Workshop Report on a Future Information Infrastructure for the Physical Sciences

The Facts of the Matter: Finding, understanding, and using information about our physical world

Hosted by the Department of Energy at
the National Academy of Sciences

May 30-31, 2000

Preface

Forty years ago it took days, weeks or even months for information regarding an interesting discovery to be communicated to the relevant community of scientists and engineers. At that time, most of us kept a collection of postcards that we used to request reprints of articles as they appeared in the journals we read. This was the situation at the time that Ted Maiman reported his results using ruby as a medium to make a laser. Some twenty years later, this time interval was shortened to days by fax machines when Müller and Bednorz revealed their experiments demonstrating high temperature superconductivity. Today, with the Internet, that interval has shrunk to seconds, minutes, at most, hours. I suspect that some astronomical observations are announced within seconds of the time it takes to type a note and launch it on the Internet to a few tens or hundreds of one's closest colleagues. This situation makes it imperative to have a system in place that allows the rapid communication of information of value to scientists and engineers who are engaged in what has become an intensely competitive research environment.

This Workshop seeks to cast light on the problem of communication and dissemination of information within the physical sciences community, and to make practical suggestions regarding steps that can be taken to improve the situation.

Alvin W. Trivelpiece
Formerly, Director of ORNL

Acknowledgement

The Department of Energy sponsors of the Workshop gratefully acknowledge the enthusiastic participation and contributions of the Workshop Chair and Panelists who developed this report.

Alvin Trivelpiece, Chair
R. Stephen Berry, University of Chicago
Martin Blume, American Physical Society
Jose-Marie Griffiths, University of Michigan
Lee Holcomb, National Aeronautics and Space Administration
Kirk McDonald, Princeton University
Krishna Rajan, Rensselaer Polytechnic Institute
Kent Smith, National Library of Medicine
Derek Winstanley, Illinois State Water Survey

Table of Contents

Executive Summary 1

The Challenge 5

The Purpose 7

Major Themes and Conclusions 7

Scope of the Initiative. 8
Information Types 8
Infrastructure, Products, and Services 9
Archiving, Preservation and Access to Information 12
Research, Education, and The Public Interest 13
Quality 13
Participation of Sectors of the Economy 14
Leadership 15
Funding 16
Timing 16
Reward Structure 16
International Diplomacy 17
A Name 17
The Biomedical Model: Validation of the Concept 17

Findings 17

Implementation 18

Doing Better At What We're Doing Now (FY 01) 18
Mobilizing For What Is Possible Tomorrow (FY02-03) 19
Realizing The Future Potential (Out Years) 19

Appendices 21

Appendix A, Workshop Panelists and Participants 21
Appendix B, Agenda 22
Appendix C, Working Group Themes 23
Appendix D, The Biomedical Model 25
Appendix E *Suggested R&D for the Future Information Infrastructure for the Physical Sciences*–Dr. Krishna Rajan 27

Executive Summary

On May 30–31, 2000, a Workshop was held at the National Academy of Sciences to address questions regarding an "Information Infrastructure for the Physical Sciences" to increase the productivity of the scientific enterprise in the United States.

Chaired by Dr. Alvin W. Trivelpiece, formerly, Director of Oak Ridge National Laboratory, the Workshop was composed of a panel of experts in science, science policy, information science, and scientific publishing. Other participants included representatives from the community of potential stakeholders in such an enterprise.

The concept of an "Information Infrastructure for the Physical Sciences" is built on a history of studies that have called for a national information resource in the sciences. The study by the President's Information Technology Advisory Committee (PITAC)[1] states a vision that "An individual can access, query, or print any ... magazine, data item, or reference document... by simply clicking a mouse..." The Loken Report [2] notes that this is a "time of true revolution in the communication of scientific information," and discussed the concept of a "worldwide Physics Information System". It references a "National Library of Science," and concludes with the recommended goal of a scientific information system from which "all the world's formal scientific literature is available, on-line, to scientific workers throughout the world, for a world scientific database." These reports served as the base to begin discussions. For well over 50 years, since the introduction of the computer, studies have called for a better focus on the information problem.

The rapid advances in information technology that are dramatically altering the nature of scientific communications were recognized. Traditional means of access to the scholarly record are no longer sufficient to meet researchers' needs and expectations or even to follow the rapid pace of scientific developments. New concepts to fill the knowledge gap are emerging. Today we are in a new and unique environment. The ability to compete is first based on our ability to know quickly. The value is not in having the knowledge, but in how it is used. Scientists have already been changing the way science is being done; institutions are ripe for change. The Internet and distributed information technologies now provide us a means to realize what were only visions in the past. These are basic assumptions that were supported by the results of this Workshop.

Scope of the Initiative

Begin with the Physical Sciences so that it is doable, but engineer it so that it is scalable.

Information Types

We need a conceptual change that allows us to better integrate information types (e.g., text, data, images, animations) as well as information at various stages in the analysis process (raw data, partially processed data, text summaries of analyzed data in varying degrees of analyses, and unreviewed or peer-reviewed documents). Such a conceptual change would facilitate the serendipity and insights that can be gained by dealing with these multiple information types and their interrelationships.

Information Products and Services

Information Infrastructure has to be more than a storage and retrieval capability. Rather, in the long term, it has to support fundamentally new ways of doing science.

[1] "Information Technology Research: Investing in Our Future," President's Information Technology Advisory Committee, February 1999, p. 13.

[2] Report of the American Physical Society (APS) Task Force on Electronic Information Systems, popularly know as "The Loken Report", published in the Bulletin of the American Physical Society, Vol. 36, No. 4, p. 1119, 1991; or available online at http://publish.aps.org/eprint/reports/lokenrep.html. *Note: APS has commissioned an update to this report.*

Archiving, Preservation and Access to Information

Archiving of data is one of the key concerns. The electronic record is much more fragile than the media that came before (print, microforms) because it comes faster, in bigger volume and is gone more quickly. There is a real need for assurance of portability, regardless of media developments.

Research, Education, and the Public Interest

The physical science information infrastructure of the future must begin by focusing on the producers of the information; i.e., the scientific community, but must recognize the potential for positively impacting education and the general public interest.

Quality

Quality in data and data processing is one of the cornerstones of success. The system must take into account the quality, the reliability, and usability aspects of the information.

Participation of Sectors of the Economy

We must continue to strive for broad and innovative collaborations. Stability comes from diversity. Though we may of necessity need to build the infrastructure with a focus on specific tasks and disciplines, we must encourage the broadest participation at the planning table to allow for future involvements.

Leadership

It is clear that we need a point of convergence and leadership to develop the common agenda and move it forward.

Funding

The government has a responsibility to disseminate the results of federally sponsored research as broadly as possible as a public good. The amount of resources needed should be benchmarked against model efforts.

Timing

The time is now, the need is now.

Infrastructure Elements

As a result of the Workshop and focused discussion, findings support the need for:

A common knowledge base that seeks in an integrated approach to provide comprehensive access and facilitate the reuse of worldwide sources of physical sciences information, regardless of where they reside, what platform(s) they reside on, or what format or data structure they employ.

A point of convergence for ensuring the awareness, availability, use, and development of information technologies and tools to facilitate information assimilation, data analyses, peer communication and collaboration, sharing of preliminary research results, remote experimentation, validation of experimental results, etc; and

An openly available source of information to serve all users, from students to scientists to concerned citizens, in a highly efficient electronic environment, with tools to assist users in their quest for information and ultimately knowledge.

