

Final Report of the Indiana University Cyberinfrastructure Research Taskforce



May 2005

*Prepared for Dr. Michael A. McRobbie,
Vice President for Research & Information Technology
Indiana University*

© *The Trustees of Indiana University*

Table of Contents

Foreword	iv
Executive Summary	1
1. Introduction	2
2. Cyberinfrastructure Investments	3
2.1 Ensure Cyberinfrastructure Foundations	3
2.2 Develop Holistic Capabilities for the Full Data Lifecycle	4
2.3 Ensure Computational Capabilities for Data Analysis and Simulation	8
2.4 Expand Visualization	9
3. Scholarly Use of Cyberinfrastructure	10
4. Partnering to Develop and Sustain Cyberinfrastructure	11
5. Conclusion	12
Appendix A: Taskforce Charge	13
Appendix B: Taskforce Membership	14
Appendix C: IT Strategic Plan	19

Foreword

On behalf of the members of the Cyberinfrastructure Research Taskforce, I am pleased to provide this report in response to the charge given by Vice President McRobbie. The full charge to the taskforce is in Appendix A, but in summary, our work was to examine Indiana University's needs for cyberinfrastructure to advance the goals set by President Herbert in his 2004 inaugural address:

In addition to increased scholarly publications, works of art, concerts and other forms of creative scholarly activity, our goal must be to double Indiana University's externally funded research grants and contracts by the end of this decade. Such a record of accomplishment will move the university even higher among the ranks of America's most distinguished institutions.

The taskforce represented a diverse set of disciplines and schools, and its members are some of the university's leading scholars. We accomplished our work through meetings of the full taskforce, subcommittees, and electronic communication. The taskforce heard from Dr. Dan Atkins, Chair of the "National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure" regarding his committee's substantial work and influential report on national cyberinfrastructure. We also heard from Dr. Sangtae Kim, Director of National Science Foundation's Shared Cyberinfrastructure Division.

While each discipline has its own unique research challenges, I was very encouraged that our first meeting surfaced so many common issues. At first blush, one might not anticipate that Ethnomusicology and the Cyclotron are working on similar technology challenges in their research agenda, yet our discussions repeatedly surfaced serendipitous commonality among fields. These conversations were instrumental in focusing the recommendations of this report on those areas that provide common cyberinfrastructure leverage points for IU scholars.

As taskforce chairman, I wish to thank the faculty members and ex-officio staff from UITS' Research and Academic Computing division, especially Craig Stewart and Tom Hacker, and the Pervasive Technology Labs for their considerable investment of time in this report (Appendix B lists their names). We have made every effort to respond to the taskforce's charge by providing a relatively short and direct report on the cyberinfrastructure needs for Indiana University.

Sincerely,



Bradley C. Wheeler, Chairman
Associate VP for Research & Academic Computing (Emeritus)
IU-Bloomington Dean of IT

Executive Summary

The Cyberinfrastructure Research Taskforce met during the 2004-05 academic year to consider Indiana University's (IU) needs for shared cyberinfrastructure investments. In particular, the charge to the taskforce asked scholars to focus on needs that could help support a doubling of IU's externally funded research by 2010-2011.

This report to the IU Vice President for Research & Information Technology conveys 10 specific recommendations. It recognizes both current progress in cyberinfrastructure development while also proposing new directions for cyberinfrastructure needs and opportunities.

In summary, the recommendations affirm a continuity of investment in the core IT infrastructure that is the foundation for advanced cyberinfrastructure. Developing deep capabilities for serving the complete research data lifecycle emerged as a clear and pervasive theme across many disciplines. The recommendations provide guidance for storage capacity; data movement across networks; collection, annotation and provenance; and data publishing, curation, and custodianship. The taskforce advocated "continuing without pause" renewed investment in IU's High Performance Computing (HPC) systems and visualization facilities and strongly advocated HPC as a competitive necessity for data-intensive scholarship.

Beyond the technology investments, the taskforce gave considerable analysis to scholars' needs in making productive use of cyberinfrastructure. The taskforce recommends investments in an array of subsidized and chargeback consulting services, complexity-hiding interfaces, and training programs that each are discipline-facing in their orientation rather than a homogenized one-size-fits-all.

Finally, developing and sustaining advanced cyberinfrastructure will be impossible with only university sources of funding. The taskforce strongly advocates aggressive partnerships and leadership at the state, national, and international levels to compete for all forms of external funding to continue incremental evolution of IU's cyberinfrastructure.

The report itself provides many more details beyond these recommendations. Diverse scholarly endeavors are evolving their use of cyberinfrastructure in different ways. Nevertheless, the themes and specific recommendations presented here represent a resounding consensus view across these disciplines for the shared cyberinfrastructure needs of IU's scholars.

1. Introduction

Cyberinfrastructure is a term used by the National Science Foundation and others to describe the information technology resources used by researchers, clinicians, engineers, and artists – collectively referred to as “scholars” – to create new knowledge. As infrastructure, it should someday become as common as the electricity, air conditioning, and plumbing that support our current research buildings. Cyberinfrastructure is a collection of instruments, sensors, high performance computational systems, massive data storage systems, data resources, and visualization facilities, tied together by high speed networks and made to work together by advanced software to accomplish goals that would not be possible by any single information technology system. Cyberinfrastructure is more than just the hardware and software of computer systems and networks. It encompasses people, processes, training, security, policies, and capabilities to sustain these over time. As noted in Vice President McRobbie’s charge to the Cyberinfrastructure Research Taskforce, many forms of research and scholarship are becoming increasingly digital endeavors.

