Written Testimony of G. Sayeed Choudhury
Associate Dean for Research Data Management
Hodson Director of the Digital Research and Curation Center
Sheridan Libraries
The Johns Hopkins University

Given before the Testimony to the Committee on Science, Space, and Technology
Subcommittee on Research
House of Representatives
Hearing on -
Scientific Integrity & Transparency
March 5, 2013

Mr. Chairman and Members of the Subcommittee, thank you for inviting me to address the following questions on data sharing, access, and preservation. I will address these questions from the perspective of infrastructure development. With prior infrastructure development (e.g., railroads, roads), there was a natural stage at which point national coordination and strategic planning moved regional systems into a cohesive national infrastructure. I believe we have reached this point with certain aspects of public access to data. The existing networks of research systems and processes at universities, scientific societies, publishers, etc. ("ecosystem") that relate to data sharing can be complemented with common, wide-scale infrastructure. The opinions expressed herein are my own and do not necessarily reflect the views of The Johns Hopkins University.

I have spent over a decade dealing with scientific data management beginning with early work associated with the Sloan Digital Sky Survey (SDSS) and continuing today through my leadership of Data Conservancy, one of the awards through the National Science Foundation's DataNet program. In addition to my experience with scientific data management, I have also had long-term experience with humanities data management, most notably through a digital manuscripts program. These diverse experiences have given me a keen appreciation for varying disciplinary needs, practices and cultures regarding data sharing but also an understanding of common infrastructure requirements that span a wide range of diverse domains and contexts. My two roles at Johns Hopkins – one related to research and development and one related to administration – allow me to focus on migration or translation of research results into operational environments.

I have led projects with funding from diverse sources including federal agencies, private foundations, corporations and a venture capital group. In addition to my experiences within the United States, I have been fortunate to work closely with colleagues and collaborators in the United Kingdom, European Union, Australia and New Zealand.

I believe that these diverse experiences, funding sources and interactions have given me a comprehensive opportunity to identify useful conditions for wide-scale implementation of data infrastructure.

Before addressing the questions directly, it is useful to consider lessons learned from historical infrastructure development. With the development of railroads within the United States, there was a period of regional railroads that served portions of the country. The recognition that a national railroad network would confer greater benefits for the transport of people and goods prompted the development of a national railroad gage that resulted in interoperability and efficiency. The evolution of automobiles reflected a process of learning and adapting from early mistakes. Eventually, we produced safer automobiles and built new roads, regional highways and eventually interstate highways. The Internet was designed and modeled with a stack model that delineated different functions and protocols, with the TCP/IP protocols being the most important.

Each of these historical infrastructure developments offers insights that are relevant when considering data infrastructure for sharing, access, and preservation, particularly relating to the balance between local versus global frameworks. The United States' investment in these earlier forms of infrastructure resulted in benefits for a range of private and public stakeholders. I believe similar investments in data infrastructure will result in benefits for scientists, the public and the private sector.

With these insights in mind, I will address the specific questions sent in advance for this hearing:

1. *What are the issues that we need to consider for wide-scale implementation of data sharing? Specifically, what are the IT infrastructure needs, including hardware, software, and technical standards, and what, if any, scientific or technical barriers to developing that infrastructure? Are there policy or non-technical barriers for sustainable digital access and preservation?*

One of the overarching issues to consider for wide-scale implementation of data sharing relates to an "ecosystem" viewpoint for infrastructure. Related to this point is the reality that all data are not alike. Scientific data comes in various levels that range from the raw, unprocessed signals generated directly by instruments (e.g., telescope, genome sequencer) to more calibrated data to highly refined, processed data cited within publications. These different levels of data possess different requirements for IT infrastructure. Additionally, the type of instrument, presence or absence of standards, community practices and other factors can result in different IT infrastructure needs (and costs as mentioned later).

Consequently, there is a need for a layered approach for data sharing, access, and preservation that includes a diversity of systems for active use of data during projects (most often directly managed by researchers); staging areas that house data for less active use (such as repositories managed by libraries; universities or data centers and cloud-based storage offered by commercial providers); data archives that preserve data but retain access and sharing provisions (nascent infrastructure that is evolving); and "dark" archives that preserve content for long-term periods without direct access. It is important to stress that these various layers of an overall infrastructure must be designed for data, which are fundamentally different than documents. Attempting to use

or re-engineer existing document management systems will result in inadequate functionality and possibly additional costs, particularly in the long-term.