Implementation

In considering implementation, the requirements for the infrastructure must deal with three time horizons:

- Doing Better at What We're Doing Now (FY01)
- Mobilizing for What is Possible Tomorrow (FY02–03)
- Realizing the Future Potential (Out Years)

Some immediate activities were cited that confirmed what coordination and leadership in the near term could accomplish. Other efforts are beginning to come on the landscape and can be moved forward now. Of utmost importance is continuation of the planning process to include other agencies and other stakeholders.

Conclusions

When comparing The National Library of Medicine's amazing success in the delivery of medical information as a benchmark, it is clear that much can be done in the physical sciences to positively impact research and practice in the physical sciences.

The overall conclusion of the workshop was an enthusiastic endorsement of a vision of a national infrastructure that benefits not just the scientific community but the national good. It could ultimately impact not only research and development (R&D), but also education and applications to our everyday lives. It would be a step to integrate the whole of science to provide a basis to improve society, the economy, and the environment.

This is the report of a Workshop on "A Future Information Infrastructure for the Physical Sciences" held at the National Academy of Sciences on May 30–31, 2000. The Department of Energy was the host for the Workshop, and it was chaired by Dr. Alvin Trivelpiece, formerly, Director of the Oak Ridge National Laboratory, with support from expert panelists representing major disciplines in the physical sciences as well as various professional roles in the scientific and technical information life cycle. The other participants in the Workshop were drawn from the community of potential stakeholders in such an enterprise. A list of the panelists and participants is provided in Appendix A. The Workshop agenda is provided in Appendix B.

The Challenge

The rapid advances in information technology have dramatically altered the nature of scientific communications. Traditional means of access to the scholarly record are no longer sufficient to meet researchers needs and expectations or even to follow the rapid pace of scientific developments. Though scientists and engineers have well-established individual patterns for information discovery, often these patterns are focused in areas of specialization and not attuned to interdisciplinary opportunities. More often than not, scientists and engineers are overwhelmed with a mass of information within their own specialty and are even more hard pressed to stay abreast of work done in other areas that could contribute to their efforts. This often results in missed opportunities or wasted resources. New concepts to fill the knowledge gap are emerging.

The current concept of a Future Information Infrastructure for the Physical Sciences, initially proposed by the Department of Energy, is built on a history of national studies that have called for a national information resource in the sciences. The study by the President's Information Technology Advisory Committee (PITAC)[3] states a vision that "An individual can access, query, or print any ... magazine, data item, or reference document... by simply clicking a mouse...". The Loken Report[4] notes that this is a "time of true revolution in the communication of scientific information," and discusses the concept of a "worldwide Physics Information System." It references a "National Library of Science," and concludes with the recommended goal of a scientific information system from which "all the world's formal scientific literature is available, on-line, to scientific workers throughout the world, for a world scientific database." These reports served as the bases to begin discussions.

For well over 50 years, since the introduction of the computer, studies have called for a better focus on the information problem. So, why now? The difference today is that we are in a new and unique environment. Our ability to compete is first based on our ability to know quickly. The value is not only in having the knowledge, but in using it. Scientists have already been changing the way science is being done; institutions are ripe for change. The Internet and distributed information technologies now provide us a means to realize what were only visions in the past. These are basic assumptions that were supported by the results of this Workshop.

[3] "Information Technology Research: Investing in Our Future, " President's Information Technology Advisory Committee, February 1999, p.13.

[4] Report of the American Physical Society (APS) Task Force on Electronic Information Systems, popularly know as "The Loken Report", published in the Bulletin of the American Physical Society, Vol. 36, No. 4, p. 1119, 1991; or available online at http://publish.aps.org/eprint/reports/lokenrep.html. *Note: APS has commissioned an update to this report.*

Our Physical World

Physical processes are all around us and the scientific research we do and the data that support it have dramatic impacts on the social and economic strength and security of our country.

Business and industry convert research results into the tools and products we often take for granted. It is essential that the linkage from research results and rate of transfer to and from the business and industry communities keep pace with the global communication processes that are evolving through the use of the Internet.

It is clear that the U.S. is losing ground in patents issued, in comparison to patents issued to foreign citizens. While the number of patents issued to U.S. citizens has risen 220 percent in the period of 1963–1998, patents issued to foreign citizens during the same period has risen 790 percent.[5]

Further, the percent change in Real Gross Domestic Product for the period of 1970–1996 shows a similar pattern. U.S. percent change during this period has shown a positive 210 percent (1996 GDP was $7.6T) while the East Asia/Pacific Rim (1996 GDP was $10.4T) has shown a positive 456 percent change in Real Gross Domestic Product.[6]

Both indicators would suggest that while we are continuing to show positive increases in both areas, our rate of improvement is not keeping pace with world competitors.

In addition, there are many incidents that can be cited where the absence of good data applied effectively has cost millions and caused significant harm. We can think of the Challenger accident where data were known about the temperature effects on the "O" ring, but they were not brought to bear in the decision making process.

Our industries are impacted by effective access and use of data. Five percent of the $135 billion in the U.S. chemical industry costs could be saved if data needs were brought to a level of completeness comparable to that of other design tools.[7]

Finally, the impact of the use of good data can be brought home to our everyday lives. For example, in 1982, $119 billion was spent in prevention or as a result of the fracture of materials in the U.S., and it is estimated that $35 billion could have been saved by use of best practices and technology. In 1975, $70 billion was spent on corrosion. Fifteen percent of this expenditure could have been avoided through available knowledge.[8] Fracture and corrosion impact the cars we drive, the buildings we live in, and the tools we use.

Our Human Resources

In 1996, there were over 3 million scientist and engineer jobs in the U.S. It is projected that this number will grow to 4.4 million by 2006, which is a 44 percent increase.[9] During the 1996–2006 time-period, the demand for scientists and engineers is expected to increase at more than three times the rate for all other occupations.[10] Yet, the growth in the number of scientists and engineers to meet this need is not apparent. Data from 1997 show 5,500 Masters degrees and 4,500 Doctoral degrees in the physical sciences awarded by U.S.

---

[5] Workshop presentation by R.L. Scott, Director for Project and Program Development, U.S. Department of Energy, Office of Scientific and Technical Information, May 30, 2000

[6] Ibid, Scott

[7] D.W.H. Roth, Jr., Allied Corporation, Morristown, NJ, "The Chemical Industry." (Presentation at workshop coordinated by U.S. House Committee on Science and Technology, Congressional Research Service, and the Numerical Data Advisory Board of the National Academy of Sciences, Towards a National S&T Data Policy, Washington, DC, April 1983.)

[8] a) R.P. Reed et al., "The Economic Effects of Fracture in the United States," (Part 1—A Synopsis of the September 30, 1982 Report to NBS by Battelle Columbus Laboratories, NBS, Spec. Publ. 647-1, U.S. Government Printing Office, Washington, DC, 1983).

b) L.H. Bennett et al., "Economic Effects of Metallic Corrosion in the United States, Part I," (A Report to the Congress by the National Bureau of Standards, NBS Spec. Publ. 511-1, U.S. Government Printing Office, Washington, DC, 1978).