The term cyberinfrastructure is a relatively young and still evolving term, yet it is drawing considerable interest from scholarly communities. Commissions, meetings, conferences, and workgroups on cyberinfrastructure have been chartered by nationally prominent agencies, organizations, and disciplines. A sample of these include

- ! *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Advisory Panel on Cyberinfrastructure*, February 2003.
- ! *Building a Cyberinfrastructure for the Biological Sciences (CIBIO) 2005 and Beyond: A Roadmap for Consolidation and Exponentiation*, Subcommittee on 21st Century Biology, NSF Directorate for Biological Sciences Advisory Committee (BIOAC), July 2003.
- ! *E-Research and Supporting Cyberinfrastructure: A Forum to Consider the Implications for Research Libraries & Research Institutions*, Association of Research Libraries (ARL), October 2004.
- ! *Commission on Cyberinfrastructure for the Humanities and Social Sciences*, American Council on Learned Societies (ACLS), in progress.

These reports and the nation’s federally-funded cyberinfrastructure efforts typically have one or all of the following three goals:

- ! Enable research, clinical service delivery, engineering design, analysis, and monitoring, and new arts forms not presently possible – often because the information technology requirements are beyond the capabilities of any single institution, facility, or project;
- ! Increase use (and the number of researchers who use) advanced IT capabilities in their research and creative activities by lowering the barriers to adoption of this technology;
- ! Equip scholars, or groups of scholars, to solve their own particular research challenges by independently and transiently assembling the required resources to create virtual organizations that may range from interdisciplinary scientific research teams to multi-continent performing arts groups.

All faculty, staff, and students make use of some common IT infrastructure, such as email, networks, enterprise software license agreements, and help desk services. Cyberinfrastructure for research provides

extended capabilities beyond those needed by everyone. New classes of “grand challenge” research problems and new approaches to existing research often press the capabilities of modern IT. Research endeavors sometimes struggle with procuring massive volumes of storage, creating data formats that can be shared and searched, and optimizing computationally-intensive programs that challenge even supercomputer class machines. Training staff to use modern IT tools, developing wizards to hide the complexity of these tools, and sustaining cutting edge use of tools as research staff and graduate students come and go are also part of the challenge.

Left to their own devices, each research project and discipline would find it necessary to spend considerable effort to develop the digital tools and resources needed to support their work. Much of this work may not be the best use of scholars’ and their staff’s time. It often leads to “re-inventing the wheel” without realizing that tools and techniques from other fields already exist that can address the challenge. Discipline specific work on what might otherwise be common cyberinfrastructure provides nearly perfect local control for a project, but this often comes at an enormous cost in lost leverage in purchasing, operating, securing, and upgrading tools over time.

Thus, the quest for understanding cyberinfrastructure to support IU’s scholars begins with understanding common classes of problems faced across research disciplines, identifying opportunities for leveraged and common solutions for those problems, and discerning effective ways to blend “leverage at the center” with “innovation at the edges” in provisioning cyberinfrastructure. The following recommendations of the taskforce are focused on common, actionable, and achievable wins for cyberinfrastructure at Indiana University while recognizing an ongoing engagement with broader national and international cyberinfrastructure developments. The report begins by affirming current investments in IT foundations and addressing four specific areas for investment. Section 3 addresses services to ensure that scholars can use cyberinfrastructure. Section 4 advocates effective partnerships to develop and sustain cyberinfrastructure, and the conclusion sets these recommendations in a broader context for scholarly success.

2. Cyberinfrastructure Investments

The taskforce recommends four specific areas of investment for the IT components of cyberinfrastructure. Each of these is explained below.

2.1 Ensure Cyberinfrastructure Foundations

The taskforce strongly affirms that the foundations provided by the 1998 IT Strategic Plan (ITSP) have been instrumental in sustaining and accelerating IU research. Common IT infrastructure and processes, such as life cycle funding for faculty computers, enterprise software licenses, IT training and support, and advances in the Digital Library Program, have benefited all members of the IU community.

The taskforce reaffirms that Recommendation 5 from the ITSP remains as true as ever:

ITSP RECOMMENDATION 5: In support of research, UITS should provide broad support for basic collaboration technologies and begin implementing more advanced technologies. UITS should provide advanced data storage and management services to researchers. The University should continue its commitment to high performance computing and computation, so as to contribute to and benefit from initiatives to develop a national computational grid.

Through the implementation of Recommendation #5 and its associated Action Items (see Appendix C), IU has developed nationally prominent strengths in supercomputing, advanced visualization, digital libraries, massive data storage, and the delivery of expert support in utilization of these facilities. These are foundations for further development of IU's cyberinfrastructure, but progress on other areas must not come at the expense of failing to sustain and advance these foundations.

In some cases, schools and departmental IT facilities have become critical to scholarly work. Many of these facilities are not budgeted for lifecycle services and replacement – especially in the arts. The taskforce affirms sustaining core IT investments and an expansion of lifecycle budgeting practices where appropriate.

CRT Recommendation #1: *Indiana University should continue investments in core IT infrastructure that is a foundation for IU's advanced cyberinfrastructure. The university should expand the successful principles of equipment life cycle budgeting as used in the ITSP to all levels (schools, departments, etc.) to ensure the long-term sustainability of the core IT infrastructure required by scholars.*

Successful use of the core IT infrastructure requires more than just providing hardware and networks. Software site licenses that provide students, faculty, and staff with the full functionality of the most widely used programs for data analysis – including SPSS, SAS, Matlab, and others – are essential to support basic scholarship. Similarly, site licensing of data sets and provisioning services to make them widely available are essential for data and computationally-intensive scholarship.

