



Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop

Scott Weidman and Thomas Arrison, Rapporteurs;
National Research Council

ISBN: 0-309-15443-X, 58 pages, 6 x 9, (2010)

This free PDF was downloaded from:
<http://www.nap.edu/catalog/12916.html>

Visit the [National Academies Press](#) online, the authoritative source for all books from the [National Academy of Sciences](#), the [National Academy of Engineering](#), the [Institute of Medicine](#), and the [National Research Council](#):

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, [visit us online](#), or send an email to comments@nap.edu.

This free book plus thousands more books are available at <http://www.nap.edu>.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

STEPS TOWARD LARGE-SCALE DATA INTEGRATION IN THE SCIENCES

Summary of a Workshop

Scott Weidman and Thomas Arrison,
National Research Council, *Rapporteurs*

Committee on Applied and Theoretical Statistics

Division on Engineering and Physical Sciences

Policy and Global Affairs Division

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

This study was supported by Contract Number N01-OD-4-2136 between the National Institutes of Health and the National Academy of Sciences, Grant Number 60NANB7D6126 from the National Institute of Standards and Technology, and Grant Number N0014-07-1-0557 from the Office of Naval Research. Any opinions, findings, or conclusions expressed in this publication are those of the authors and do not necessarily reflect the views of the agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-15442-0

International Standard Book Number-10: 0-309-15442-1

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313; Internet, <http://www.nap.edu>.

Copyright 2010 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

**PLANNING COMMITTEE FOR THE WORKSHOP ON
OVERCOMING POLICY AND TECHNICAL BARRIERS
TO LONG-TERM DATA INTEGRATION**

MICHAEL STONEBRAKER (*Chair*), Adjunct Professor of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge
JOSEPHINE CHENG, IBM Fellow and Vice President, IBM Almaden Research Center, Almaden, California
TIMOTHY FRAZIER, Senior Architect, National Ignition Facility, Lawrence Livermore National Laboratory, Livermore, California
CARL KESSELMAN, Professor of Industrial and Systems Engineering, University of Southern California, Los Angeles
CLIFFORD LYNCH, Director, Coalition for Networked Information, Washington, D.C.
RAGHU RAMAKRISHNAN, Chief Scientist, Audience and Cloud Computing; Fellow and Vice President, Yahoo! Research, Santa Clara, California

Principal Project Staff

SCOTT WEIDMAN, Study Director
THOMAS ARRISON, Study Director
BARBARA WRIGHT, Administrative Assistant
BETH COBB DOLAN, Financial Manager

Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We wish to thank the following individuals for their review of this report:

Michael Goodchild, University of California, Santa Barbara,
Laura Haas, IBM Almaden Research Center,
Arie Shoshani, Lawrence Berkeley National Laboratory, and
Alex Szalay, Johns Hopkins University.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the report's conclusions, nor did they see the final draft of the report before its release. The review of this report was overseen by Jeff Dozier of the University of California, Santa Barbara. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures

and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the rapporteurs and the institution.

We also thank National Research Council staff members Jon Eisenberg and Paul Uhlir for their constructive comments on an earlier draft of this report.

Contents

1	INTRODUCTION	1
2	THE CURRENT STATE OF DATA INTEGRATION IN SCIENCE	6
	Data Integration Goals, 6	
	Size and Other Characteristics of Some Scientific Data Sets, 8	
	Complexity of Data Sets, 10	
	Distributed Nature of the Data, 11	
	Metadata, 12	
	Data-Integration Tools, 13	
	Crosscutting Discussion, 15	
3	IMPROVING CURRENT CAPABILITIES FOR DATA INTEGRATION IN SCIENCE	18
	Federators, 21	
	Resource Description Framework, 24	
	MapReduce and Its Clones, 26	
	Data Management for Scientific Data, 28	
4	SUCCESS IN DATA INTEGRATION	31
	Freebase, 32	
	Melbourne Health, 32	
	Science Commons and NeuroCommons, 33	
	Bio2RDF, 34	

<i>x</i>		<i>CONTENTS</i>
5	WORKSHOP LESSONS	35
	REFERENCES	38
	APPENDIXES	
A	Workshop Agenda	43
B	Workshop Participants	47

1

Introduction

This report summarizes a National Research Council (NRC) workshop to identify some of the major challenges that hinder large-scale data integration in the sciences and some of the technologies that could lead to solutions. The workshop was held August 19-20, 2009, in Washington, D.C. The charge to the planning committee was as follows:

To 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 nontrivial, 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 NRC Committee on Applied and Theoretical Statistics organized the activity, with the original impetus coming from discussions of the NRC's Government-University-Industry Research Roundtable.

Advances in information technology have resulted in enormous

increases in the amount of data available to science and engineering researchers. This includes not only data from experiments and observations but also data generated by computer simulations. It is becoming common for research groups to quickly gather or generate terabytes of data, and a number of programs are accumulating petabytes of data. (One terabyte equals 10^{12} bytes and 1 petabyte equals 10^{15} bytes.) Data integration must overcome the challenge of finding disparate, distributed sources of data, which is often referred to as “data discovery,” and the challenge of effectively utilizing the collective information in those sources to produce new insight—a process known as “data exploitation.” The workshop on which this report is based did not try to characterize comprehensively the various ways in which data integration is useful or necessary for the advance of science.

The term “data integration” first emerged in connection with the need for organizations to provide data users “with a homogeneous logical view of data that is physically distributed over heterogeneous data sources” (Ziegler and Dittrich, 2004). The concept of data integration used here is a broad one, encompassing 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 the concept’s scope.

All too often, data discovery depends on word of mouth: A researcher happens to have heard about a data set that might be useful in his or her own research or makes inquiries of colleagues in order to find relevant data. In fields where there are a limited number of large facilities (for example, high-energy physics and astronomy) or a predictable administrative structure for data storage (for example, national weather bureaus), the challenge may be manageable, although meeting it still often depends on a haphazard, serendipitous process. But in research fields where small groups can accumulate and store large amounts of data, valuable data sets can exist in many places. In particular, useful data might be held by someone who is outside the network of a researcher who is seeking those data. More problematic still are instances where a researcher seeks to integrate data from very different communities, such as geospatial data with sociological, medical, and other overlays. Such creative merging of knowledge can lead to very novel insights, but it is hindered by the data discovery challenge.

Once data sources have been found, data exploitation presents another set of challenges. A researcher must develop a clear understanding of the meaning of each of the data sets. Achieving such an understanding is difficult, because documentation of the conditions under which the data were collected can be spotty. Simple aspects such as the units of measure must be known definitively, and more subtle aspects such as environmen-

tal conditions, equipment calibrations, preprocessing algorithms, and so on can also be important. If data are being used for research outside the field for which they were collected, the risk of misinterpretation is severe, because research communities can have unstated assumptions about what to document or what to assume, and these assumptions can be overlooked during the integration process.

There are technical and policy challenges associated with the actual aggregation of data. If some data were collected with privacy guarantees, how should those guarantees be interpreted if only a subset of the data, or a summary of it, is used for a secondary analysis? There are also technical challenges in translating disparate data sets so that they can be merged: for example, putting maps into the same coordinate system, aligning data that were collected on different sampling grids, correcting for systematic differences among equipment, and so on.

For the purposes of the workshop, “large-scale data integration” was taken to refer to the aggregation of data sets that are so large that searching or moving them is nontrivial, a technical challenge that is becoming ever more common as it becomes easy to produce and store terabytes. Workshop participants were also aware that a growing number of opportunities require the aggregation of large numbers of modest-size datasets, and some of the workshop discussion reflects the challenges associated with those situations. To bound the discussion and produce the most useful outcomes, the workshop planning committee decided to focus on issues related to integrating scientific research data.¹ The particular disciplines discussed include physics, biology, chemistry, Earth sciences, satellite imagery, astronomy, geospatial data, and research medical data. By and large, these are all structured data—that is, records of fairly rigidly formatted information. In contrast, many data integration efforts outside scientific research deal more with unstructured data (text) and semistructured data (want ads, personnel records, and so on). Unstructured data and the needs of nonresearch users with an interest in data integration were not a focus of the workshop. Of course, there is a substantial gray area. For example, even when one is seeking and aggregating structured scientific data, tools designed for unstructured data might be necessary because structure may not be readily recognizable.

Michael Marron of the National Institutes of Health (NIH), a co-sponsor of the workshop, explained NIH’s interest in the topic. The long-

¹ The statement of task and original work plan for the project documented in this report presumed two workshops and a committee consensus report. The project was scaled back to one workshop and a rapporteur-authored summary in order to align with available resources. The workshop planning committee decided that focusing the subject matter coverage on scientific research data and related communities would allow for the most productive discussion of issues and possible solutions during a single two-day workshop.

time predictions about the data deluge have come to pass: Many fields of science now have more data than they know what to do with. The amounts of data being collected are increasingly important to biomedical research. In addition, more and more research is now built on the analysis of data that were not collected by the researchers themselves, and many of the extant data have not been utilized to their full potential. Alex Szalay of Johns Hopkins University reported that analogous changes are underway in astronomy, with the collection of data increasingly separated from its subsequent analysis, which is a disruption from the way science has been practiced over the centuries. Increasingly, the connection between data and their analysis is facilitated through data archives and different sorts of federation services. This represents a new way of doing science, and the infrastructure must be able to support it.