From a hardware perspective, there is a need to consider enhancements to existing storage systems particularly from a data preservation perspective. Over the last three years, my colleagues at Johns Hopkins have learned firsthand regarding the issues of storage hardware and software as we have managed the data from the Sloan Digital Sky Survey. Examples include storage system block size being too large compared to smallest unit of data, inadequate methods for generating fixity (machine generated code to verify data integrity), and performance issues related to throughput (volume of data processed in a particular unit of time). Our current engagements with storage companies indicate that they view development of new capabilities as a business opportunity between the private sector and universities.

From a standards perspective, it is important to note that many scientific communities have existing standards for data sharing and access. Even in these cases, developing infrastructure and mechanisms (e.g., semantic Web) for sharing across disciplines or communities remains a challenge. It may be possible to span across two disciplines or communities through bilateral agreements. However, this approach does not scale for multiple disciplines or communities. While it is possible to develop common denominator standards for discovery of data, there remain fundamental research problems to address interdisciplinary or cross-disciplinary data sharing and access. Federal agency funding to support this type of research with the goal of developing working systems or infrastructure would be helpful.

One of the most important non-technical barriers for sustainable digital access and preservation relates to a lack of awareness regarding comprehensive data management. Terms such as storage, archiving, preservation and curation are often used interchangeably and inappropriately. My colleagues and I from the Data Conservancy have developed a data management layer stack model that conceptualizes the concepts of storage, archiving, preservation and curation. This model is not intended to be definitive, but rather reflective of our lessons learned. For this model, storage describes bits on disk, tape or in the cloud with backup and restore services. Archiving focuses on persistent identification and data protection through actions such as generating and verifying fixity and maintaining or tracking multiple copies. The term “preservation” is perhaps most often mentioned loosely or vaguely. For our model, preservation involves providing enough representation, context, metadata, fixity, and provenance information such that someone -- or some machine -- other than the original data producer can use and interpret the data. Provenance can be defined simply as whom or what machine handled the data and what did they do with the data. Finally, curation refers to adding value to foster discovery, access and re-use of data.

Researchers do not always realize the full extent of sustainable digital access and preservation. Educating researchers and changing their data management practices and behavior represents an important social component of infrastructure development. This type of behavioral change is not unlike the process that automobile drivers went through in the United States. Drivers have changed their behavior over time as we have gained greater understanding regarding safe driving and greater willingness to introduce safety through seat belts, speed limits, laws, etc. This type of social or cultural change represents an important aspect of the social-technical dimension of infrastructure development.

2. *What are the most important factors to consider in the economics of digital data access and preservation? What funding models have proven effective and how scalable are they? What should be the role of federal science agencies in supporting and preserving accessible databases? What should be the role of the private sector and of universities? How can all three work together to minimize costs and maximize benefit to the scientific community?*

The economics of digital data access and preservation require greater examination of both costs and benefits. There has been relevant work for cost models in the UK and even recent application of those models for scientific data. The Australian National Data Services has developed a business plan.

Within the US, there is a need to conduct more analyses in the full accounting sense of costs including hardware, software, human labor, utilities, etc. Furthermore, cost estimates must consider the long-term implications. For example, referring to the previous discussion about the data management layer stack model (storage, archiving, preservation, curation), some cost models account for storage only. As mentioned previously, not all data are alike so there is a need to consider cost issues according to data levels, types, presence of standards, etc. For example, a terabyte of data produced from a single instrument according to well defined standards and a single processing pipeline will probably require less cost for access and preservation than a terabyte of data produced by a single investigator using multiple instruments and within a discipline without well defined community standards.

One of the most important costs that are often unconsidered relates to data center operating costs. The power and cooling requirements for these data centers can be significant. Technologies that use less power and space will reduce these escalating costs.