CRT Recommendation #2: *Indiana University should continue its investment in site licenses for software and datasets as scholarly tools for data analysis and interpretation. Licensing arrangements should include both personal and university-owned workstations, and whenever feasible, include the entire university.*

Beyond sustaining and enlarging the core infrastructure, the university should support and invest in advanced cyberinfrastructure to support scholarly endeavors. Investments should adapt to changes in research needs that have developed since the IU Information Technology Strategic Plan was written and approved. The following recommendations address the clearest and most common needs for advanced cyberinfrastructure at Indiana University.

2.2 Develop Holistic Capabilities for the Full Data Lifecycle

A clear, resounding, and immediate theme that emerged from the taskforce is concern that Indiana University does not have the requisite facilities, tools, and support mechanisms to adequately manage data generated from scholarly activities. Many scholars also lack the tools and skills to help transform massive data sets into information, and information into knowledge.

The concerns expressed by the task force members fell into several broad areas:

- ! Lack of sufficient storage capacity
- ! Unmet needs to quickly move data between instruments, storage systems, and computational resources across campus and across the nation
- ! Lack of facilities to provide for the long term annotation, curation, provenance management, and archiving of data
- ! Lack of a common authentication mechanism for all storage and computational resources

This section examines each of these concerns in detail, and recommends actions to address the whole of the scholarly data lifecycle.

A) Data Storage Capacity

The acceleration of rates at which all areas of intellectual and artistic endeavor create data of diverse forms has been widely noted – generally without the investment and thoughtful creation of facilities that constitute an adequate response to address the challenge. The growth in the rate of data production is driven by the increasing integration of sensor data, knowledge databases, and computational simulation and analysis models. In the humanities, new uses of information technology are similarly generating TeraBytes of multimedia data at an accelerating rate.

Even though the university's current storage capacity is on par with leading institutions, current and planned research activities require more data storage capacity than currently exists at Indiana University. IU researchers are developing new facilities that will accelerate our ability to generate new data. For example, Indiana University's analytical chemistry group has developed advanced spectrometry instrumentation that is currently capable of producing data faster than IU's cyberinfrastructure can absorb, transport, or analyze it. Future developments will change the situation from one in which a lab might generate TeraBytes of data per year – which IU's cyberinfrastructure is not currently capable of managing – into tens of TeraBytes per year.

This problem is a general one spanning many areas of the natural and social sciences, humanities, and arts. Digital library programs convert video and audio content into digital form for instruction and research use at a rate of several TeraBytes per year. The EVIA Digital Archive Project¹ is converting ethnomusicology video and audio content into digital form for instruction and research use at a rate of three TB per year for 100 hours of video. Right now the scale of the digital data sets so generated has compelled the EVIA project to use a compression and encoding scheme (MPEG-2) that results in the loss of data. Better, lossless storage formats exist but are impractical because the amount of storage required would be beyond the capabilities of IU's current storage facilities.

Concerns regarding an explosion of scholarly data have long been voiced in many forums². This imminent flood of scholarly data along with the issues of annotation, provenance, and data curation are viewed as critical impediments to scientific progress. One example of this is the ATLAS particle physics project in which Indiana University is a partner through the US-ATLAS test bed project. When the ATLAS project goes into production in 2006, experiments will produce around 10 TB per year, and by 2015, particle physicists are projected to require ExaBytes³ of storage. The San Diego Supercomputing Center's May 2005 announcement of adding 1.1 PetaBytes of spinning disk storage to serve data-intensive sciences provides relevancy to these concerns⁴.

The Taskforce recommends that IU substantially increase its overall storage infrastructure in preparation for these research trends. The data storage architecture should be extensible to allow incremental additions of storage capacity over time as storage densities increase and costs decrease. Storage must be supported by a networking infrastructure that can facilitate access among IU researchers and their

¹ Ethnomusicological Video for Instruction and Analysis, <http://www.indiana.edu/~eviada>.

² Hey, A. J. G. and Trefethen, A. E. (2003) 'The Data Deluge: An e-Science Perspective', in Berman, F., Fox, G. C. and Hey, A. J. G., Eds. *Grid Computing - Making the Global Infrastructure a Reality*, 809-824. Wiley and Sons.

³ ExaByte is 10^{18} or roughly a quintillion Bytes or a million TeraBytes.

⁴ "SDSC Achieves Petabyte Milestone," <http://www.taborcommunications.com/dsstar/04/0323/107693.html> (accessed 12 May 2005).

collaborators. This infrastructure will need to be budgeted with ongoing life cycle funding to provide timely equipment and software upgrades to keep up with storage demands, and to maintain a competitive edge for grant submissions.

CRT Recommendation #3: *Indiana University should continue to execute and accelerate an incremental and extensible strategy that enhances its overall storage infrastructure from online storage to long-term archival storage. Dependable archival storage must include a commitment to ongoing and periodic data validation and maintenance of software for reading and migrating the data to newer formats. (see Section 2.2.D, below).*

B) Data Movement

High performance computer networks are necessary to move large amounts of data between storage and computational elements across short, regional, and long distances. Backbone networks, such as Abilene and I-Light, are well provisioned with high-speed fiber optic links. For some labs and intensive use purposes, however, intra and intercampus networking within the university can be a limiting factor in moving massive data sets for computation, analysis, or backup. IU's core campus network has great capacity between buildings, but some of the challenges are connecting and provisioning this capacity to individual workstations or departmental computing systems. In some cases, it may take days to transport massive files from a lab connection to national networks in Indianapolis or from one part of campus to another. The complexity of modern networks and interfacing with them means that actual end-to-end performance is often far below a network's apparent capabilities.

