Overcoming the Technical and Policy Constraints That Limit Large-Scale Data Integration

Revised Proposal from The National Academies

Summary
An NRC-appointed committee will plan and organize a cross-disciplinary public workshop to explore alternative visions for achieving large-scale data integration in fields of importance to the federal government. Large-scale data integration refers to the process of aggregating data sets that are so large that searching or moving them is non-trivial or of drawing selected information from a collection (possibly large, distributed, and heterogeneous) of such sets. The workshop will address the following questions:

- What policy and technological trajectories are assumed by some different communities (climatology, biology, defense, and others to be decided by the committee) working on large-scale data integration?

- What could be achieved if the assumed policy and technological advances are realized?

- What are the threats to success? Who is working to address these threats?

The committee will develop a consensus summary based on the workshop to convey the various perspectives with regard to these questions and identify major opportunities and threats that cut across multiple communities.

Estimate Cost: $215,000
Project Duration: 10 Months
Anticipated Workshop Date: Fall/Winter 2009

Background
Advances in information and communications technology (ICT) continue to transform most areas of human endeavor, including those of particular interest to the federal government such as scientific and engineering research, intelligence gathering and analysis, and protecting public health. One important aspect of this ICT-enabled transformation is the growing data intensity of these activities. The collapsing unit costs of digital data collection and storage capabilities are making it possible to address new research questions, to develop new approaches to solving a variety of problems, and to create new services.

However, storage costs are declining faster than processing power is growing, with the result that the amount of data to be handled is growing more rapidly than our ability to process it. This presents serious challenges to the communities that generate, curate, and use data collections, including:

- improving the technology for discovering, accessing, integrating, and analyzing data;
- ensuring that the infrastructure for long-term preservation and access to data is sustained;

and

- balancing the advantages of open access with the imperative to accurately track provenance, protect integrity and, in some areas, maintain the privacy and/or security of data.

Large-scale data integration as the process of aggregating data sets that are so large that searching or moving them is non-trivial or of drawing selected information from a collection (possibly large, distributed, and heterogeneous) of such sets. There are generally both syntactic and semantic differences to overcome, with those terms being defined broadly enough to cover the challenge of integrating, say, tabular, text, and imagery data. The need for large-scale data integration arises in a variety of settings. For example, biomedical scientists might be faced with the challenge of finding and combining relevant data about a certain disease and its genetic or environmental underpinnings, even though that information might be in widely disparate free-text and structured databases. The dynamic integration of data from various sources, including data generated through simulations as well as images and data collected from sensors, sometimes combining information from disparate fields of inquiry, is opening new avenues for insight and productivity. Our use of the term “data integration” may be broader than how the term is generally understood by some experts. We consider as within scope any technology, process, or policy that affects a scientist or engineer’s ability to find, interpret, and aggregate/mine/ analyze distributed sources of information. Data interoperability and knowledge discovery are both intended to be within scope.

Data integration, when done well, facilitates the subsequent data search and analysis, and so developments in data integration should be informed by an understanding of the emerging capabilities in search and analysis—they are the tools that will operate on the integrated data. And in some ways search technology aims to take on some tasks of data integration, in the sense of aggregating data from disparate sources rather than presenting the user with disconnected searches. To the extent that search technology succeeds in that direction, it will change the nature of the data integration challenges. Hints of what might be possible in the future can be seen in the more innovative web application hybrids, or mashups.

As might be imagined, most of the applied research and product development in the area of data integration is proprietary and targeted at business customers. In scientific and engineering applications areas, individual disciplines do not represent a large potential market for commercial entities. As a result, researchers in these disciplines often struggle in relative isolation to develop the ICT tools and techniques required for their increasingly data-driven work. This is frequently the case not only in data integration, but also in other areas of cyberinfrastructure. Deep familiarity with the physical processes and research questions central to a given discipline is essential for developing effective cyberinfrastructure tools and techniques. But this need for domain experts to drive the process often means that cutting-edge approaches in computer science, statistics, and other relevant areas are not incorporated, except perhaps in those disciplines large enough to have developed a critical mass of people with deep domain AND informatics expertise.

