***Toward a Strategy of Research Priorities in Data-Intensive Science:***

***Understanding Science Drivers, Challenges, Infrastructure Needs and Best Practices***

**A Workshop Proposal**

**Board on Research Data and Information**

**Policy and Global Affairs**

**National Academy of Sciences**

**in collaboration with**

**Earth Science Information Partnership (ESIP) Federation**

**Draft Prospectus**

We live in a time of unprecedented collection of and access to scientific data. It is world of opportunity for advances in scientific understanding. At the same time, scientists are increasingly overwhelmed by data management tasks that not only waste resources, but can result in bad research. Data sets exist but cannot be discovered. Data sets gathered at great (and often public) expense are at risk of being lost for technological or organizational reasons. Funding for data management is meager and generally must be found within existing budgets. Scientific results are being scrutinized and questioned, and occasionally retracted due to problems in data management. A lack of sufficient documentation and understanding of data can mean that the data are used incorrectly or not at all.

The volume of data is greatly increasing as well, compounding the existing problems. With improvements in sensor technologies, new data sources are constantly appearing. Important research questions can span projects, discipline domains, and other boundaries. Data sets are being used for unexpected purposes far from the original point of collection. The nature of science is evolving, with more open science, open publications, and changes to the nature of peer review and data "publication". Data-intensive, or computational science, has been identified as a new research paradigm. Federally funded projects are expected to make their data open and accessible to everyone, if possible. Scientific progress is thus ever more dependent on good data management practices and policies.

On February 22, 2013, the Office of Science and Technology Policy issued a memorandum that "government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable." For a variety of reasons only alluded to above, current data practices are insufficient to resolve these pressures and too often are reactive in nature. Though a variety of organizations are making efforts to improve research data practices, they lack a unifying, long-term vision and articulation of expectations and a research strategy to address them. Additionally, the OSTP memorandum required that executive departments and agencies develop plans to meet the data management challenges, which gives substantial impetus to a study of these problems.

In light of these developments, members of the Federation of Earth Science Information Partnership (ESIP, esipfed.org) and of the National Research Council’s Board on Research Data and Information (BRDI, www.nas.edu/brdi) are collaborating to determine whether to initiate a strategy of research priorities in data intensive-science practices and if so, identify what some of the questions might be that need to be examined in detail. Such a strategy could address overarching issues and research priorities in scientific data management and stewardship, and improve the public return on investment. Better technologies and practices in this area could ultimately enhance scientific knowledge by increasing the meaningful availability of higher quality data. It would redirect time and resources previously required by scientists for data discovery, acquisition, and formatting to performing actual science. When data sets can be easily analyzed and combined in novel ways then new scientific insights are more likely to occur and more quickly. Such a strategy could thus address at the broadest level gaps in data management knowledge and practices that hold back scientific progress, and recommend ways to address them.

The workshop would be held pursuant to the following preliminary statement of task:

1. What could be the major goals and elements of developing a strategy for research priorities in data-intensive science practices established over the next decade across the federal government? What are the advantages and disadvantages of doing that now? If such a strategy is deemed desirable, what could be the discipline scope and the pros and cons of such option(s)?
2. How might such a strategy be formed? What are the relative merits of the options identified?
3. What institutions in the federal government, academic, commercial and nongovernmental sectors ought to be involved in both developing and implementing such a strategy? What are the external factors that need to be considered?

A steering committee of approximately 6 experts would be appointed according the procedures of the National Research Council. The steering committee would plan the one-day workshop of about 40 experts. The workshop would be held on Tuesday, January 7, 2014, in conjunction with the Winter ESIP Meeting, which is already scheduled to be held in Washington, DC. In consultation with the sponsor(s) we would invite relevant experts to make presentations and discuss in a panel format the topical areas identified above in the statement of task. Such experts would include some of the principal investigators of the sponsoring agency(s). We also would invite the sponsor(s) and other federal agency representatives to discuss these issues with the expert speakers and panelists.

The meeting would be structured in several sessions that would follow the questions posed in the statement of task, along with some introductory and concluding remarks. Prof. Bill Michener would be a facilitator of the workshop and the entire discussions would be recorded and transcribed.

The meeting background will be openly posted on the BRDI website along with links to relevant resources. The agenda will also be posted with links to the presentation slides and bios of the expert speakers. An audio file of the proceedings will be openly archived on the BRDI website as well. Finally, a lightly edited written transcript will be made available to the workshop sponsor(s).

A separately funded summary report is expected to be prepared by a consultant of the Earth Science Information Partnership Federation. The summary report would be published openly online as well.

The principal investigator for the project would be:

Paul F. Uhlir, J.D.

Director, Board on Research Data and Information

National Academy of Sciences

Washington, DC

puhlir@nas.edu