The current network challenge for data movement is at the "edges" of the network near offices, laboratories, and specific classrooms where research data is used for instruction. Network infrastructure at the campus and building level is too expensive to continually upgrade to keep up with advances in the backbone network, and the need for very high capacity networks is, at present, specialized to specific types of research and facilities. Moreover, the requisite knowledge to performance tune applications, operating systems, and networks to fully exploit the network capabilities is becoming greater over time (the "wizard gap"). Thus, the mismatch between the abilities of the storage/computation facilities and the abilities of computer networks can impede effective scholarly use of these advanced facilities.

CRT Recommendation #4: *Indiana University should enhance its networks through optimized engineering or capacity growth to include much faster end-to-end network capabilities from specific points of need (laboratories, offices, classrooms) to IU's central computing facilities and national and international research networks.*

C) Data Collection, Annotation, and Provenance

The current state of understanding – knowledge – stems from interpretation of data in light of theory. Data are, or at least should be, continually contemplated, reinterpreted, reviewed, and diffused to broader communities as our theories and assumptions change. A static set of data is brittle and cannot adequately represent the dynamic aspects of knowledge. In this context, the taskforce identified three broad, unmet problem areas for managing and using data: collection, annotation, and provenance.

Data is increasingly being collected via real-time sensor arrays or simulations that generate massive streams of data for analysis. Better techniques and facilities for ensuring unobtrusive and simultaneous development of metadata are needed to provide for reuse of these data streams.

An example of the problem is in the maintenance of genomic data in initiatives like the Indiana Genomics Initiative or MetaCyt⁵. The data are of little value without annotation, and genomic data are constantly being updated. This creates two problems – 1) version control, and 2) synchronization of data and annotation. It is essential that scholars know which version of a particular data set is being used and have mechanisms for version control. Likewise, synchronization is also essential so that modifications to the data are annotated, and that annotations are modified in concert with data (or at least that annotations are marked as being potentially incorrect due to data updates). This problem will become more and more pervasive as the trend toward ‘big data science’ and ‘multi-media humanities’ results in many similar situations across many disciplines.

Another issue related to annotation is the problem of establishing and maintaining the provenance of the data. If a scholar reaches a conclusion based upon data of uncertain origin or lineage, the strength of that conclusion is harmed by the “weak link” of questionable source data. It should be possible for information about the source of data to be tracked back through version control and annotation to the scholars who created and maintained the data.

The university has several opportunities to improve early lifecycle data activities and establish leadership. IU should establish a service to provide for the maintenance and electronic publishing of datasets. This service would provide mechanisms to allow a community of scholars to securely maintain annotation and provenance through peer review mechanisms – analogous to current academic journal and conference publication processes.

IU should engage in developing metadata management across disparate knowledge domains. Robust systems for rich data lifecycle management do not yet exist and are a fertile area for research. The university has researcher and technical strengths to explore proofs-of-concept and full development of reusable metadata management systems using eXtensible Markup Language and other tools. Thus, effective solutions for data annotation, provenance over time, and efficacious data publishing to serve a full data lifecycle will require both pioneering development and effective blending of off-the-shelf systems. .

CRT Recommendation #5: *Indiana University should research, develop, acquire, and implement new capabilities for the collection, annotation, and provenance management of data generated by IU researchers. Development of these capabilities should provide for annotation and management of massive streams of data, facilities for metadata management and reusability (such as XML- and standards-based data annotation), and management of data provenance.*

D) Data Publishing, Archive Curation, and Custodianship

Maintaining data for the long term in a form in which they can be used by researchers other than those who create the data is a critical aspect of custodianship and curation. The difference between custodianship and curation is subtle, but important: custodians protect, curators provide care and manage access. Similar to the custodian and curation aspects of a bank, there is a need for an entity that can act as a repository for the collection and storage of data for research, teaching, and public service uses. An example that demonstrates this type of service is the Indiana University Geographic Information Service (<http://www.indiana.edu/~gis/>), which provides access to valuable GIS data and software for students, faculty, and the public. The service acts as a repository for GIS data, adding value to the data by organizing available GIS software and data and providing access via the Indiana Spatial Data Portal (<http://www.indiana.edu/~gisdata/>) to the public.

⁵ Indiana Metabolomics Cytomics Initiative, <http://metacyt.indiana.edu>.

Critical aspects of curation include engagement in all parts of the data life cycle; ensuring that the data do not physically degrade over time which includes maintaining data integrity and archiving or maintaining the software used to access data; providing a secure and reliable means of archiving data; and actively working with research communities to judiciously cull data from archives that are no longer useful.

Scholars need assistance for managing these activities for later stages in the data life cycle. Librarians' historical work with long-term physical data and their work on many digital formats and processes gives them a procedural knowledge head start on assisting with these challenges. A data publishing service could provide a common set of leveraged processes for the long-term collection, archiving, and curation of research data. This service would also address technology insertion over the years as data must be migrated to newer storage media and formats.