A recent NASA workshop ("2nd NASA Data Mining Workshop: Issues and Applications in Earth Science," May 23-24, 2006, Pasadena, Calif.) illustrates this problem, documenting the significant gap between the state of the art in data mining and statistical methods for data analysis and what is currently being used in the earth sciences. That workshop also explored the barriers and opportunities to progress that exist in the earth sciences field, including cultural issues such as the structure of recognition and rewards. One anticipated benefit of the proposed activity will be to improve leveraging opportunities across fields of science and engineering. Even federal agencies in application areas with significant resources for research and development in data integration (e.g., the intelligence community) feel the need for greater exposure to developments in the commercial world, and recognize the potential for more strategic leverage across the federal government.

During 2005, the National Academies' Government-University-Industry Research Roundtable (GUIRR) and Committee on Applied and Theoretical Statistics (CATS) launched an informal working group aimed at exploring current and projected challenges in large-scale data integration and the potential opportunities open to user communities if challenges can be addressed. CATS has a longstanding interest in bringing cutting-edge statistical approaches to bear on problems involving large-scale data (CATS 1992, 1999). In pursuing its mission to strengthen the U.S. research enterprise, GUIRR has grappled with the implications of ICT advances for the future of the research university. An NRC study on the topic (PGA, 2002) and the subsequent Forum on Information Technology and Research Universities were incubated in GUIRR discussions. Engaging the challenges of large-scale data represents a continuation of GUIRR's interest in the impacts of ICT on the research enterprise. As the proposed study moves forward, CATS and GUIRR staff will also involve experienced staff from the Academies' Computer Science and Telecommunications Board and the U.S. National Committee for Data (CODATA).

The GUIRR-CATS working group included participation from industry, a range of federal agencies (NSF, NIH, NOAA, NASA, NARA, NSA, and DOD), and academia, representing a number of disciplines. The goals were to examine the technical and policy issues raised by large-scale data, to exchange information about existing efforts, and to identify issue areas where additional work is needed, particularly those areas where the exchange of perspectives from different sectors and disciplines would be valuable.

The working group took note of extensive efforts underway to address several critical aspects of the large-scale data challenge, including those in which the National Academies play an important role. For example, CSTB has provided strategic advice to NARA, the lead agency in the federal government's own data management activities, in the development of its Electronic Records Archive (CSTB, 2005). Also, the U.S. National Committee for Data (CODATA) is deeply involved in international efforts to preserve and promote the public domain and open access to scientific and technical data, with particular attention to the needs of developing countries (CODATA, 2004).

In addition, the working group exchanged information on efforts being undertaken in various disciplines to deal with large-scale data at the technical and policy levels. For example, in the earth sciences several U.S. agencies are engaged in the international initiative to build a Global Earth Observation System of Systems (GEOSS), which aims to provide worldwide access to an unprecedented amount of environmental information, integrated into new data products.

Finally, growing federal interest in the challenges of large-scale data across a range of disciplines is reflected in several recent reports and initiatives. These include the National Science Board's 2005 report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (dealing with long-term stewardship of digital data collections), NSF's *Cyberinfrastructure Vision for 21st Century Discovery* (in which data, data analysis and data visualization constitute a major component of this evolving strategic plan), and recent discussions under the auspices of the National Science and Technology Council.

The GUIRR-CATS working group undertook a series of telephone and face-to-face discussions during 2005 and 2006. Over time, it became clear that addressing the interrelated technical and policy issues related to data integration was of critical importance to all those engaged in the discussion, particularly to the federal agencies.

The NRC study described here is the product of these discussions, and reflects this interest. The project is designed to explore and compare the goals, approaches, and barriers to progress in data integration across various application areas of interest to the federal government and to develop options for possible new collaborations between agencies, sectors, and disciplines that would provide significant leverage in facilitating broad advances. As the above discussion suggests, the technical barriers to progress are tightly intertwined with policy and cultural issues. This is why the proposed NRC activity would add value above what is accomplished in the technical conferences that already take place.