[9] Ibid, Scott

[10] Ibid, Scott

academic institutions.[11] This number of post secondary degrees in the physical sciences has remained fairly constant over the last 30 years, despite increases in population and increasing reliance on new technology as a key component of our economy. Given this possible shortage in talent, more efficient tools and processes to collect, organize, and synthesize physical science information must be used to optimize the human resources we will have available.[12]

It is estimated that a scientist spends between 35 and 50 percent of his or her productive time using and producing scientific and technical information.[13] Given the dramatic improvements in information technology, it should be possible to reduce this time and to make it far more productive. Yet, most agencies allocate far less than a percent of the federal investment in R&D to ensure that information is produced, obtained and utilized productively.

The Purpose

The purpose of the Workshop was to "obtain input from the scientific community regarding the merits of the concept of a 'Future Information Infrastructure for the Physical Sciences' that would offer both a comprehensive collection of scientific and technical information in the physical sciences and services that would facilitate scientific communication and increase the productivity of the scientific enterprise in the United States. The Infrastructure would impact science methods and science education as well as the scientific record as a public good."

The Workshop consisted of general sessions and other discussions. A table of themes produced by the working groups is provided in Appendix C.

Major Themes and Conclusions

The Workshop addressed questions in several key areas regarding an information infrastructure for the physical sciences. How could scientists benefit from such an initiative? What scientific information/communication needs could such an initiative serve? What kinds of information should be included? Are there any useful infrastructure models available that would facilitate the concept? What mechanisms might exist or be developed for securing future scientific community input? The overall conclusion of the workshop was an enthusiastic endorsement of a vision of a national infrastructure that benefits not just the scientific community but the national good. It could ultimately impact not only R&D, but also education and applications to our everyday lives. It would be a step to integrate the whole of science to provide a basis to improve society, the economy, and the environment. The following text includes conclusions that address these and other questions, along with key points and examples, which collectively endorse and expand upon the overall concept.

[11] Ibid, Scott

[12] Ibid, Scott

[13] Wood, Fred, Office of Technology Assessment, "Helping America Compete: The Role of Federal Scientific & Technical Information," Report to US Congress, 1990.

Scope of the Initiative

Begin with the Physical Sciences so that it is doable, but engineer it so that it is scalable

The boundaries between physical sciences and other sciences in terms of advancing our understanding and solving scientific problems are increasingly less defined. The interfaces and the interactions among disciplines are often where the important discoveries are made. Ultimately a national information infrastructure for all of science is needed. We should not be looking at disciplines like stovepipes. Rather we should look across disciplines to determine what related or contributory work has been done or is underway that supports a researcher's disciplinary focus. Interdisciplinary issues like information technology, nano-technology or biotechnology are current examples. Equally recognized is the importance of bringing in stakeholders from other fields at an early stage.

Starting with selected areas within the physical sciences is a scalable approach that allows the effort to gain experience as it draws in stakeholders. From the opening talk of the Workshop, the question of why limit this to the physical sciences was raised. The need to start with a defined scope that is doable but scalable was recognized.

Information Types

We need a conceptual change that allows us to better integrate information types (e.g. text, data, images, animations) as well as information at various stages in the analysis process (raw data, partially processed data, text summaries of analyzed data in varying degrees of analyses, and unreviewed or peer-reviewed documents). Such a conceptual change would facilitate the serendipity and insights that can be gained by dealing with these multiple information types and their interrelationships.

The information community has developed significantly advanced systems to provide access to corpora of published documentation. Especially using Internet technology, science communities have created an explosion of new types of information to be created and exchanged. Much of this had been done by spontaneous efforts of interested parties. Figure A provides examples of information exchange that the DOE physicist community has created. The long-range challenge in access to information is with non-published text, with data and particularly images. The challenge is to build on the spontaneous efforts and support their integration through an accessible information infrastructure. A few research projects have been leaders in plowing this ground.[14] Much more work needs to be done and government-funded initiatives must continue to be supported.

Even by the physics community, science is not as systematic as some would like to think. In addition to non-textual information, scientists also need types of information other than the highest levels of peer-reviewed journals and the commercially published record. They need more gray literature, they need prepublications (preprints), and they need to know where data may be wrong or ideas are half-filtered. We need to have systems that can help us with lessons learned from others' experiences. These can be both successes and failures. In some cases, failures are at least as important to know about as successes.

The project to build a neutrino factory, a possible future DOE particle accelerator facility, is a case in point. This project involves over 200 physicists worldwide, and has produced hundreds of documents of which less than 10 have been published in peer-reviewed journals. Exchange of information among the geographically distributed workers is essential for success of such a project and has resulted in the creation of several special-purpose, Internet-based information archives. More desirable would be an information infrastructure beyond that of particular projects, but which is flexible enough to provide easy creation of and access to "libraries" of technical

[14] See summary of dli1 and dli2 at http://www.dli2.nsf.gov.

Figure A. Examples of Information Exchange in Physics

The well-known LANL archive must be regarded as a prototype for many key aspects of the proposed Information Infrastructure Initiative. This archive is, however, limited to documents prepared according to a specified electronic standard. The URL is:

http://xxx.lanl.gov/

Two information search engines that are fairly specific to elementary particle physics are SLAC SPIRES and the CERN Library catalogue:

http://www.slac.stanford.edu/spires/
http://weblib.cern.ch//Home/Library_Catalogue/

Another kind of useful information is catalogs of Conferences:

http://physics.web.cern.ch/Physics/Events/#otherlists

Research Institutions:

http://physics.web.cern.ch/Physics/HEPWebSites.html

Particle Accelerators:

http://www-elsa.physik.uni-bonn.de/accelerator_list.html

A Catalog of Catalogs:

http://www.hep.net/

Special-purpose, public-domain computer codes for the creation of new information. Some of the scattered sites are:

http://www.beamtheory.nscl.msu.edu/cosy/
http://www-ap.fnal.gov/MARS/
http://laacg1.lanl.gov/laacg/services/parmela.html
http://laacg1.lanl.gov/laacg/services/possup.html
http://wwwinfo.cern.ch/asd/geant4/geant4.html

A text-based survey of this problem is

http://wwwslap.cern.ch/collective/bibliography/bib2html/codes.html

but a better solution is needed for the 21st century!

documents related to a specific project. Examples of what might be integrated in a national infrastructure is included in Figure B.

Increasingly, scientists want access to knowledge resources ranging from raw data to published information to interpretations of information, all through a common desktop interface. Systems should also access processing theories so the user can have the whole paradigm at the desktop.

In addition to primary databases, metadata and derivative databases are also needed to help organize the way data are identified and, therefore, retrieved and as indicators of data quality and data structures.

Infrastructure, Products, and Services

Information Infrastructure has to be more than a storage and retrieval capability. Rather, in the long term, it has to support fundamentally new ways of doing science.

Information infrastructure can totally change the way we do research and education and the way we bring things to market. We need to go beyond traditional products and perspectives. Researchers want access to a spectrum of knowledge resources and a range of all types of digital objects. They want access to other researchers. Scientists also want desktop access to

Figure B. Examples of Information Archives for the Neutrino Factory

The URL for the entire neutrino factory project is:

http://www.cap.bnl.gov/mumu/mu_home_page.html

The Fermilab-based neutrino-factory document archive is:

http://www-mucool.fnal.gov/htbin/mcnote1LinePrint

The CERN-based neutrino-factory document archive is:

http://molat.home.cern.ch/molat/neutrino/nfnotes.html

A panel member, Dr. K. McDonald, maintains photos (and one video clip) as well as links to text documents:

http://www.hep.princeton.edu/~mcdonald/mumu/
http://www.hep.princeton.edu/~mcdonald/mumu/target
http://www.hep.princeton.edu/~mcdonald/nufact/

Special-topic bibliographies include:

http://www.hep.princeton.edu/~mcdonald//mumu/nuphys/
http://www.hep.princeton.edu/~mcdonald/mumu/physics/

Another example of a very useful kind of Internet-based information archive is:

http://www.hep.anl.gov/ndk/hypertext/nu_industry.html

This provides access to theory and results related to neutrino physics, but also gives links to the experiments themselves for those interested in grubbier details.

facilities from the local research laboratories to those at the North Pole so that instrument data acquisition and processing can be easily integrated in the analysis and collaboration processes.