CRT Recommendation #6: *Indiana University should provide a service for maintaining and publishing of digital datasets within and beyond the university. This service should enable scholars to securely maintain annotation and provenance through appropriate review mechanisms – analogous to journal and conference publication processes used today in the academic community – and provide for the ongoing re-use of IU's scholarly data.*

2.3 Ensure Computational Capabilities for Data Analysis and Simulation

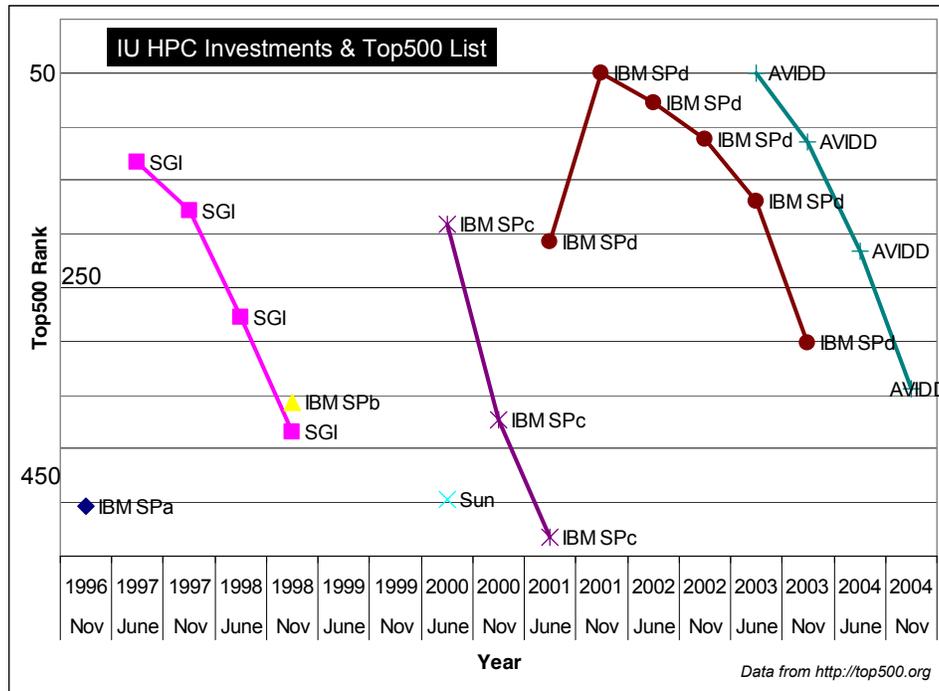
The capabilities to store large amounts of data are of limited value in the absence of appropriate statistical routines, expertise, and computing systems that permit analysis and interpretation of these data⁶. The university has sustained substantial investments in High Performance Computing (HPC) environments since the IT Strategic Plan, and deans and department chairs assert that IU's computing environment is often a factor in recruiting and retaining the best faculty. Policies that enable any member of the IU research community to obtain and use accounts on the university's supercomputers without cost recovery have enabled relatively broad adoption.

Both the IBM SP supercomputer and the AVIDD⁷ system were, when first announced, listed in 50th place on the list of the world's 500 most powerful (unclassified) supercomputers (www.top500.org). The following chart illustrates the sustained pace of technological innovation in HPC in the ever increasing capabilities of supercomputers. IU's IBM SP is no longer powerful enough, relative to current standards, to appear on the Top500 list. AVIDD lingers in the bottom third (#344) of the November 2004 list.

It is not just relative to peers that IU's systems seem small. IU's supercomputers are now chronically oversubscribed, and scholars analyzing large data sets and running large simulations are experiencing frustrations and delays in completing their research. Furthermore, as the diversity and sophistication of IU's supercomputer users increase, the need for systems is felt in many ways. Some scholars need access to many processors with large memory configurations; some need simply as many processors as they can possibly apply to a particular task. Other scholars require specialized hardware that is suitable for only particular tasks (such as astronomical or dynamical systems simulation) but which are unparalleled in their capabilities for these particular types of analyses.

⁶ *Statistics: Challenges and Opportunities for the Twenty-First Century*, Report of the National Science Foundation workshop on the Future of Statistics, May 2002.

⁷ Analysis and Visualization of Instrument-Driven Data – distributed multi-TeraFLOPS Linux clusters designed specifically for managing and analyzing very large data sets.



Thus, the necessity of maintaining computational cyberinfrastructure for data analysis and simulation requires continued investment. Expansion of IU's high performance computing environment should proceed in a fashion that supports the use, management, analysis, and distribution of extremely large scientific data sets from a variety of disciplines as described in Section 2.2. Even as computing systems of a given level of capability become cheaper every year, the systems required to advance cutting edge research grow ever more costly.

CRT Recommendation #7: *Indiana University should continue without pause its substantial investments in high performance computing systems and supercomputers, addressing the diverse needs of IU researchers, clinicians, engineers, and artists. IU should in particular focus on high performance computing for data-intensive scholarship.*

2.4 Expand Visualization

Beyond networks, visualization is perhaps the most broadly utilized and useful aspects of cyberinfrastructure. Scientific research commonly creates data sets and simulation results so large that visualization is the only means by which they can be understood and apprehended. Visualization facilities have created new art forms, and scholars in Fine Arts are among the leaders in developing this approach. IU's Department of Theater is a leader in the application of advanced visualization in theater design. IU librarians, musicians, folklorists, and scholars in many areas of the humanities are leaders in the visualization of music and still and moving images. Indeed, the annotation and markup of video images is one of the many projects pressing on (and pressed by) IU's capabilities in mass storage systems as previously mentioned.

CRT Recommendation #8: *Indiana University should continue to invest in a variety of distributed visualization facilities that broadly impact the scholarly and creative efforts of IU.*

3. Scholarly Use of Cyberinfrastructure

The capabilities promised by cyberinfrastructure are indeed tremendous, yet the complexities of the technology often impede its effective scholarly use. The ability to create digital data exceeds our collective ability to transform data into information and to use information to create knowledge and artistic works. One taskforce member observed that we can “generate GigaBytes of data in minutes, while real knowledge is only created at the rate of a few KiloBytes per month.”