Dr. Marron expressed concern about science's abilities to share, manage, and curate data; correct errors; and map the provenance of data. In short, he is concerned about all of the factors that go into ensuring the reliability of data and enabling their exploitation. Thus, NIH is exploring where to make investments in building those capabilities and generally developing parts of the information infrastructure. He pointed to the Biomedical Informatics Research Network (BIRN) as an example of an NIH investment in information. He said it is not *the* solution, but that it is an important contribution to an infrastructure that will help facilitate sharing data and tools. The Cancer Bioinformatics Grid (CaBIG), a similar infrastructure for cancer-related research, is another example.

Dr. Marron said that it is far from clear how one can find and access data. He noted the common hope for a capability that would be as useful as Google and other search engines but that could also perform more of the exploration and filtering that is now left to researchers. This hoped-for tool could work with multidimensional data and could find not only the data that are deliberately made available ("published" or placed in repositories) but also the huge amounts of data that are less readily identified but nevertheless of value to people other than those who collected them. It would also have to have the ability to recognize data sets that are similar, redundant, or overlapping.

Ed Seidel of the National Science Foundation (NSF) explained that the tendency to collaborate is increasing in every single area of science. He gave the example of modeling the effects of hurricanes and storm surges, which requires bringing together a wide range of models and data, including satellite observations, atmospheric models, storm-surge models, wave models, levee models, traffic flow models, and so on. This increase in the prevalence of collaboration calls for cyberinfrastructure to support distributed teams of researchers who collaborate through sharing data.

The workshop examined a collection of scientific research domains, with application experts explaining the issues in their disciplines and current best practices. This approach allowed the participants to gain insights about both commonalities and differences in the data integration challenges facing the various communities. In addition to hearing from research domain experts, the workshop also featured experts working on the cutting edge of techniques for handling data integration problems. This provided participants with insights on the current state of the art. The goals were to identify areas in which the emerging needs of research communities are not being addressed and to point to opportunities for addressing these needs through closer engagement between the affected communities and cutting-edge computer science.

The workshop also discussed policy barriers to widespread data sharing, considering the pros and cons of various ways forward.

2

The Current State of Data Integration in Science

The workshop opened with a series of presentations about data integration challenges and approaches in several areas of science.

DATA INTEGRATION GOALS

Carl Kesselman of the University of Southern California presented some examples of how data integration provides value to biomedical research and shared his vision of important goals. He noted that there is a generic shift in biomedical research, from advances being based on a new understanding of fundamental biological mechanisms to advances being driven by patterns in data. That is, insights are arising from connections and correlations found between diverse types of data acquired from various modalities. An example is the use of biomarkers—the finding of connections between data that suggest predictors or indicators of various disease profiles, diagnostic procedures, and so on. Related to this is the carrying out of retrospective studies to discern patterns that suggest mechanisms that might be investigated. These trends are driving the need for data integration. Dr. Kesselman gave several examples of research results that required identifying and retrieving data from distributed locations.

Alex Szalay of Johns Hopkins University observed that many fields of science are becoming data intensive, and thus reliant on cyberinfrastructure. An example is the use of virtual observatories in astronomy, in which the database serves as a sort of laboratory in which an astronomer

can make “observations.” The Large Hadron Collider (LHC) and human genome research are other examples of data-intensive science. These efforts require sizeable investments in software. Dr. Szalay estimated that the Sloan Digital Sky Survey (SDSS) allots some 30 percent of its budget to software, and the Large Synoptic Survey Telescope (LSST) project is planning to allocate 50 percent to software. The LHC’s data-management elements constitute a major part of the overall operation. Tim Frazier of Lawrence Livermore National Laboratory added that a large amount of hardware is also required if one plans to move data into and out of a large repository: many network switches and high-speed networks. If computations can be performed within the repository, they can be carried out faster and more efficiently.

Michael Stonebraker of the Massachusetts Institute of Technology (MIT) observed that a virtual observatory requires a global schema,¹ a concept that has not worked very well in most enterprises. There have been numerous efforts to develop global schema, but anticipating the many questions that might be posed of the data constitutes a significant barrier. Orri Erling of OpenLink Software, Inc., suggested that there might be a need for a framework that enables a globally evolving schema. Dr. Kesselman said he has had a reasonable amount of success by aiming for a point somewhere between the notion of a global standard and total chaos. He saw something similar in the workshop’s presentation on data integration at the National Oceanic and Atmospheric Administration (NOAA), which suggested defining limited communities of interest in order to constrain the problems of data interoperability. He saw the challenge as providing the infrastructure to support an exchange of information within communities of practice that are connected but not global.

Clifford Lynch of the Coalition for Networked Information pointed out that there are two kinds of data reuse and felt the notion suggested by Kesselman is not a complete solution. One kind of reuse is reexamination of data for a compilation or a meta-analysis, in conjunction with similar data and carried out for purposes that are not too far afield from those that drove the collection originally. The other kind is reuse of data outside the disciplinary frameworks within which they were collected. Some examples of reuse are very interdisciplinary and jump the fences between science, social science, and humanities in unpredictable ways. These latter types of reuse make it difficult to know which kinds of life cycle should be assumed. No one has a good understanding of the kinds of metadata that facilitate reuse of data in a new context. In contrast, we have a much better understanding of the kinds of metadata that facilitate

¹ A global schema is a single structure that can be used to organize all the data stored by a specified field.

incremental or predictable kinds of reuse, such as meta-analyses and compilation. In addition, we normally conceive of metadata as documenting a data set in isolation. But if data sets are to be integrated, it might be important to include metadata that inform the integration process, such as metadata about commonalities or disparities that reduce or increase uncertainties when those sets are aggregated.

Dr. Szalay asked how the value of data is established, because that would guide planning for data reuse. Making data accessible for reuse requires resources, but in general there is no clear business model for who should pay and how much. Dr. Lynch pointed out that we do not generally understand the cost-benefit trade-offs of metadata: how much it costs to create metadata and how much they improve discovery and usability. By “metadata,” Dr. Lynch meant more than just the documentation purposely attached to a data set. He said that reuse could also require other documentation about, say, the technologies used for the data collection or generation, but in general we have a very imprecise understanding of what to retain. We also do not know when it is worthwhile to hold onto data with fairly deficient metadata in the hope that someone who cares enough will figure it out. In some cases, data with deficient documentation can be as useless as no data at all or as dangerous as corrupt data.

In a related question, Dr. Lynch asked who should take charge of these data for the longer term. These investments are a necessary part of scientific research, but they are not routinely accounted for in budgets and plans. He proposed that one of the most compelling problems is how to give concrete guidance to the research communities about what is good practice in handing off data at the end of a project so that they can be curated and made available for reuse.

SIZE AND OTHER CHARACTERISTICS OF SOME SCIENTIFIC DATA SETS

The sheer amount of data available in many fields of science is well known. Dr. Szalay reported that astronomy is experiencing a doubling of data every year. The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) will soon contain more than a petabyte (PB) of data, and the Visible and Infrared Survey Telescope for Astronomy (VISTA) will reach the same level about a year later. The Large Synoptic Survey Telescope (LSST) project might accumulate hundreds of petabytes of data over the next decade, while the proposed Square Kilometer Array (SKA) would be receiving a terabit of data per second from the radio antenna. The Sloan Digital Sky Survey ended up with more than 18 TB of data. Frazier reported that the National Ignition Facility is producing some

5 PB a day. Dr. Kesselman added that genome-sequencing machines can produce 1 TB per week of information, and Dr. Marron observed that this rate will probably grow to terabytes per day soon. In addition, simulations produce enormous amounts of data and will soon also be operating at the petascale.

Keith Clarke of the University of California, Santa Barbara, observed that new data sources are dramatically improving the resolution of geospatial data. For example, it was common until recently to find databases of terrain elevations with 30-m spatial resolution, but now 10- or 3-m resolution is common. Similarly, elevations may now be measured more precisely: For example, Dr. Clarke is mapping his own campus at 2-cm resolution with terrestrial scanning lidar (light detection and ranging, a laser-based means of measuring distances). Many streams are coming directly from GPS devices that record 2 to 3 readings per second. When there are multiple paths, time and space stamps are recorded at submeter accuracy. These developments call for new methods when integrating geospatial data—when registering one data set with another, for example.

Disciplines such as astronomy and high-energy physics rely on a limited number of large-scale data sources. In that situation, plans and protocols can be put in place to manage the data and their reuse, and these steps facilitate the finding of data. More difficult are disciplines where enormous data sets can be produced by many laboratories and research groups. Dr. Marron pointed to genome sequencing as an example. With emerging capabilities, a biology laboratory will be able to produce over 100 billion base pairs a day, which begins to rival the 150 billion base pairs produced in all of 2007 by the Human Genome Institute. A laboratory that produces 100 billion base pairs per day will never be able to fully analyze those data, so researchers from the broader community will have to be called on. First, however, they will have to be able to find the data, then download them, and then have a mechanism for analyzing them, all of which are challenging.

Dr. Frazier pointed out that as data volumes get larger and larger, at least the first round of analysis is most efficient if it is done where the data reside rather than importing enormous portions of data into an analysis tool. He suggested that certain data sets for climate research clearly exceed that threshold, and spoke of a colleague who works with a 350 PB database of climate information. At that scale, it is impractical to query anything out of the database without some initial analysis to cut the size of what is returned. The efficient approach—the only one for massive sets—is to ship analysis code to the database and begin the computation where the data reside.

COMPLEXITY OF DATA SETS

Thomas Karl of NOAA's Climate Data Center gave some indication of the complexity of climate data. Data are collected from a number of observing systems making atmospheric measurements from space and on land and the ocean. Different research communities collect in the different domains, and the communities could be even further segregated according to their collection modality (such as radar or infrared). These communities tend to have their own practices for formatting data, so it is necessary to have software that can translate data sets into compatible formats. In NOAA alone, some 50 different formats have been identified. The subcommunities in climate research have different ways of reporting uncertainty: Some include confidence intervals with the data, while others report best estimates. More generally, there are silos in modeling systems, measurement systems, and knowledge systems, and even different concepts of what to include in the metadata. Integrating these disparate data streams to create a whole-system view is far from trivial.