They want remote computational tools where a problem is launched on the net and it finds its solution that can be anywhere. Information must become a more active tool in science. The peer review process should take place as part of the work in progress. Further, scientists need an electronic laboratory notebook to be shared "on the fly." Tools for value added processing of data such as data mining and recombining need to be focused, applied to the advancement of the scientific process, and provided more pervasively as a tool for research. There needs to be capabilities to visualize data from 2D to immersion. Scientists should be able to query for what they don't know, not what they know.

The following are some examples of experimental efforts and prototypes. The technologies are being developed through projects around the world. The information infrastructure needs to provide a focal point to help catalog these initiatives and bring them together with the information sources.

- Today, we do remote experiments through Internet connections with instruments. Lehigh University taps this research process allowing real time interaction. Students should be able to observe this scientific process in action as part of the learning experience.
- The University of Tennessee developed a system that finds the best available computing facility nation-wide to solve very complex parallel processing/ supercomputing algorithm problems. The decision is made automatically and transparently to the researcher.
- The Space and Aeronomy Research Collaboratory (SPARC) [http://www.windows.umich.edu/sparc/] at the University of Michigan uses remote instrumentation in collaboratories, which can move through theoretical models with visualizations, observing both empirical and theoretical results. Theorists and experimentalists interact on models in real time. This has been changing the way scientists see their science.

Scientists need systems that integrate simulation data with empirical and theoretical data. In the area of materials sciences, scientists have phenomenological databases in such areas as crystallography, materials properties, and thermodynamics. Some began as early as the 1920s. Each of these databases are independent and created for a certain use, but not for repurposing and reuse. They often don't physically or mathematically interrelate. If new materials are desired, the scientist must interrelate them. The paradigm must change. There can be great temptation to use such data in contexts where they may be inapplicable. As these databases become available to wider circles of users, it becomes increasingly important to supplement them with metadata describing their ranges of tested validity, and with tests to determine their validity for use in new contexts. An example: model multidimensional potential surfaces developed for interpreting spectra may contain serious deviations from reality that only become apparent if the surfaces are used for other purposes; notably, to study scattering processes.

Synthetic organic chemistry is a field where a shift in approach has proven enormously successful. In the past, the synthesis of each new chemical was an independent invention. The major change in organic synthesis came when computers began to predict the rational paths to follow for syntheses—to predefine processes to pursue. This model needs expansion to other disciplines, and innovative deployment of information technology is the solution to this integration.

The infrastructure further needs to encourage a diversity of scientific and technical information supplied by the publishing community. The coordinating layer needs to help account for security, legal and property rights, and provenance and quality of the suppliers. The Internet has forever changed the dynamic of intellectual property rights. There have been dramatic changes in the relationship of the producer, publisher, user and vendor. Licensing has superseded the first sale doctrine. Balances must be struck between the nature of scientific information as a public good and information as a commodity. There will need to be ways to access both toll roads and freeways on the information highway for scientists.

One must go from specific content products to linking to create a knowledge environment. And the validation of content from the front end is critical to this. The front end must provide for selectivity and quality indicators to help find the right information from the increasing volume of data available. The National Library of Medicine has developed such a model for the biomedical sciences, which is also needed in the physical sciences.

The infrastructure will provide opportunity for others to actively participate with new and better value-added services and capabilities. It is important that the need for open standards be a continuing responsibility and a point of convergence for the information infrastructure. The need for open standards ranges from technology to content. In the best interests of its research enterprise, the government should continue to actively promote open systems for the exchange of scientific data.

It is important that information about given topics can be associated through time even as jargon and vocabularies change. This use of terminology needs a leadership vision and can have substantial impact on the progress of scholarship. To address such a need in biomedicine, the National Library of Medicine developed the Unified Medical Language. Such a system requires a substantial commitment of resources to develop, but once done, opens up connections through time and disciplines where vocabulary was once a barrier to knowledge discovery and retrieval.

To build the information infrastructure, a strong information science community for the physical sciences is needed. Library and information scientists must gain knowledge of the physical sciences so they can help envision what is possible, and the physical scientists need to be information and information technology literate to provide the applications perspective. The advent of the Internet and many associated information technologies has brought the communities together, and the infrastructure needs this cooperation.

Archiving, Preservation and Access to Information

Archiving of data is one of the key concerns. The electronic record is much more fragile than the media that came before (print, microforms) because it comes faster, in bigger volume and is gone more quickly.[15] There is a real need for assurance of portability, regardless of media developments.

The critical need for a reliable, stable archive, which is trusted by all parties, is a primary concern in information management today. For government information, there is a national responsibility to protect the taxpayers' investment. With regard to other information that is produced with government funds, it is important to address the reliability of the private sector as a custodian of such resources.

There has been an active discussion in the information industry of roles in archiving. At one time, publishers looked to libraries for archiving. Now there is economic value in the collection; so publishers are taking on archival functions, or at least controlling the use of electronic archives in ways not done with paper backfiles. Libraries in turn are concerned about long-term preservation. In recent negotiations between publishers and the government, the publishers are looking at economic models where they are willing to allow third party archives since most of their revenues are coming through current issues. New models are evolving, but we must have interim measures and work rapidly to see long-term options before valuable knowledge is lost. One example of an innovative approach is that of the American Geophysical Union, which has set up a trust fund for data migration to aid in long-term preservation.

Preservation is a prerequisite but not equal to access. To be fully useful, the massive amounts of information must be accessible from anywhere. The concept of a distributed archive or repository is under development by the Corporation for National Research Initiatives (CNRI) with National Science Foundation funds under the leadership of Dr. Robert Kahn. CENDI[16] (an interagency group of Scientific and Technical Information managers of the nine major science and technology agencies in the federal government) is working with CNRI on concepts of federated repositories for government information. This work may be critical to the open systems standards that will allow for the massive and distributed scientific information archives of the future. Many of the agencies that are involved in this effort would also be key participants in the information infrastructure for the physical sciences, implying that future coordination should be facilitated by current relationships.

In addition to electronic information, there is a critical need for stewardship of all collections, regardless of medium. It is important to address issues of legacy information, that is, information that is not in useable digital form. Old data are still in significant demand. The Defense Technical Information Center reports that 11.4 percent of their 62,100 requests in the last year was for material over 25 years old. The National Technical Information Service reports similar demands for older material.

All of these initiatives just scratch the surface of capturing the knowledge base in the physical sciences. There are lessons learned, there are laboratory notebooks and there are experiences that are not documented. There are also new forms of digital objects such as visualizations, results of simulations, and processes of collaborations that are important to the scholarly scientific record that we have not yet begun to effectively address in terms of capture and preservation. Leadership is needed to explore the new issues raised by new information technologies and new ways of doing science. The coordinating role of the infrastructure initiative should assert leadership in exploring these archiving issues.