The state of cyberinfrastructure skills varies by both discipline and researcher. Even disciplines with decades of history and tradition in the use of advanced information technology resources find the current state of cyberinfrastructure daunting. For example, chemists and physicists struggle with the complexity of national cyberinfrastructure efforts, such as the TeraGrid, while other scholars wrestle with complexities of evolving desktop tools.

This same cyberinfrastructure is intended for use by artists, library scientists, and scholars from many other disciplines that begin with a more modest history, tradition, and collective disciplinary knowledge of advanced IT tools. One of the key problems with cyberinfrastructure as it exists today is that it is simply too complex to use. Stanley Ahalt, Director of the Ohio Supercomputer Center, points out that advanced information technology is being used effectively by just a small fraction of the scholars (particularly artists) for whom it could be potentially beneficial⁸. Even for those disciplines that have strong traditions in using cutting-edge IT, the increased complexity of each component and subcomponent of particular research endeavors is becoming overwhelming. The need for diverse and sophisticated skills means that many advances in scholarly endeavors increasingly will depend on collaborative teams where each person brings a particular set of expertise.

The taskforce identified three approaches to accelerate the use of cyberinfrastructure:

- 1) Subsidizing modest help and consulting while also offering full cost recovery outsourcing (e.g., planned budget line items in grant applications) for particular technical needs would be a valuable shared service for scholars. Provision of these services would require a blend of deep technical expertise and discipline knowledge in the application of cyberinfrastructure. UITS' Research and Academic Computing division already uses this model of both subsidized and chargeback services. Taskforce members affirmed the value of a blended approach serving both modest needs and full outsourcing.
- 2) Developing tools and interfaces that hide cyberinfrastructure complexity is viewed as a highly valuable strategy for advancing cyberinfrastructure use. As one taskforce member remarked, “I would rather we eliminate the need to know eXtensible Markup Language (XML) than be trained to understand it.”

Human-Computer Interaction experts some time ago developed the concept of “patterns” – commonly performed tasks that may have a number of options. Software tools commonly referred to as “software wizards” are the most widely used example of this concept. To date, there seem to be no good examples of wizards in HPC or cyberinfrastructure applications. This has been a powerful technique, exploited most effectively perhaps by Microsoft, Inc., and development of this approach to easing the challenges of utilizing cyberinfrastructure would be of great benefit to IU scholars.

⁸ “Blue Collar Computing: HPC for the Rest of Us.” Stanley C. Ahalt and Kathryn L. Kelley. *ClusterWorld* 2(11), 2004.

Scholarly portals that provide intuitive, web-based user interfaces to advanced IT facilities are another example of hiding complexity. Such portals are already in use by the Bioinformatics Program and the Center for Medical Genomics of the Indiana Genomics Initiative in Indianapolis and the Center for Genomics and Bioinformatics in Bloomington. Other scientific portals are also being developed.

3) Providing education and training that is suitable to the particular needs of individuals in particular areas of research, clinical, engineering, and artistic pursuit. Scholars need reliable assistance to develop their skills to use cyberinfrastructure, and effective training and help programs must recognize that disciplines are developing at differing rates and face different challenges in using cyberinfrastructure.

In each approach, consulting, complexity-hiding interfaces, and training must be discipline-facing in their orientation. A one size fits all approach is unlikely to be responsive to the blend of technical and domain knowledge required for increased scholarly outputs.

CRT Recommendation #9: *Indiana University should foster effective use of cyberinfrastructure through an array of consulting services, complexity-hiding interfaces, and training that will enable scholars to be more innovative and productive (and thus more competitive for grants) through the use of cyberinfrastructure.*

4. Partnering to Develop and Sustain Cyberinfrastructure

The previous recommendations represent considerable investments of university strategy, finance, and political will. As is the nature of cyberinfrastructure, many of these are best pursued as leveraged investments through partnerships within and beyond the university. As the vice president himself recently noted,

Cyberinfrastructure is truly global, it knows no national boundaries and in many fields the best research can only be done in a global context. Thus TransPAC2 is a vital component of the global cyberinfrastructure as it links research and education networks in the Asia-Pacific - the fastest growing economic region in the world, to research and education networks in the US⁹.

To develop and sustain the university's cyberinfrastructure, IU must aggressively and effectively pursue external sources of funding. These include extramural grants and public/private partnerships to leverage university investments. I-Light (and soon I-Light2) demonstrates considerable leverage for cyberinfrastructure investments when the State of Indiana, IU, and Purdue University collaborate. I-Light demonstrates a founding investment in developing a statewide cyberinfrastructure that is helping enhance the State's high-tech economy and knowledge ecology. IU must continue to engage, and when possible, establish a leadership position in developing cyberinfrastructure at the state, national, and international levels.

Taskforce members also acknowledge that each cyberinfrastructure gain can pay increasing returns for future grants. The university's existing IT infrastructure has contributed to obtaining grants, and a rapid development of advanced university cyberinfrastructure is expected to further differentiate IU in intense grant competition. The taskforce encourages vigilance in brokering internal and external partnerships at every opportunity that can help develop and sustain both IU's cyberinfrastructure and the advancement of knowledge creation through the use of cyberinfrastructure.

⁹ Michael A. McRobbie, TransPAC2 Inauguration Ceremony, Saturday, April 2, 2005, in Tokyo, Japan.

CRT Recommendation #10: *Indiana University should continue to lead and participate in leveraged efforts to develop, deploy, and make use of cyberinfrastructure for the State of Indiana, at national, and international levels.*

5. Conclusion

The taskforce members collectively affirm these recommendations as essential directions for development of IU's cyberinfrastructure. These recommendations represent the consensus views on cyberinfrastructure across many fields of scholarly endeavor.