Proposed Plan of Work

The National Research Council (NRC) will appoint a broadly cross-disciplinary committee of about 7 experts to carry out this activity. The committee will be composed of prominent experts from data-intensive fields of science and engineering. We will seek industrial, defense, and academic experts in computer science and information technology, encompassing expertise in data integration, databases, statistics, middleware, digital libraries, machine learning, earth and life sciences, physical sciences, and national defense. Following its normal practice, the NRC will solicit suggestions from a wide variety of sources, with the Computer Science and Telecommunications Board, Committee on Applied and Theoretical Statistics, Government-University-Industry Research Roundtable, and U.S. National Committee on Data leading the canvassing.

The study committee will carry out the following charge:

> *Plan and organize a cross-disciplinary public workshop to explore alternative visions for achieving large-scale data integration in fields of importance to the federal government. Large-scale data integration refers to the challenge of aggregating data sets that are so large that searching or moving them is non-trivial, or to the challenge of drawing selected information from a collection (possibly large, distributed, and heterogeneous) of such sets. The workshop will address the following questions:*
>
> - *What policy and technological trajectories are assumed by some different communities (climatology, biology, defense, and others to be decided by the committee) working on large-scale data integration?*

- What could be achieved if the assumed policy and technological advances are realized?
- What are the threats to success? Who is working to address these threats?

The committee will develop a consensus summary report based on the workshop to capture diverse perspectives on these questions and identify major opportunities and threats that cut across multiple communities dealing with the challenges of large-scale data integration.

The committee will hold a series of conference calls to adjust the plans sketched out below and identify up to 15 speakers for the public workshop, and it will also hold a short meeting in conjunction with the workshop. The workshop will include speakers with academic, industrial, and government experience. In addition to invited speakers, many federal scientists and engineers from data-intensive fields will be invited to participate in the workshops. The estimate of costs includes funding for a speaker from Europe so that European perspectives can be heard.

The workshop will elicit the policy and technological visions underlying several communities' efforts at large-scale data integration. Speakers from the chosen communities will be asked to do the following:

- Give an overview of their community's current efforts, capabilities, and limitations with regard to data integration;
- Articulate their community's "stretch goals" relating to data integration;
- Estimate the opportunities associated with being able to integrate vast amounts of data and mine it for secondary purposes and the opportunity costs for failing to create that capability;
- Identify cultural, policy, technological, and economic issues that are limiting their community's progress, such as mismatched expectations about data sharing, weak incentives for documentation or for long-term access to data products, mismatches between the needs of scientists and engineers and commercial software producers, and so on.

Workshop attendees will identify common challenges and shared visions, and generally parse the landscape of issues that are limiting progress in large-scale data integration.

Meeting after the workshop adjourns, the committee will identify the key perspectives that were raised and collect its preliminary thoughts about commonalities and important distinctions across communities. Subsequently, it will work through conference calls and email exchanges to draft summary thoughts on the following questions:

- What challenges and opportunities associated with large-scale data integration are common across multiple fields/agencies?
- What are the highest-priority policy issues identified at the workshop?
- What suggestions were given about how to address these policy issues?

A transcript of the workshop will be edited and used as the basis for the committee’s summary report. After the committee has finished its drafting, the report will undergo peer review and editing according to standard Academies practice. It will be released as a published report of the

National Academies Press and posted on its Website. The report will be widely disseminated among policy leaders, the research community, agency officials, and others working in large-scale data integration. The committee chair, or another member, will be available to present the results to federal agencies with an interest in large-scale data integration.

Overview of the Technical and Policy Issues

The technical and policy issues to be explored in the workshop can be clustered into three general areas:

1. How to develop and effectively apply enabling technologies

Some of the key enabling technological advances needed for large-scale data integration include:

- tools and services that automatically capture or create metadata in real time;
- wrappers;
- algorithms and their implementation for search, statistical sampling, and visualization; and
- visual analytics to synthesize content and allow insights to be drawn.

Progress is already being made in these technologies, but significant further advances are necessary.