---

[15] *"Digital Electronic Archiving: The State of the Art and The State of the Practice,"* Carroll, Bonnie C. and Gail M. Hodge. Report sponsored by International Council for Scientific and Technical Information [http://www.icsti.org] and CENDI *(see footnote 14),* April 1999.

[16] CENDI members are: Commerce, Energy, EPA, NASA, National Libraries of Medicine, Agriculture, and Education, Defense, and Interior. For more information the URL is www.dtic.mil/cendi/.

Research, Education, and The Public Interest

The physical science information infrastructure of the future must begin by focusing on the producers of the information, i.e. the scientific community, but must also recognize the potential for positively impacting education and the general public interest.

Audiences for the information infrastructure for the physical sciences exist at every level of technical sophistication. Prioritization, however, is a must if the project is to have any realistic chance of success. From the practical perspective, the early adopters and the audience that will allow infrastructure development to proceed initially is the scientific community.

However, the physical science information infrastructure should, in the very early stages, consider effective ways to benefit both the education community and the public. It should focus on higher education, high-end initiatives such as Internet2 (a high-end consortium of universities, with a mission to advance telecommunications technologies in support of academic research), and particularly on methods for involving diverse populations.

In the global economy of the 21st century, educators must have the information they need to optimize their performance in both the classroom and the laboratory. In a 1998 report, it was identified that U.S. science and math achievement for grade 12 students falls substantially below what we would expect when compared to other countries. In science, we rank 16th, with a score of 480 and in math, we rank 19th, with a score of 461, significantly below the world average of 500.[17] It is the middle school years that are critical in preparing for a strong and diverse pool of scientists and engineers for the future. In order for the seed corn of education to take root, children must become interested in science and technology at an early age, and we must strive to hold their interest by making it come alive and by providing access to age-appropriate information. It is one thing to be computer literate as a "gamer;" it is quite another to use the tool to expand the intellectual capacity of a child directed toward productive ends.[18] How do we keep them interested? The infrastructure should provide for actively involving these groups.

We need to find ways to get science to the public. NASA ensures that every program must have outreach and education. The information infrastructure should become a viable tool for supporting this outreach.

At the bottom line, science needs to be publicized and democratized. From the education perspective, a physical science infrastructure with resources for educational use presents wonderful opportunities. The advances in the infrastructure will be supportive of and paralleled by advances in education, learning, and data simulations in ways we can hardly imagine today.

Quality

Quality in data and data processing is one of the cornerstones of success. The system must take into account the quality, the reliability, and usability aspects of the information.

The infrastructure has to help be selective in the quality of information that is made accessible. Given that the volume of data being produced today out paces our ability to absorb it, we need to ensure we have quality discriminators built into the information infrastructure.

It was noted that the irony of the information age is that it gives credibility to uninformed opinion. The value of data lies in its use and we must insure against "garbage in-garbage out" or, even worse, "garbage in, gospel out," or "garbage in, garbage at greater speeds." In medicine, this can easily and directly mean the difference between life and death. In the application of other sciences, it can also directly impact our personal lives as we note failures in the physical environment we have created.

[17] Ibid, Scott

[18] Ibid, Scott

New models for data quality are emerging as early experiments with new paradigms show results. For example, the LANL preprint server (http://xxx.lanl.gov/) was a "quasi spontaneous change" and is now becoming a trusted source for physical science information, even with the caveats that limit peer review. Some researchers in the physical sciences have shown they value unrefereed articles. The information infrastructure for the physical sciences must understand the dynamics of these changes and capitalize on them.

While we generate and manipulate data ever more easily, we must not lose the knowledge about what the data mean. High-quality metadata are critical in knowing the genesis of the information content. Through metadata and other means, we must try to ensure that the data and systems are transparent. We must work toward making the scientific expertise involved with the creation and manipulation of data properly captured in the interpretation and application processes. Attention to this issue will require national leadership.

Participation of Sectors of the Economy

We must continue to strive for broad and innovative collaborations. Stability comes from diversity. Though we may of necessity need to build the infrastructure with a focus on specific tasks and disciplines, we must encourage the broadest participation at the planning table to allow for future involvements.

Given that we now live in an information economy and one whose economic models of publishing and communication have been dramatically shaken by the new information technologies, the development of the physical sciences information infrastructure must be sensitive to the value-added roles of all the participants in the information industry. There are different parts of the industry, e.g. primary versus secondary publishing, which have very different issues as the life cycle of scientific information becomes more electronic and more distributed.

Historically, there has been a symbiosis between publishers and scientists. As the diversity and complexity of the information industry has grown, this relationship has changed through the increased use of the Internet and the tools available. The government role in this, because it is a major funder of scientists, has been changing.

The Department of Energy's recent development of PubSCIENCE is an excellent example of an innovative and effective collaboration between a government agency and the private sector publishing industry to bring science information to a worldwide user community. This new tool provides increased visibility of scientific article citations while providing revenue potential for the publishing community through expanded subscriptions and pay-per-view full text services.

Another government-funded initiative is that sponsored by the National Institutes of Health (NIH). NIH is planning to provide a publishing infrastructure for original articles. It will be opened to all publishers. NIH's original thinking about this system was to provide better access to government-generated information. The functionality of the system has undergone a number of changes as the public and private sectors have expressed their interests and concerns.

The compression of product cycle time is a key issue in the relationship among the sectors. In the past, certain developments in information systems were pre-commercial, and prototypes were developed within government. In the rapidly enabling information technology environment of today, the government must move rapidly and innovatively forward in providing systems to meet the needs of its scientists and engineers. It has no motivation for duplicating what already is done well in the private sector. However, the government has a responsibility to move forward and be good stewards of its information capital.

It is difficult to generalize regarding when a product duplicates an offering by the private sector. There are issues of A-76 studies and the Economy Act of 1932, which focus on the roles of the pubic sector and

agencies in getting the job of government done and place different demands on agencies. OMB A-130, which focuses on information policy, also has many conflicting directions (e.g., calling for diversity of sources, requiring the government to go to electronic information management and dissemination, and restricting pricing to the incremental costs of dissemination). Government requirements are complex, but understanding them is a prerequisite to good relations among the stakeholder communities. Similarly, the private sector, both for profit and not for profit (which both have significant roles in scientific and technical information), has its own inherently distinct interests. Finally, the academic sector also has its own unique and changing nature and must be a major player in any scientific enterprise. Clearly, an education process must be an integral part of partnership building. To facilitate communication and cooperation, clear and open notification of developments and plans are essential.

At the bottom line, the worst course is to close off paths of action or cooperation. We should focus on how to construct an information system to encourage broad participation. We must look for innovative partnerships between all sectors and collaboration will be a cornerstone of success.

Leadership

It is clear that we need a point of convergence and leadership to develop the common agenda and move it forward.

Given that the U.S. government spends over $12 billion in research and development in the physical sciences, we must ensure that the results of taxpayers' investment is properly mobilized in the national interest. Given that the volume of information, the "information explosion," is increasing at unprecedented rates and outstripping our ability to deal with it in human terms,[19] we must ensure quality information is delivered in a useful and timely manner. And given the potential to take advantage of the power of new information technologies, we must have infrastructural resources to meet researchers' needs and expectations and add new value to the advancing processes of science and engineering.

We must actively shape the future. There is a need to have a point of convergence and leadership to mobilize the stakeholder communities to provide their value-added contributions in the physical sciences particularly and in science in general.

