seem advisable that digital materials submitted for copyright protection should be deposited in full, which is not now always the case. This would require that copyright authorities articulate which version of the work—the *best edition*—is registered for copyright; enforce demand deposit when necessary; and scale up copyright deposit system capacities to meet the challenges posed by the enormous scale of content creation.