Dr. Szalay observed that the astronomy community and its data are similarly disaggregated. For example, most infrared data are stored in Pasadena, x-ray data in Boston, and optical ultraviolet data in Baltimore.

Dr. Seidel observed that, in addition to developing standards that would allow disparate data streams to interoperate, there is also a need to specify the properties of the algorithms that produced them, to observe how they interact with one another, and to understand how uncertainties and errors might propagate from one to the other. Dr. Szalay pointed out that, with the rapidly increasing abilities to collect data and produce simulations, algorithms have to stay bounded, scaling perhaps as $O(n \log n)$. N-squared or N-cubed algorithms are not useful in these data-intensive fields, or at least not for long. Dr. Clarke added that there are large inherent uncertainties within geospatial data more generally, arising in particular from measurement errors and from mismatches between data collected at different times or through different means. Research is just beginning to explore how to deal with the uncertainties visually and statistically.

Dr. Clarke also noted that, with geospatial data, the goal is often more complex than producing single map layers (showing, for example political boundaries, geography, or roads), although that can be extremely valuable. Rather, one might produce multiple map layers and then coregister them to allow detecting differences in each image. These differences can be happening at different timescales—minutes (in the case of some military imaging) to years. Imagine climate research that might need to examine changes that take place over decades or even centuries, such as fluctuations in glaciers or the urbanization of terrain.

Each geospatial data set is associated with one of the 39 or 40 reference systems that can be used for determining Earth's average shape, depend-

ing on the accuracy desired. Some systems convert place names to coordinates and others deal with problems of tiling, mosaicing, registration, edge matching, conflation, temporal inconsistencies, and so on. Many choices are made in converting raw data to their final form, and these choices must be accounted for when data are integrated.

Dr. Kesselman offered an example of a data-integration challenge from the life sciences, where the data from one functional magnetic resonance imaging (fMRI) machine need to be calibrated with those from another before any integration can be performed. The correct metadata must be associated with each data set so that the integration can proceed properly. He pointed to a research study that examined whether there are positive symptoms in schizophrenics associated with severe temporal gyrus dysfunction. Answering that question required integrating results from multiple data sources, from multiple sites, and multiple imaging modalities. The fMRI data came from multisite studies that were distributed across 15 different scanners at 15 different sites. But these data sets had to be integrated in order to answer the question.

Dr. Kesselman said that biomedical research is increasingly dealing with new types of data, such as genomics, blood proteins, and new imaging modalities. Clinical observations and clinical data become a critical part of doing biomedical research in a lot of settings. The diversity of data types means we have to look at questions such as whether we can do analytics or queries across genomics and brain structure and imaging, and the research must look at all of those things simultaneously. Lastly, data integration in the biomedical sciences is distinct in that there are fairly severe privacy and data anonymization issues because the work often involves the use of information about individuals and may include identifying information.

Dr. Lynch ended the discussion by saying that the semantic complexity of a data set is not really closely correlated to its size. While a really massive data set is very likely to cause technical problems, difficult challenges can arise with small sets, too. While they do not necessarily cause enormous technical problems in terms of the computational, storage, or communications capacity required to work with them, they can be semantically complex. They can also be hard to characterize, and they can be quite difficult to integrate or reuse because of that.

DISTRIBUTED NATURE OF THE DATA

Some areas of science tend naturally to have data sources that are distributed geographically. Biomedical research is one of these: Its research groups tend to be associated with universities or hospitals as opposed to being clustered at single large facilities. Thus, according to Dr. Kesselman, data integration is a common challenge in that field. Climate research, too,

often involves data integration, Dr. Karl observed, not only because data collected or compiled at different sites and with different types of instruments need to be integrated, but also because the simulation centers distributed around the globe are increasingly central to progress. For example, the 2007 report of the Intergovernmental Panel on Climate Change required coordinating many different model simulations and developing an archive of their outputs. This was a major nontrivial advance: For the first time, a researcher could analyze 25 different models and attempt to develop error bars for their integrated projections.

As noted above, there are a limited number of large data repositories in astronomy, so researchers tend to know where to find the biggest data sets. But it is much more difficult to learn about the many smaller data sets, said Dr. Szalay, and they are just as important for many lines of investigation. Because the threshold for properly publishing smaller, more specialized sets of data and adding the necessary metadata might be too high for an individual or small group. Dr. Szalay believed there would be value in a repository service that enables users everywhere to obtain a single view of collections all over the world.

METADATA

Because information is gathered when a researcher is first granted access to the equipment at a controlled-access experimental facility, it is possible to capture at least some of the metadata automatically. Dr. Frazier noted that records of every experiment run on the National Ignition Facility show who worked on the machine, when they were working on it, and what they were doing. There is detail about precisely what was installed in the machine at the time the experiment was run, how it was calibrated, and what happened for each of those parts. Such records are not captured automatically in most experimental settings, however.

Laura Haas of the IBM Almaden Research Center noted that the database community has been integrating work flows more generally, to include not just metadata but also information about the subsequent analysis, such as which data were selected, which analyses were performed, and what methods and software were used, all of which could apply to situations such as the one mentioned by Dr. Frazier. Dr. Haas suggested that the VisTrails work by Juliana Freire at the University of Utah could be useful in that regard. VisTrails keeps track of the details of experimental setups and the resulting data provenance, all in an XML database.

Unfortunately, the compilation of metadata is often given the lowest priority and assigned to the most junior people in a research group, in Dr. Karl's experience. This can mean that metadata are suboptimal for some or many secondary applications of the data.

Philip Bernstein of Microsoft Research asked the physical and life scientists at the workshop whether there are any incentives for them to make their own data reusable. If, for example, it takes three times as much the effort to make data reusable as to simply create them for one's own use (which he thought was probably a good estimate) do scientists get rewarded for that effort, or is this just a labor of love? If, on the other hand, a researcher looking at an important question has the choice of either pursuing a large grant that will purchase a new instrument and support other people or reusing existing data, how will he or she decide? Are institutions and reward systems biased in favor of the former course of action?

Dr. Szalay noted that astronomy has seen a substantial sociological shift over the last 10 years in this regard. Before, people were not sharing data very much—they kept their tapes in their desks. Now the community has almost reached the point where a researcher would be questioned if they did not find and reuse existing data. It is key, here, that the researcher(s) who originally collected the data and made them available for reuse be acknowledged and receive scholarly credit akin to that received when one of their publications is subsequently cited.

Dr. Seidel suggested that data-management plans might be made a requirement for proposals, with reviewers being instructed to take that part of the proposal seriously. For example, since 2003 NIH has required that all proposals involving direct cost expenditures of more than \$500,000 per year must include a data-sharing plan. Dr. Szalay added that it would also be useful to have supplemental funds available for data management just as supplemental funds sometimes are available for educational elements of a research proposal. If the reviewers conclude that the data-management plan is a good one, then the researcher would receive the supplement rather than having to pay for data management and scientific research from the same pool of funds.

DATA-INTEGRATION TOOLS

Dr. Clarke pointed out that geospatial researchers have created the beginnings of a data-integration policy through the adoption of the National Spatial Data Infrastructure. Other countries have developed their own counterpart standards of practice. Early federal standards were top-down and not as successful as those that have emerged through consensus.

Dr. Kesselman gave an example of tools called the human imaging database (HID), which was developed in conjunction with the fMRI research mentioned above. One of the functions implemented in the HID is the ability to do distributed database query and the ability to do data integration. More generally, the biomedical field has shown a lot of interest in ontologies and defining vocabularies and dictionaries. That interest has

led others to explore the federation of the semantic descriptions of the data. Dr. Kesselman's own group has been looking at whether there are reusable data integration tools that can be applied in order to avoid creating data federations and data environments for specific uses.

Dr. Frazier said that for NIF data, which are intended to be held for 30 years, unique identifiers are generated. These are better than surrogate keys in databases, because their value is not lost if the data are at some point migrated into a new schema with new surrogate keys. These identifiers can be applicable throughout the long life of the data.

He also observed that scientists are not generally afraid of new technologies; however, they certainly are afraid of interrupting their work while someone creates specialized software or when software has failure modes that the scientist cannot fix on the spot. Scientists also are unwilling to invest in new technologies unless they know that long-term support exists. Finally, most prefer to control their data and analyses personally, so they lean toward the use of methods and software that they understand and can run themselves. There can be resistance to methods and software that are less familiar or that require control to be transferred to another person.

In contrast, though, Dr. Szalay pointed to the increasing presence of federations with astronomy data. For example, the National Virtual Observatory was set up with the explicit understanding that it would own the massive data sets and manage them as they evolve. Dr. Lynch observed that this delegation of data stewardship is a general trend for other scientific projects that generate vast data streams. Such endeavors usually operate at a scale large enough that someone has the mandate to plan for information management. Usually, information management is factored into the budgets, and Dr. Lynch sees some willingness among funders to also provide some money for data stewardship.

However, with medium- to small-scale science, any kind of data management or information management is often an ad hoc process. Frequently it is relegated to graduate students because the enterprise is not large enough to support specialists in data management and information technology. Both the funders of scientific research and the researchers themselves have over the past decade come to recognize that data are a very important part of the output of research, one that deserves management in its own right. But data sharing and data stewardship fall into two different timescales. The former often happens on timescales that are similar to those of a research project, perhaps extending a few years longer. But data stewardship operates on timescales that are more familiar to data archivists and research librarians, which are longer than the active professional life and interests of many of the researchers involved in the project that produced the data (and certainly longer than the tenure of a graduate student). More important for the question of data integration, stewardship timescales may exceed the lifetimes

of the experimental or computational environments that created the data, making it difficult to interpret the data because tacit knowledge erodes as the people involved move on and the corporate memory is lost.