The extended discussions of the taskforce also make clear that investments in cyberinfrastructure are essential, but not sufficient, for a continued growth of IU's externally funded research. Development of advanced cyberinfrastructure is itself a research endeavor for the computer sciences and related technology fields. Accurate data analysis and knowledge creation often requires ready access deep statistical expertise in appropriate techniques. Developing cyberinfrastructure use in the arts and humanities requires effective engagement within their culture. Physical space for research projects remains a deep and pervasive challenge across the entire university. These observations – and many others of enormous importance to specific disciplines – were not addressed in this consensus report for shared cyberinfrastructure.

The taskforce members provide this advice in response to the charge received from the IU Vice President for Research & Information Technology. It is the taskforce's sincere hope that these recommendations will be a basis of action within IU, and will provide for an ongoing dialogue to meet the cyberinfrastructure needs of the scholarly community.

Appendix A: Taskforce Charge

Dear Colleagues,

As you are no doubt aware, research is becoming almost completely digital. In this world of e-research, instruments and measuring devices are producing vast amounts of data from sites throughout the world. This data is stored in repositories from which it can be rapidly retrieved and analyzed on supercomputers, the results of which are then visualized on large-scale imaging devices. High-performance networks connect these instruments, the data repositories, and data analysis environments that are an essential part of global, digital science. Collectively such infrastructure is known as cyberinfrastructure, and a recent important NSF report calls for a coordinated approach to the construction of national cyberinfrastructure.

President Herbert in his Inaugural Address stated: "In addition to increased scholarly publications, works of art, concerts and other forms of creative scholarly activity, our goal must be to double Indiana University's externally funded research grants and contracts by the end of this decade. Such a record of accomplishment will move the university even higher among the ranks of America's most distinguished institutions."

Hence the continued development of IU's cyberinfrastructure will play an essential role in helping to achieve the President's goal of doubling IU's external research funding. The IT Strategic Plan and numerous grants have provided a very strong foundation for IU's cyberinfrastructure. However, we need to ensure IU is both on the right course in the development of its cyberinfrastructure and that it will fully support the President's goal.

Therefore, I am commissioning a Cyberinfrastructure Research Taskforce (CRT) with some of IU's leading researchers for whose disciplines cyberinfrastructure is essential to advise me regarding the University's needs in this regard. The CRT will conduct its work during the 2004-05 academic year, and I ask for its full report by mid spring 2005. The Taskforce will be chaired by Brad Wheeler, Associate VP for Research and Academic Computing and Dean for IT at IU Bloomington. I will attend as many of the CRT meetings as possible.

I am very mindful that your time is precious. I sincerely hope that you will accept this invitation to help shape IU's research technology infrastructure.

Yours sincerely,

Michael McRobbie
Vice President for Research and Information Technology

Appendix B: Taskforce Membership



Michael McRobbie, Vice President for Research & Information Technology

Sponsor

[Homepage](#)

vpit@indiana.edu



Brad Wheeler, Chairman

IU Assoc VP for Research & Academic Computing, IUB Dean of IT

Associate Professor of Information Systems, Kelley School of Business

[Homepage](#)

bwheeler@iu.edu



Hasan Akay

IUPUI Chair, Department of Mechanical Engineering

[Homepage](#)

hakay@iupui.edu



M. Pauline Baker

IUPUI Associate Professor, Informatics

[Homepage](#)

baker@indiana.edu



Randall Barry Bramley

IU Associate Professor, Department of Computer Science

[Homepage](#)

bramley@indiana.edu



David E. Clemmer

IUB Chairman, Analytical Chemistry

[Homepage](#)

[Homepage2](#)

clemmer@indiana.edu



Alex R. Dzierba
IUB Professor of Physics
[Homepage](#)
dzierba@indiana.edu



Howard J. Edenberg
IUMed Director, Center for Medical Genomics & Professor, Biochemistry & Molecular
Biology & Medical and Molecular Genetics
[Homepage](#)
edenberg@iupui.edu



Lawrence Einhorn
IUMed Distinguished Professor of Medicine
[Homepage](#)
leinhorn@iupui.edu



Geoffrey C. Fox
IUB Professor of Computer Science, Informatics, Physics & Director of Community Grids
Laboratory
[Homepage](#)
gcf@indiana.edu



Dennis B. Gannon
IUB Professor of Computer Science & Science Director for the Indiana Pervasive
Technology Labs
[Homepage](#)
gannon@indiana.edu



Jeffrey Hass
Director of the Center for Electronic and Computer Music, IU School of Music
URL:
hassj@indiana.edu



John Huffman
IUB Senior Scientist in Chemistry, Adjunct Professor of Informatics, Director Indiana
University Molecular Structure Center, Director Informatics Research Institute
[Homepage](#)
huffman@indiana.edu



Thom Kaufman
IUB Distinguished Professor of Genetics
[Homepage](#)
kaufman@bio.indiana.edu



J. Scott Long
IUB Chancellor's Professor, Department of Sociology
[Homepage](#)
julong@indiana.edu



Bruce A. Molitoris
IUMed Professor of Medicine
[Homepage](#)
bmolitor@iupui.edu



Sean David Mooney
IUPUI Assistant Professor, Center for Computational Biology and Bioinformatics,
Department of Medical and Molecular Genetics
[Homepage](#)
sdmooney@iupui.edu



William Royall Newman
IUB Professor, History and Philosophy of Science
[Homepage](#)
wnewman@indiana.edu