The Department of Energy (DOE), because it is the largest funder of research and development in the physical sciences, because it has already invested in the information infrastructure and coordination role, and because it is a clear champion of the vision, should be tasked to continue to move the agenda forward.

The structure of the information infrastructure that is envisioned will make access to the comprehensive collection of content and services from all sectors seamless. The nature of the infrastructure will go beyond the mechanics of access to the cooperative, distributed model in which governance is coordinated and certain services are offered centrally, but implementation is distributed.

Given the size of the DOE commitment to physical science research and the leadership that DOE's Office of Scientific and Technical Information (OSTI) has played, it is desirable for them to take the lead, but other organizations or groups should be encouraged to participate from the earliest stages.

From the top and from the beginning, there is a need to operate with a structured means of obtaining ongoing senior scientific advice. The National Library of Medicine has a Board of Regents engaging the top medical and informatics people in the country, and this has proven invaluable to their planning processes. In the interests of time and to begin the process in the physical sciences, it will be good to go through an existing organizational infrastructure for senior scientific input. There are several committees dealing with physical science coordination. Within DOE, the most appropriate advisory body is the Office of Advanced Scientific Computing and Research Advisory Committee.

[19] "Beyond Databases and E-mail," Science, Vol. 261, 13 August 1993, p.841.

Funding

The government has a responsibility to disseminate the results of federally sponsored research as broadly as possible as a public good. The amount of resources needed should be benchmarked against model efforts.

R&D is funded federally to produce a public good. As an integral part of the R&D process, information is both an input and an output. The value of this process comes from the use of the knowledge output and is enhanced by the effective use of the existing knowledge base. The Web makes it possible for information to be searched at essentially zero or marginal cost once the information is made available. However, loading the information on a server, making it searchable, and providing the gatekeeper function to promote reliability all require capital investment. Adding in the tools to gain the added processing and collaboration technologies also requires coordination and investment. Managing the coordination and providing leadership requires staffing investments. And finally, there is a role to advance the state of the art and practice in physical science informatics. These are some of the key roles of the information infrastructure initiative, and they require a national commitment to their continuity.

Although it was beyond the scope of the Workshop to estimate the cost of initiating and developing the concept proposed, other benchmarks were reviewed and a budget adequate to create an information infrastructure to serve the size of the R&D enterprise is required. The NLM, a model for an area of science has a FY2000 budget of over $214 million. It supports the research interests of the $17.8 billion NIH budget. This ratio provides a loose parameter. The budget for physical science research in the U.S. is over $12 billion.

Timing

The time is now, the need is now!

The time is right for such an initiative because the challenge is a national one, the need for leadership and convergence is pervasively felt, technology is enabling, and the positive impact on the public good would be significant.

For well over 50 years, since the introduction of the computer, studies have called for better management of the nation's scientific and technical information resources.

So why now? Because the window of opportunity for the scientific community to influence the growth of the Internet and the Web is still open as these are used more and more for commerce and entertainment. That window will shrink (and may even close) if the next generation Internet does not come into being.

So why not now? Progress and change are incredibly rapid today. We have an economy that is based on information. Time to market can be measured in days. We need an information support structure where the impetus is in tune with the rate of change and the conduct of research. Science is global and the competition is intense. In an electronic world, the facts of the matter are no longer conveyed on paper or material but rather by electrons. These are not retrievable by human senses (as is the printed word) without the help of the information technology infrastructure. Once missed, they are gone. The most current information and the best technology services can now be in the same desktop toolbox. Our scientists need these tools to continue to compete in this increasingly competitive marketplace.

Reward Structure

Any analysis, economic or social, that might be developed to guide the construction of a scientific information system must recognize and accommodate the reward system of the sciences. The immediate rewards of doing science come from dissemination and acceptance of ideas, not from immediate monetary recompense. This occurs at all levels, from the apprenticeship of graduate school and postdoctoral work through the journeyman stage of pre-tenured faculty to the master craftsman level of tenured faculty. Understanding this value system is central to effective design of the future information infrastructure for the physical sciences.

International Diplomacy

The proposed information infrastructure for the physical sciences can be an asset or tool in international diplomacy. For example, the United States is spending considerable sums of money to work with retraining the Russian scientific establishment from military to civilian applications. An information system could be a tool in facilitating that development.

Today sending data across the ocean can be a problem. International cooperation, including mirror sites, becomes a tool for United States as well as foreign researchers.

A Name

The name of the initiative is important and it should reflect the key nature of what needs to be achieved. The initiative is more than a library, more than a physical infrastructure, and must help to advance the state of the art of physical science informatics. The concept of "institute" serves as an umbrella for advancing comprehensiveness of content, the infrastructure, and the enabling tools. Institute for Physical Sciences Information (IPSI) was selected as the operating moniker.

The Biomedical Model: Validation Of The Concept

Throughout the Workshop, the life sciences model was held up as a target. Clearly, the National Library of Medicine, focused on the electronic delivery of information to the desktop, provides an excellent model from the life sciences for the overall initiative. So what makes this model so successful? An analysis of NLM's success factors with current points of reference in the physical sciences is provided in Appendix D.

Findings

As a result of the Workshop and focused discussion, the infrastructure envisioned by the Department of Energy was in fact confirmed and findings support the establishment of the following:

A common knowledge base that seeks in an integrated approach to provide comprehensive access and facilitate the reuse of worldwide sources of scientific information, initially in the physical sciences, regardless of where they reside, what platform(s) they reside on, or what format or data structure they employ.

A point of convergence for ensuring the awareness, availability, use, and development of information technologies and tools to facilitate information assimilation, data analyses, peer communication and collaboration, sharing of preliminary research results, remote experimentation, validation of experimental results, and other uses not yet envisioned.

An openly available source of information to serve all users, from students to scientists to concerned citizens, in a highly efficient electronic environment, with tools to assist users in their quest for information and ultimately knowledge.[20]

[20] Ibid, Scott; (Note that openly available is not necessarily free of charge. As access to open government information gets increasingly integrated with private sector resources, we will need to find ways that toll roads and freeways can comfortably coexist and the user will determine the best path for his/her needs.)

Implementation

The requirements for the Institute for Physical Sciences Information (IPSI) (the operating moniker) must deal with three time horizons, each of which presents its own challenges:

- Doing better at what we're doing now (FY01)
- Mobilizing for what is possible tomorrow (FY02–03)
- Realizing the future potential (Out Years)

Development cannot effectively proceed without resources and an implementation plan should be proposed based on the concepts outlined in the May 2000 Concept Paper, modified by the findings of the Workshop.[21] The amount of resources needed to accomplish the development of the infrastructure should be benchmarked against model efforts of other disciplines and initiatives. This report of the Workshop provides a vision and some goals. Form here, we must flesh out the strategies and specific actions to reach these goals.

Doing Better At What We're Doing Now (FY 01)

Great progress has already been made in the delivery of textual information to the desktop.[22] A Department of Energy system provides access to the full text of over 1,000 journals from 24 publishers through a bibliographic front end that allows effective searching and identification of articles of interest. Another system provides desktop access to the gray literature produced by the Department of Energy. In fact, as a result of the Workshop itself, an understanding in principle was reached between NASA, DOE, and DOD to create a Gray Literature[23] (GrayLIT) Network. It will be a combination of the full text gray literature that NASA, DOD, and DOE make available on the web. DOE's Office of Scientific and Technical Information will overlay a distributed search tool on top of already existing databases at the three Agencies. There is a preprint network that provides access to over 1,000 preprint servers, including the leader that was developed for physical sciences by Los Alamos National Laboratory. Now that there is proof of concept and a critical mass, many new sites are being offered for connection.