Dr. Stonebraker asked how much of the data-integration problem would be solved if software were developed to address the technical challenges mentioned so far in this section. Dr. Szalay responded that such software would solve a lot of the issues, especially if it were scalable to the tens or hundreds of petabytes. But it would not solve the problem of very large, dispersed data, which requires figuring out what to do when a petabyte of data must be moved on demand across the Internet, or how to avoid that. Dr. Szalay said such movement is possible now for sets up to tens of terabytes of data, but it will not be possible for at least 5 or 10 years or more with petabytes of data.

CROSSCUTTING DISCUSSION

Michael Brodie of Verizon Communications brought up some more general issues of standards across enterprises. Establishing and maintaining standards in a very large community, whether a scientific community or an enterprise, is difficult because there are few general principles to help one decide how well a given standard will suit a particular data set, particularly when that data set is innovative or might be subject to novel reuse sometime in the future. It is unclear how to assess whether an existing standard can be extended, or whether a new standard should be developed from scratch.

Alon Halevy of Google, Inc., suggested that improvements to search tools could be a productive way to improve data-integration capabilities. For certain types of science, the hardest part of data integration may be finding the few data sources that are relevant to the research task. The integration will often be ad hoc, done for one task and then finished, and that is fine. So the real bottleneck is the ability to find the necessary databases among the thousands or millions of data sets that might be relevant. In such cases, the key enabler might be including metadata that allow the data set to be uncovered by a search engine. Similarly, it would be very useful if search engines could find and index the many data transforms that various groups have developed for a wide range of integrations.

Dr. Marron raised another topic having to do with data sharing and access. He thought that several workshop speakers were suggesting that the solution was for funding agencies to just require everybody to share data. Certainly there have been instances where that has been done. At NIH this has been given serious consideration, and some programs have requirements for data sharing.

But there is also the other aspect of that, the motivation, which needs to be investigated. Dr. Marron suggested that more widespread use of registries would be very helpful here. Properly designed and managed registries are not only able to facilitate the reuse of data, but they might also improve the incentives for sharing them. Registries could affect the incentives by designating the entry of data as equivalent to publication. Then, by tracking reuse, registries could provide tenure and review committees with credible statistics about how widely the data were used or the value of their publication, measures that are analogous to those associated with paper publications and citation indexes. In addition, registries could provide a means of enforcing certain rules about metadata, because to register the data one must include appropriate metadata. It is important, though, that the registries be supported over the long term or else researchers will be wary of investing the time and effort to contribute to them.

Repositories are not the same as registries: New, large, centralized repositories are very unlikely to receive funding in Dr. Marron's view because, once an agency gets involved in supporting a database, it is a never-ending process. NIH supports a number of large centralized databases, and their cost has increased dramatically. He thinks the better model is distributed databases and distributed costs of maintaining them.

David Maier of Portland State University raised the question of how to train people to work effectively with shared data. The best technology solutions require some domain knowledge along with knowledge from computer science. Dr. Maier is not sure if it is better to start with people from a domain of science and try to give them data management skills, or start with people from a computer science background and try to give them domain skills. In his personal experience, computer science students do not get nearly enough statistics training to do the kinds of analyses they might be called upon to do, and their training in databases rarely exposes them to the challenges of working with other people's preexisting databases.

Dr. Stonebraker observed that there are two main ways in which scientific data today differ noticeably from scientific data of, say, a decade ago:

- *Scale.* Data sets are rapidly becoming larger. For example, it used to be commonplace for satellite imagery to tile the Earth into 100-m squares. Now the technology supports 5-m squares. Satellite imagery data sets have thus become larger by a factor of 400. This increase in data set size is expected to continue for the foreseeable future.
- *Number and type.* The numbers and types of scientific data sets appear to be increasing exponentially. For example, sensor tagging technology is making it possible to tag everything of value and

have it report interesting data on a real-time basis. Temperature data are available not only from traditional sources but from cell phones, car navigation systems, portable GPS devices, and the like. These are just a few examples of expanding range of the disparate types of data of possible interest to the scientific community.

In Dr. Stonebraker's view, this dramatic increase in data availability calls for the following four capabilities:

- *Locate data sets more effectively.* Scientists must be able to discover data sets of interest much more easily than they can today.
- *Convert data sets easily to a usable format.* It should be much easier for scientists to reformat data sets than is currently the norm.
- *Integrate multiple data sets.* Since data on the same phenomenon often come from many sources, a scientist needs to readily discover the syntax and semantics of data sets and to convert them to be syntactically and semantically comparable.
- *Process larger data sets.* As noted above the scale of scientific data is increasing rapidly.

The benefits of integrating large volumes of data, multiple data sets from different sources, and multiple types of data are enormous, and this integration will enable science to advance more rapidly and in areas heretofore outside the realm of possibility.

3

Improving Current Capabilities for Data Integration in Science

Any new direction or method of scientific inquiry starts out with a few visionary scientists blazing the path. All are focused on getting results from their research, and invariably they invent new data formats and semantics. This behavior leads to rapid innovation by each individual group but greater difficulty in sharing data across groups, or even across projects in a single group. In the early days of any domain, this state of affairs is a good thing because it maximizes the rate of early innovation, and a similar situation holds as new directions and innovative methods are explored even in mature disciplines.

However, there are drawbacks to this state, and these were noted by workshop participants. Usually, data are available only haphazardly from these early projects—that is, they are not well documented or curated and are not always easily accessible. Individual groups have little incentive to publish data, which slows the progress of the broader field. A new researcher in the domain is presented with a daunting data-discovery problem. And when the data are finally found, they may not be in a usable format. It is common, in this stage, for data to be transmitted to a requester as a bundle of code and data, such that the code is required in order to read the data. But getting code to run in a new environment can be far from trivial because of differences in operating systems, compilers, search paths for libraries, and so on, so that a researcher attempting to reuse the data might spend a good deal of time just getting to the point of being able to read the incoming data. Because most of the areas of scientific research discussed at the workshop are still in this stage with regard to data integration, the researchers share these challenges.

Further, there are multiple ways in which reuse might be hindered. The structure selected by the original researcher to organize the data might be inconvenient for a subsequent user—for example, they might be stored as geographic images, one for each time step, whereas the new researcher needs a time series for each spatial location. Or an underlying choice that was not even explicitly considered by the original researcher—perhaps the projection that was used to map the data from Earth's surface onto two dimensions—might not be suitable for the reuse context. (Even for research areas that have matured, such challenges can arise whenever data are applied in unanticipated ways.) The parameters that characterize the projection, or even the units, might not be clear because of incomplete metadata. Lastly, the second researcher's software tools may not be able to handle the individual data elements. To massage the data into correct format and organization may pose a tedious data-manipulation problem. It can take weeks or more of effort to convert data into a form suitable for reuse. Many new researchers give up before they get to this stage. In short, it is often just too difficult to reuse data gathered by other researchers.

It is crucial to focus on this transformation problem. Several workshop participants noted that it is not difficult to write clear transforms if the relevant metadata are available. Most popular transforms have been written multiple times by multiple labs, which is, of course, inefficient. Workshop participants said it was rarely easy to locate existing transformation software of interest, and some suggested that an online service to share transforms could be established. Such a service would allow scientists to avoid having to reinvent tools, but it would require publishing and documenting transforms in a systematic way so that others could locate them.

In our Internet-savvy world, one should be able to locate data sets and transforms of interest using the Web. At present this is a hopeless task. Workshop participants identified four steps that would make this task possible:

- *Repositories.* Several participants noted the need for domain-specific (as well as general) repositories where scientific data sets can be archived. Because data decay over time and require periodic maintenance, such repositories must be staffed with professionals who can do such maintenance as well as assist scientists trying to use data sets in the repository. Good search tools are needed so the contents of a repository can be easily browsed and objects of interest located. Lastly, curation facilities are also needed so that the precise semantics of data sets can be documented. Obviously, the curation cannot be such an onerous human task that the repository will not

be used. Curation information must be easy to locate, browse, and understand. Dr. Stonebraker suggested that Genbank and the Sloan Digital Sky Survey are examples of data repositories with effective search tools and good curation, but he said that many more such facilities are needed.

- *Web-based search.* It is nearly impossible to locate structured data using current text-oriented search engines. Moreover, there seems to be little incentive on the part of the search engine companies to provide this capability. Thus, targeted research will be necessary to enable locating structured data on the Web. Ideas for doing this include a science-oriented tagging system—that is, a system that makes assumptions about the content of a file based on some knowledge of the field of science—and storing science data in hypertext markup language (HTML), which would make them visible to search engines. The latter idea is only feasible for small data and is not a general approach.
- *Community-driven information extraction.* Given how much information is now available on the Web, the ability to interpret and integrate relevant Web content can have huge benefits. However, search alone can be a tedious means of collecting data from disparate sources. Webscale information extraction, assisted by an automated tool, represents a bottom-up complement to top-down approaches like the Semantic Web.¹ Another approach is to provide a suite of extraction tools to enable communities of interest on the Web to collaborate in creating and curating integrated datasets in domains they care about. This seems particularly promising for scientific domains, given that scientists are technically sophisticated and willing to collaborate.
- *Locating transforms.* As noted above, several workshop participants suspected that the data transforms they need at any given time have probably already been written at least once, but cannot be found, leaving individual researchers and groups to write their own. The same is true for all sorts of data manipulations, with similar kinds of code modules appearing over and over among

¹ The Semantic Web is an ambitious dream of deploying interlinked information via the resource description framework (RDF) throughout the Web. It encompasses a wide variety of philosophies, goals, and technologies. In general, it would rely on the establishment of ontologies and tools to help those who publish data to mark their content in terms that can be recognized semantically. Many of the Semantic Web technologies are proving to be useful, especially RDF, SPARQL, and OWL. Because the Semantic Web per se does not provide any particular set of standard entity names (URIs) or any particular approach to semantics, leaving these to particular application layers, any practical system for data integration must add these.

different research groups. Effort is wasted in writing such transforms many times and in maintaining such code as circumstances change. Obviously, it would be best to have a system that allows for reuse of common transforms; such a system might also support the development of more robust transforms. Perhaps one or more repositories (something like SourceForge) could be established to store such code. Another option would be for science funding agencies to form their own code repositories.