J. M Overhage
IUMed Associate Professor of Medicine
[Homepage](#)
joverhag@iupui.edu



James H. Patterson
IUB Professor of Operations and Decision Technologies & Chairman, Technology Policy
Committee (Bloomington Faculty Council) & Co-Chairman, Technology Policy Committee
(University Faculty Council)
[Homepage](#)
pattersj@indiana.edu



Catherine A. Pilachowski

IUB Professor of Astronomy and Daniel Kirkwood Chair in Astronomy

[Homepage](#)

cpilacho@indiana.edu



Beth A. Plale

IUB Assistant Professor, Department of Computer Science

[Homepage](#)

plale@indiana.edu



Robert A. Shakespeare

IUB Professor, Lighting Designer, Head of Design and Technology; Department of Theatre and Drama

URL:

shakespe@indiana.edu



Paul E. Sokol

IUB Director and Professor of Physics

[Homepage](#)

[Homepage2](#)

pesokol@indiana.edu



Ruth M. Stone

IUB Laura Boulton Professor and Chair, Department of Folklore and Ethnomusicology

[Homepage](#)

stone@indiana.edu



Wendy Ching-Wen Chang, ex-officio

IUE Vice Chancellor for Information Technology & Associate Professor of Computer Science

[Homepage](#)

wchang@indiana.edu



Tom Hacker, ex-officio

Associate Director for Research & Academic Computing, UITS

URL:

hacker@indiana.edu



Carol Kegeris, ex-officio
Project Coordinator, OVPIT
URL:
ckegeris@iupui.edu



Gustav Meglicki, ex-officio
Senior Technical Advisor, OVPIT
[Homepage](#)
gustav@indiana.edu



Craig Stewart, ex-officio
Director for Research and Academic Computing & Director, Indiana Genomics Initiative,
Information Technology Core & Special Assistant for the Life Sciences, OVPIT & Adjunct
Associate Professor, Department of Medical and Molecular Genetics, IUPUI; Department of
Biology, IUB.
[Homepage](#)
stewart@indiana.edu



Steve Wallace, ex-officio
Director and Chief Technologist, Advanced Network Management Lab
URL:
ssw@anml.iu.edu

Appendix C: IT Strategic Plan

The 1998 Indiana University IT Strategic Plan included two recommendations and a total of 15 specific action items that deal specifically with scholarly research and artistic activities, as follows:

“RECOMMENDATION 5: In support of research, UITS should provide broad support for basic collaboration technologies and being implementing more advanced technologies. UITS should provide advanced data storage and management services to researchers. The University should continue its commitment to high performance computing and computation, so as to contribute to and benefit from initiatives to develop a national computational grid.

- ! ACTION 27: UITS should launch an aggressive program to systematically evaluate and deploy across the University state-of-the-art tools and infrastructure that can support collaboration within the University, nationally and globally.
- ! ACTION 28: UITS should explore and deploy advanced and experimental collaborative technologies within the University's production information technology environment, first as prototypes and then if successful, more broadly.
- ! ACTION 29: In order to maintain its position of leadership in the constantly changing field of high performance computing, the University should plan to continuously upgrade and replace its high-performance computing facilities to keep them at a level that satisfies the increasing demand for computational power.
- ! ACTION 30: The University needs to provide facilities and support for computationally and data-intensive research, for non-traditional areas such as the arts and humanities, as well as for the more traditional areas of scientific computation.
- ! ACTION 31: The University should plan to evolve its high performance computing and communications infrastructure so it has the features to be compatible with and can participate in the emerging national computational grid.
- ! ACTION 32: The University should evaluate and acquire high-capacity storage systems, capable of managing very large data volumes from research instruments, remote sensors, and other data-gathering facilities.
- ! ACTION 33: The University through UITS should provide support for a wider range of research software including database systems, text-based and text-markup tools, scientific text processing systems, and software for statistical analysis. UITS should investigate the possibilities for enterprise-wide agreements for software acquisitions similar to the Microsoft Enterprise License Agreement.
- ! ACTION 34: UITS should participate with faculty on major research initiatives involving information technology, where it is appropriate and of institutional advantage. Further, UITS should provide proactive encouragement and supportive services that create opportunities where faculty from diverse disciplines might come together on collaborative projects involving information technology.”

“RECOMMENDATION 9: The University should build upon and expand its digital library program, and develop the digital library infrastructure needed to support research, teaching and learning.

- ! ACTION 59: The University should develop a program of digital library research, and engage in national initiatives, to address the issues of user services, creation and management of digital collections, the federation of distributed digital libraries, and the design of digital library systems.

- ! ACTION 60: The University should develop a digital library infrastructure that will provide a common technical and organizational base for new and ongoing digital library programs.
- ! ACTION 61: The University Libraries, with UITS, should provide students, faculty and staff at all campuses with convenient and reliable access to a comprehensive and coordinated collection of electronic information resources, on the campuses and off.
- ! ACTION 62: The University should develop within its digital library program an "electronic reserve" service so that faculty can assemble and make available content in all media and formats: text, image, audio, or video; published or unpublished; digitized representation or original digital artifact; etc.
- ! ACTION 63: The University should establish sound funding for existing digital library initiatives (including Variations, LETRS, IMDS, others), and should provide support for other digital library projects of merit that are advanced in the years ahead.
- ! ACTION 64: UITS, in partnership with the University Archives, Internal Audit, the Committee of Data Stewards, and others should develop a program to assure preservation of electronic institutional records.
- ! ACTION 65: UITS, in partnership with the University Libraries, University Archives, and others should evaluate technologies and propose methods and standards to protect digital materials against media deterioration and technological obsolescence."