These are some of the immediate activities that can be cited that confirm what coordination and leadership in the near term can accomplish.

Other immediate opportunities for enhancement include increased comprehensiveness of journal and gray literature. As was suggested at the Workshop, we need to go further and begin to build databases on the fractal model: begin with one and extend and connect and extend and connect. We need to optimize our current systems with the goal that we can reach relevant content within three clicks. As we extend our systems and the heterogeneity of resources and participants involved, we need to be able to assure that our interfaces do not lead to dead ends and false paths.

There are other efforts that are beginning to come on the landscape that can be moved forward now. For example, joining papers from different journals in different disciplines in such topics as nanoscale technology. The American Association for the Advancement of Science (AAAS) is already working in this area, and these efforts should be visibly supported. Effectively, this builds virtual journals of interdisciplines, especially for hot topics. The American Physical Society would welcome participation in such experimental efforts.

---

[21] *"Future Information Infrastructure for the Physical Sciences: Concept Paper, May 2000,"* White paper distributed at Workshop, May 30-31, 2000, 5 pages.

[22] Examples of DOE systems were given in the handout package and by Walter Warnick in his talk. These are further described at http://www.osti.gov.

[23] Gray literature is defined by GreyNet, the Grey Literature Network Service (Farace, Dominic,1997) as, *"that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers."*

And, finally, we must continue the planning process and be inclusive. In the first stage of the plan, DOE should have a call for participation from other agencies and other stakeholders. This Workshop was the first step.

Mobilizing For What Is Possible Tomorrow (FY02-03)

We need to create a capability for testbeds so that the system will remain involved in the cutting developments and can evolve as new experiments are tried and new ways of advancing science with the knowledge bases are developed.

The Institute for Physical Sciences Information should be set up ready for change, with the opportunity to adapt built in to the model. To move ahead at the forefront of informatics opportunities, we need to build on what leaders and centers of excellence are doing in other disciplines. In particular, the physical sciences should find areas where there are good intersections with the biomedical sciences. Joint testbeds should be initiated to test and deploy the results of cutting-edge information technology R&D programs.

There was a generally recognized need that the process must be kick started and some additional resources need to be brought to bear.

As a rapidly evolving area, devising specific agendas in the mid-term is difficult. Under such dynamic circumstances, it is imperative that particular attention be paid to guidance from the scientific community.

Realizing The Future Potential (Out Years)

As part of the IPSI, we must have both intra-and extra-mural research capabilities in the developing fields of disciplinary informatics. We must also focus on a commitment to extend the boundaries of the infrastructure to interface with other science areas' information resources, including the biomedical and environmental sciences.

Agencies such as DARPA and NSF fund the basic research in these areas. The mission agencies such as DOE, NASA or DOD can bring the results forward through prototyping and testing.

We need R&D to invent this; we need to build testbeds to prove this; and we need the loop back to education to fuel this. Appendix E provides some suggestions for the necessary R&D from a Panel member.

And we must actively complete the linkages with the infrastructures in the other sciences to realize the ultimate vision of a science information infrastructure for the U.S. Other fields like biology and medicine will lead in some areas. The physical sciences will innovate in other areas. Throughout the development of all of these systems, interoperability and standards compatibility will be a long-term goal. Within the next four to six years, it should become a reality.

Appendices

Appendix A

Workshop Panelists and Participants

Panelists

Alvin Trivelpiece, Chair
Formerly, Director, Oak Ridge National Laboratory

R. Stephen Berry
The University of Chicago

Martin Blume
American Physical Society

Jose-Marie Griffiths
University of Michigan

Lee Holcomb
NASA

Kirk McDonald
Princeton University

Krishna Rajan
Rensselaer Polytechnic Institute

Kent Smith
National Library of Medicine

Derek Winstanley
Illinois State Water Survey

Participants

Fran Buckley
Government Printing Office

Eileen Collins
National Science Foundation

Blane Dessy
Department of Justice

Denise Diggin
DOE Energy Library

Susan DiMattia
Special Libraries Association

Gail Feldman
Archive.org

Eleanor Frierson
National Agricultural Library

Laura Garwin
Nature Magazine

Daniel Greenstein
Digital Library Federation

Howard Harris
University of Maryland

Larry Lannom
Corporation for National Research Initiatives

C. Diane Martin
National Science Foundation

Kurt Molholm
Defense Technical Information Center

Donna Scheeder
Special Libraries Association

Bill Sittig
Library of Congress

Mike Spinella
American Association for the Advancement of Science

Paul Uhlir
National Research Council

Greg Wood
Internet2

Staff Facilitator/Editor

Bonnie C. Carroll
Information International Associates, Inc.

Appendix B

Workshop for a Future Information Infrastructure for the Physical Sciences Agenda

Day 1, Tuesday, May 30, 2000
2:00 PM–5:15 PM

Welcome and Workshop Objectives	Al Trivelpiece Formerly, Director of ORNL
A Future Information Infrastructure for the Physical Sciences: Concept and Assumptions	Walt Warnick Department of Energy
Internet 2 and the Changing Nature of Scientific Communication	Greg Wood Internet2
A Future Information Infrastructure for the Physical Sciences: Partner and User Considerations	RL Scott Department of Energy
The National Library of Medicine: A Case Study	Kent Smith National Library of Medicine
Discussion Panel: Views and Opportunities	Jose-Marie Griffiths University of Michigan
	Krishna Rajan Rensselaer Polytechnic Institute
	Lee Holcomb NASA
	Kirk McDonald Princeton University

Day 2, Wednesday, May 31, 2000
9:00 AM–12:00N

9:00–9:45 Plenary Discussion
Present the list of considerations with opened discussion.

9:45–11:00 Discussion Sections

Why

- Understanding the Need
- Success Factors

What

- Content
- Services

Working Group Leaders: Stephen Berry, University of Chicago
Derek Winstanley, Illinois State Water Survey

11:00–12:00
Reports of the Working Groups and Discussion

Wrap-up and Next Steps

Appendix C
Working Group Themes

What: Content and Services

Working mission statement—provide an infrastructure for the physical sciences with standards, links, etc. . . . to facilitate resource location and access across sources and media for the purpose of generating new knowledge through coordination and integration.