Workshop participants discussed some of the tools that have been produced by the database community that could help with data integration in the sciences, for both structured and semistructured data. The four subsections that follow provide a sampling of the approaches covered. The workshop was not designed to prioritize the potential value of database tools to scientific research data, and so this sample should not be construed as being more than just illustrative. Other critical techniques for data integration in some contexts—such as parallel processing and data indexing, which are very important when working with very large sets of data—are not covered here.

FEDERATORS

Dr. Haas's presentation provided an overview of how federators can be used to integrate data. She covered technical federation techniques, not the use of federation as a management or governance concept. Federation engines present users with a virtual repository of information. The users can manipulate information as if it were stored together in a single place with a single interface whereas it may actually be stored in multiple, possibly heterogeneous places. Federation engines come in different flavors, each presenting a different interface to users. The most common interface is that of a relational database management system (DBMS), effected through methods such as the Open Database Connectivity (ODBC) method, the Structured Query Language (SQL), and the relational data model. However, some federation engines present an extensible markup language (XML) interface (supporting some variant of XPath or XQuery, typically), and others might act like an object-oriented database or even a content repository. Besides having different interfaces, the capability of federation engines also varies, from "gateway" systems that allow simple queries against one source at a time while providing a common interface to all sources, to systems that allow users to leverage the full power of their query language to gather or correlate information from multiple diverse sources with a high level of query function.

To illustrate the potential of federation, Dr. Haas described the example of a pharmaceutical company with four main research sites in different countries. Each site has many data sources, including these:

- A special-purpose store for chemical compound information, searchable by chemical structure;
- A relational database holding results from various assays; and
- A literature source linking drugs to diseases and symptoms.

Data sizes range from hundreds of thousands of compounds to billions of test results. The four sites focus on different diseases and, for the most part, different compounds, locally storing the information they produce and use. However, as a scientist forms hypotheses about a compound, he or she might need to ask a coworker to find a compound with a structure similar to the one he or she is working with that has been associated with asthma and that has assay scores on test X within range [A,B]. Such a query might need data from all of the sites.

In this example, federation allows the scientist to pose the query without worrying about the geographic distribution of the data or about the different interfaces for the chemical stores, relational databases, and literature sources. The federation engine bridges this heterogeneity and drives the execution of the query across the different sources, reporting the results to the waiting scientists.

The architecture of IBM's InfoSphere Federation Server (IFS) illustrates how federation works. IFS has two main components: a query engine that supports either SQL or SQL/XML and a set of wrappers that connect the engine to a wide variety of data sources. A wrapper is a code module that handles four main functions:

- It handles the connection to the data source and transaction management.
- In response to requests from the query engine it drives the data source to produce the required result and retrieve the data.
- The wrapper also provides a mapping from the data model and functions in the underlying source into the relational model. If the underlying database is relational, this is straightforward. But in the case of the chemical store described earlier, the chemical similarity search and the chemical structure must be mapped to relational constructs.
- The wrapper participates in query planning, providing estimates of the costs of various operations to allow the query processor to identify a feasible and efficient plan for the query.

The query processor is an extended relational query processor. When a query arrives, it is parsed, the table and column names are resolved, the query is rewritten into a canonical form and optimized, and a run-time plan is produced and executed. Each phase of query processing after name resolution is modified to deal with wrappers and distributed data. In addition, a new phase analyzes the rewritten query and looks for opportunities to push work down to the remote data sources. Complex queries can be handled, and all improvements to the basic query processor—for example, new execution strategies or better optimizations—are immediately available for dealing with distributed, heterogeneous data.

Federation has been used for many purposes. It is often used to extend an existing database with heterogeneous, hard-to-convert data that are separately owned or that will rarely be used. This usage saves maintenance or creation costs for the warehouse. Federation is also frequently used to build a view across multiple organizational units, as in the four research labs in the example. This is an appealing use case, but the query workload must be watched carefully, as it is easy for complex queries to be generated that are challenging to optimize and may lead to unacceptable performance in some circumstances. Portals are more easily built on top of a federation engine, rather than hand-coding access to different data sources. Another common use of federation is as a prototyping environment for data-intensive applications. Even if a large materialized store must eventually be built, federation is easy to set up, and it allows testing of queries and early examination of the data.

Federation is a powerful tool for data integration, but it is not a panacea. Federation integrates data lazily, as it is needed. It is appropriate when data sets are not too large or when the queries are selective enough that only a small fraction of the data will ever be returned. It works well when the data do not need too much preprocessing or cleansing or when the data change frequently and up-to-date results are desired. The extract, transform, and load (ETL) paradigm, which is commonly used in business, is an alternative approach for the integration of primarily structured data. Its first step is to extract data from various sources, which includes conversion into some common format. The collected data are then transformed through a series of rules to prepare them for use. Transformations might include filtering, sorting, cleaning, and translating individual records for consistency, and other such operations. Finally, ETL loads the resulting data into the system where it will be warehoused and used. Generally speaking, ETL has strengths and weaknesses that are complementary to those of federation.

Dr. Haas suggested the following as potential steps for improving federation technology:

- Federation engines need continued work to minimize data movement, exploit multiprocessors, and leverage caching and even indexing to reduce response times for complex queries and large volumes of data.
- Other work is needed to extend the engines' capabilities. Today, entity resolution (figuring out when two data elements refer to the same real-world object) and data cleansing (discovering and correcting errors in the data) are typically batch operations. When those steps are necessary, federation cannot be used. Dynamic algorithms for these tasks would enable federation.
- Most federation engines today work only on traditional structured or semistructured data, though they can also return some uninterpreted fields, such as images. As the ability to extract information from unstructured data is improved, federation engines will need to grow to handle these new types.
- Finally, understanding where data come from is critical to many scientific endeavors. Hence, mechanisms for tracking provenance must be extended to function in a federated environment.

RESOURCE DESCRIPTION FRAMEWORK

Orri Erling gave an overview of the resource description framework (RDF) and linked data principles for science data and metadata. Using RDF as the data model for these metadata has numerous advantages. Sometimes, especially in the life sciences, data themselves are also represented in the RDF model. For other domains, such as those involving large arrays of instrument data, RDF is not a convenient format for the bulk of the data but is still appropriate for annotation. From the viewpoint of processes, data and metadata should go hand in hand, but different sizes and modeling characteristics often necessitate different representations for data and metadata.

RDF has several advantages for science metadata. To begin with, data are self-describing, and all entities and terms used have universal resource identifiers (URIs). The term "linked data" is used to mean a set of RDF triples where the URIs representing the entities, classes, and properties thereof are dereferenceable via HTTP. In addition, there is a constantly growing body of reusable ontologies, which provide the conceptual bases for RDF. Reusing terminology and modeling metadata has obvious advantages over reinventing the metadata schema for each application. Also, RDF is inherently schemaless—that is, not all entities of a class need have the same properties, and properties can be attached to data instance by instance without any database-wide schema alteration. This makes RDF less cumbersome than, say, relational database manage-

ment systems (RDBMSs) for highly variable or sparse data. Further, there is a constantly developing set of tools for harvesting, exchanging, storing, and querying RDF. Finally, the RDF model has well-defined semantics for inferencing and many features for facilitating mapping between ontologies and instance data sets. Classes, properties, and instances can be declared to be the same for purposes of a query. Scalar values may be typed value by value—for example, denoting a unit of measure. Thus, both issues of different identifiers for the same entities and different units of measure can be made explicit value by value in RDF.

Scalability of RDF storage is no longer a major problem, with billions of RDF statements being stored per server and with scale-out clustering available from at least OpenLink, Systap, and Garlic for larger scales. Also, data compression for RDF continues to advance, leading to further improvement of scalability. With the next generation of RDF storage, the performance penalty that RDF suffers when compared to RDBMSs for the same workload is likely to be substantially reduced through use of techniques such as adaptive indexing and caching of intermediate results. Task-specific relational schemas will probably continue to have some performance advantage for applications where the schema and workload are stable and known in advance, according to Dr. Erling.

Relational databases can also be mapped into RDF without storing the data in RDF. This is possible with tools such as Virtuoso or D2RQ. Thus, if science metadata are already in relational form, the RDF conversion for data interchange and integration can be done declaratively and on demand. A World Wide Web Consortium (W3C) working group aimed at developing standards for such mapping was launched in October 2009.

As an example, Dr. Erling described a harvesting model used for media metadata, which could be easily adapted to science metadata. The site bbc.openlinksw.com publishes metadata about programs of the BBC. The bbc.openlinksw.com server periodically crawls this content and presents it for search and structured querying via SPARQL, the SQL equivalent for RDF. Additionally, this server, if used as a proxy for accessing other RDF content, caches this content and allows querying over the BBC data and other cached data. For example, one can combine data from the BBC, LastFM, Musicbrainz, and other sources, all of which contain information about a musical artist. For the content producer, publishing the metadata is as simple as exposing RDF files for HTTP access. These files can be generated within the pipeline for content production.

