# 1. Introduction

New digital technologies are transforming the practice of science. Science is now increasingly computational, data-intensive, and collaborative because digital technologies provide new ways for scientists to both create scientific information, and to communicate, replicate, and reuse scientific knowledge and data. Two key elements of this transformation are access to data and access to knowledge. Digital technology can make data openly accessible to scientists, reducing data-management burdens, formalizing generalizable and replicable science, and enabling new kinds of data-driven science. Digital technology can similarly facilitate the dissemination and transmission of knowledge by making information widely available electronically.

Institutional and professional barriers limit both data and knowledge access, however. For example, the costs associated with data access, including storage, documentation, and dissemination are not uniformly supported. Further, the current system of scientific attribution does not capture the complexity of contributions to scientific knowledge.

International funding agencies have an important opportunity to change policies to reduce these barriers. They could use digital technologies to promote scientific collaboration, and to foster the replication and reuse of scientific information, thereby changing the conduct of science. They could identify technical solutions for developing and maintaining open data platforms to promote collaboration and cooperation, foster the replication of scientific research, and ensure attribution for the intellectual contributions of researchers (National Science Foundation 2010b).[1]

To examine how these barriers might be overcome by international funding agencies and organizations supporting research, the U.S. National Science Foundation (NSF) held a workshop titled "*Changing the Conduct of Science in the Information Age"* on November 12, 2010. The workshop brought together members of the research community, computer and information scientists, as well as behavioral and social scientists, to identify guiding principles and approaches that could help inform organizations that fund research, scientific research organizations, and publishing houses.

The workshop placed questions into three categories:

- *Technical constructs*—What are the most important digital technologies that could be used to facilitate access to data and knowledge? To what extent is progress already being made, and how can progress be accelerated? What role might the private sector play in facilitating change?

[1] References given in parentheses are listed at the end of this report.

- *Social constructs*—What incentives are necessary to engage scientists in making data accessible and shared with the broader community? What are the appropriate business models necessary to promote connecting publications to data? How might private-sector participants be engaged in the effort? What are the social barriers to adopting and using unique researcher numbers?

- *The Pragmatic Experience*—What lessons have been learned from Brazil's experience with the Lattes platform? What opportunities are possible as a result of the establishment of the ORCID (Open Researcher and Contributor ID) project? What can be learned from data preservation, libraries and other coordinated data and publication efforts? What can be learned from domain-specific successes?

## 2. Data Access

Access to data generated by the "data deluge" is crucial.[2] Research reproducibility is critical (Hirsh 2010; Donoho 2010; Donoho et al. 2009), as "[r]eplicability is a hallmark of science" (Börner 2010), but research is only reproducible if the underlying data are accessible (Stodden 2010) and reliable. The challenge is daunting: one workshop speaker noted that the scientific community now generates more data each year than the entire sum of data produced in all prior years combined (Seidel 2010). Much data are inaccessible because of the dramatic increase in the amount of "information which is 'off the records' of science, not available to peer reviewers, [and] in many cases not even recorded in formal lab notebooks or laboratory information management systems" (Pfeiffenberger 2010). A recent NSF/Office of Cyberinfrastructure (OCI) Grand Challenges Task Force Report identified reproducibility of computational results as an

> ***Exemplar*: Sloan Digital Sky Survey (SSDS)**
> The SDSS is a map of the universe that was compiled from 1991 to 2008. It has generated 850 million web hits in 9 years by 1,000,000 distinct users (globally there are only 15,000 professional astronomers). SDSS tops the astronomy citation list and has delivered more than 100 billion rows of data. It has facilitated both remote collaborations and discoveries by amateur scientists (http://www.sdss.org/).

[2] Recently held workshops and reports devoted to data access include the European Commission's High level Expert Group on Scientific Data, *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*, October 2010, available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf; Yale Law School, *Data and Code Sharing Roundtable*, November 21, 2009. See also http://www.law.yale.edu/intellectuallife/codesharing.htm, and "A Special Report on Managing Information: Data, Data Everywhere," *The Economist*, February 25, 2010, available at http://www.economist.com/node/15557443.