On the benefits side, there is a greater need for understanding the demand for accessing, re-using and preserving data. There are potential organizations from both the private sector and university environment that would provide highly useful information case studies for costs and benefits. Examples from my own experience include the National Snow and Ice Data Center and Inter-university Consortium for Political and Social Research, both of which have successful, long-term track records with providing access to and preserving scientific data. These case studies could lead to the development of business models and eventually economic models that could be applied in a scalable manner.

In this context, it is worth mentioning that archival principles such as appraisal and intrinsic value are important, particularly as they relate to unanticipated use. There are cases where re-use of the data or secondary uses by individuals other than the original data producer generates unforeseen benefits. There is evidence that some astronomers use data archives even more often than new telescopes and that some use of high-performance computing facilities relates to re-use of existing data.

The development of wide-scale IT infrastructure for data sharing, access, and preservation is multi-faceted in that there are reinforcing roles for the federal agencies, the private sector, and

universities or national laboratories. The case for preservation highlights the possible delineation and coordination of roles. Preservation of data ensures persistent use and re-use for scientific and commercial reasons. However, it is likely that preservation for public access by itself is not a profitable activity and therefore possible that the private sector would not develop relevant capacity and service. Universities, libraries and national laboratories – which have established relationships with researchers, the public, and the private sector – have developed nascent infrastructure for data preservation but require additional resources for further development. Federal agencies could develop contracts with universities, libraries and national laboratories to further develop data preservation infrastructure that supports a range of scientific and commercial uses.

3. *What federal policies are necessary to maximize data sharing and access? Do you have any recommendation with respect to current science agency data management policies at NSF or at other agencies?*

The recent memorandum for the Heads of Executive Departments and Agencies from the Office of Science and Technology Policy's Director John Holdren offers a useful framework for considering federal policies to maximize data sharing and access, including potential extensions to existing data management policies. It reinforces the benefits of public access to data for "the public, industry, and the scientific community."

The memorandum acknowledges that federal agencies need flexibility in developing and implementing plans for data sharing, access, and preservation given the diverse set of disciplines, missions and approaches. However, the memorandum also identifies some uniform guidelines. To the extent possible, federal agencies should coordinate their responses to this memorandum and their associated plans to minimize burden and costs associated with compliance.

It is encouraging to note that each federal agency's response and plan must "ensure appropriate evaluation of the merits of submitted data management plans." In order to meet this condition, reviewers will need guidelines for effective evaluation of data management plans. My colleagues at Johns Hopkins have developed such guidelines based on our experience to date with data management plans and reviewers' responses to those plans. Many other universities and libraries can collect such information to develop community-based guidelines that federal agencies might use to inform their proposal reviewers.

The memorandum also asks federal agencies to "develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan." In this regard, it is worth examining the recent workshop report from US CODATA and the Board on Research Data and Information (BRDI) "**For Attribution—Developing Data Attribution and Citation Practices and Standards.**"

This report discusses and outlines examples of persistent identifiers—a long-lasting reference to a digital object consisting of a single file or a set of files. As an analogy, often when a webpage is not found, one encounters a "404" error and little other information to resolve the problem. Persistent identifiers mitigate this problem by assigning a permanent reference that tracks the movement of the associated digital object.

The persistent identifier is a key piece of infrastructure that demonstrates the value of using systematic approaches for data citation or identification that can be used for sharing, access and preservation. A balanced approach between local and global dimensions would include a requirement that researchers use persistent identifiers for data without prescribing the specific choice of identifier. Even though different communities will probably choose different identifier schemes initially, doing so represents progress analogous (in a rough sense) to regional railroads. As communities choose and adopt persistent identifiers, the opportunity to consider cross-community or global approaches becomes possible similar to the equivalent of TCP/IP in the Internet model.

I hope that my testimony has provided background, context and recommendations that can advance the development of data infrastructure within the United States. Such infrastructure, developed through partnership of the public and private sectors, would result in benefits for science, industry and the public. While there remain important research and social issues to consider, there are practical steps we can take now to advance our scientific enterprise especially in light of the recent OSTP memorandum related to public access to data.

Thank you again Mr. Chairman and Members of this Subcommittee for the opportunity to address these questions.