This harvesting example is low-cost, incremental integration that does not require a priori agreement on schema and can accommodate any future data without schema alteration by a database administrator. Query-time inference can be used for identifying different names for the same entity and presenting the union of properties associated with each

identifier. More complex matching and inference can be done as an ELT transformation step without altering the source data.

The broader utilization of RDF might have positive impacts on the metadata publishing practices of scientific communities over time. Since every element has a URI, many of which can be dereferenced over HTTP, both schema and instance data identifiers point to their source, which provides a means of implicit attribution. Since data and their schema are thus objects of attribution and citation, there is an incentive for publishing data and schemas of high quality.

If many RDF data sets are kept in a common repository, it is easy to see which identifiers, ontologies, or taxonomies are in the broadest use. This ease of discovery will drive convergence of terminology. While very complex, centrally administrated ontologies exist, the ones enjoying the fastest adoption are lightweight ones developed through a bottom-up community process.

MapReduce AND ITS CLONES

MapReduce² and the accompanying Google File System³ were developed at Google to solve the problem of massive explosion in data by leveraging cheap hardware for both storage and processing. They are designed to scale to thousands of commodity servers, which means that failure is assumed to be not an exception but more of a rule. Hence, many design decisions within these systems are biased toward fault-tolerance, scalability, and agility as opposed to performance. Apache Hadoop⁴ is the open-source implementation of MapReduce, and it has the sister technology Hadoop Distributed File System (HDFS). At the workshop, Amr Awadallah of Cloudera Computing described a popular example that illustrates the scalability of Hadoop for economically storing large amounts of scientific data: the Large Hadron Collider Tier 2 site at the University of Nebraska-Lincoln, which currently stores 400 TB of data.⁵ As scientific data sets continue to grow at exponential rates, the need is paramount for scalable, fault-tolerant systems that can both store and process data economically. MapReduce and its clones represent an option for addressing that need for some types of scientific data.

The MapReduce model is a programming paradigm for processing large data sets; it makes it easy to scale execution linearly over a large

² See <http://labs.google.com/papers/mapreduce.html>.

³ See <http://labs.google.com/papers/gfs.html>.

⁴ See <http://hadoop.apache.org/core>.

⁵ Details of this example may be found at <http://www.cloudera.com/blog/2009/05/01/high-energy-hadoop>.

number of servers. In its simplest form, the developer specifies a map function that does the first stage of processing. The output data from the mappers are consistently hash-sorted then pulled by the reducers in what is known as the shuffle stage. Finally, the reducers perform the postprocessing of the results from the mappers. The origins of the MapReduce programming model come from functional languages such as LISP.

The MapReduce programming model is available in many shapes and forms. In fact, many of the traditional RDBMS vendors (for example, Teradata, Oracle, Greenplum) support MapReduce indirectly through user-defined functions (for mappers) and user-defined aggregates (for reducers).

The power of the overall MapReduce system (the distributed scheduling system that executes MapReduce jobs) comes from its ability to (1) automatically distribute/schedule the jobs and (2) transparently handle failures without requiring the jobs to be reexecuted from scratch (which would be very frustrating for multihour jobs processing large amounts of data). The system also allows the number of servers to be dynamically scaled up or down while jobs are running, so a number of additional servers can be thrown into the processing pool and jobs will begin using them transparently. The system is also designed to run a large number of data-processing jobs with various operating requirements. Some of these jobs can be operational jobs with high priority, so the system will automatically kill (preempt) the mapper or reducer tasks of lower-priority jobs to make room for the operational jobs. The jobs that have been preempted are resumed once the system has available resources for them. Furthermore, the system has optimizations to detect partial failure. For example, if one of the mappers executing a part of the job is running slowly compared with the rest of the mappers (maybe that node has unreliable disks), the system automatically starts a redundant mapper on a separate server, and whichever one finishes first wins. The MapReduce system is storage-system independent: It can read data from a normal file system, a distributed file system, an in-memory, key-value store, or even a traditional RDBMS.

Dr. Awadallah presented a list of MapReduce scientific examples and presentations that was assembled by members of the NSF Cluster Exploratory program. The list includes the following:

- Florida International University's Indexing Geospatial Data with MapReduce,
- University of Washington's Scaling the Sky with MapReduce and Interactive Visualization of Large Data,
- University of Maryland's Commodity Computing in Genomics Research,

- Carnegie Mellon University's Cluster Computing for Statistical Machine Translation,
- University of California, Irvine, Large-Scale Automated Data Cleaning, and
- University of California, Santa Barbara, Scalable Graph Processing.

Dr. Awadallah believed that MapReduce is most suitable for batch data-processing jobs. This would include ETL jobs that process original raw data into their relational form (because MapReduce does not require a predefined schema to be able to process data) and complex data transformations that are difficult to express in SQL (e.g., optical correction algorithms for astronomical images). MapReduce also has the ability to process data from multiple heterogeneous systems, such as those that exist in federations, through simple reader and writer functions. For example, one can have a MapReduce job that fetches input data from the distributed file system then joins them with data from a RDBMS. This allows the MapReduce system to run on top of data sources that range from unstructured (for example, collections of text, video streams, or satellite images), to semistructured (for example, XML, JSON, or RDF-like data), to relationally structured data (for example, tables with predefined column schemas).

DATA MANAGEMENT FOR SCIENTIFIC DATA

Dr. Maier's workshop presentation covered data management concepts that are of use for scientific data. Most commercial data-integration solutions are based on the relational model, with a few using XML as a target model. Such offerings are not likely to be of great help for integrating scientific data sets because there is not much support for some data types common to science, such as sequences, time series, and multidimensional arrays. Commercial relational DBMSs offer support for some scientific data types, most often time series and spatial objects, such as are used in geographic information systems (GISs). However, such support is supplied either by an encoding into the underlying relational model or through an extension of an abstract data type (ADT). In either case, the data types are not part of the core model of the system, and there is limited understanding of the types in query and storage-management layers.

Many scientific data types exhibit some form of order or, more generally, topology (a notion of adjacent elements and neighborhoods). This structure arises from the organization of the underlying physical world, such as chains of nucleotides or amino acids (ordered sequences) or discretized versions of continuous spaces arising from sensing or simulation

(multidimensional arrays, finite-element meshes). The desired operations on these data types are often order- or neighborhood-sensitive: examples include pattern matching, image filtering, and regridding.

Dr. Maier said that it has long been recognized that relational models and languages lack support for ordered types. While it is possible to encode ordered structures into the relational model, the associated operations can be hard to express, and optimization opportunities are obscured. Over the years, query languages for array and mesh data types have been suggested, such as AQL (Libkin, Machlin, and Wong, 1996), Array Manipulation Language (Marathe and Salem, 2002), and GridFields (Howe and Maier, 2005). However, no full-featured DBMS based on these languages is currently available.

Because of the limitations of relational DBMSs for supporting arrays, Maier reported that many scientific data end up in files using array data formats, such as NetCDF⁶ and HDF.⁷ While such formats support multidimensional arrays directly and appropriate access methods, they offer a file-per-dataset model and limited operations and hence are far from a full DBMS. They support interfaces to languages popular in scientific domains (C++, Fortran, Python) and to multiple data-analysis environments (R, Matlab, Octave). There are libraries of utilities for common operations available on some platforms, but there is no automatic optimization over groups of operators.

There are also approaches that layer support for scientific data types over existing storage managers, usually a DBMS. Maier stated that the following are the two main approaches:

- *Array Model and Query Language.* This approach provides an array data model and query language and performs some optimization and evaluation natively in that model, with the underlying storage system managing persistent storage, and possibly some degree of support for memory management, access methods, and query execution. Raster Data Manager (RasDaMan) (Baumann et al., 1998) is the most mature example of this approach. RasDaMan is an open-source system supporting an array data model and query language, with commercial support and extensions available. It provides its own query optimization, query evaluation, and main-memory management, using the underlying system (usually a relational DBMS) as a “tile store” for fragments of arrays. A more recent example is the RAM research project (van Ballegooij et al., 2003), which provides an array model and query facility that has

⁶ See <http://www.unidata.ucar.edu/software/netcdf/>.

⁷ See <http://www.hdfgroup.org/>.

been layered over various back ends, notably MonetDB. RAM performs query normalization, simplification, and optimization within its array model before translating into queries on the underlying relational engine. That layer can perform further optimization in the relational model before executing the queries.

- *Secondary-Storage Extensions to Data-Analysis Environments.* The second approach to layering uses a DBMS to provide relatively seamless access to secondary storage from a data analysis environment. The type system of the environment thus effectively becomes the data model, usually providing vectors, matrices, and higher-dimensional arrays. There is no special query language in this approach—disk-resident data are manipulated with the same functions used for in-memory data. It is up to the underlying interface to the DBMS to determine when functions can be performed in the database and when data need to be retrieved for main-memory manipulation. Ohkawa (1993) used this approach with the New S statistical package and an object-oriented DBMS. The RIOT prototype (Zhang et al., 2009) supports the R data-analysis environment using a relational DBMS. To create optimization opportunities in the underlying DBMS, both systems use lazy evaluation techniques. An operation on a secondary-storage object merely creates an expression that represents the application of the operation. Repeated deferral allows accumulating operations into one or more expression trees. Such trees are evaluated only when their result is to be output to the user, at which point they may be optimized before processing.