- Private vs. Public: Accessibility among sources should be collaborative, not exclusionary
- Portal that shows reliability
- Long term stability
- Assurance of portability and accessibility
- Access to older literature — with fully searchable text — making adjustments for changing vocabularies through time
- Profile based automatic notification of information (individualized)
- Means to identify others who have common interests who wish to be identified (privacy issues acknowledged) — community of practice
- Easy passage among different kinds of information — papers (text) to primary data to simulations and back again (through different categories)
- Function of scientific unity: meta-tools development, e.g. synthetic organic chemistry — traditionally, each scientist figured out a unique synthesis path. Now computer programs allow systematic development of pathways. Knowledge based program to conduct synthesis. Need to have systematic approach to looking for such tools.
- Testbeds to demonstrate applications of information infrastructures (education as an easy model)
- Access to cross cutting fields — subservice of access to different vocabularies in different fields
- Visualization tools — how to make accessible to support scientific discovery
- Reliable stewardship of unique bodies of information — content and systems for accessibility
- Need for helpdesks for finding and interpretation and analysis of information
- Security of data, characterization of consistency, reliability (including authenticity), validation information including later testing applied retroactively (forward and backward referencing)
- Flexible infrastructure to allow for the continued evolution of services that are today not apparent
- Current customers and data providers envision new services; information science can envision future possibilities — need to develop the informatics of the physical sciences to move the agenda forward
- Adaptive Management Model

(Appendix C continued)

Why: Understanding the Need and Success Factors

Needs of the Communities

Funders

- Facilitate resource location and access, across sources and media, for the purpose of generating new knowledge through co-operation, co-ordination and integration
- Increase national competitive advantage
- Maximize taxpayers' investment

Customers

- Verify reliability and quality
- Access currently unavailable information
- Archive reliably information and data

Industry

- "Pre-competitive" approach

Success Factors

- Create strong advocacy groups (professional groups, policy makers, general public)
- Identify champions
- Network of collaborators
- Integral part of the research process
- Quality
- Develop comprehensive plan
- The planning process is key. It is the source of major new initiatives.
- Organization and structure

Appendix D

The Biomedical Model

Throughout the Workshop the biomedical sciences was held up as a target. So what is it that makes this model so successful? And what already exists in the physical sciences that could be ingredients in a successful information infrastructure for the physical sciences? The following table provides some points of reference:

NLM'S Success Factors	Points of Reference in the Physical Sciences
Rich History of Achievement	DOE has produced the world's most comprehensive database in Energy Science and Technology, now containing over 5.5 million records. The combination of journal literature (PubSCIENCE), gray literature (Information Bridge), preprint literature (PrePRINT Network), and EnergyFiles begins to lay the infrastructure for the near term base for the IPSI.
Strong Congressional Mandate	The DOE-enabling legislation (Atomic Energy Act of 1954), states "The dissemination of scientific and technical information relating to atomic energy should be permitted and encouraged so as to provide that free interchange of ideas and criticism which is essential to scientific and industrial progress and public understanding and to enlarge the fund of technical information." Although this is the basis for past work, it would need significant strengthening to commit to the national leadership needed for the IPSI.
National Network of Libraries of Medicine	There is a strong network of science and technology libraries, with outstanding resources in all of the National Laboratories. There is no exact equivalent to the Medical Library Association, but there are divisions of the Special Library Association that help to support library networking.
Applying the latest in computer and communication technologies to health problems	DOE has a strong history and record as a leader in these technologies applied to the physical sciences including ESNet and DOE high performance computing. Most of the National Laboratories are centers of high performance capabilities, particularly for scientific computing.
Being an integral part of the research process	Within the DOE Office of Science, there is the Office of Scientific and Technical Information. This office has been a champion of the vision for the IPSI.
Dynamic Planning Process	There needs to be more systematic involvement of the best and brightest in the physical sciences for this initiative just like the role the Board of Regents plays for NLM.
Strong Advocacy Groups	The major Library Associations and the Federal Depository Library Program are just the most rudimentary beginning of a strong advocacy coalition. This Workshop was a step in focusing on this issue.
Effective Publicity Program	Top DOE management commitment is becoming increasingly visible and the media are recognizing the initiatives that have been taken. There is now a rather famous picture of the Secretary of Energy doing the ribbon cutting for PubSCIENCE.
Good Organizational Structure	DOE has the structure under the Office of Science with a major information facility in Oak Ridge, Tennessee.
Garner Necessary Resources—Dollars, People, and Facilities	Fiscal budgets have been good for raw infrastructure such as high performance computing and networking. Application specifically to the management of content has substantially lagged behind both this infrastructure and the funding of research and development.
Quality	Historically, the information products from the DOE have had worldwide recognition and use. Bibliographic products have been enhanced and transitioned to full text delivery mechanisms. New collaborative initiatives are partnering DOE with the private publishing industry. And the DOE has been recognized as a model agency in working with the GPO for electronic public access to scientific and technical information.

(Appendix D continued)

Two additional factors were added to the original list presented by NLM. First, NLM began with peer reviewed scientific journals in the highest quality bibliographic database in the industry. In keeping up with the pace of the times they were the pioneers in adding citation linking to their bibliographic database in PubMed (their major web-based information service); they have added free public access to PubMed[24]; they have added new types of content to their journal literature including information on clinical trials in process; and they have linked to websites to direct the public to quality information about health issues (PubMedPlus). Most recently they have entered the publishing arena with PubMedCentral as a direct reaction to the changing nature of scientific publishing and as a reflection of the enabling Internet technologies.

Second, NLM has made a substantial investment in metadata standards and ontologies that help to structure knowledge. Unified Medical Language allows consistency across time and perspective (i.e., doctor, researcher, nurse, medical records manager, pharmacist).

[24] DOE/OSTI has followed the NLM PubMed lead with PubSCIENCE by adding citation linking to their bibliographic database and working with the Government Printing Office to provide public access to this web-based information product.

Appendix E

Suggested R&D for the Future Information Infrastructure for the Physical Sciences

Dr. Krishna Rajan, Professor of Materials Engineering
Rensselaer Polytechnical Institute

Information Infrastructure: Managing Scientific Complexity for Knowledge Discovery

Traditionally, computers have been used as storage mediums for large volumes of data and as tools for carrying out extensive numerical computations and simulations. Recently, computers have started to take a more active role in guiding the scientist through the research and discovery process with the help of data mining methods for automatic discovery of patterns in large volumes of data. The challenge is now to make these data mining methods ubiquitous and an integral part of the data collection and verification process. The physical sciences and engineering offer a unique challenge in data mining due to the variety of data types, and their complex interconnections. For instance during the engineering design process, there is a need to integrate multiple, heterogeneous databases to reach new and even unexpected conclusions as well as to use databases actively to design new processing strategies. This complex coupling of data models, data analysis methods and physical methods offer a unique computing challenge that has not yet been addressed sufficiently in information technology research. Some issues to focus on may include:

- **How do we combine heterogeneous databases so that we can discover interesting patterns from them?** These databases contain data from different lengths of scale, from simulations as well as experiments. For instance, in the field of materials science, how does information on phase stability extracted from thermochemical data compare with information derived from electronic structure calculations? These databases need to be combined with information on microstructure and physical properties derived from experiment and simulations. From these independently organized databases, one can compare and search for associations and patterns that can lead to ways of relating information among these different datasets.
- **What are the most interesting patterns that can be extracted from the existing data?** The new patterns to be discovered should reflect the complex relationships that exist in both spatial and temporal dimensions. One needs to explore datasets on engineering performance. These may include for example, chemistry, crystal structure, microstructure, diffusion behavior, processing methodologies among others. Such a pattern search process can potentially yield associations between seemingly disparate data sets as well as establish possible correlations between parameters that are not easily studied experimentally in a coupled manner.
- **How can we use mined associations from large volumes of data to guide future experiments and simulations?** Data mining methods should be incorporated as part of design and testing methodologies to increase the efficiency of material application process. One of the challenges in mining the results of different experiments and simulations is the fact that results of many of them describe engineering problems at different scales and different forms (from visual, such as the form of microstructure, to numerical, like material properties). Coupled to this is the fact that properties and structure are time dependent. These temporal based data sets will not only come from reported experimental data but also from simulations which will permit us to explore a wider set of “virtual” data. This in turn requires knowledge of time dependent behavior at different length scales ranging from atomic motion to morphological changes in microstructure.