The workshop will explore whether technical capabilities developed in proprietary settings or in particular science and engineering communities are appropriately shared elsewhere in science, engineering, and intelligence. If not, what are policy options for improving the diffusion of insights and tools that might be applicable outside their own domains? Among the policy issues to be consider are options for eliminating anticipated technical bottlenecks. For example, should we consider incentives for companies to pass along toolsets they have developed—where the market is too small to deploy commercially—to research or other federal users? Are federal funding streams aligned with opportunities and risks?

Another technical issue to be explored is comparison of emerging ideas for knowledge discovery of non-text information. The workshop will also examine the state of methods for analyzing distributed information without pulling all the data to a central location.

2. How to support the development of open, stable, and flexible standards

Many of the standards relevant to data integration and interoperability are developed at the community level. In some fields, particularly those in which "big data" has long been central to addressing cutting-edge research questions, organizations and institutions are well established. In other fields, communities that are used to setting their own standards are realizing a need to collaborate closely with other communities. This raises questions such as the following:

- How is the larger technical and policy context for data standards evolving, and what are the implications? For example, what impact will the emergence of the semantic web have on incentives for making data available, and the adoption of standards that support open access?
- What are the important differences in "data cultures" between fields, and how do they act as barriers?

- What are the most effective organizational models for developing standards? Specific areas for ongoing development of standards relevant to data integration and interoperability include ontologies and vocabularies, and standards for metadata.
- Are additional incentives needed to encourage communities to work together?
- How can the federal government most effectively use its leverage to encourage the development of open, flexible, stable standards?

3. What are the prospects for wider availability and openness of data?

In fields where effective data integration/interoperability holds the promise of significantly advancing a community's mission, special questions arise over how to deal with requirements for protecting the security or privacy of data. Some combination of regulations, standards, and new technologies will be required to realize this promise. For instance:

- To what extent (and in what timeframe) can technical fixes (e.g., software that automatically "scrubs" elements from data that would identify individuals) be expected to address privacy issues in using personal data in areas such as biomedical research, epidemiology, and public health? What are the issues and questions that need to be addressed in developing policy frameworks and standards for using such data?
- In intelligence and national/homeland security contexts, it may be necessary or desirable to integrate classified data elements in a way where results can be openly shared, or to determine when results of integrating open data sources needs to be classified. What combination of technical progress, organizational capability, and new policy is needed to realize the potential of data interoperability in these fields?

Consumer privacy is of course a very large policy issue connected with greater openness of data. The study committee will decide what aspects of consumer privacy need to be considered in the course of this study as part of its process of focusing the workshops.

Prior National Academies Work of Relevance

Commission on Physical Sciences, Mathematics, and Applications, 1995, *Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data*
Committee on Applied and Theoretical Statistics, 1992, *Combining Information: Statistical Issues and Opportunities for Research*
Committee on Applied and Theoretical Statistics, 1999, *Record Linkage Techniques—1997*
Computer Science and Telecommunications Board, 2005, *Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for a Long-Term Strategy*
U.S. National Committee on Data (CODATA), 2004, *Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium*
Policy and Global Affairs Division, 2002, *Preparing for the Revolution: Information Technology and the Future of the Research University*

Current and Pending Proposals Under NIH Task Order Contract No. N01-OD-4-2139

As of the date of the submission of this proposal, we are unaware of any current or pending proposals for a project under Task Order No. N01-OD-4-2139 on the same or closely related topics

Current Support Under NIH Task Order Contract No. N01-OD-4-2139

There is one on-going study funded by NIH Task Order Contract No. N01-OD-4-2139 that has some overlap with the proposed work: Task Order #131 provides $125,000 as partial support of the National Research Council's study on "Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data." That study addresses a particular policy aspect of data integration, which is not a major focus of the activity described by this proposal. If funded, the project that is the subject of the current proposal will incorporate results from Task Order #131 as available and will invite a member of the study committee funded by that Task Order to speak at the proposed workshop.

None of the other currently on-going studies funded by NIH under Task Order No. N01-OD-4-2139 are on topics that are closely related to the proposed work and which would require coordination with this activity should an award be received as a result of this proposal.