According to Dr. Maier, the SciDB project (Cudré-Mauroux et al., 2009) has recently begun development of an open-source database with fully native support for an array model, including an array-aware storage manager. In addition to a data model and algebra for multi-dimensional arrays, SciDB will support history and versioning of arrays, provenance, uncertainty annotations, and parallel execution of queries. If successful, it should provide a suitable platform for integrating extremely large scientific datasets.

4

Success in Data Integration

As a domain becomes more mature, more scientists begin to develop interest in it and progress starts to depend on the sharing of data. In the beginning such sharing is quite difficult, so a domain must develop ways to facilitate sharing as it matures. This includes the setting of standards, which may slow progress in individual groups to achieve a greater good for all. While the discussions recounted above evinced skepticism about any global schema, there are places where standards have been quite successful, some of which are described below. The most successful standards tend to occur bottom-up. In other words, individual scientists recognize the need and work to build consensus standards. Other standards are imposed top-down by some sort of dominant force in an enterprise. Top-down appears to work only rarely, and bottom-up approaches have a much better chance of success, according to several workshop participants. However, standards are also facilitated if there is a dominant player in a domain, as pointed out by Dr. Stonebraker. In enterprise data, for example, Walmart has so much influence that it can specify standards and force all of its suppliers to conform if they wish to sell goods to Walmart. Google also has this sort of influence in the Web search space. In domains where there is a dominant player, standards are much easier to achieve.

The successes of the Sloan Digital Sky Survey and Genbank in sharing astronomy data and genomic data are well known in the scientific community. The National Spatial Data Infrastructure (NSDI), mentioned by Dr. Clarke, has been emulated worldwide as the global spatial data infra-

structure and is another example of success. The NSDI was prompted by an Executive Order issued by President Clinton in 1994, which also called for “development of a National Geospatial Data Clearinghouse, spatial data standards, a National Digital Geospatial Data Framework and partnerships for data acquisition.”¹ The NSDI enables sharing of geographical information, elimination of redundancies, and other significant benefits. Some other success stories, perhaps less well known, are presented here.

FREEBASE

Freebase is a large, collaboratively edited database of crosslinked data developed by Metaweb Technologies. Freebase has incorporated the contents of several large, openly accessible data sources, such as Wikipedia and Musicbrainz, allowing users to add data and build structure by adding metadata tags that categorize or connect items.² To date, most of the information in Freebase relates to people and places, though it can accommodate a wide range of data types, including research data.

Freebase is intended to be an important component of the Semantic Web, allowing automation of many Web search functions and communication between electronic devices (*New York Times*, 2007). However, Freebase has quality issues, omissions, errors, and redundant information—most of its information is not truly integrated. While Freebase is a success in some respects (community contributions have led to large volumes of information and it is possible to get useful answers to some queries), it cannot guarantee accurate and complete answers. Overall, Freebase demonstrates a novel mechanism for data aggregation, but it has not yet solved many of the challenges of information integration.

MELBOURNE HEALTH

Melbourne Health, a healthcare provider in Melbourne, Australia, envisions building a generic informatics model for beneficial collaboration across organizations and expansion to other research areas (Bihammar and Chong, 2007). Melbourne Health’s original goal was to link the databases from seven hospitals and two research institutes for multiple disease research. The challenges in this work come from the large amount of data, the paucity of data standards, poor interoperability between databases, and the need to ensure compliance with ethical, privacy, and regulatory norms.

¹ Quoted from http://www.fgdc.gov/nsdi/policyandplanning/executive_order. Accessed May 5, 2010.

² Available at <http://freebase.com>. Accessed October 23, 2009.

Medical documents and research data come from files, Excel spreadsheets, and databases. The hospitals and clinics may use different systems. The HL7 Clinical Document Architecture (CDA), an XML-based markup standard intended to specify the encoding, structure, and semantics of clinical documents for exchange, is used. According to the IDC case study (Bihammar and Chong, 2007), Melbourne Health has linked research databases in 16 organizations, allowing them to collaborate.

SCIENCE COMMONS AND NEUROCOMMONS

Science Commons (<http://sciencecommons.org>), launched in 2005, is an offshoot of Creative Commons, a not-for-profit organization that develops and disseminates free legal and technical tools to facilitate the finding and sharing of creative content (Garlick, 2005). It also focuses on lowering barriers that researchers face to sharing data, publications, and materials.

The goals are to expand sharing, interoperability, and reuse of data, but these goals are hampered by legal and cultural barriers. Although research data are not subject to copyright protection, the arrangement of data and the structure of databases may be protected (for a discussion of the legal context for sharing and accessing research data, see NRC, 2009). Specific rights to reuse or integrate data may be unclear, and integrating data collected under different jurisdictions may be problematic. Researchers in some fields might take proprietary approaches to data or might lack the motivation to make their data available proactively.

Science Commons has developed several programs and tools to lower these barriers. The Protocol for Implementing Open Access Data allows researchers to mark their data for machine-readable discovery in the public domain so that their databases can be legally integrated with others, including those collected in other jurisdictions.³

The NeuroCommons project, under the auspices of Science Commons, is developing an open-source knowledge management platform for biological research. The goal is to make all knowledge sources—including articles, knowledge bases, research data, and physical materials—interoperable and uniformly accessible by computational agents. NeuroCommons is a prototype framework for creating information artifacts that can provide lessons for future communities, particularly in reaching community consensus around technical standards and curation processes. The NeuroCommons framework utilizes URIs and RDF, making it part of the Semantic Web.⁴

³ Information drawn from <http://www.sciencecommons.org>. Accessed October 23, 2009.

⁴ Information drawn from <http://neurocommons.org>. Accessed October 23, 2009.

To apply this idea to scientific information artifacts, one creates a set of conventions for syntactic and semantic compatibility among components and a standard packaging mechanism to make selecting and installing components easy. One starts with the primary sources (databases, knowledge bases, and the like), applies a script to do the normalization, and comes up with a packaged component. The resulting “binary” may or may not be collected with others to make a distribution. Someone creating a local installation optimized for local query obtains needed components from one or more distributions and installs those into their own environment.

Some two dozen components have been created and collected in the NeuroCommons framework. The components are independent and the architecture is open, so that anyone may pick and choose the ones they like without having to take all of them. One may create new components and either add them to the distribution (subject to quality control), create a new distribution, or just use them privately. Currently the NeuroCommons distribution is accomplished either through a set of RDF files or a database dump.

Bio2RDF

Bio2RDF (<http://bio2rdf.org>) is an open-source project that aims to facilitate biomedical knowledge discovery using Semantic Web technologies. Bio2RDF is an important contributor to the Linked Data Web, offering the integration of over 30 major biological databases with content ranging from biological sequences (such as are stored in UniProt, Genbank, RefSeq, Entrez Gene), structures (from the Protein Data Bank), pathways and interactions (cPATHs), and diseases (OMIM), to community-developed biomedical ontologies (OBO).

This project builds on W3C standards for sharing information over existing Web architecture and representing biomedical knowledge using standardized logic-based languages. Powered by open-source tools, Bio2RDF enables scientists to not only explore manually curated and computed aggregated knowledge about biological entities but to also link their data and enable all scientists to ask fairly sophisticated questions across distributed, but integrated, biomedical resources. Bio2RDF-linked data are available today as N3 files, indexed Virtuoso databases, and SPARQL endpoints across three mirrors located in Canada and Australia.

With interest growing in the Bio2RDF data and services beyond the initial developers, the group is fielding requests to add more than 50 additional data sources in the areas of yeast and human biology, toxicogenomics, and drug discovery.

5

Workshop Lessons

At the end of the workshop, Michael Stonebraker presented the following list of messages that he thought were brought out by the discussions:

- Many research groups leave the task of developing data integration software to science postdoctoral students, which is wasteful of the students' time and can lead to inadequate results. Good DBMSs are difficult to write and take many person-years of effort. A better idea is to apply computer science expertise early in the process. A partnership of equals between computer scientists and natural scientists can pay off admirably. The successful collaboration between Alex Szalay and Jim Gray is a prime example.
- It is impossible to build a complete software stack quickly. The best way to progress is to specify modest short-term goals and get them accomplished. Once something is working, one can build the next phase. In other words, one should take "baby steps," always going from something that works to something that continues to work. What often kills projects is the desire to take a giant leap in functionality, without having intervening milestones.
- Funding agencies can help scientists establish the capability for data integration by steps encouraging (or, indeed, requiring) the researchers they support to publish and curate their data. Agencies should strengthen the incentives for scientists to preserve their

data in reusable form, such as by giving special consideration to proposals that include plans for careful data publication.

- Moreover, funding agencies can encourage the establishment and maintenance of data repositories and work to improve the tools available for data curation and sharing.
- An open-source tool-kit to assist with data transformations would be of immense value. This is something that agencies can budget for, solicit proposals for, and fund.
- An open-source science-oriented DBMS would also be of immense value. Again, this is something that agencies can budget for, solicit proposals for, and fund.

Dr. Stonebraker offered his own thoughts on how to improve the software that enables data integration. Noting that scientists often build the entire software stack for each new project, he pointed out how this limits, even precludes, the reuse of software modules and the leveraging of well-established tools. Building afresh was followed by the Mission to Planet Earth a decade ago as well as more recently by the Large Hadron Collider project. In contrast, the Sloan Digital Sky Survey (SDSS) made data available in an SQL server database and allowed astronomers to run a collection of queries of interest.

Dr. Stonebraker suggested a number of ways to improve the common state of practice:

- Send the query to the data, not the other way around. Currently, publication schemes typically send data sets to scientists who load these data into their favorite software system and then further reduce them to find actual data of interest. In effect, a central system sends data to scientists who query them locally to discover items of interest. This approach is an inefficient use of bandwidth, because large data sets are sent over networks only to then be reduced two or three orders of magnitude. It would be much more efficient to reduce the data upstream in response to a request and save the bandwidth.¹ An alternate approach for saving bandwidth, which is sometimes practiced today, is to store the data in a processed form, so that their transmittal is easier. But this has the shortcoming that requesting scientists have different needs, so any

¹ An anonymous reviewer pointed out that, in general, this approach may not scale, as some centralized stores will have to support an ever increasing number of queries. A complementary approach is to have replication on demand, where subsets of the data are replicated to secondary sites based on local demand. A form of this approach was taken by the LHC with its predetermined tier structure.

given processing will not be optimal for everyone. To facilitate the flexibility that scientists need, one may have to make available the raw data and not just a highly processed derived data set.

- Put the raw data in a DBMS and then run the processing inside the DBMS engine. The only feasible way to allow a scientist to insert his or her own components into the processing pipeline is to make the processing a collection of DBMS tasks. Otherwise, the complexity of altering the pipeline is just too daunting.
- Record the provenance (lineage) of the data carefully, with an automated system. This is necessary for the raw data, of course, but it is also crucial to precisely record the semantics of any derived data, thus carefully maintaining the provenance of those data sets. This is not something that current application code or system software is good at. Also, anything that requires human effort is not going to be widely used, and so systems are needed that record provenance as a side effect of natural science inquiry and processing, not an additional step. One of the big advantages of a DBMS is that it can record provenance automatically by recording every query and update that has been run.
- A better DBMS is obviously needed for science applications, one of the challenges called for in Chapter 2. Scientists who spoke at the workshop did not like current relational DBMSs, which were built for business data processing, because they do not work well, if at all, on science data. The six messages presented at the beginning of this chapter are unlikely to be successful with current commercial DBMSs. Self-documenting data sets, via RDF with reference to code systems, will be needed, along with separation of the data from the application/analysis software.
- At present, most fields of science do not have systematic means for a scientist to make data available. They do not have public repositories in which to insert data, standards for provenance to describe the exact meaning of data sets, or easy ways to search the Internet looking for data sets of interest. In addition to data repositories, repositories of standards and translators are also needed.

While there was some discussion of these ideas at the workshop, no attempt was made to capture the range of opinions, and the thoughts presented in this chapter do not necessarily represent a consensus of the workshop participants.

References

- Baumann, P., A. Dehmel, P. Furtado, R. Ritsch, and N. Widmann. "The multidimensional database system RasDaMan." *SIGMOD* (1998), Seattle, Washington.
- Bihammar, Patrik, and Chris Chong. "IDC case study: Melbourne health realizes research efficiency with information integration." Doc. AP753033P (2007). Available at <http://www-07.ibm.com/innovation/au/ideas/healthcare/pdfs/IDC+Melbourne+Health+2-07.pdf>. Accessed February 28, 2010.
- Cudré-Mauroux, Philippe, Hideaki Kimura, Kian-Tat Lim, et al. "A demonstration of SciDB: A science-oriented DBMS." *Proceedings of the VLDB Endowment* (PVLDB) 2(2): 1534-1537 (2009).
- Garlick, Mia. "A review of Creative Commons and Science Commons." *EDUCAUSE Review* 40(5): 78-79 (2005). Available at <http://www.educause.edu/EDUCAUSE+Review/EDUCAUSE+ReviewMagazineVolume40/AReviewofCreativeCommonsandSci/158002>.
- Howe, B., and D. Maier. "Algebraic manipulation of scientific datasets." *The VLDB Journal* 14(4): 397-416 (2005).
- Intergovernmental Panel on Climate Change. *Climate Change 2007—The Physical Science Basis*. Cambridge University Press, Cambridge (2007).
- Libkin, Leonid, Rona Machlin, and Limsoon Wong. "A query language for multidimensional arrays: Design, implementation, and optimization techniques." *Proceedings of the ACM SIGMOD International Conference on Management of Data* (June 1996): 228-239.

- Marathe, Arunprasad P., and Kenneth Salem. "Query processing techniques for arrays." *The VLDB Journal* 11(1): 68-91 (2002).
- National Research Council. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data*. The National Academies Press, Washington, D.C. (2009).
- New York Times*. "Start-up aims for database to automate web searching" (2007). Available at <http://www.nytimes.com/2007/03/09/technology/09data.html>.
- Ohkawa, H. "Object-oriented database support for scientific data management: A system for experimentation." Ph.D. thesis, Oregon Graduate Institute (1993).
- van Ballegooij, A.R., A.P. de Vries, and M.L. Kersten. "RAM: Array processing over a relational DBMS." Technical Report INS-R0301, Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands (2003).
- Zhang, Y., H. Herodotou, and J. Yang. "RIOT: I/O-efficient numerical computing without SQL." *Proceedings of the 4th Biennial Conference on Innovative Data Systems Research (CIDR '09)*, Asilomar, California (2009).
- Ziegler, Patrick, and Klaus R. Dittrich. 2004. "Three decades of data integration—All problems solved?" University of Zurich. Available at <http://www.ifi.uzh.ch/stff/pziegler/papers/ZieglerWCC2004.pdf>.

Appendixes

A

Workshop Agenda

**NATIONAL ACADEMIES
WASHINGTON, D.C.**

Wednesday, August 19, 2009

8:00 am Continental breakfast available

8:30 Chair's opening remarks
Michael Stonebraker, MIT

*Data Integration Stretch Goals, Technical Needs, and Policy Issues—Views
From Various Domains*

9:00 Geospatial data
Keith Clarke, University of California, Santa Barbara

9:20 Life sciences
Carl Kesselman, University of Southern California

9:40 Physics
Tim Frazier, Lawrence Livermore National Laboratory

10:00 Astronomy
Alex Szalay, Johns Hopkins University

44 *STEPS TOWARD LARGE-SCALE DATA INTEGRATION IN THE SCIENCES*

- 10:20 Earth sciences
 Tom Karl, National Oceanic and Atmospheric
 Administration
- 10:40 Research libraries
 Clifford Lynch, Coalition for Networked Information
- 11:00 Break
- 11:20 Open discussion

Working Lunch: Agency Perspectives

- Noon 10-15 minutes apiece from sponsors and other agencies
 James St. Pierre, National Institute of Standards and
 Technology
 Michael Marron, National Institutes of Health
 Ed Seidel, National Science Foundation

State of the Art in Data Integration—Structured Data

- 1:20 pm Data federations
 Laura Haas, IBM Almaden Research Center
- 1:40 Data type conversion and ETL technology
 Lee Scheffler, IBM Information Integration Solutions
- 2:00 Automatic conversion
 Michael Siegel, MIT
- 2:20 The SciDB approach
 David Maier, Portland State University
- 2:40 Linked open data
 Orri Erling, OpenLink
- 3:00 Microsoft approach to data conversion
 Phil Bernstein, Microsoft
- 3:20 Break

State of the Art Data Integration Solutions—Semistructured Data

- 3:40 Google Approach
Alon Halevy, Google
- 4:00 Yahoo! Approach
Raghu Ramakrishnan, Yahoo!
- 4:20 MapReduce/Hive/Pig paradigm
Amr Awadallah, Cloudera

Policy Perspectives

- 4:40 Policy perspective
Michael Nelson, Georgetown University
- 5:00 Policy perspective
Christopher Greer, National Coordination Office for
Networking and Information Technology R&D
- 5:20 Lessons from a large-scale information integration ecosystem
Michael Brodie, Verizon
- 5:40 Business perspective
Josephine Cheng, IBM Almaden Research Center
- 6:00 Develop organization for Day Two
- 6:30 Working dinner

Thursday, August 20, 2009

Open Brainstorming

- 8:30 am Reflections on Day One
- 9:00 Identify three areas for detailed and moderated discussion

Topic Area One

- 9:15 Open discussion
- 10:00 Break

Topic Area Two

10:30 Open discussion

Topic Area Three

11:15 Open discussion

Noon Adjourn public workshop

12:30 pm Planning committee lunch/executive session

3:00 Planning committee adjourns

B

Workshop Participants

Thomas Arrison, National Research Council
Amr Awadallah, Cloudera Computing
John Bates, National Oceanic and Atmospheric Administration (NOAA)
Philip Bernstein, Microsoft Corp.
Michael Brodie, Verizon
Josephine Cheng, IBM Almaden Research Center
Keith Clarke, University of California, Santa Barbara
David Dean, Department of Energy
Orri Erling, OpenLink Software, Inc.
Timothy Frazier, Lawrence Livermore National Laboratory
John Gardner, Food and Drug Administration
Christopher Greer, National Coordination Office for Networking and
Information Technology R&D
Laura Haas, IBM Almaden Research Center
Alon Halevy, Google, Inc.
Jeffrey Huskamp, University of Maryland
Thomas Karl, NOAA
Carl Kesselman, University of Southern California
Subhash Kuvelker, National Research Council
Clifford Lynch, Coalition for Networked Information
David Maier, Portland State University
Michael Marron, National Institutes of Health
Michael Nelson, Georgetown University
Raghu Ramakrishnan, Yahoo! Research

Yrjänä Rankka, OpenLink Software, Inc.
James St. Pierre, National Institute of Standards and Technology
Lee Scheffler, IBM
H. Edward Seidel, National Science Foundation
Michael Siegel, Massachusetts Institute of Technology
Michael Stonebraker, Massachusetts Institute of Technology
Alex Szalay, Johns Hopkins University
Scott Weidman, National Research Council